Relative Flatness and Generalization

Henning Petzka[*]
Lund University, Sweden
henning.petzka@math.lth.se

Michael Kamp[*]
CISPA Helmholtz Center for Information Security,
Germany and Monash University, Australia
michael.kamp@monash.edu

Linara Adilova
Ruhr University Bochum, Germany
and Fraunhofer IAIS

Cristian Sminchisescu
Lund University, Sweden
and Google Research, Switzerland

Mario Boley
Monash University, Australia

Abstract

Flatness of the loss curve is conjectured to be connected to the generalization ability of machine learning models, in particular neural networks. While it has been empirically observed that flatness measures consistently correlate strongly with generalization, it is still an open theoretical problem why and under which circumstances flatness is connected to generalization, in particular in light of reparameterizations that change certain flatness measures but leave generalization unchanged. We investigate the connection between flatness and generalization by relating it to the interpolation from representative data, deriving notions of representativeness, and feature robustness. The notions allow us to rigorously connect flatness and generalization and to identify conditions under which the connection holds. Moreover, they give rise to a novel, but natural relative flatness measure that correlates strongly with generalization, simplifies to ridge regression for ordinary least squares, and solves the reparameterization issue.

1 Introduction

Flatness of the loss curve has been identified as a potential predictor for the generalization abilities of machine learning models [6, 10, 11]. In particular for neural networks, it has been repeatedly observed that generalization performance correlates with measures of flatness, i.e., measures that quantify the change in loss under perturbations of the model parameters [4, 8, 16, 21, 34, 39, 41, 44]. In fact, Jiang et al. [14] perform a large-scale empirical study and find that flatness-based measures have a higher correlation with generalization than alternatives like weight norms, margin-, and optimization-based measures. It is an open problem why and under which circumstances this correlation holds, in particular in the light of negative results on reparametrizations of ReLU neural networks [5]: these reparameterizations change traditional measures of flatness, yet leave the model function and its generalization unchanged, making these measures unreliable. We present a novel and rigorous approach to understanding the connection between flatness and generalization by relating it to the interpolation from representative samples. Using this theory we, for the first time, identify conditions under which flatness explains generalization. At the same time, we derive a measure of *relative flatness* that simplifies to ridge/Tikhonov regularization for ordinary least squares [36], and resolves

[*]equal contribution

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

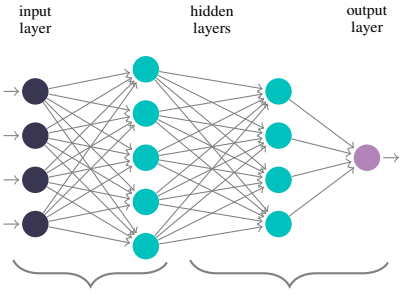

Figure 1: Decomposition of f = into a feature extractor and a model for neural networks.

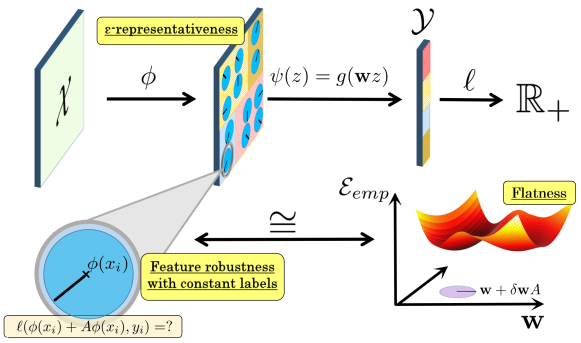

Figure 2: Overview: We theoretically connect a notion of representative data with a notion of feature robustness and a novel measure of flatness of the loss surface.

the reparametrization issue for ReLU networks [5] by appropriately taking the norm of parameters into account as suggested by Neyshabur et al. [25].

Formally, we connect flatness of the loss surface to the *generalization gap* $E_{gen}(f; S) = E(f) - E_{emp}(f; S)$ of a *model* $f : X \to Y$ from a *model class* H with respect to a twice differentiable *loss function* $\ell : Y \times Y \to \mathbb{R}_+$ and a finite sample set $S \subset X \times Y$, where

$$E(f) = \mathbb{E}_{(x;y)\sim\mathcal{D}}\left[\ell(f(x); y)\right] \quad \text{and} \quad E_{emp}(f; S) = \frac{1}{|S|}\sum_{(x;y)\in S} \ell(f(x); y) \ .$$

That is, $E_{gen}(f; S)$ is the difference between the *risk* $E(f)$ and the *empirical risk* $E_{emp}(f; S)$ of f on a finite sample set S drawn iid. according to a *data distribution* D on $X \times Y$. To connect flatness to generalization, we start by decomposing the generalisation gap into two terms, a *representativeness* term that quantifies how well a distribution D can be approximated using distributions with local support around sample points and a *feature robustness* term describing how small changes of feature values affect the model's loss. Here, feature value refers to the implicitly represented features by the model, i.e., we consider models that can be expressed as $f(x) = (\mathbf{w}; \phi(x)) = g(\mathbf{w}\ \phi(x))$ with a feature extractor and a model (which includes linear and kernel models, as well as most neural networks, see Fig. 1). With this decomposition, we measure the generalization ability of a particular model by how well its interpolation between samples in feature space fits the underlying data distribution. We then connect feature robustness (a property of the feature space) to flatness (a property of the parameter space) using the following key identity: Multiplicative perturbations in feature space by arbitrary matrices $A \in \mathbb{R}^{m\times m}$ correspond to perturbations in parameter space, i.e.,

$$(\mathbf{w}; \phi(x) + A\ \phi(x)) = g(\mathbf{w}(\phi(x) + A\ \phi(x))) = g((\mathbf{w} + \mathbf{w}A)\ \phi(x)) = (\mathbf{w} + \mathbf{w}A; \phi(x)) \ . \quad (1)$$

Using this key equation, we show that feature robustness is approximated by a novel, but natural, loss Hessian-based *relative flatness* measure under the assumption that the distribution can be approximated by locally constant labels. Under this assumption and if the data is representative, then flatness is the main predictor of generalization (see Fig. 2 for an illustration).

This offers an explanation for the correlation of flatness with generalization on many real-world data distributions for image classification [14, 21, 26], where the assumption of locally constant labels is reasonable (the definition of adversarial examples [35] even hinges on this assumption). This dependence on locally constant labels has not been uncovered by previous theoretical analysis [26, 37]. Moreover, we show that the resulting relative flatness measure is invariant to linear reparameterization and has a stronger correlation with generalization than other flatness measures [14, 21, 26]. Other measures have been proposed that similarly achieve invariance under reparameterizations [21, 37], but the Fisher-Rao norm [21] is lacking a strong theoretical connection to generalization, our measure sustains a more natural form than normalized sharpness [37] and for neural networks, it considers only a single layer, given by the decomposition of f (including the possibility of choosing the input layer when $\phi = id_{\mathcal{X}}$). An extended comparison to related work is provided in Appdx. A.

The limitations of our analysis are as follows. We assume a noise-free setting where for each $x \in X$ there is a unique $y = y(x) \in Y$ such that $P_{x;y\sim D}(y|x) = 1$, and this assumption is also extended to

the feature space of the given model, i.e., we assume that $(x) = (x')$ implies $y(x) = y(x')$ for all $x; x' \in X$ and write $y(x) = y(\ (x))$. Moreover, we assume that the marginal distribution $D_{\mathcal{X}}$ is described by a density function $p_{\mathcal{D}}(x)$, that $f(x) = (\mathsf{w};\ (x)) = g(\mathsf{w}\ (x))$ is a local minimizer of the empirical risk on S, and that g, , are twice differential. Quantifying the representativeness of a dataset precisely is challenging since the data distribution is unknown. Using results from density estimation, we derive a worst-case bound on representativeness for all data distributions that fulfill mild regularity assumptions in feature space (X), i.e., a smooth density function $p_{(\mathcal{D})}$ such that $\int_{z\in(\mathcal{X})} r^2\ p_{(\mathcal{D})}(z)\|\|^2\ dz$ and $\int_{z\in(\mathcal{X})} p_{(\mathcal{D})}(z)/\|\|^m\ dz$ are well-defined and finite. This yields a generalization bound incorporating flatness. In contrast to the common bounds of statistical learning theory, the bound depends on the feature dimension. The dimension-dependence is a result of the interpolation approach (applying density estimation uniformly over all distributions that satisfy the mild regularity assumptions). The bound is consistent with the no-free-lunch theorem and the convergence rate derived by Belkin et al. [2] for a model based on interpolations. In practical settings, representativeness can be expected to be much smaller than the worst-case bound, which we demonstrate by a synthetic example in Sec. 6. Generally, it is a bound that remains meaningful in the interpolation regime [1, 3, 24], where traditional measures of generalization based on the empirical risk and model class complexity are uninformative [22, 42].

Contribution. In summary, this paper rigorously connects flatness of the loss surface to generalization and shows that this connection requires feature representations such that labels are (approximately) locally constant, which is also validated in a synthetic experiment (Sec. 6). The empirical evaluation shows that this flatness and an approximation to representativeness can tightly bound the generalization gap. Our contributions are: (i) the rigorous connection of flatness and generalization; (ii) novel notions of representativeness and feature robustness that capture the extent to which a model's interpolation between samples fits the data distribution; and (iii) a novel flatness measure that is layer- and neuron-wise reparameterization invariant, reduces to ridge regression for ordinary least squares, and outperforms state-of-the-art flatness measures on CIFAR10.

2 Representativeness

In this section, we formalize when a sample set S is representative for a data distribution D.

Partitioning the input space. We choose a partition $\{V_i \mid i = 1; \ldots; |S|\}$ of X such that each element of this partition V_i contains exactly one of the samples x_i from S. The distribution can then be described by a set of densities $p_i(x) = \frac{1}{\ _i}\ p_{\mathcal{D}}(x)\ \ 1_{V_i}(x)$ with support contained in V_i (where $1_{V_i}(x) = 1$ if $x \in V_i$ and 0 otherwise) and with normalizing factor $\ _i = \int_{V_i} p_{\mathcal{D}}(x)dx$. Then the risk decomposes as $E(f) = \sum_{i=1}^{|S|}\ _i\ \ \mathsf{E}_{x\sim p_i}[\ell(f(x); y(x))]$: Since $x_i \in V_i$ for each i, we can change variables and consider density functions $\ ^*_i(\ ) = p_i(x_i + \ )$ with support in a neighborhood around the origin of X. The risk then decomposes as

$$E(f) = \sum_{i=1}^{|S|}\ _i\ \ \mathsf{E}_{\ \sim\ _i}[\ell(f(x_i + \ ); y(x_i + \ ))]\ . \tag{2}$$

Starting from this identity, we formalize an approximation to the risk: In a practical setting, the distribution $p_{\mathcal{D}}$ is unknown and hence, in the decomposition (2), we have unknown densities $\ ^*_i$ and unknown normalization factors $\ _i$. We assume that each neighborhood contributes equally to the loss, i.e., we approximate each $\ _i$ with $\frac{1}{|S|}$. Then, given a sample set S and an $|S|$-tuple $= (\ _i)_{1\le i\le|S|}$ of "local" probability density functions on X with support $\mathrm{supp}(\ _i)$ in a neighborhood around the origin $0_{\mathcal{X}}$, we call the pair $(S;\ )$ *-representative for* D with respect to a model f and loss ℓ if $|E_{Rep}(f; S;\ )| \le$, where

$$E_{Rep}(f; S;\ ) = E(f) - \sum_{i=1}^{|S|}\frac{1}{|S|}\ \ \mathsf{E}_{\ \sim\ _i}[\ell(f(x_i + \ ); y(x_i + \ ))]\ . \tag{3}$$

If the partitions V_i and the distributions $\ _i$ are all chosen optimal so that the approximation $\ _i = \frac{1}{|S|}$ is exact and $\ _i = \ ^*_i$, then $E_{Rep}(f; S;\ ) = 0$ by (2). If the support of each $\ _i$ is decreased to the origin so that $\ _i = \ _0$ is a Dirac delta function, then $E_{Rep}(f; S;\ ) = E_{gen}(f; S)$ equals the

generalization gap. For density functions with an intermediate support, the generalization gap can be decomposed into representativeness and the expected deviation of the loss around the sample points:

$$E_{gen}(f;S) = E_{Rep}(f;S;\ ) + \sum_{i=1}^{jSj} \frac{1}{jSj} \quad E \quad _i [`(f(x_i + \ ); y(x_i + \ )) \quad `(f(x_i); y_i)]$$

The main idea of our approach to understand generalization is to use this equality and to control both representativeness and expected loss deviations for a suitable tuple of distributions .

From input to feature space. An interesting aspect of representativeness is that it can be considered in a feature space instead of the input space. For a model $f = (\ ) : X \to Y$, we can apply our notion to the feature space (X) (see Fig. 1 for an illustration). This leads to the notion of representativeness in feature space defined for a tuple $= (\ _i)_{1 \ i \ jSj}$ of densities on (X) by replacing x_i with (x_i) in (3), which we denote by $E_{Rep}(f;S;\ )$. By measuring representativeness in a feature space, this becomes a notion of both data and feature representation. In particular, it assumes that a target output function $y(\ (x))$ also exists for the feature space. We can then decompose the generalization gap $E_{gen}(f)$ of $f = (\ )$ into

$$E_{Rep}(f;S;\ ) + \left(\frac{1}{jSj} \sum_{i=1}^{jSj} E \quad _i [`(\ (\ (x_i) + \ ); y(\ (x_i) + \ )) \quad `(f(x_i); y_i)] \right)$$

The second term is determined by how the loss changes under small perturbations in the feature space for the samples in S. As before, for $_i = \ _0$ the term in the bracket vanishes and $E_{Rep}(f;S;\ ) = E_{gen}$. But the decomposition becomes more interesting for distributions with support of nonzero measure around the origin. If the true distribution can be interpolated efficiently in feature space from the samples in S with suitable $_i$ so that $E_{Rep}(f;S;\ ) \ 0$, then the term in the bracket approximately equals the generalization gap and the generalization gap can be estimated from local properties in feature space around sample points.

3 Feature Robustness

Having decomposed the generalisation gap into a representativness and a second term of loss deviation, we now develop a novel notion of feature robustness that is able to bound the second term for specific families of distributions using key equation (1). Our definition of feature robustness for a model $f = (\ ) : X \to Y$ depends on a small number > 0, a sample set S and a feature selection defined by a matrix $A \in R^{m \ m}$ of operator norm $jjAjj \ 1$. With feature perturbations $_A(x) = (I + A) \ (x)$ and

$$E_F(f;S;A) := \frac{1}{jSj} \sum_{i=1}^{jSj} \left[`(\ (\ _A(x_i)); y[\ _A(x_i)]) \quad `(f(x_i); y_i) \right] ; \qquad (4)$$

the definition of feature robustness is given as follows.

Definition 1. Let $` : Y \quad Y \to R_+$ denote a loss function, and two positive (small) real numbers, $S \quad X \quad Y$ a finite sample set, and $A \in R^{m \ m}$ a matrix. A model $f(x) = (\ )(x)$ with $(X) \quad R^m$ is called $((\ ;S;A);\ )$-feature robust, if $E_F(f;S;\ A) \quad$ for all $0 \quad$: More generally, for a probability distribution A on perturbation matrices in R^m, we define

$$E_F(f;S;A) = E_{A \ A} \left[E_F(f;S;A) \right] ;$$

and call the model $((\ ;S;A);\ )$-feature robust on average over A, if $E_F(f;S;\ A) \quad$ for 0 .

Given a feature extractor, feature robustness measures the performance when often feature values are perturbed (with constant feature extractor). This local robustness at sample points differs from the robustness of Xu and Mannor [40] that requires a data-independent partitioning of the input space. The matrix A in feature robustness determines which feature values shall be perturbed. For

each sample, the perturbation is linear in the expression of the feature. Thereby, we only perturb features that are relevant for the output for a given sample and leave feature values unchanged that are not expressed. For mapping into an intermediate layer of a neural network, traditionally, the activation values of a neuron are considered as feature values, which corresponds to a choice of A as a projection matrix. However, it was shown by Szegedy et al. [35] that, for any other direction $v \in \mathbb{R}^m$; $\|v\| = 1$, the values $\langle h(x), v \rangle$ obtained from the projection of $h(x)$ onto v, can be likewise semantically interpreted as a feature. This motivates the consideration of general feature matrices A.

Distributions on feature matrices induce distributions on the feature space. Feature robustness is defined in terms of feature matrices (suitable for an application of (1) to connect perturbations of features with perturbations of weights), while the approach exploiting representative data from Section 2 considers distributions on feature vectors (cf. (3)). To connect feature robustness to the notion of ε-representativeness, we specify for any distribution on matrices $A \in \mathbb{R}^{m \times m}$ and $|S|$-tuple $\varepsilon_A = (\varepsilon_i)$ of probability density functions ε_i on the feature space $\mathbb{R}^m$ with support containing the origin. Multiplication of a feature matrix with a feature vector $h(x_i)$ defines a feature selection $A\ h(x_i)$, and for each $z \in \mathbb{R}^m$ there is some feature matrix A with $h(x_i) + z = h(x_i) + A\ h(x_i)$ (unless $h(x_i) = 0$). Our choice for distributions ε_i on $\mathbb{R}^m$ are therefore distributions that are induced via multiplication of feature vectors $h(x_i) \in \mathbb{R}^m$ with matrices $A \in \mathbb{R}^{m \times m}$ sampled from a distribution on feature matrices A. Formally, we assume that a Borel measure ε_i is defined by a probability distribution ε_A on matrices $\mathbb{R}^{m \times m}$. We then define Borel measures ε_i on $\mathbb{R}^m$ by $\varepsilon_i(C) = \varepsilon_A(\{A \mid A\ h(x_i) \in C\})$ for Borel sets $C \subseteq \mathbb{R}^m$. Then ε_i is the probability density function defined by the Borel measure ε_i. As a result, we have for each i that

$$E_{A\,\varepsilon_A}\left[\ell(f(A\,h(x_i)), y(f(A\,h(x_i))))\right] = E_{z\,\varepsilon_i}\left[\ell(f(h(x_i)+z), y(f(h(x_i)+z))\right]$$

Feature robustness and generalization. With this construction and a distribution ε_A on the feature space induced by a distribution A on feature matrices, we have that

$$E(f) = E_{emp}(f;S) + E_{Rep}(f;S;\varepsilon_A) + E_F(f;S;A) \tag{5}$$

Here, A can be any distribution on feature matrices, which can be chosen suitably to control how well the corresponding mixture of local distributions approximates the true distribution. The third term then measures how robust the model is in expectation over feature changes for A. In particular, if $E_{Rep}(f;S;\varepsilon_A) \approx 0$, then $E_{gen}(f;S) \approx E_F(f;S;A)$ and the generalization gap is determined by feature robustness. We end this section by illustrating how distributions on feature matrices induce natural distributions on the feature space. The example will serve in Sec. 5 to deduce a bound on $E_{Rep}(f;S;\varepsilon_A)$ from kernel density estimation.

Example: Truncated isotropic normal distributions are induced by a suitable distribution on feature matrices. We consider probability distributions $k_{\|\cdot - h(x_i)\|}$ on feature vectors $z \in \mathbb{R}^m$ in the feature space defined by densities $k_{\|\cdot - h(x_i)\|}(0;z)$ with smooth rotation-invariant kernels, bounded support and bandwidth h:

$$k_h(z_i;z) = \frac{1}{h^m}\, k\left(\frac{\|z_i - z\|}{h}\right) 1_{\|z_i - z\| < h} \tag{6}$$

with $1_{\|z_i - z\| < h} = 1$ when $\|z - z_i\| < h$ and 0 otherwise, and such that $\int_{z \in \mathbb{R}^m} k_h(z_0;z)\, dz = 1$ for all z_0. An example for such a kernel is a truncated isotropic normal distribution with variance $h^2\ ^2 I$, $k_h(z_i;z) = \mathcal{N}(z_i; h^2\ ^2)(z)$. The following result states that the densities (6) can indeed be induced by distributions on feature matrices, which will enable us to connect feature robustness with ε-representativeness.

Proposition 2. Let $S = \{h(x_i) \mid x_i \in S\}$ be a set of feature vectors in $\mathbb{R}^m$. With k_h defined as in (6), let $\varepsilon_i(z) = k_{\|\cdot - h(x_i)\|}(0;z)$ define an $|S|$-tuple ε of densities. Then there exists a distribution A on matrices in $\mathbb{R}^{m \times m}$ of norm less than ε such that for each $i = 1, \ldots, |S|$,

$$E_{A\,\varepsilon_A}\left[\ell(f(A\,h(x_i)), y(f(A\,h(x_i))))\right] = E_{\varepsilon_i}\left[\ell(f(h(x_i)+\cdot), y(f(h(x_i)+\cdot))\right]$$

The technical proof is deferred to the appendix, but we describe the distribution A on matrices for later use: The desired distribution is defined on the set of matrices of the form ε for a real number

and an orthogonal matrix O (i.e. $OO^T = O^T O = I$) as a product measure combining the (unique) Haar measure on the set of orthogonal matrices $O(m)$ with a suitable distribution on $\mathbb{R}$. The Haar measure on $O(m)$ induces the uniform measure on a sphere of radius r via multiplication with a vector of length r [17], and we choose a measure on $\mathbb{R}$ to match the radial change of the kernel k_σ.

4 Relative Flatness of the Loss Surface

Flatness is a property of the parameter space quantifying the change in loss under small parameter perturbations, classically measured by the trace of the loss Hessian $\mathrm{Tr}(H)$, where H is the matrix containing the partial second derivatives of the empirical risk with respect to all parameters of the model. In order to connect feature robustness (a property of the feature space) to flatness, we present how key equation (1) translates to the empirical risk: For a model $f(x; w) = \ell(w; \phi(x)) = g(w\ \phi(x))$ with parameters $w \in \mathbb{R}^{d\ \ m}$ and $g : \mathbb{R}^d \to Y$ a function on a matrix product of parameters w and a feature representation $X \to \mathbb{R}^m$ and any feature matrix $A \in \mathbb{R}^{m\ \ m}$ we have that

$$\begin{aligned} E_{\mathrm{emp}}(w + wA;\ \phi(S)) &= \frac{1}{|S|}\sum_{i=1}^{|S|} \ell(\ (w + wA;\ \phi(x_i)); y_i) \\ &= \frac{1}{|S|}\sum_{i=1}^{|S|} \ell(\ (w;\ \phi(x_i) + A\ \phi(x_i)); y_i) = \frac{1}{|S|}\sum_{i=1}^{|S|} \ell(\ (w;\ _A(x_i)); y_i) \end{aligned} \tag{7}$$

Subtracting $E_{\mathrm{emp}}(w;\ \phi(S)) = \frac{1}{|S|}\sum_{i=1}^{|S|} \ell(\ (w;\ \phi(x_i)); y_i)$, we can recognize feature robustness (4) on the right side of this equality when labels are constant under perturbations of the features, i.e. $y(\ _A(x_i)) = y_i$. In other words, flatness $E_{\mathrm{emp}}(w + v;\ \phi(S))\ \ E_{\mathrm{emp}}(w;\ \phi(S))$ describes the performance of a model function on perturbed feature vectors while holding labels constant. We proceed to introduce a novel, but natural, loss Hessian-based flatness measure that approximates feature robustness, given that the underlying data distribution satisfies the assumption of locally constant labels.

With $w_s = (w_{s;t})_t \in \mathbb{R}^{1\ \ m}$ denoting the s-th row of the parameter matrix w, we let $H_{s;s'}(w;\ \phi(S)) \in \mathbb{R}^{m\ \ m}$ denote the Hessian matrix containing all partial second derivatives of the empirical risk $E_{\mathrm{emp}}(w;\ \phi(S))$ with respect to weights in rows w_s and $w_{s'}$, i.e.

$$H_{s;s'}(w;\ \phi(S)) = \left(\frac{\partial^2 E_{\mathrm{emp}}(w;\ \phi(S))}{\partial w_{s;t}\,\partial w_{s';t'}}\right)_{1\ \ t;t'\ \ m}. \tag{8}$$

Definition 3. For a model $f(x; w) = g(w\ \phi(x))$, $w \in \mathbb{R}^{d\ \ m}$, with a twice differentiable function g, a twice differentiable loss function ℓ and a sample set S, relative flatness is defined by

$$\mathrm{Tr}_\phi(w) := \sum_{s;s'=1}^{d} \langle w_s; w_{s'}\rangle\ \ \mathrm{Tr}(H_{s;s'}(w;\ \phi(S))), \tag{9}$$

where Tr denote the trace and $\langle w_s; w_{s'}\rangle = w_s w_{s'}^T$ the scalar product of two row vectors.

Properties of relative flatness (i) Relative flatness simplifies to ridge regression for linear models $f(x; w) = wx \in \mathbb{R}$ ($X = \mathbb{R}^d$, $g =$ id and $\phi =$ id) and squared loss: To see this, note that for any loss function ℓ, the second derivatives with respect to the parameters $w \in \mathbb{R}^d$ computes to $\frac{\partial^2 \ell}{\partial w_i \partial w_j} = \frac{\partial^2 \ell}{\partial (f(x;w))^2} x_i x_j$. For $\ell(y; y') = \|y - y'\|^2$ the squared loss function, $\frac{\partial^2 \ell}{\partial y^2} = 2$ and the Hessian is independent of the parameters. In this case, $\mathrm{Tr}_{\mathrm{id}} = c\ \|w\|^2$ with a constant $c = \sum_{x \in S} 2\mathrm{Tr}(xx^T)$, which is the well-known Tikhonov (ridge) regression penalty.

(ii) Invariance under reparameterization: We consider neural network functions

$$f(x) = w^L\ (\cdots\ (w^2\ (w^1 x + b^1) + b^2)\cdots) + b^L \tag{10}$$

of a neural network of L layers with nonlinear activation function. By letting $^l(x)$ denote the composition of the first $l - 1$ layers, we obtain a decomposition $f(x; w^l) = g^l(w^l\ ^l(x))$ of the network. Using (9) we obtain a relative flatness measure $\mathrm{Tr}_{^l}(w)$ for the chosen layer.

For a well-de ned Hessian of the loss function, we require the network function to be twice differentiable. With the usual adjustments (equations only hold almost everywhere in parameter space), we can also consider neural networks with ReLU activation functions. In this case, Dinh et al. [5] noted that the network function —and with it the generalization performance— remains unchanged under linear reparameterization, i.e., multiplying layer with > 0 and dividing layer by , but common measures of the loss Hessian change. Our measure xes this issue in relating atness to generalization since the change of the loss Hessian is compensated by multiplication with the scalar products of weight matrices and is therefore invariant under layer-wise reparameterizations [cf. 26]. It is also invariant to neuron-wise reparameterizations, i.e., multiplying all incoming weights into a neuron by a positive number and dividing all outgoing weights by [23], except for neuron-wise reparameterizations of the feature layer. Using a simple preprocessing step (a neuron-wise reparameterization with the variance over the sample), our proposed measure becomes independent of all neuron-wise reparameterizations.

Theorem 4. Let $_i$ denote the variance of the i-th coordinate of (x) over samples $x \in S$ and $V = \mathrm{diag}\left({}_1; \ldots; {}_{n_{l-1}} \right)$. If the relative atness measure $_{\mathrm{Tr}}$ is applied to the representation

$$f(x) = w^L \left(\ldots \left(w^l V \left(V^{-1} w^{l-1} \left(\ldots \left(w^1 x + b^1 \right) \right) \ldots \right) + V^{-1} b^{l-1} \right) + b^l \right) \ldots \Big) + b^L$$

then $^l_{\mathrm{Tr}}$ is invariant under all neuron-wise (and layer-wise) reparameterizations

We now connect atness with feature robustness: Relative atness approximates feature robustness for a model at a local minimum of the empirical risk, when labels are approximately constant in neighborhoods of the training samples $(x); y) \in (S)$ in feature space.

Theorem 5. Consider a model $f(x; w) = g(w \ (x))$ as above, a loss function and a sample set S, and let $O_m \quad R^{m \ m}$ denote the set of orthogonal matrices. Let be a positive (small) real number and $w = ! \in R^{d \ m}$ denote parameters at a local minimum of the empirical risk on a sample set S. If the labels satisfy that $y(\ _A(x_i)) = y(\ (x_i)) = y_i$ for all $(x_i; y_i) \in S$ and all $jjAjj \quad 1$, then $f(x; !)$ is $((\ ; S; O_m); \ )$-feature robust on average over O_m for $= \frac{2}{2m} \ _{\mathrm{Tr}}(!) + O(\ ^3)$.

Applying the theorem to Eq. 5 implies that if the data is representative, i.e. $E_{\mathrm{Rep}}(f; S; \ _A) \quad 0$ for the distribution A of Prop. 2, then $E_{\mathrm{gen}}(f(\ ; !); S) \ . \ \frac{2}{2m} \ _{\mathrm{Tr}}(!) + O(\ ^3)$. The assumption on locally constant labels in Thm. 5 can be relaxed to approximately locally constant labels without unraveling the theoretical connection between atness and feature robustness. Appendix B investigates consequences from even dropping the assumption of approximately locally constant labels.

5 Flatness and Generalization

Combining the results from sections 2–4, we connect atness to the generalization gap when the distribution can be represented by smooth probability densities on a feature space with approximately locally constant labels. By approximately locally constant labels we mean that, for small , the loss in $jj \ (x_i)jj$-neighborhoods around the feature vector of a training sample x_i is approximated (on average over all training samples) by the loss for constant $y(x_i)$ on these neighborhoods. This and the following theorem connecting atness and generalization are made precise in Appendix D.4.

Theorem 6 (informal). Consider a model $f(x; w) = g(w \ (x))$ as above, a loss function and a sample set S, let m denote the dimension of the feature space de ned by , and let be a positive (small) real number. Let denote a local minimizer of the empirical risk on a sample set S. If the distribution D has a smooth density p_D on the feature space R^m with approximately locally constant labels around the points $x \in S$, then it holds with probability 1 over sample sets S that

$$E_{\mathrm{gen}}(f(\ ; !); S) \ . \ jSj^{\frac{2}{4+m}} \left(\frac{ \ _{\mathrm{Tr}}(!)}{2m} + C_1(p_D; L) + \frac{C_2(p_D; L)}{p} \right)^{!}$$

up to higher orders in jSj^{-1} for constants $C_1; C_2$ that depend only on the distribution in feature space p_D induced by , the chosen jSj-tuple and the maximal loss .

To prove Theorem 6 we bound both representativeness and feature robustness in Eq. 5. For that, the main idea is that the family of distributions considered in Proposition 2 has three key properties: (i) it

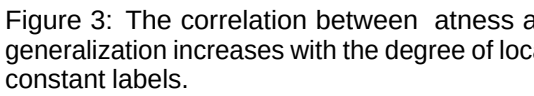

Figure 3: The correlation between atness and generalization increases with the degree of locally constant labels.

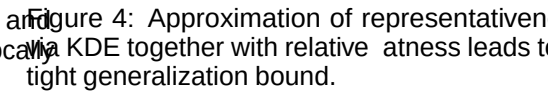

Figure 4: Approximation of representativeness via a KDE together with relative atness leads to a tight generalization bound.

provides an explicit link between the distributions on feature matrices A used in feature robustness and the family of distributions of -representativeness (Proposition 2) (ii) it allows us to bound feature robustness using Thm. 5; and (iii) it is simple enough that it allows us to use standard results of kernel density estimation (KDE) to bound representativeness.

Our bound suffers from the curse of dimensionality, but for the chosen feature space instead of the (usually much larger) input space. The dependence on the dimension is a result of using KDE uniformly over all distributions satisfying mild regularity assumptions. In practice, for a given distribution and sample set S, representativeness can be much smaller, which we showcase in a toy example in Sec. 6. In the so-called interpolation regime, where datasets with arbitrarily randomized labels can be t by the model class, the obtained convergence rate is consistent with the no free lunch theorem and the convergence rate derived by Belkin et al. [2] for an interpolation technique using nearest neighbors.

A combination of our approach with prior assumptions on the hypotheses or the algorithm in accordance to statistical learning theory could potentially achieve faster convergence rate. Our herein presented theory is instead based solely on interpolation and aims to understand the role of atness (a local property) in generalization: If the data is representative in feature layers and if the distribution can be approximated by locally constant labels in these layers, then atness of the empirical risk surface approximates the generalization gap. Conversely, Equation 7 shows that atness measures the performance under perturbed features only when labels are kept constant. As a result, we offer an explanation for the often observed correlation between atness and generalization: Real-world data distributions for classi cation are benign in the sense that small perturbations in feature layers do not change the target class, i.e., they can be approximated by locally constant labels. (Note that the de nition of adversarial examples hinges on this assumption of locally constant labels.) In that case, feature robustness is approximated by atness of the loss surface. If the given data and its feature representation are further -representative for small 0, then atness becomes the main contributor to the generalization gap leading to their noisy, but steady, correlation.

6 Empirical Validation

We empirically validate the assumptions and consequences of the theoretical results derived above[2]. For that, we rst show on a synthetic example that the empirical correlation between atness and generalization decreases if labels are not locally constant, up to a point when they are not correlated anymore. We then show that the novel relative atness measure correlates strongly with generalization, also in the presence of reparameterizations. Finally, we show in a synthetic experiment that while representativeness cannot be computed without knowing the true data distribution, it can in practice be approximated. This approximation—although technically not a bound anymore—tightly bounds the generalization gap. Synthetic data distributions for binary classi cation are generated by sampling 4 Gaussian distributions in feature space (two for each class) with a given distance between their means (class separation). We then sample a dataset in feature space S, train a linear classi er

[2] Code is available at https://github.com/kampmichael/relativeFlatnessGeneralization .

on the sample, randomly draw the weights of a 4-layer MLP and generate the input data as $S = (f^{-1}(S_x); S_y)$. This yields a dataset S and a model f such that $\hat{\phi}(S)$ has a given class separation. Details on the experiments are provided in Appdx. C.

Locally constant labels: To validate the necessity of locally constant labels, we measure the correlation between the proposed relative flatness measure and the generalization gap for varying degrees of locally constant labels, as measured by the class separation on the synthetic datasets. For each chosen class separation, we sample 100 random datasets of size 500 on which we measure relative flatness and the generalization gap. Fig. 3 shows the average correlation for different degrees of locally constant labels, showing that the higher the degree, the more correlated flatness is with generalization. If labels are not locally constant, flatness does not correlate with generalization.

Approximating representativeness: While representativeness cannot be calculated without knowing the data distribution, it can be approximated from the training sample S by the error of a density estimation on that sample. For that, we use multiple random splits of S into a training set S_{train} and a test set S_{test}, train a kernel density estimation on S_{train} and measure its error on S_{test}. Again, details can be found in Appx. C. The lower the class separation of the synthetic datasets, the harder the learning problem and the less representative a random sample will be. For each sample and its distribution, we compute the generalization gap and the approximation to the generalization bound. The results in Fig. 4 show that the approximated generalization bound tightly bounds the generalization error (note that this approximation is technically not a bound anymore). Moreover, as expected, the bound decreases the easier the learning problems become.

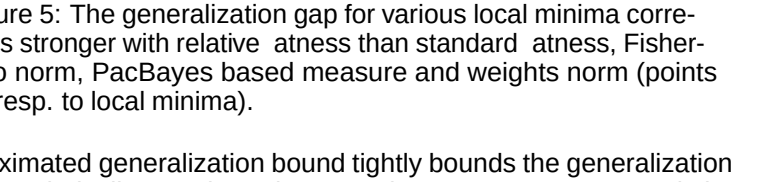

Figure 5: The generalization gap for various local minima correlates stronger with relative flatness than standard flatness, Fisher-Rao norm, PacBayes based measure and weights norm (points corresp. to local minima).

Relative flatness correlates with generalization: We validate the correlation of relative flatness to the generalization gap in practice by measuring it for 1100 different local minima—achieved via different learning setups, such as initialization, learning rate, batch size, and optimization algorithm—of LeNet5 [19] on CIFAR10 [18]. We compare this correlation to the classical Hessian-based flatness measures using the trace of the loss-Hessian, the Fisher-Rao norm [21], the PACBayes flatness measure that performed best in the extensive study of Jiang et al. [14], and the L_2-norm of the weights. The results in Fig. 5 show that indeed relative flatness has higher correlation than all the competing measures. Of these measures, only the Fisher-Rao norm is reparameterization invariant but shows the weakest correlation in the experiment. In Appdx C we show how reparameterizations of the network significantly reduce the correlation for non-reparameterization invariant measures.

7 Discussion and Conclusion

Contributing to the trustworthiness of machine learning, this paper provides a rigorous connection between flatness and generalization. As to be expected for a local property, our association between flatness and generalization requires the samples and its representation in feature layers to be representative for the target distribution. But our derivation uncovers a second, usually overlooked condition. Flatness of the loss surface measures the performance of a model close to training points when labels are kept locally constant. If a data distribution violates this, then flatness cannot be a good indicator for generalization.

Whenever we consider feature representations other than the input features, the derivation of our results makes one strong assumption: the existence of a target output function $y(\phi(x))$ on the feature space (X). By moving assumptions on the distribution from the input space to the feature space, we achieve a bound based on interpolation that depends on the dimension of the feature layer instead of the input space. Hence, we assume that the feature representation is reasonable and does not lose information that is necessary for predicting the output. To achieve faster convergence rates independent of any involved dimensions, future work could aim to combine our approach of interpolation with a prior-based approach of statistical learning theory.

Our measure of relative atness may still be improved in future work. Better estimates for the generalization gap are possible by improving the representativeness of local distributions in two ways: The support shape of the local distributions can be improved and their volume-parameter can be optimally chosen. Both improvements will affect the derivation of the measure of relative atness as an estimation of feature robustness for the corresponding distributions on feature matrices. Whereas different support shapes change the trace to a weighted average of the Hessian eigenvalues, the volume parameter can provide a correcting scaling factor. Both approaches seem promising to us, as our relative measure from De nition 3 already outperforms the competing measures of atness in our empirical validation.

Acknowledgements

Cristian Sminchisescu was supported by the European Research Council Consolidator grant SEED, CNCS-UEFISCDI (PN-III-P4-ID-PCE-2016-0535, PN-III-P4-ID-PCCF-2016-0180), the EU Horizon 2020 grant DE-ENIGMA (688835), and SSF.

Mario Boley was supported by the Australian Research Council (under DP210100045).

We would like to thank Julia Rosenzweig, Dorina Weichert, Jilles Vreeken, Thomas Gärtner, Asja Fischer, Tatjana Turova and Alexandru Aleman for the great discussions.

References

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

[2] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Over tting or perfect tting? risk bounds for classi cation and regression rules that interpolate. In Advances in Neural Information Processing Systems, pages 2300–2311, 2018.

[3] 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.

[4] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In Proceedings of the International Conference of Learning Representations, 2017.

[5] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1019–1028. JMLR. org, 2017.

[6] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. arXiv, 2017.

[7] Gintare Karolina Dziugaite and Daniel M Roy. Data-dependent pac-bayes priors via differential privacy. In Advances in Neural Information Processing Systems, pages 8430–8441, 2018.

[8] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for ef ciently improving generalization. In Proceedings of the International Conference on Learning Representations, 2021.

[9] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256. PMLR, 2010.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in Neural Information Processing Systems*, pages 529–536, 1995.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[12] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *34th Conference on Uncertainty in Artificial Intelligence*, 2018.

[13] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

[14] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *Proceedings of the International Conference on Learning Representations*, 2020.

[15] MC Jones, IJ McKay, and T-C Hu. Variable location and scale kernel density estimation. *Annals of the Institute of Statistical Mathematics*, 46(3):521–535, 1994.

[16] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *Proceedings of the International Conference on Learning Representations*, 2017.

[17] Steven G. Krantz and Harold R. Parks. *Geometric integration theory*. Springer Science and Business Media, 2008.

[18] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[19] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. Technical report, AT&T Labs, 2010.

[20] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pages 396–404, 1990.

[21] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[22] Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, pages 11611–11622, 2019.

[23] Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *Advances in Neural Information Processing Systems*, volume 28, 2015.

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

[25] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015.

[26] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

[27] Roman Novak, Yasaman Bahri, Daniel A Abola a, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. Proceedings of the International Conference on Learning Representation, 2018.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 32, pages 8024–8035, 2019.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[30] Henning Petzka, Linara Adilova, Michael Kamp, and Cristian Sminchisescu. A reparameterization-invariant flatness measure for deep neural networks. Workshop on Science meets Engineering of Deep Learning at NeurIPS 2019, 2019.

[31] Akshay Rangamani, Nam H. Nguyen, Abhishek Kumar, Dzung T. Phan, Sang H. Chin, and Trac D. Tran. A scale invariant flatness measure for deep network minima. arXiv preprint arXiv:1902.02436, 2019.

[32] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. Workshop contribution at the International Conference on Learning Representation, 2018, 2019.

[33] Bernard W. Silverman. Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability. Chapman and Hall, 1986.

[34] Xu Sun, Zhiyuan Zhang, Xuancheng Ren, Ruixuan Luo, and Liangyou Li. Exploring the vulnerability of deep neural networks: A study of parameter corruption. arXiv preprint arXiv:2006.05620, 2020.

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[36] Andrey Nikolayevich Tikhonov, A. Goncharsky, V. V. Stepanov, and Anatolij Grigorevic Yagola. Numerical Methods for the Solution of Ill-Posed Problems. Mathematics and Its Applications. Springer Netherlands, 1995.

[37] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. In Proceedings of the 37th International Conference on Machine Learning, pages 9636–9647, 2020.

[38] Huan Wang, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. Identifying generalization properties in neural networks. arXiv preprint arXiv:1809.07402, 2018.

[39] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In Advances in Neural Information Processing Systems, volume 33, pages 2958–2969, 2020.

[40] Huan Xu and Shie Mannor. Robustness and generalization. Machine learning, 86(3):391–423, 2012.

[41] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. Advances in Neural Information Processing Systems, volume 32, 2019.

[42] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In Proceedings of the International Conference on Learning Representations, 2017.

[43] Chiyuan Zhang, Qianli Liao, Alexander Rakhlin, Brando Miranda, Noah Golowich, and Tomaso Poggio. Theory of deep learning iib: Optimization properties of sgd. arXiv preprint arXiv:1801.02254, 2018.

[44] Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. arXiv preprint arXiv:2010.04925, 2020.

Organization of the Appendix The appendix is organized as follows:

A – Related work contains an extended discussion on related work.

B – The effect of local label changes discusses consequences for the association of flatness and generalization for general output functions $f(x)$ without the assumption of locally constant labels.

C – Details on the Empirical Validation contains a detailed description of the experiments.

D – Proofs contains the full proofs to all statements. In detail:

- D.1: Proposition 2,
- D.2: Theorem 4,
- D.3: Theorem 5,
- D.4: Theorem 6.

E – Relative flatness for a uniform bound over general distributions on feature matrices defines a variant of relative flatness that uniformly bounds feature robustness over all feature matrices.

A Related Work

It has long been observed that algorithms searching for flat minima of the loss curve lead to better generalization [10, 11]. More recently, an association between flatness and low generalization error has also been validated empirically in deep learning [16, 27, 38]. Here, flatness is measured by the Hessian of the empirical loss evaluated at the model at hand. Indeed, in their recent extensive empirical study of generalization measures, Jiang et al. [14] found that measures based on flatness have the highest correlation with generalization.

For models trained with stochastic gradient descent (SGD), this could present a (partial) explanation for their generalization performance, since the convergence of SGD can be connected to flat local minima by studying SGD as an approximation of a stochastical differential equation [13, 43]. However, while large and small batch methods appear to converge in different basins of attraction, the basins can be connected by a path of low loss, i.e., they can actually converge into the same basin [32]. Moreover, as Dinh et al. [5] remarked, classical flatness measures—which are based only on the Hessian of the loss function—cannot theoretically be related to generalization: For deep neural networks with ReLU activation functions, there are linear reparameterizations that leave the network function unchanged (hence, also the generalization performance), but change any measure derived only from the loss Hessian. Novel measures related to flatness have been proposed that are invariant to linear reparameterizations [21, 31, 37]. Rangamani et al. [31] measure flatness in the quotient space of a suitable equivalence relation, and Liang et al. [21] utilize the Fisher-Rao metric, but the theoretical connection of these two measures to generalization is not well-understood. Neyshabur et al. [26] noted that the reparamterization-issue can in general be resolved by balancing a measure of flatness with a norm on the parameters, which is the way that normalized flatness [37], Fisher-Rao metric [21] and our proposed relative flatness become reparameterization-invariant. However, the solution proposed in Neyshabur et al. [26] necessitates data-dependent priors [7] as for related approaches, which "adds non-trivial costs to the generalization bounds" [37].

The question arises in which way the loss Hessian and parameter norm should be combined. A simple scaling of the full Hessian with the squared parameter norm does not provide a reparamterization-invariant measure. Doing so for each layer independently and summing up the results provides a measure that is only invariant under layer-wise reparameterizations. Similarly, only considering a single feature layer yields a measure that is layer-wise reparameterization invariant [30]. While the resulting measure can also be analyzed within our framework to obtain a bound on feature robustness, our proposed measure yields a tighter bound and is also invariant under neuron-wise reparameterizations.

Tsuzuku et al. [37] derive a flatness measure that scales an approximation to the loss Hessian by a parameter-dependent term. Their proposed measure correlates well with generalization and is theoretically connected to it via the PAC-Bayesian framework. However, this connection requires the assumption of Gaussian priors and posteriors and is not informative with respect to conditions under which this connection holds. Moreover the measure is impractical, since computing it requires

solving an optimization problem for every layer that can be numerically unstable. (Tsuzuku et al. [37] propose a solution to the numerical instability at the cost of losing the reparameterization-invariance.) Instead, relative flatness can be computed directly and takes only parameters of a specific layer into account—although combining relative flatness of all layers by simple summation is possible.

A series of recent papers studies flatness by minimizing the loss at local perturbations of the parameters considering $\min_a E_{emp}(f(w+a), S)$ [8, 34, 39, 44]. Regularization techniques enforcing these notions of flatness during training in classification tasks lead to better generalization. These empirical results follow earlier works by Chaudhari et al. [4] and Izmailov et al. [12] that similarly obtained better generalization by enforcing flatter minima. Their observations are well-explained by our theory: Low error at perturbations $E_{emp}(f(w+a), S)$ lead to good generalization around training samples. This requires that the underlying distribution has (approximately) locally constant labels (using key equation (1)), which is reasonable for the image classification tasks they consider.

Xu and Mannor [40] propose a notion of robustness over a partion of the input space and derive generalization bounds based on it. However, their notion requires the choice of a partitioning of the input space before seeing any samples. Thus, robustness over the partition can be hard to estimate for a model that depends on a sample set S. Our notion of feature robustness is measured around a given sample set and thus does not require a uniform data-independent partitioning. Such a sample-dependent notion of robustness is necessary to connect it to the flatness of the loss surface, since flatness is a local property around training points.

Novak et al. [27] find that robustness to input perturbation as measured by the input-output Jacobian correlates well with generalization on classification tasks. This is in line with our findings applied to $= \mathrm{id}_X$ chosen as the identity (for neural networks this means considering the input layer as features): it follows from Equation 1 that robustness to input perturbations directly relates to flatness. Therefore, these findings give additional empirical evidence to the correlation between flatness and generalization. Yao et al. [41] study the Hessian with respect to the input $x \in X$ and also find that robust learning tends to converge to minima where the input-ouptut Hessian has small eigenvalues.

B The effect of local label changes

For classification tasks with one-hot vectors as labels, the assumption of locally constant labels, i.e., locally constant target output function $y(x)$, seems reasonable since we would not expect the class label to change under (infinitesimally) small changes. One could nonetheless consider a smooth output function with values encoding class probabilities for classification, which may change locally around the training points. For regression tasks, the assumption of locally constant output function is rather unrealistic or at the very least restrictive.

Taking the term definition feature robustness (4) as a starting point, we investigate its connection to flatness when the output function $y(x)$ is a smooth function. In the usual setting of machine learning, this information is unknown. We will show that label changes can contribute stronger to the loss in neighborhoods around training samples than (relative) flatness.

To investigate the label dependence, we use the same trick as (7) to transfer perturbations in the input x to perturbations in parameter space w. To simplify the analysis, we apply feature robustness to the input space (i.e., we only consider $T = \mathrm{id}_X$ here). Let $f(x; w) = f(wx)$ be a model composed of a matrix multiplication of x with w and a differentiable predictor function f.

$$\begin{aligned} E_F(f; S; A) &= \frac{1}{n} \sum_{i=1}^{n} \left(\ell(f(x_i + Ax_i; w), y[x_i + Ax_i]) - \ell(f(x_i; w), y_i) \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\ell(f(x_i; w + wA), y[x_i + Ax_i]) - \ell(f(x_i; w), y_i) \right) \end{aligned}$$

Defining a function

$$\ell_i(\cdot) = \ell(f(x_i; w + wA), y[x_i + Ax_i]); \tag{8}$$

we have that $E_F(f; S; A) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\cdot)$. For each ℓ_i we use Taylor approximation in $\cdot$. In the following, we write $\ell_w(x_i; w; y_i)$ for the first derivative of the loss with changes in w at $x_i; y_i = y[x_i]$ and w, and we write $\ell_y(x_i; w; y_i)$ for the first derivative of the loss with changes of the output y at $x_i; y_i = y[x_i]$ and w. Similarly, we consider second derivatives ℓ_{ww}, ℓ_{yy} and ℓ_{wy}. Finally, we denote the derivative of $y(x)$ with respect to x by y_x and the second derivative by y_{xx}. Then,

$$\ell_i'(0) = \ell_w(x_i; w; y_i) \cdot (w \cdot A) + \ell_y(x_i; w; y_i) \cdot (y_x(x_i) - Ax_i) \tag{9}$$

and

$$\ell_i''(0) = (w \cdot A)^T \ell_{ww}(x_i; w; y_i)(w \cdot A) + (y_x(x_i) - Ax_i)^T \ell_{yy}(x_i; w; y_i)(y_x(x_i) - Ax_i) \tag{10}$$

$$+ \sum_{\text{labels } c} \ell_{y_c}(x_i; w; y_i) \cdot (Ax_i)^T (y_c)_{xx}(x_i) \cdot (Ax_i) + 2(y_x(x_i) - Ax_i)^T \ell_{yw}(x_i; w; y_i)(w \cdot A) \tag{11}$$

At a critical point we have that $\sum_i \ell_w(x_i; w; y_i) = 0$, but since we do not know how the target output function $y(x)$ changes locally, we do not necessarily[3] enforce that $\sum_i \ell_y(x_i; w; y_i) = 0$ at a local optimum. In that case $E_F(f; S; A) = \sum_i \ell_y(x_i; w; y_i) \cdot + O(^2)$ has a non-zero term of first order in $$ and flatness only contributes as a term of order two. Similarly, other terms in (10) can be non-zero, further reducing the influence of relative flatness to a bound on feature robustness.

As an interesting special case, we note that for one-hot encoded labels in classification and letting the output function $y(x)$ describe a parameter vector of a conditional label-distribution given x, we have $y_x(x_i) = 0$ (recall that we supposed $y(x_i) = y_i$) as each vector component is either 1 or 0 and must be a local extreme point ($y(x)$ cannot contain values larger than 1 or smaller than 0 by assumption),

We leave a detailed investigation of the consequences of label changes as future work, but identify the implicit assumption of locally constant labels in loss Hessian-based flatness measures as a possible limitation: Flatness can only be descriptive if optimal label changes are approximately locally constant. The fact that a strong correlation between flatness and generalization gap has been often observed points to the fact that distributions in practice satisfy this implicit assumption.

[3] This depends on the loss function in use.

C Details on the Empirical Validation

Here we provide additional details on the empirical evaluation. Jupyter notebooks containing the experiments are available at https://github.com/kampmichael/relativeFlatnessGeneralization, ensuring reproducibility, together with an implementation of the relative flatness measure in pytorch [28].

C.1 Synthetic Experiments

The experiments on locally constant labels and approximating representativeness use a synthetic sample in feature space. The schema for both experiments is to

1. create a synthetic dataset in feature space S and test set T,
2. create a model $f = \ldots$,
3. derive input data as $S = \ldots^{-1}(S_x); S_y$, $T = \ldots^{-1}(T_x); T_y$
4. compute relative flatness (or other measures) of f on S,
5. and estimate its generalization gap by computing the empirical risk of f on S, and computing the test error on the test set T to estimate the risk.

1) To create S with a given class separation c, we randomly sample 4 cluster centroids from a hypercube in $\mathbb{R}^6$ and scale them so that their distance is c. We then sample a random covariance matrix for each cluster and sample points from a Gaussian $N(\mu, \Sigma)$. Furthermore, we create two redundant features that are a random linear combination of the informative features. We obtain labels by assigning two clusters to class 1 and the other two to class 1.

2) We create the model f by first training a linear model on S using ridge regression from scikit-learn [29]. We then sample a random 4-layer MLP (with architecture 784-512-128-16-8, tanh activation, and Glorot initialization [9]) that we use as feature extractor $\ldots$. With this, we obtain the 5-layer MLP $f = \ldots$ by adding an 8 2 layer with weights obtained from $\ldots$.

3) We obtain input data S by reverse propagation of samples in feature space S_x through the 4-layer MLP $\ldots$. This is an approximation to the inverse feature extractor $\ldots^{-1}$. For the output of each layer z, we first compute $z' = \tanh^{-1}(z)$, i.e., the inverse of the activation function. We then solve $Wz + b = x$, where W, b are the weights and bias of that layer, and x is the corresponding input we want to compute. This yields $S_x = \ldots^{-1}(S_x)$. Note that this reverse propagation of samples introduces a small error. To keep experiments realistic, we discard $\ldots$ after this step and use only the input dataset S and model f in our computations.

4) We compute relative flatness as in Def. 3 (an implementation in pytorch is available on github, see above).

5) We compute the empirical risk of f on S and estimate the risk on a test set. For the experiments on locally constant labels, generate 5000 samples, use a training set of size 500, a test set of size 4500 (to ensure an accurate estimate of the risk), and repeat the experiment 100 times for each class separation c. For the experiment on approximating representativeness, we use a sample of size 600 and perform 3-fold cross-validation.

Locally constant labels: For classification, labels are locally constant if in a neighborhood around each point the label does not change. They are approximately locally constant, if this holds for most points. By increasing the distance between the means of the Gaussians, we decrease the likelihood of a point within a neighborhood having a different label. For a finite sample, this means that the likelihood of observing two points close by with different labels decreases. Thus, by increasing the class separation parameter, we increase the degree of locally constant labels.

Approximating representativeness: A finite random sample as described in 1) has a higher chance of being representative when the means of the Gaussians have a high distance, because each individual Gaussian can be interpolated easily. Of course, the actual representativeness of a sample at hand can vary. Note that this is a very simple form of generating datasets with varying "difficulty". It will be interesting to further explore the impact of the choice of data distribution on (an approximation to) representativeness.

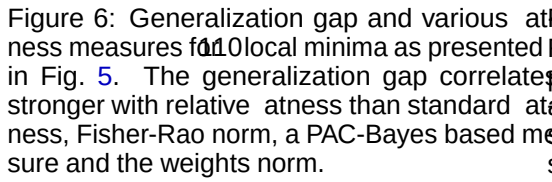

Figure 6: Generalization gap and various atness measures for 1100 local minima as presented in Fig. 5. The generalization gap correlates stronger with relative atness than standard atness, Fisher-Rao norm, a PAC-Bayes based measure and the weights norm.

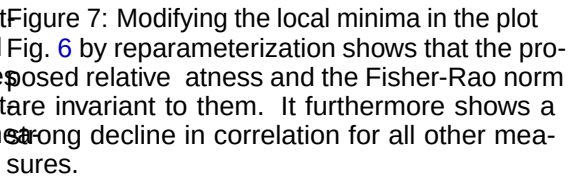

Figure 7: Modifying the local minima in the plot Fig. 6 by reparameterization shows that the proposed relative atness and the Fisher-Rao norm are invariant to them. It furthermore shows a strong decline in correlation for all other measures.

Experiments on the synthetic datasets are run on a laptop with Intel Core i7 and NVIDIA GeForce GTX 965 M 2 GB GPU. The code of the experiments is provided as a jupyter notebook so that they can be easily reproduced.

C.2 Relative Flatness Correlates with Generalization

In this experiment, we validate that the proposed relative atness correlates strongly with generalization in practice. For that, we measure relative atness (as well as classical atness measured by the trace of the loss Hessian, the Fisher-Rao norm, a PAC-Bayes based measure[4], and the weight norm) together with the generalization gap for various local minima.

To obtain model parameters at various local minima, we train networks (LeNet5 2[0]) on the CIFAR10 dataset until convergence (measured in terms of achieving a loss of less than 0:1 during an epoch, which has been used as a criteria for convergence in similar experiments 1[4]) with varying hyperparameters. In accordance to works studying the impact of hyperparameters on generalization [14, 16, 27, 31, 38], we vary learning rate, mini-batch size, initialization, and optimizer. We vary the mini batch size in 64; 128; 256; 512; 1024, and the learning rate in 0:0001; 0:02; 0:05, running 10 randomly initialized training rounds for each setup. We use SGD, ADAM, and RMSProp as optimizers. We only use combinations that lead to convergence. The experiments were conducted on a cluster node with 4 NVIDIA GPU GM200 (GeForce GTX TITAN X). As discussed in Sec. 6, relative atness has the highest correlation with generalization from all measures we analyzed.

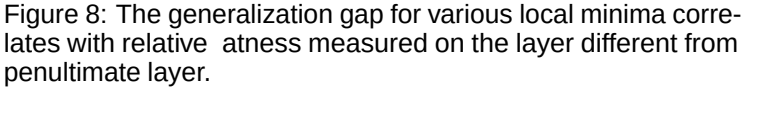

Figure 8: The generalization gap for various local minima correlates with relative atness measured on the layer different from penultimate layer.

[4]The implementation of PAC-Bayes based atness measure is taken from https://github.com/nitarshan/robust-generalization-measures/blob/master/data/generation/measures.py

Figure 9: Layer1-based relative atness for MNIST experiment. Layer1 is the topmost layer.

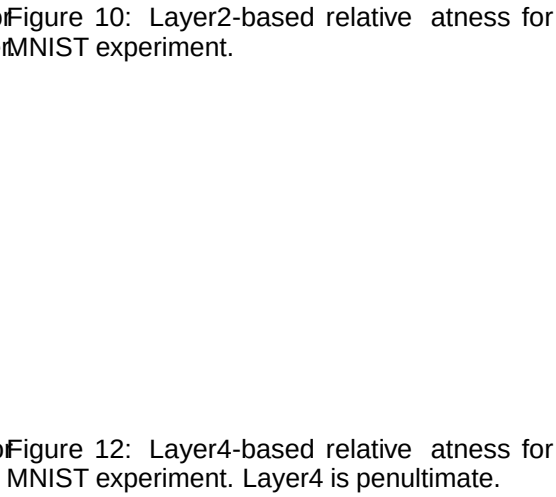

Figure 10: Layer2-based relative atness for MNIST experiment.

Figure 11: Layer3-based relative atness for MNIST experiment.

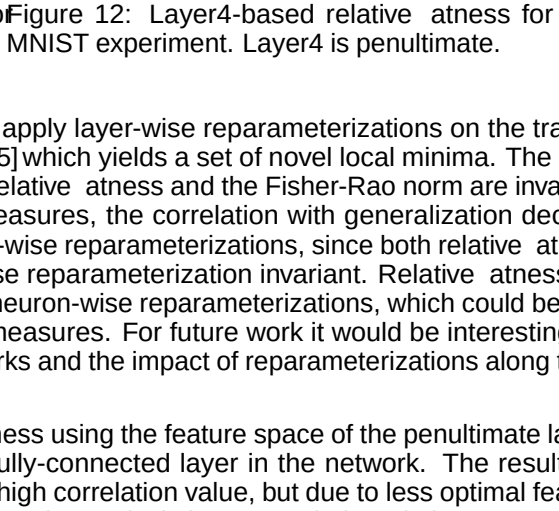

Figure 12: Layer4-based relative atness for MNIST experiment. Layer4 is penultimate.

To study the effect of reparameterization, we apply layer-wise reparameterizations on the trained network using random factors in the interval $[a, \frac{1}{a}]$ which yields a set of novel local minima. The results in Fig. 7 show that both our proposed relative atness and the Fisher-Rao norm are invariant to these reparameterization. For all other measures, the correlation with generalization declines substantially. The same would hold for neuron-wise reparameterizations, since both relative atness and the Fisher-Rao norm are also neuron-wise reparameterization invariant. Relative atness and the Fisher-Rao norm are also invariant under neuron-wise reparameterizations, which could be used to further break the correlation for the other measures. For future work it would be interesting to investigate further symmetries in neural networks and the impact of reparameterizations along these symmetries on atness measures.

In addition to the calculation of the relative atness using the feature space of the penultimate layer, we also performed calculations for another fully-connected layer in the network. The resulting correlation can be seen in Fig. 8. It keeps the high correlation value, but due to less optimal feature space we observe smaller number, than in the previous calculation. Nevertheless, it demonstrates that any - separation allows to compute relative atness.

For checking deeper the viability of the relative atness computed in different feature representations, we ran a similar experiment with a fully-connected $(784, 50, 50, 50, 30, 10)$ neural network trained on MNIST dataset (Fig. 9, 10, 11, 12). We varied parameters of the optimization (batch size in $1000, 2000, 4000, 8000$ and learning rate in $0.02, 0.04, 0.08, 0.16$ in order to keep the ratio between them constant) and trained each network with SGD for 500 epochs. Only the networks that achieved training loss lower than 0.07 are used for the plots. The observed correlation with generalization gap is high for each of the representations in the network.

D Proofs

D.1 Proof of Proposition 2

Proof. Let K_h denote probability distribution defined by a rotational-invariant kernel k_h as in (6) with $k_h(0,z) = \frac{1}{h^m}\, k\left(\frac{\|z\|}{h}\right) 1_{\|z\|<h}$ and let $\ell_i(z) = k_{\|\cdot - (x_i)\|}(0,z)$. Let L denote a continuous function on $\mathbb{R}^m$ and O_m the set of orthogonal matrices in $\mathbb{R}^{m\times m}$. We show that there exists a probability measure on a set M of matrices of norm smaller than, defining a probability distribution A, and a probability measure on the product space $(0,\ ] \times O_m$ such that for each $z \in \mathbb{R}^m \setminus \{0\}$:

$$E_{A\sim\ }\left[L(z+Az)\right] = E_{(r,O)\sim\ }\left[L(z+rOz)\right] = E_{K_{\|z\|}}\left[L(z+\ )\right] \tag{12}$$

Applying this result for each $h = 1,\dots,|S|$ to $L_i(z) = \ell(\ (w,z), y_i[z])$ at $z = \ (x_i)$ completes the proof. For all the standard measure-theoretic concepts used in the proof, we refer the reader to [17].

Fix some $\ _0$ in $\mathbb{R}^m$ with $\|\ _0\| = 1$. We consider the Haar measure on the set of orthogonal matrices O_m. By [17, Proposition 3.2.1] and the change of variables formula, we have for each $r \in (0,\ ]$

$$\int_{O\in O_m} L(z + r\|z\| O\ _0)\, d\ (O) = \frac{1}{\mathrm{Vol}(S^{m-1})} \int_{\in S^{m-1}} L(z + r\|z\|\ )\, d$$

where S^{m-1} is the $(m-1)$-sphere. We multiply both sides by $\frac{\mathrm{Vol}(S^{m-1})}{m} k\left(\frac{r}{-}\right) r^{m-1}$, integrate over $r \in (0,\ ]$ to obtain

$$\begin{aligned}
&\frac{\mathrm{Vol}(S^{m-1})}{m} \int_{r=0}^{} \int_{O\in O_m} L(z + r\|z\| O\ _0) k\left(\frac{r}{-}\right) r^{m-1} dr\, d\ (O) \\
&= \frac{1}{m} \int_{r=0}^{} \int_{\in S^{m-1}} L(z + r\|z\|\ ) k\left(\frac{r}{-}\right) r^{m-1}\, dr\, d \\
&= \frac{1}{m} \int_{\|\ \|\|z\|}^{} L(z + \|z\|\ ) k\left(\frac{\|\ \|}{-}\right) d \\
&= \int_{\|\ \|\|z\|}^{} L(z + \ ) \frac{1}{(\|z\|)^m} k\left(\frac{\|\ \|}{\|z\|}\right) d
\end{aligned}$$

Introducing the product measure $:= \frac{\mathrm{Vol}(S^{m-1})}{m} (k\left(\frac{r}{-}\right) r^{m-1} dr\ )$ on $(0,\ ] \times O_n$, this implies that

$$E_{(r,O)\sim\ }\left[L(z + r\|z\| O\ _0)\right] = E_{K_{\|z\|}}\left[L(z+\ )\right] \tag{13}$$

The measure can be pushed forward to a measure on matrices of norm $\|A\|$. For this, consider the homeomorphism

$$H : (0,\ ] \times O_n \to \{\ rO \mid r \in (0,\ ], O \in O_n\} =: M \subset \{\ A \in \mathbb{R}^{n\times n} \mid \|A\|\ \}$$

given by $H(r,O) = rO$. We use the inverse of H to push forward the measure to a measure on M and obtain from (13) that

$$E_{A\sim(M,\ )}\left[L(z + \|z\| A\ _0)\right] = E_{K_{\|z\|}}\left[L(z+\ )\right]$$

Finally, there exists an orthogonal matrix O such that $O\|z\|\ _0 = z$. Since $(A) = (AO^{-1})$ by definition of and since $MO = M$, we get for any z that

$$\begin{aligned}
E_{K_{\|z\|}}\left[L(z+\ )\right] &= E_{A\sim(M,\ )}\left[L(z + A\|z\|\ _0)\right] \\
&= E_{A\sim(MO^{-1},\ )}\left[L(z + AO\|z\|\ _0)\right] \\
&= E_{A\sim(M,\ )}\left[L(z + Az)\right]
\end{aligned}$$

Hence, the probability distribution A on matrices with norm bounded by defined by the probability measure with support on M, and the space $(0,\ ] \times O_m$ equipped with $! = \frac{\mathrm{Vol}(S^{m-1})}{m} (k\left(\frac{r}{-}\right) r^{m-1} dr\ )$ give the desired probability distributions satisfying (12). □

D.2 Proof of Theorem 4

We rephrase Theorem 4 split into Theorem 7 and a subsequent corollary that specify the reparameterizations under consideration. Let $f = f(w^1, b^1, w^2, b^2, \dots, w^L, b^L)$ denote a ReLU network function parameterized by parameters $w^k = w^k_{s,t}$ and biases $b^k = b^k_s$ of the k-th layer given by

$$f(x) = w^L\,(\dots\,(w^l\,(w^{l-1}\,(\dots\,(w^1 x + b^1))\dots) + b^{l-1})\dots) + b^L.$$

Recall that we let $\varphi^l(x)$ denote the composition of the first $(l-1)$ layers so that we obtain a decomposition $f(x; w^l) = g^l(w^l\,\varphi^l(x))$ of the network. Using (9) we obtain a relative flatness measure $\mathrm{Tr}_l(w)$ for the chosen layer.

A layer-wise reparameterizaton multiplies all weights in a layer k with a positive number λ and divides the weights of another layer $k' \neq l$ by the same λ. Due to the positive homogeneity of the ReLU activation, this does not change the network function. By a neuron-wise reparameterization, we mean the operation that multiplies all weights into a neuron by some positive λ and divides all outgoing weights of the same neuron by λ. Again, the positive homogeneity of the activation function implies that this operation does not change the network function. A layer-wise reparameterization is simply the parallel application of neuron-wise reparameterization for all neurons of one layer with the same reparameterization parameter $\lambda > 0$.

Theorem 7. Let $f = f(w^1, b^1, w^2, b^2, \dots, w^L, b^L)$ denote a neural network function parameterized by parameters $w^k = w^k_{s,t}$ and biases $b^k = b^k_s$ of the k-th layer. Suppose there are positive numbers $\lambda^k_{s,t}$ such that the parameters $\tilde{w}^k, \tilde{b}^k$, obtained from multiplying $w^k_{s,t}$ at matrix position (s,t) in layer k by $\lambda^k_{s,t}$ and b^k_s by $\lambda^k_{(s;0)}$, satisfy that $f(w^1, b^1, w^2, b^2, \dots, w^L, b^L) = f(\tilde{w}^1, \tilde{b}^1, \tilde{w}^2, \tilde{b}^2, \dots, \tilde{w}^L, \tilde{b}^L)$. If for the layer with index l it holds that $\lambda^l_{(s,t)} = \lambda^l_{(s,t')}$ for each s, t and t', then $\mathrm{Tr}_l(w) = \mathrm{Tr}_l(\tilde{w})$ for the notion of relative flatness from Definition 3.

Corollary 8. Let σ_i denote the variance of the i-th coordinate of $\varphi^l(x)$ over samples $x \in S$ and $V = \mathrm{diag}(\sigma_1, \dots, \sigma_{n_{l-1}})$. If the relative flatness measure Tr_l is applied to the representation $f = f(w^1, b^1, \dots, V^{-1} w^l V^{-1}, V^{-1} b^l, w^{l+1} V, b^{l+1}, \dots, w^L, b^L)$, i.e.,

$$f(x) = w^L\,(\dots\,(w^l V\,(V^{-1} w^{l-1}\,(\dots\,(w^1 x + b^1))\dots) + V^{-1} b^{l-1}) + b^l)\dots) + b^L,$$

then Tr_l is invariant under all neuron-wise (and layer-wise) reparameterizations

Proof. We are given a neural network function $f(x; w^1, b^1, \dots, w^L, b^L)$ parameterized by parameters w^k and bias terms b^k of the k-th layer and positive numbers $\lambda^1_{(s,t)}, \dots, \lambda^L_{(s,t)}$ such that the parameters $\tilde{w}^k$ obtained from multiplying weight $w^k_{(s,t)}$ at matrix position (s,t) in layer k by $\lambda^k_{(s,t)}$ and b^k_s by $\lambda^k_{(s;0)}$ satisfies that

$$f(x; w^1, b^1, w^2, b^2, \dots, w^L, b^L) = f(x; \tilde{w}^1, \tilde{b}^1, \tilde{w}^2, \tilde{b}^2 \dots, \tilde{w}^L, \tilde{b}^L)$$

for all w^k, b^k and all x.

For fixed layer l, we denote the s-th row of w^l by w^l_s before reparameterization, and we denote the s-th row of w^l by $\tilde{w}^l_s$ after reparameterization. For simplicity of the notation, we will collect all bias terms in terms $b, \tilde{b}$ before and after reparameterization respectively. Let

$$F(u) := \sum_{i=1}^{|S|} \ell(f(x_i; w^1, w^2, \dots, [w^l_1, \dots, w^l_{s-1}, u, w^l_{s+1}, \dots w^l_d], \dots, w^L, b), y_i)$$

denote the loss as a function on the parameters of the s-th neuron in the l-th layer (encoded in the s-th row of w^l) before reparameterization and

$$\tilde{F}(u) := \sum_{i=1}^{|S|} \ell(f(x_i; \tilde{w}^1, \tilde{w}^2, \dots, [\tilde{w}^l_1, \dots, \tilde{w}^l_{(s-1)}, u, \tilde{w}^l_{(s+1)}, \dots \tilde{w}^l_d], \dots, \tilde{w}^L, \tilde{b}), y_i)$$

denote the loss as a function on the parameters into the s-th neuron in the l-th layer (encoded in the s-th row of w^l) after reparameterization.

For the same layer l, we define a linear function $\rho_s : \mathbb{R}^m \to \mathbb{R}^m$ by

$$\rho_s(u) = \rho_s(u_1, u_2, \ldots, u_m) = (u_1 \alpha_{(s,1)}, u_2 \alpha_{(s,2)}, \ldots, u_m \alpha_{(s,m)}).$$

By assumption, we have that $\tilde{F}(\rho_s(w_s^l)) = F(w_s^l)$ for all w_s^l. By the chain rule, we compute for any coordinate u_t of u,

$$\frac{\partial F(u)}{\partial u_t}\Big|_{u=w_s^l} = \frac{\partial \tilde{F}(\rho_s(u))}{\partial u_t}\Big|_{u=w_s^l}$$
$$= \sum_k \frac{\partial \tilde{F}(\rho_s(u))}{\partial \rho_s(u)_k}\Big|_{\rho_s(u)=\rho_s(w_s^l)} \frac{\partial \rho_s(u)_k}{\partial u_t}\Big|_{\rho_s(u)=\rho_s(w_s^l)}$$
$$= \frac{\partial \tilde{F}(v)}{\partial v_t}\Big|_{v=w_s^l} \alpha_{(s,t)}^l.$$

Similarly, for

$$G(u, u') := \sum_{i=1}^{|S|} \ell(f(x_i; w^1; w^2; \ldots; [w_1^l; \ldots; w_{s-1}^l; u; w_{s+1}^l; \ldots; w_{s'-1}^l; u'; w_{s'+1}^l; \ldots w_d^l]; \ldots$$
$$\ldots; w^L;; b); y_i)$$

denoting the loss as a function on the parameters of the s-th and s'-th neuron in the l-th layer (encoded in the s-th and s'-th row of w^l) before reparameterization and for

$$\tilde{G}(u, u') := \sum_{i=1}^{|S|} \ell(f(x_i; w^1; w^2; \ldots$$
$$\ldots; [w^l{}_1; \ldots; w^l{}_{(s-1)}; u; w^l{}_{(s+1)}; \ldots; w_{s'-1}^l; u'; w_{s'+1}^l; \ldots w_d^l]; \ldots$$
$$\ldots w^L;; b); y_i)$$

we have $\tilde{G}(\rho_s(w_s^l), \rho_{s'}(w_{s'}^l)) = G(w_s^l, w_{s'}^l)$. For all s, s', t, t' we obtain second derivatives

$$\frac{\partial^2 G(u,u')}{\partial u_t \partial u'_{t'}}\Big|_{u=w_s^l; u'=w_{s'}^l} = \alpha_{(s,t)}^l \alpha_{(s',t')}^l \frac{\partial^2 \tilde{G}(u,u')}{\partial u_t \partial u'_{t'}}\Big|_{u=w_s^l; u'=w_{s'}^l}.$$

Consequently, the Hessian $H(w^l; S)$ of the empirical risk before reparameterization and the Hessian $\tilde{H}(w^l; S)$ after reparameterization satisfy at the position corresponding to $w_{s,t}$ and $w_{s',t'}$ that

$$H_{s,s'}(w^l; S)_{(t,t')} = \alpha_{(s,t)}^l \alpha_{(s',t')}^l \tilde{H}_{s,s'}(w^l)_{(t,t')}.$$

Assuming that $\alpha_s^l := \alpha_{(s,t)}^l = \alpha_{(s,t')}^l$ for all s, t and t', then we get that

$$\tilde{l}_{Tr}(w) = \sum_{s,s'=1}^{d} \langle w_s^l, w_{s'}^l \rangle \, Tr(H_{s,s'}(w^l; S))$$
$$= \sum_{s,s'=1}^{d} \langle \frac{w_s^l}{\alpha_s^l}, \frac{w_{s'}^l}{\alpha_{s'}^l} \rangle \, Tr(\alpha_s^l \alpha_{s'}^l \tilde{H}_{s,s'}(w^l; S))$$
$$= \sum_{s,s'=1}^{d} \langle w_s^l, w_{s'}^l \rangle \, Tr(\tilde{H}_{s,s'}(w^l; S))$$
$$= \tilde{l}_{Tr}(w^l)$$

This proves Theorem 7.

To show the corollary, we first observe that all layer-wise reparameterizations are covered by the theorem. To see this, we only need to check that the condition $\alpha_{(s,t)}^l = \alpha_{(s,t')}^l$ holds for each s, t and t'. For layer-wise reparameterizations, we even have that $\alpha_{(s,t)}^l = \alpha^l$ for all s, t, since all weights of one layer are multiplied by the same scalar, and $\alpha_{(s,t)}^l = \alpha_{(s,t')}^l$ is easily seen to hold true.

Note further, that any neuron-wise reparameterization given by multiplying all weights into a neuron in a layer $\ell - 1$ by $\alpha > 0$ and dividing all outgoing weights by α is also covered by the theorem. Hence, the only neuron-wise reparameterization that can change the relative flatness measures is the one multiplying some row of $W^{\ell-1}$ by some $\alpha > 0$ and dividing the corresponding column of W^{ℓ} by the same α. However, by multiplying both $W^{\ell-1}$ and W^{ℓ} with V^{-1} and V from the left and right respectively, we perform an explicit neuron-wise reparameterization that chooses a unique representative and therefore removes the dependence on such reparamerizations. □

D.3 Proof of Theorem 5

In this section, we prove Theorem 5. For clarity, we repeat the assumptions and the statement we prove in this section:

We consider a model $f(x;w) = g(w\ (x))$, a loss function ℓ and a sample set S, and let $O_m \subset R^{m\ m}$ denote the set of orthogonal matrices. Let $\ $ be a positive (small) real number and $w = ! \ 2\ R^d$ denote parameters at a local minimum of the empirical risk on a sample set S. If the output function satisfies that $y[\ _A(x_i)] = y[\ (x_i)] = y_i$ for all $(x_i; y_i) \ 2\ S$ and all matrices $jjAjj \ \ 1$, then we want to show that $f(x;!)$ is $((\ ;S;O_m);\ )$-feature robust on average over O_m for $\ = \frac{^2}{2m}\ Tr(!) + O(\ ^3)$, i.e.,

$$E_F(f;S;\ \ A) \ \ \frac{^2}{2m}\ Tr(!) + O(\ ^3) \text{ for all } 0$$

Proof. Writing $z_i = \ (x_i)$ and $E_{emp}(w;S) = E_{emp}(f(w;x);S)$ and using the assumption that $y[\ _A(x_i)] = y_i$ for all $(x_i; y_i) \ 2\ S$ and all $jjAjj \ \ 1$, we have for any 0 ,

$$\begin{aligned} E_F(f;S;\ A) + E_{emp}(w;S) &= \frac{1}{jSj}\sum_{i=1}^{jSj} `(\ [w;\ _A(x_i)];\ y[\ _A(x_i)]) \\ &= \frac{1}{jSj}\sum_{i=1}^{jSj} `(\ (w; z_i + \ Az_i); y_i) \\ &= \frac{1}{jSj}\sum_{i=1}^{jSj} `(\ (w + \ \ wA; z_i); y_i) \\ &= E_{emp}(w + \ \ wA;S) \end{aligned} \tag{14}$$

The latter is the empirical error $E_{emp}(w + \ \ wA;S)$ of the model f on the sample set S at parameters $w + \ \ wA$. If $\ $ is sufficiently small, then by Taylor expansion around the local minimum $w = !$, we have up to order of $O(\ ^3)$ that

$$\begin{aligned} E_{emp}(! \ + \ !A;S\ ) &= E_{emp}(!;S\ ) + \frac{^2}{2}\sum_{s;t=1}^{d} (!_s A)\ \ H_{s;t}(!;\ \ (S))\ \ (!_t A)^T \\ &\ \ E_{emp}(!;S\ ) + \frac{^2}{2}\sum_{s;t=1}^{d} (!_s A)\ \ H_{s;t}(!;\ \ (S))\ \ (!_t A)^T \end{aligned} \tag{15}$$

where $!_s$ denotes the s-th row of $!$.

We consider the set of orthogonal matrices O_m as equipped with the (unique) normalized Haar measure. (For the definition of the Haar measure, see e.g. [17].) We need to show that $E_{A\ O_m}\ E_F(f;S;\ A\ ) \ \ \frac{^2}{2m}\sum_{s;t} hw_s; w_t i\ \ Tr(H_{s;t})$ for all 0 with $E_F(f;S;A)$ defined as in Eq. 4. Using (14) and (15) we get

$$E_{A\ O_m}\ \Big[\ E_F(f;S;\ A\ )\Big] \ \ E_{A\ O_m}\ \Big[\frac{^2}{2}\sum_{s;t=1}^{d} (!_s A) H_{s;t}(!;S\ )(!_t A)^T\Big] + O(\ ^3)$$

Using the unnormalized trace $Tr([m_{s;t}]) = \sum_s m_{s;s}$ we compute with the help of the so-called Hutchinson's trick:

$$\begin{aligned} Tr(E_{A\ O_m}\ (!_t A)^T(!_s A)\ ) &= E_{A\ O_m}\ Tr((!_t A)^T(!_s A)\ ] \\ &= E_{A\ O_m}\ Tr((!_s A)(!_t A)^T\ ] \\ &= E_{A\ O_m}\ Tr(!_s !_t^T\ ] \\ &= h!_s; !_t i \end{aligned}$$

We can interchange two vector coordinates by multiplication of a suitable orthogonal matrix. Since the Haar measure is invariant under multiplication of an orthogonal matrix, the diagonal of $E_{A \sim O_m}[(!_t A)^T (!_s A)]$ must contain a constant value. This value along the diagonal must then equal $\frac{1}{m}\langle !_s, !_t \rangle$. Further, we can multiply one vector coordinate by -1 via multiplication by an orthogonal matrix, and hence the off-diagonal entries of $E_{A \sim O_m}[(!_t A)^T (!_s A)]$ must be zero, giving that

$$E_{A \sim O_m}[(!_t A)^T (!_s A)] = \frac{\langle !_s, !_t \rangle}{m} I.$$

Therefore

$$\begin{aligned}
E_{A \sim O_m}[(!_s A) H_{s,t} (!_t A)^T] &= Tr\left(E_{A \sim O_m}[(!_s A) H_{s,t} (!_t A)^T]\right) \\
&= E_{A \sim O_m}[Tr((!_s A) H_{s,t} (!_t A)^T)] \\
&= E_{A \sim O_m}[Tr(H_{s,t} (!_t A)^T (!_s A))] \\
&= Tr(H_{s,t}\, E_{A \sim O_m}[(!_t A)^T (!_s A)]) \\
&= Tr\left(H_{s,t}\, \frac{\langle !_s, !_t \rangle}{m} I\right) \\
&= \frac{\langle !_s, !_t \rangle}{m} Tr(H_{s,t})
\end{aligned}$$

Putting things together, we have for the local optimum $w = !$ that

$$\begin{aligned}
E_{A \sim O_m}\left[E_F(f, S, A)\right] &\leq \frac{\varepsilon^2}{2} \sum_{s,t=1}^{d} E_{A \sim O_m}\left[(!_s A) H_{s,t} (!_t A)^T\right] + O(\varepsilon^3) \\
&= \frac{\varepsilon^2}{2m} \sum_{s,t=1}^{d} \langle !_s, !_t \rangle\, Tr(H_{s,t}) + O(\varepsilon^3) \\
&= \frac{\varepsilon^2}{2m}\, Tr(!) + O(\varepsilon^3)
\end{aligned}$$

□

We can further generalize Theorem 5 to more complex labels by introducing a notion of approximately locally constant labels. The following de nition frees us from the strong assumption of locally constant labels, i.e. $y[\ _A(x_i)] = y[\ (x_i)] = y_i$ for all $(x_i, y_i) \in S$ and all matrices $\|A\| \leq 1$, while still restricting label changes to be one order smaller than the contribution of atness.

De nition 9. Let D be a data distribution on a labeled sample space $X \times Y$ and S a nite iid sample of D. Let $f = \ $ be a model composed into a feature extractor and predictor . We say that D has approximately locally constant labels of order three around the points $(x, y) \in S$ in feature space , if there is some constant C such that

$$\frac{1}{|S|} \sum_{i=1}^{|S|} \ell(\ (\ (x_i) + \ _i), y[\ (x_i) + \ _i]) - \ell(\ (\ (x_i) + \ _i), y_i) \leq C \varepsilon^3 \text{ for } \|\ _i\| \leq \|\ (x_i)\|$$

Corollary 10. Consider a model $f(x; w) = (w; (x)) = g(w\ (x))$ as above, a loss function and a sample set S, and let $O_m \subset R^{m \times m}$ denote the set of orthogonal matrices. Let be a positive (small) real number and $w = ! \in R^{d \times m}$ denote parameters at a local minimum of the empirical risk on a sample set S. If D has approximately locally constant labels of order three around the points $(x, y) \in S$ in feature space, then $f(x; !)$ is $(\ (\ ; S, O_m), \ )$-feature robust on average over O_m for $= \frac{\varepsilon^2}{2m}\, Tr(!) + O(\varepsilon^3)$.

Proof. As before, we abbreviate (x_i) by z_i. We only need to modify (14) to account for the strictly weaker assumption on the labels. For this, we perform Taylor approximation with respect to the

labels at $y[\ (x_i)] = y_i$ to obtain

$$
\begin{aligned}
E_F(f;S;\ A\ ) + E_{emp}(w;S) &= \frac{1}{jSj}\sum_{i=1}^{jSj} `(\ [\ _A\ (x_i)];\ y[\ _A\ (x_i)]) \\
&\overset{\text{Def 9}}{=} \frac{1}{jSj}\sum_{i=1}^{jSj} `(\ (w;z_i + \ Az\ _i);y_i) + O(\ ^3) \\
&= \frac{1}{jSj}\sum_{i=1}^{jSj} `(\ (w + \ \ wA;z_i);y_i) + O(\ ^3) \\
&= E_{emp}(w + \ \ wA;S) + O(\ ^3)
\end{aligned}
$$

The rest of the proof follows the arguments used to show Theorem 5.

□

D.4 Proof of Theorem 6

To prove Theorem 6, we will require a proposition that bounds ϵ-representativeness for the local densities from Proposition 2. This is achieved in Proposition 12 below uniformly over all distributions D that satisfy mild regularity assumptions necessary for a well-defined kernel density estimation. We first compose the proof to Theorem 6 and subsequently show the arguments leading to the required proposition.

The main idea to prove Theorem 6 is that the family of distributions considered in Proposition 2 (a) provides an explicit link between ϵ-representativeness and feature robustness (Proposition 2 and Equation 5), (b) allows us to approximately bound feature robustness by relative flatness (Theorem 5), and (c) allows us to apply a kernel density estimation to uniformly bound ϵ-representativeness (Proposition 12).

Theorem 6 is the informal counterpart to the following version.

Theorem 11. Consider a model $f(x;w) = g(w\phi(x))$, a loss function ℓ, a sample set S, and let m denote the dimension of the feature space defined by ϕ and let δ be a positive (small) real number. Let $\omega \in \mathbb{R}^{d\times m}$ denote a local minimum of the empirical risk on an iid sample set S.

Suppose that the distribution D has a smooth density p_D on the feature space $\mathbb{R}^m$ such that $\int_z \nabla^2\left(p_D(z)\|z\|^2\right) dz$ and $\int_z \frac{p_D(z)}{\|z\|^m} dz$ are well-defined and finite. Then for sufficiently large sample size $|S|$, if the distribution has approximately locally constant labels of order three (see Definition 9), then it holds with probability $1-\delta$ over sample sets S that

$$E_{gen}(f(\cdot;\omega);S) \lesssim |S|^{-\frac{2}{4+m}}\left(\frac{\kappa^{\phi}_{Tr}(\omega)}{2m} + C_1(p_D;L) + \frac{C_2(p_D;L)}{\sqrt{\delta}}\right)$$

up to higher orders in $|S|^{-1}$ for constants C_1, C_2 that depend only on the distribution in feature space p_D induced by ϕ and the chosen $|S|$-tuple Λ_δ as in Proposition 2 and the maximal loss L.

Proof. The proof combines Equation 5 with Proposition 12 and Theorem 5. At first we use Equation 5 to split the generalization gap into $E_{gen}(f) = E_{Rep}(f;S;\Lambda_A) + E_F(f;S;A)$. For the family of distributions Λ_δ from Proposition 2, we have by Proposition 12 that

$$|E_{Rep}(f;S;\Lambda_\delta)| \le \left(C_1(p_D;L) + \frac{C_2(p_D;L)}{\sqrt{\delta}}\right)|S|^{-\frac{2}{4+m}} + O(|S|^{-\frac{3}{4+m}})$$

when $\delta = |S|^{-\frac{1}{4+m}}$. With $E_F(f;S;A) = E_{A\sim\mathcal{A}}\left[E_F(f;S;A)\right]$, we use (the proof to) Proposition 2 to see that this can be written as

$$E_{A\sim\mathcal{A}}\left[E_F(f;S;A)\right] = E_{0}\left[E_{A\sim\mu_{(O_m;\,)}}\left[E_F(f;S;A)\right]\right]$$

where $O_m \subseteq \mathbb{R}^{m\times m}$ denote the set of orthogonal matrices and μ the Haar measure on this set. Finally, Theorem 5 bounds the latter by $\frac{|S|^{-\frac{2}{4+m}}}{2m}\kappa^{\phi}_{Tr}(\omega)$ up to higher orders in $|S|^{-1}$. □

We finally prove that the bound on ϵ-representativeness in the proof to the preceding Theorem indeed holds true.

Proposition 12. Consider a model $f(x;w) = \psi(w,\phi(x))$, a loss function ℓ and let $S \subseteq X\times Y$ be a finite sample set. With $x_i \in S$, let $\lambda_i(z) = k_{\|\phi(x_i)\|}(0;z)$ define an $|S|$-tuple Λ of densities as in Proposition 2 and assume that the loss function is bounded by L. Suppose that the distribution D has a smooth density p_D on a feature space $\mathbb{R}^m$ such that $\int_z \nabla^2\left(p_D(z)\|z\|^2\right) dz$ and $\int_z \frac{p_D(z)}{\|z\|^m} dz$ are well-defined and finite. Then there exist constants $C_1(p_D;L), C_2(p_D;L)$ depending on the distribution and the maximal loss such that, with probability $1-\delta$ over possible sample sets S, Λ_δ-interpolation is bounded for $\delta = |S|^{-\frac{1}{4+m}}$ by

$$|E_{Rep}(f;S;\Lambda_\delta)| \le \left(C_1(p_D;L) + \frac{C_2(p_D;L)}{\sqrt{\delta}}\right)|S|^{-\frac{2}{4+m}} + O(|S|^{-\frac{3}{4+m}})$$

Proof. We let

$$\hat{\rho}(z) = \frac{1}{|S|} \sum_{i=1}^{|S|} k_{\| \cdot - x_i \| \| \cdot - x_i \|}(x_i, z)$$

With $\alpha_i = k_{\| \cdot - x_i \| \| \cdot - x_i \|}(0, z)$ we have

$$E_{Rep}(f; S; \quad) = E(f) \quad \frac{1}{|S|} \sum_{i=1}^{|S|} E_{\quad i}[\ell(\quad(\quad(x_i) + \quad); y[\quad(x_i) + \quad]]$$

$$= \int_z p_D(z) \quad \ell(\quad(z); y(z))\, dz \quad \frac{1}{|S|} \sum_{i=1}^{|S|} \int_z k_{\| \cdot - x_i \| \| \cdot - x_i \|}(x_i, z) \quad \ell(\quad(z); y(z))\, dz \qquad (16)$$

$$\int_z (p_D(z) \quad E_S[\hat{\rho}(z)])\ \ell(\quad(z); y(z))\, dz}_{(I)} + \underbrace{\int_z (E_S[\hat{\rho}(z)] \quad \hat{\rho}(z))\ \ell(\quad(z); y(z))\, dz}_{(II)}$$

For the further analysis, we make use of Jones et al. [15] and combine it with the generalization to the multivariate case in Chp. 4.3.1 in Silverman [33]. A Taylor approximation with respect to the bandwidth of the kernel yields

$$(I) = \frac{\quad^2}{2}\ _2 \int_z r^2 \quad p_D(z) \| z \|^2 \quad \ell(\quad(z); y(z))\, dz + O(\quad^3)$$

where

$$_2 = \int_z \| z \|^2 k_1(0, z)\, dz.$$

For (II) we consider the random variable $Z = \int_z \hat{\rho}(z) \ell(\quad(z); y(z))\, dz$ as a function on the set of possible sample sets of a fixed size. Applying Chebychef's inequality to Z we get that

$$Pr\left[\ Z \quad E_S[Z] > \quad_{est}\right] \quad \frac{Var(Z)}{\quad^2_{est}} =: \quad :$$

Solving for $\quad_{est}$ yields that with probability $1 \quad$ we have

$$(II) = |Z \quad E_S[Z]| \quad \sqrt{\frac{Var(Z)}{\quad}}$$

Further, the variance of Z can be bounded by

$$Var(Z) = E_S\left[(Z \quad E_S[Z])^2\right]$$

$$= E_S\left[\left(\int_z \hat{\rho}(z) \ell(\quad(z); y(z))\, dz \quad E_S\left[\int_z \hat{\rho}(z) \ell(\quad(z); y(z))\, dz\right]\right)^2\right]$$

$$= E_S\left[\int_z \left(\hat{\rho}(z) \quad E_S[\hat{\rho}(z)]\right) \ell(\quad(z); y(z))\, dz\right]^2$$

$$E_S\left[\int_z \left(\hat{\rho}(z) \quad E_S[\hat{\rho}(z)]\right)^2 dz\right] \int_z \ell(\quad(z); y(z))^2\, dz$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{(III)} \quad \underbrace{\int_z \qquad\qquad}_{L^2 Vol(\quad(D))}$$

It follows from Eq. (2.3) in Jones et al. [15] together with Eq. 4.10 in Silverman [33] for (III) that for small $\quad$ and large sample size $|S|$ the term (III), i.e., the variance of $\hat{\rho}$, is given by

$$(III) = \quad |S|^{\quad 1 \quad m} + O(|S|^{\quad 2}),$$

where $\quad = \int_z \frac{p_D(z)}{\| z \|^m} dz$ and $\quad = \int_z k_1(0, z)^2\, dz$. Putting things together gives

$$E_{Rep}(f; S; \quad) \quad L \frac{\quad^2}{2}\ _2 \int_z r^2 \quad p_D(z) \| z \|^2 \quad dz + L \sqrt{\quad} \sqrt{Vol(\quad(D))}\, |S|^{\quad \frac{1}{2}} \quad^{\frac{m}{2}}$$

$$+ O(|S|^{\quad 2}) + O(\quad^3) :$$

Choosing the bandwidth as $= jSj^{\frac{1}{4+m}}$ gives

$$jE_{Rep}(f;S; \quad)j \; j \; Sj^{\frac{2}{4+m}} \quad {}_2L \int_z r^2 \; p_D(z)jjzjj^2 \quad dz + \frac{p}{p}\frac{L}{=} \; p \; \overline{Vol(\ (D))}$$

$$+ \; O(jSj^{\frac{3}{m+4}}) \; :$$

The result follows from setting

$$C_1 = {}_2L \int_z r^2 \; p_D(z)jjzjj^2 \quad dz$$

$$C_2 = \frac{p}{L} \; p \; \overline{Vol(\ (D))} \; :$$

□

E Relative flatness for a uniform bound over general distributions on feature matrices

This article based its consideration on the specific distribution on feature matrices of Proposition 2, since this distribution allows to use standard results of kernel density estimation in the proof to Theorem 6. However, the decomposition of the risk in Equation 5 holds for any distribution on feature matrices A and induced distributions on feature space $\varphi(A)$. To allow maximal flexibility in the choice of a distribution P_A on feature matrices of norm $\|A\| = 1$, we define another version of relative flatness based on the maximal eigenvalues of partial Hessians instead of the trace.

Definition 13. For a model $f(w; x) = g(w \varphi(x))$ with a twice differentiable function g, a twice differentiable loss function ℓ and a sample set S we define maximal relative flatness by

$$\Phi(w) := \sum_{s=1}^{d} \|w_s\|^2 \lambda_{\max}(H_{s;s}(w; \varphi(S))) \tag{17}$$

where $\lambda_{\max}$ denotes the maximal eigenvalue of a matrix and $H_{s;s}$ the Hessian matrix as in (8).

The analogue to Theorem 5 for maximal relative flatness shows that maximal flatness bounds feature robustness uniformly over all feature matrices of norm $\|A\| = 1$.

Theorem 14. Consider a model $f(x; w) = g(w \varphi(x))$ as above, a loss function ℓ and a sample set S, and let $O_m \subset \mathbb{R}^{m \times m}$ denote the set of orthogonal matrices. Let ε be a positive (small) real number and $w = \omega \in \mathbb{R}^{d \times m}$ denote parameters at a local minimum of the empirical risk on a sample set S. If the labels satisfy that $y[\varphi_A(x_i)] = y[\varphi(x_i)] = y_i$ for all $(x_i; y_i) \in S$ and all $\|A\| = 1$, then, for each feature selection matrix $\|A\| = 1$ the model $f(x; \omega)$ is $((\varepsilon; S; A); \varepsilon)$-feature robust for $= \frac{\varepsilon^2 d}{2} \Phi(\omega) + O(\varepsilon^3)$

Proof. Writing $z_i = \varphi(x_i)$ and $E_{emp}(w; S) = E_{emp}(f(w; x); S)$ and using the assumption that $y[\varphi_A(x_i)] = y_i$ for all $(x_i; y_i) \in S$ and all $\|A\| = 1$, we have by the first part of the proof of Theorem 5 that for any η ,

$$E_F(f; S; \varphi_A) + E_{emp}(w; S) = E_{emp}(w + \eta w A; S) \tag{18}$$

and

$$E_{emp}(\omega + \omega A; S) \leq E_{emp}(\omega; S) + \frac{\eta^2}{2} \sum_{s;t=1}^{d} (\omega_s A) H_{s;t}(\omega; \varphi(S)) (\omega_t A)^T + O(\eta^3) \tag{19}$$

at a local minimum ω, where ω_s denotes the s-th row of ω.

Note that for $\|A\| = 1$ and a row vector w_s it holds that $\|w_s A\| \leq \|w_s\|$. Further, since the full Hessian matrix $H(\omega; S) = (H_{s;t}(\omega; S))_{s;t}$ is a positive semidefinite matrix at a local minimum ω, it holds for each row vectors w_s, w_t that

$$w_s H_{s;t}(\omega; S) w_t^T \leq \frac{1}{2} \left(w_s H_{s;s}(\omega; S) w_s^T + w_t H_{t;t}(\omega; S) w_t^T \right); \tag{20}$$

We therefore get that for any feature matrix A with $\|A\| \leq 1$,

$$
\begin{aligned}
E_F(f;S;A) &\leq \max_{\|A\| \leq 1} E_F(f;S;A) \\
&\overset{(18);(19)}{=} \max_{\|A\| \leq 1} \frac{\epsilon^2}{2} \sum_{s,t=1}^{d} (\omega_s A)\, H_{s,t}(\omega;S)\, (\omega_t A)^T + O(\epsilon^3) \\
&\overset{(20)}{=} \max_{\|A\| \leq 1} \frac{\epsilon^2 d}{2} \sum_{s=1}^{d} (\omega_s A)\, H_{s,s}(\omega;S)\, (\omega_s A)^T + O(\epsilon^3) \\
&\leq \frac{\epsilon^2 d}{2} \sum_{s=1}^{d} \max_{\|z\| \leq \|\omega_s\|} z H_{s,s}(\omega;S) z^T + O(\epsilon^3) \qquad (21) \\
&= \frac{\epsilon^2 d}{2} \sum_{s=1}^{d} \max_{\|z\|=1} \|\omega_s\|^2 z H_{s,s}(\omega;S) z^T + O(\epsilon^3) \\
&= \frac{\epsilon^2 d}{2} \sum_{s=1}^{d} \|\omega_s\|^2 \max(H_{s,s}(\omega;S)) + O(\epsilon^3) \\
&= \frac{\epsilon^2 d}{2} \rho(\omega) + O(\epsilon^3)
\end{aligned}
$$

where we used the identity that $\max_{\|x\|=1} x^T M x = \lambda_{\max}(M)$ for any symmetric matrix M.

□

With this, the analogue to Theorem 6 (or its version Theorem 11 in the appendix) allows maximal flexibility to choose A (and $\epsilon > 0$) to bound representativeness. This leads to the following generalization bound.

Theorem 15. Consider a model $f(x;w) = g(w\ \phi(x))$, a loss function ℓ, a sample set S, and let m denote the dimension of the feature space defined by ϕ and let ϵ be a positive (small) real number. Let $\omega \in \mathbb{R}^{d \times m}$ denote a local minimum of the empirical risk on an iid sample set S.

Let Φ be the set of all $|S|$-tuple of distributions ϕ_A on feature vectors induced by a distribution on feature matrices of norm smaller than A as in Section 3. Then it holds that

$$
E_{gen}(f(\cdot;\omega);S) \leq \inf_{A} \left(E_{Rep}(f;S;\phi_A) + \frac{\epsilon^2 d}{2} \sum_{s=1}^{d} \rho(\omega) + O(\epsilon^3) \right).
$$

Proof. Part (i) follows from combining Equation 5 with Theorem 14. First, we use (5) to split the generalization gap into $E_{gen}(f) = E_{Rep}(f;S;\phi_A) + E_F(f;S;A)$. Then, Theorem 14 shows that $E_F(f;S;A) \leq \frac{\epsilon^2 d}{2} \rho(\omega) + O(\epsilon^3)$ as $E_F(f;S;A) \leq \frac{\epsilon^2 d}{2} \rho(\omega) + O(\epsilon^3)$ for all $\|A\| \leq 1$. □

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Yes, we describe limitations and assumptions in the introduction, as well as at the appropriate positions in the text and discuss broader properties of the results in this paper in the discussion.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This work contributes to the theoretical understanding of machine learning, in particular deep learning. As such, it has no direct ethical or societal impact. However, this understanding contributes to the trustworthiness of machine learning. Thus, it indirectly impacts the acceptance of automatic decisions made with such models and might facilitate their adoption in critical fields, such as medical applications and autonomous driving.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have carefully read the ethics review guidelines and confirm that this paper conforms to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We state all assumptions precisely in the introduction and provide further details, interpretations, and all proofs in the appendix.
   (b) Did you include complete proofs of all theoretical results? [Yes] The proofs can be found in Appdx D.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Details are in the supplemental material, an implementation is published in a repository at `https://anonymous.4open.science/r/RelativeFlatnessAndGeneralization-B175` (anonymized).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Details are specified in the appendix and the published code contains all training details to reproduce the results.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] In our experiments, we report the individual results of multiple runs and capture variations by reporting the correlation.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The details are provided in appendix C.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite existing assets we used in our experiments (CIFAR-10 dataset, LeNet5 network, PacBayes-measure implementation).
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]