Sharp minima can generalize for deep networks

Dinh, L. and Pascanu, R. and Bengio, S. and Bengio, Y.
ICML 2017

Matteo Gamba

2nd October 2018

Loss landscape

Goal: study properties of the loss surface and optimisation

- theoretical studies
  - given strong assumptions on the network architecture or data
  - characterise the minima found by gradient descent
  - hypothesis: sgd brings optimisations in regions with good minima (global under strong assumptions)
- empirical studies
  - study the geometry of the loss landscape to achieve better optimisers
  - attempt to drive optimisation towards wide basins of the error surface

Loss landscape

Not covered in this presentation:

- "*Flat minima*". Hochreiter, and Schmidhuber. 1997.
- "*The loss surface of multilayer networks*". Choromanska, Henaff, Mathieu, Ben Arous, LeCun. 2015.
- "*Entropy-SGD: biasing gradient descent into wide valleys*". Chaudhari et al. 2016.
- "*On large-batch training for deep learning: generalisation and sharp minima*". Keskar et al, 2016.
- "*The loss surface of deep and wide neural networks*". Nguyen, Q. and Hein, M. 2017.
- "*Theoretical insights into the optimisation landscape of over-parameterised shallow neural networks*". Soltanolkotabi, Javanmard and Lee. 2018.

Loss landscape

Theoretical studies

Not covered in this presentation: "*The loss surface of deep and wide neural networks*". Nguyen, Q. and Hein, M. 2017.

- square activation, square loss, rank of weight matrices maximal in each critical point
- every critical point is a global minimum

"*Theoretical insights into the optimisation landscape of over-parameterised shallow neural networks*". Soltanolkotabi, Javanmard and Lee. 2018.

- square loss, single layer network, synthetic data
- all local minima are global, independent of labelling of data

Loss landscape

Empirical studies

Not covered in this presentation:

- "*Entropy-SGD: biasing gradient descent into wide valleys*". Chaudhari et al. 2016.
- "*On large-batch training for deep learning: generalisation and sharp minima*". Keskar et al, 2016.

Based on: "*Flat minima*". Hochreiter, and Schmidhuber. 1997.

- in a flat minimum, the loss varies slowly in a relatively lare neighbourhood.

Flat minima

Study the loss function in a neighbourhood of a solution.
Conjecture:

- “flat” minima are robust to perturbations in the parameter space and numerical errors (Hochreiter and Schmidhuber, 1997).
- information theory: low-precision weights encode less information from the training data
- hence they favour a “simpler” model and tend to generalise well in practice

Flat minima

Motivation:

- (Chaudhari, 2016) intuitive: wide valleys of the loss contain close to optimal minima
- (Keskar, 2016) small batch SGD finds wider basins, observed to generalise better than wider basins found by large batch methods
- (Hinton and Vancamp, 1993) bayesian argument.

Sharp minima can generalise

Bayesian argument is non-parametric, while the definitions of flatness proposed in the literature depend on the parametrisation of the weight space

Any measure expressing how the weights should change, for a given unit change in the model behaviour, would depend on the highly non-linear geometry induced on the parameter space by the network architecture.

Sharp minima can generalise

Outline of the presentation:

- definitions of flatness
- theoretical setting
- methodology
- theoretical results
- connection to empirical studies
- open questions

Sharp minima can generalise

Interesting insights:

- geometry induced by ReLU on the parameter space
- how to control eigenvalues without altering the behaviour of the model
- for a given minimum with good generalisation power, there exist another minimum with the same generalisation performance but arbitrary measure of flatness.

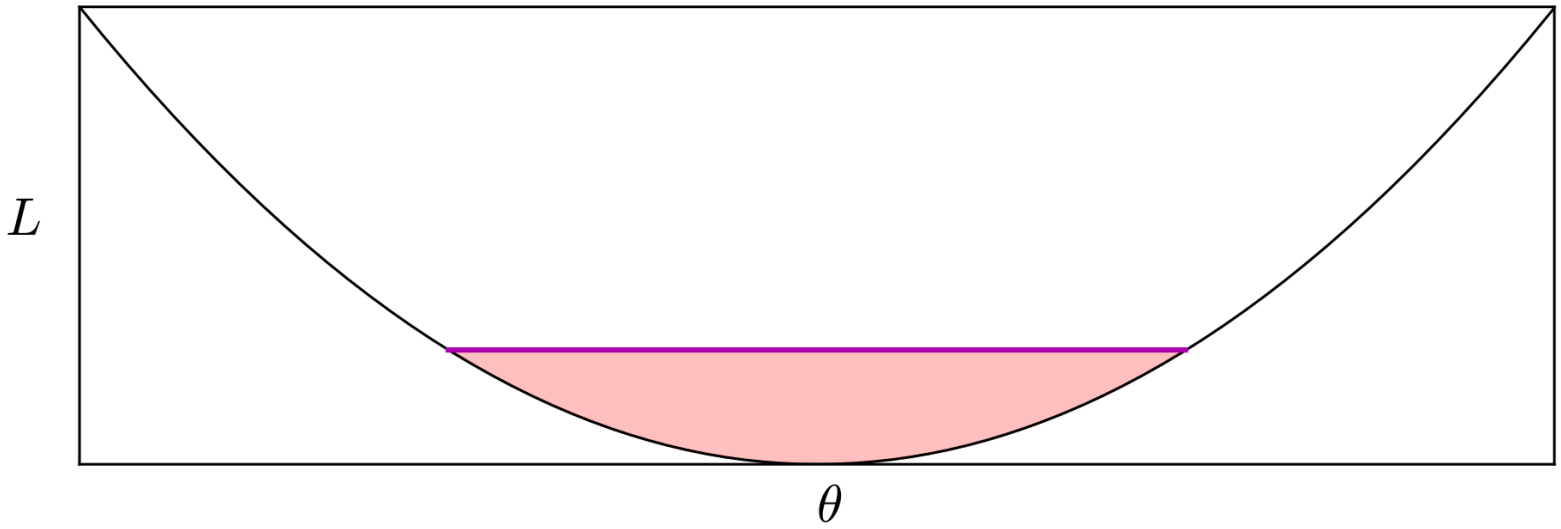
Definitions of flatness

All definitions are given in a neighbourhood of a minimum $\theta \in \Theta$.

- ϵ-flatness (Hochreiter and Schmidhuber. 1997)
- ϵ-sharpness (Keskar et al. 2017)
- Second-order measures (Chaudhari et al. 2017)
  - Eigenvalues of the Hessian (Chaudhari et al. 2017 and Keskar et al. 2017)
  - spectral norm and trace of the Hessian

Definitions of flatness

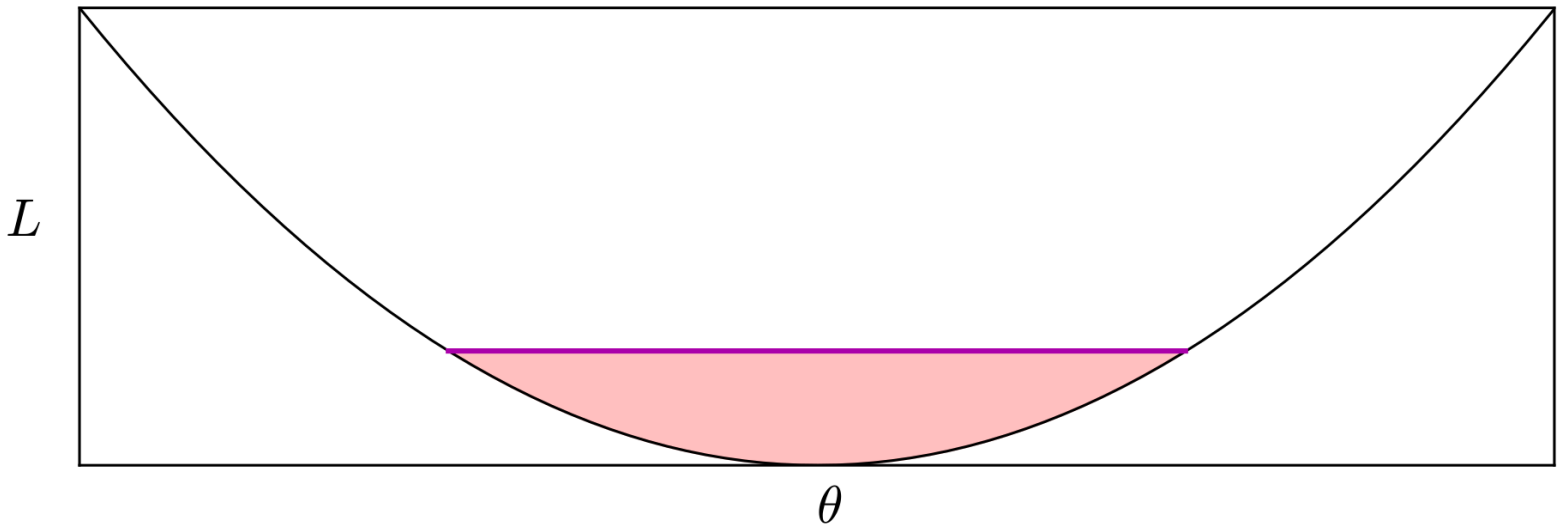
ϵ-flatness



For $\epsilon > 0$, flatness is defined as the largest connected region in the param space for which the error is approximately constant (up to a factor of ϵ).

Definitions of flatness

ϵ-sharpness



For $\epsilon > 0$, $B_2(\epsilon, \theta)$, $\dfrac{\max_{\theta' \in B_2(\epsilon,\theta)}(L(\theta') - L(\theta))}{1 + L(\theta)}$

Definitions of flatness

Hessian based measures

- characterise flatness with the eigenvalues of the hessian at θ
- spectral radius ($\propto$ largest eigenvalue)
- trace norm ($\propto$ sum of eigenvalues)

Setting

Framework:

- study deep rectifier networks
- supervised learning setting
- scalar output (scalar loss)
- network learns a scalar function $f_\theta : \mathcal{X} \to \mathbb{R}$, parametrised by $\theta \in \Theta$

Setting

- loss assumed non-negative
- with continuous second order partial derivatives (in a neigh. of a minimum θ).
- depth-K rectifier network with linear output layer formalised as:

$$y = \phi_{relu}(\phi_{relu}(\cdots \phi_{relu}(x \cdot \theta_1) \cdots) \cdot \theta_{K-1})\theta_K$$

Methodology

- Loss $L(f_\theta(x))$ as a function of the weights only, $L(\theta)$.
- Symmetries induced by non-negative homogeneity of ReLU:

$$\forall \alpha > 0, \quad \phi_{relu}(\alpha \cdot x) = \alpha \cdot \phi_{relu}(x)$$

e.g. for a 2-layer network

$$\phi_{relu}(\theta_1 \cdot x) \cdot \theta_2 = \phi_{relu}(x \cdot (\alpha\theta_1)) \cdot (\alpha^{-1}\theta_2)$$

- parameter space: $\Theta : (\theta_1, \theta_2) = (\theta_1^1, \ldots, \theta_{n_1}^1, \theta_1^2, \ldots, \theta_{n_2}^2)$

Methodology

Observational equivalence:

- $(\theta, \theta') \in \Theta^2$ are *observationally equivalent* if

$$f_\theta(x) = f_{\theta'}(x) \quad \forall x \in \mathcal{X}$$

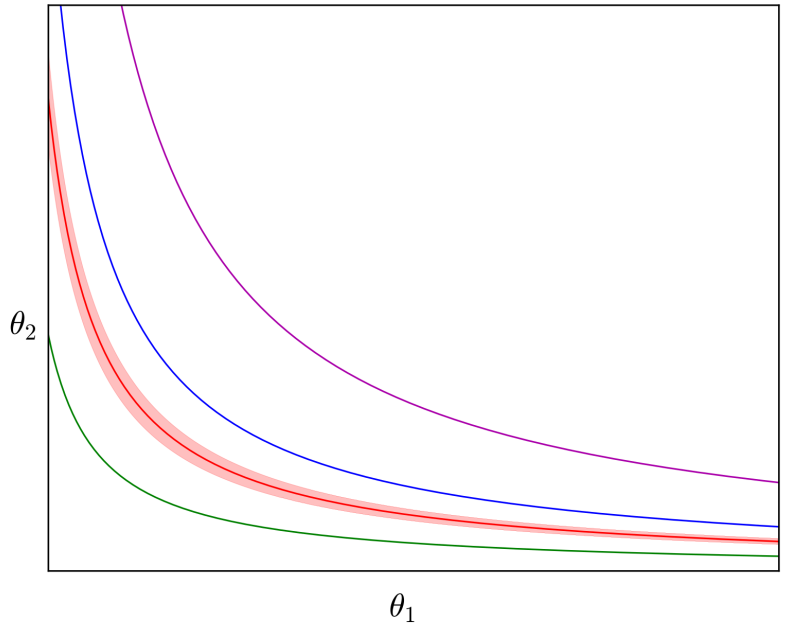
- w.r.t ReLU (θ_1, θ_2) and $(\alpha\theta_1, \alpha^{-1}\theta_2)$ are equivalent.

Methodology

Parameter transformations:

- $T_\alpha : (\theta_1, \theta_2) \mapsto (\alpha\theta_1, \alpha^{-1}\theta_2)$
- produce obs. equivalent parameters
- the behaviour of the prediction function is not altered (same output of the network)

Navigating the parameter space



Same loss value, network architecture and input x

$\implies$ same generalisation error

$\not\Rightarrow$ same flatness

The definitions depend on the parameter θ

Results

Strategy:

- exploit symmetries induced on the parameter space by the architecture
- equivalence of norms for finite spaces
- control the proposed measures and make them diverge

Move along level curves of the loss function in Θ to control the flatness around a minimum

ϵ-flatness

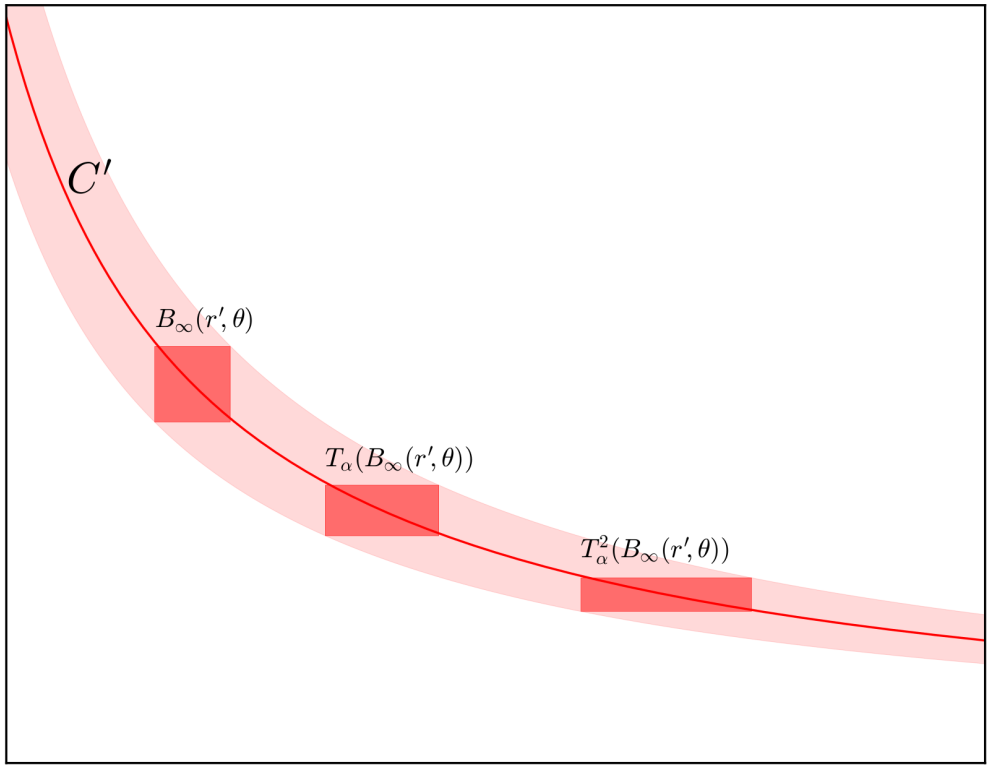
Proof provided $\forall \epsilon > 0$, for a 2-dimensional NN

$$y = \phi_{relu}(x \cdot \theta_1) \cdot \theta_2$$

sketch of the proof:

- consider the largest connected region C' of Θ containing θ, where the loss remains approx. constant
- show that C' can be lower-bounded by one with infinite volume

ϵ-flatness



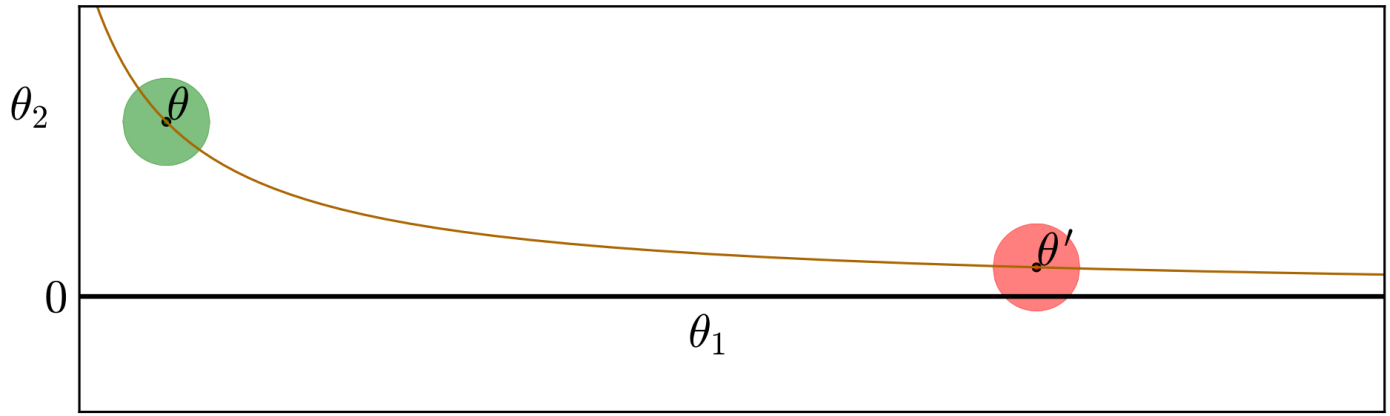
Every minimum is contained in a region of infinite volume with approximately constant loss.

Hence, all minima are equally ϵ-flat.

ϵ-sharpness

Proof provided $\forall \epsilon > 0$, for a 2-dimensional NN.

- for each minimum θ it is possible to find another minimum θ' with *high* ϵ-sharpness
- for θ', the maximum loss in $B_2(\epsilon, \theta')$ is as high as that of the degenerate model $y \equiv 0$, which is assumed to be high.



Second order measures

Preliminary results

The loss and hessian depend on the parameter transformation T_α.

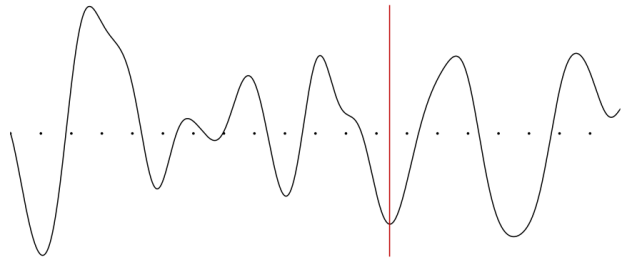
by differentiating both sides of $L(\theta_1, \theta_2) = L(\alpha\theta_1, \alpha^{-1}\theta_2)$

- $(\nabla L)(\alpha\theta_1, \alpha^{-1}\theta_2) = (\nabla L)(\theta_1, \theta_2) \begin{pmatrix} \alpha^{-1}\mathbb{I}_{n_1} & 0 \\ 0 & \alpha\mathbb{I}_{n_2} \end{pmatrix}$
- $(\nabla^2 L)(\alpha\theta_1, \alpha^{-1}\theta_2) =$

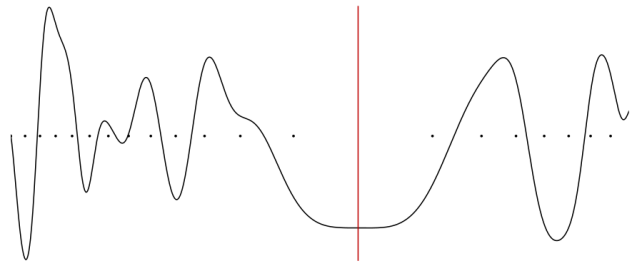
$$\begin{pmatrix} \alpha^{-1}\mathbb{I}_{n_1} & 0 \\ 0 & \alpha\mathbb{I}_{n_2} \end{pmatrix} (\nabla^2 L)(\theta_1, \theta_2) \begin{pmatrix} \alpha^{-1}\mathbb{I}_{n_1} & 0 \\ 0 & \alpha\mathbb{I}_{n_2} \end{pmatrix} \quad (1)$$

Second order measures

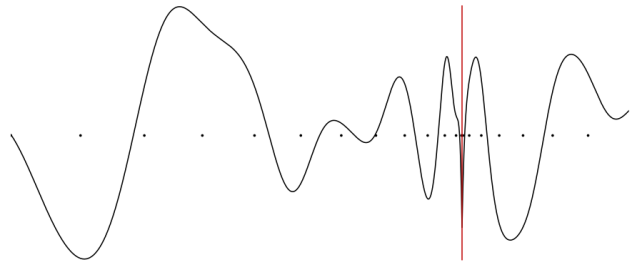
Given a minimum with non degenerate Hessian, the transformations allow to find an obs. equivalent minimum with arbitrarily large spectral norm (and thus, trace norm and spectral radius).



(a) Loss function with default parametrization



(b) Loss function with reparametrization



(c) Loss function with another reparametrization

Full eigenspectrum

Wide valleys

To move in the parameter space along multiple directions while exploiting the geometry induced by ReLU:

- depth-K rectifier networks are considered
- for $\alpha_k > 0$: $T_\alpha : (\theta_1, \ldots, \theta_K) \mapsto (\alpha_1\theta_1, \ldots, \alpha_K\theta_K)$ so that $\prod_{k=1}^{K} \alpha_K = 1$.

Full eigenspectrum

- then, eq. 1 becomes

$$(\nabla^2 L)(T_\alpha(\theta)) = D_\alpha (\nabla^2 L)(\theta) D_\alpha$$

- where D_α is the inverse of the Jacobian $J(T_\alpha(\theta))$:

$$D_\alpha = \begin{pmatrix} \alpha_1^{-1}\mathbb{I}_{n_1} & 0 & \dots & 0 \\ 0 & \alpha_2^{-1}\mathbb{I}_{n_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_K^{-1}\mathbb{I}_{n_K} \end{pmatrix}$$

Full eigenspectrum

- If the hessian matrix in a minimum θ in the full space has rank r
- the authors provide a sufficient condition to make $r - n_K$ eigenvalues arbitrarily large.
- i.e. they can control all but n_k eigenvalues ($k \leq K$ is chosen arbitrarily).
- let $n := \sum_k n_k$, no control over the $n - r$ eigenvalues that are zero.

Hessian matrix in practice

Summary:

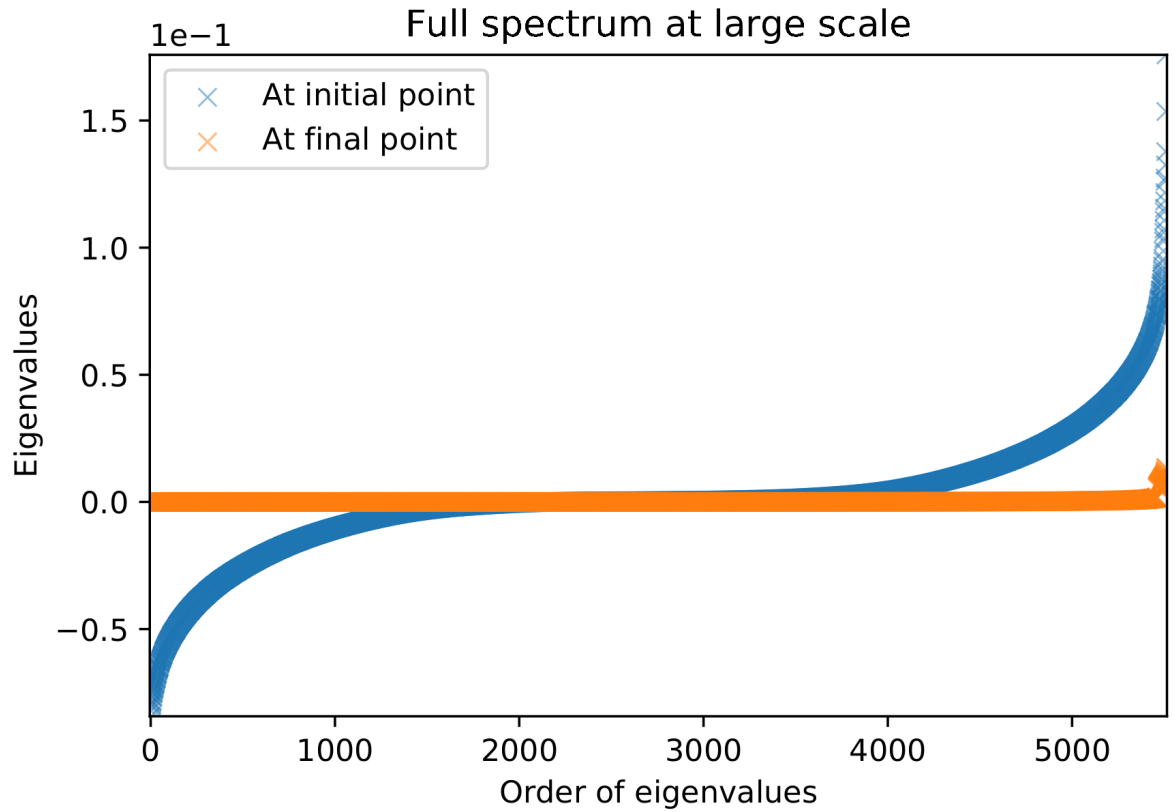
- Assuming a positive, semidefinite Hessian, up to $(r - n_k)$ eigenvalues can be made arbitrarily large.
- How does the Hessian at a minimum look in practice?

Hessian matrix in practice

"*Empirical analysis of the Hessian of over-parametrised neural networks*". Sagun, L. and Evci, U. and Güney, V. and Dauphin, Y. and Bottou, L. ICLR 2018 Workshop paper.

- Empirical study of the Hessian matrix before and after optimisation.
- Interesting case: depth-2 fully connected rectifier network with scalar output.

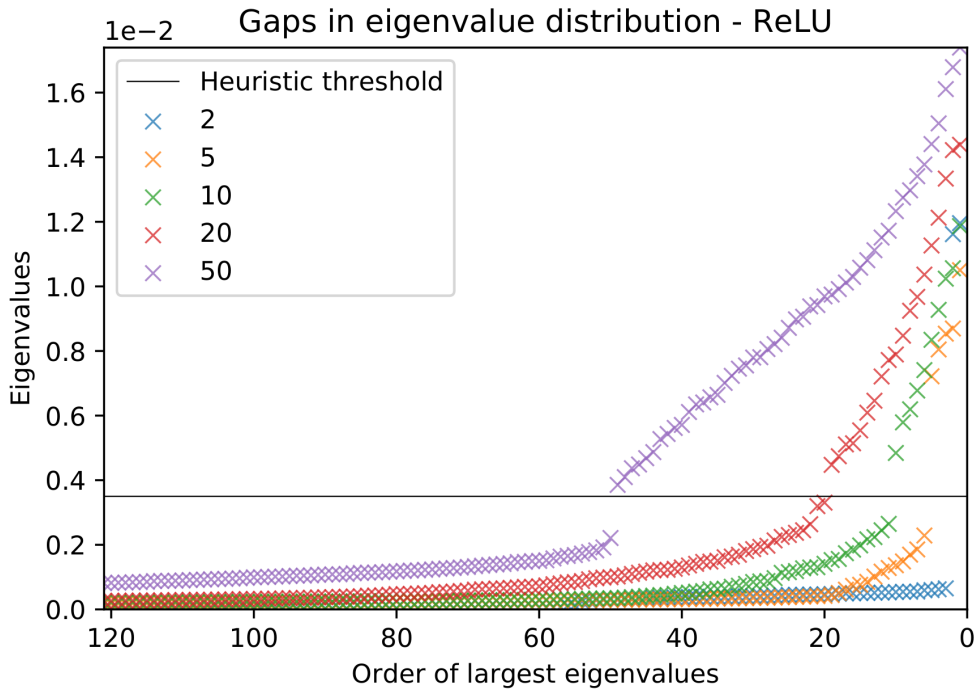
Hessian matrix in practice



Two hidden-layer ReLU network ($\approx$ 5000 params). Eigenvalues before (random initial point) and after (minimum?) training.

Hessian matrix in practice

Data vs eigenvalues

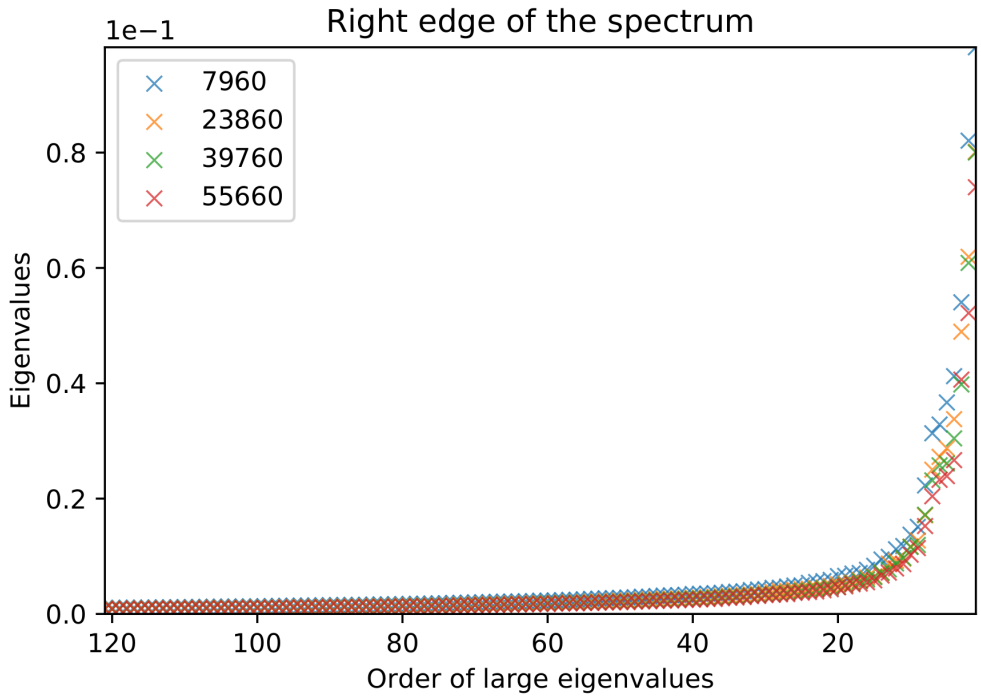


Two hidden-layer ReLU network ($\approx 4 - 5000$ params), k outputs, synthetic data (k clusters).

The number of large eigenvalues (above threshold) matches k.

Hessian matrix in practice

Number of parameters vs eigenvalues

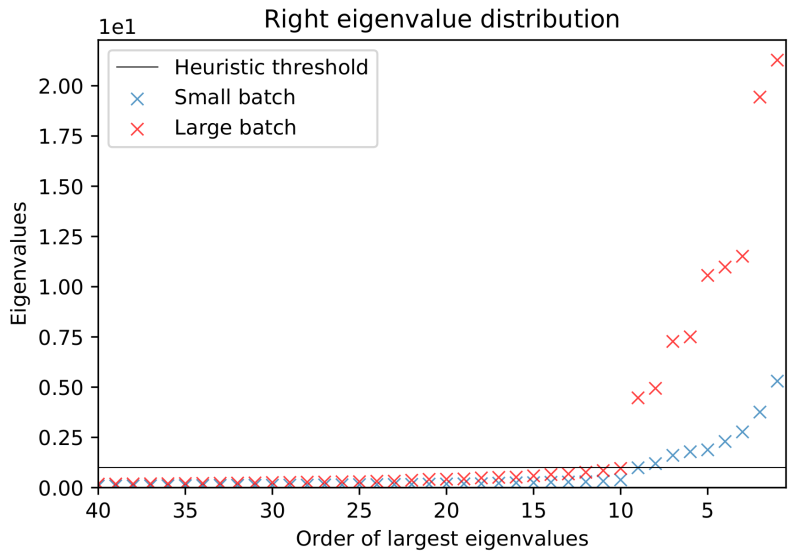


Two hidden-layer ReLU network trained on subset of MNIST.

Top 120 eigenvalues. Almost no change in size and number of eigenvalues.

Hessian matrix in practice

Batch size vs eigenvalues



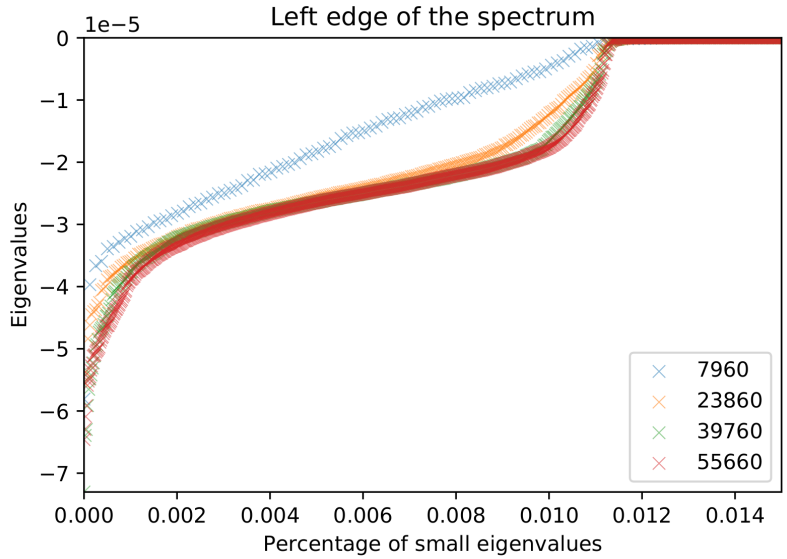
CNN with ReLU and maxpooling + 2 FC layers, trained on subset of MNIST.

Batch size of 10 vs 512.

Top 40 eigenvalues. Large batch produces larger positive eigenvalues.

Hessian matrix in practice

Negative eigenvalues



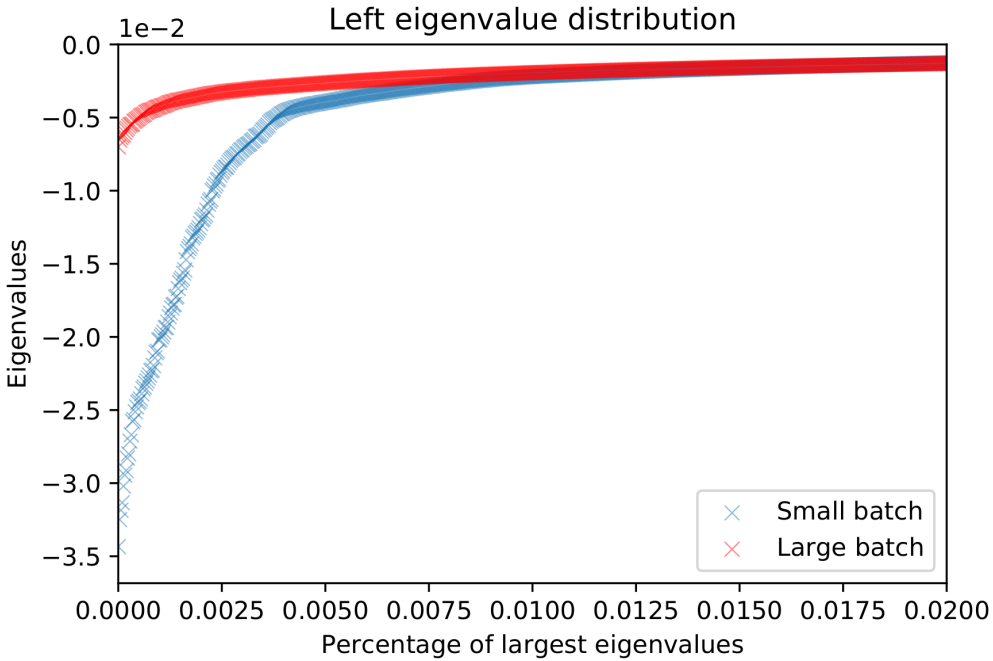
Two hidden-layer ReLU network trained on subset of MNIST.

Negative eigenvalues at the end of training have smaller magnitude than the positive ones.

x-axis: percentage of small eigenvalues.

Hessian matrix in practice

Negative eigenvalues



CNN with ReLU and maxpooling + 2 FC layers, trained on subset of MNIST.

Batch size of 10 vs 512.

x-axis: percentage of small eigenvalues.

Open questions

- What is the geometry induced on the parameter space the the size of each layer?
- How does overparametrisation of a hidden layer affect the Hessian at a minimum?
- Does the gradient always find a minimiser?
- Is ϵ-sharpness a well defined measure when averaged over random subspaces of the parameter space (Keskar et al. 2017)?

Sharp minima can generalize for deep networks

Questions?

ϵ-flatness

proof

One-layer rectifier network:

$$y = \phi_{relu}(x \cdot \theta_1) \cdot \theta_2$$

If $\theta \in \Theta$ minimum s.t. $\theta_1 \neq \mathbf{0}, \theta_2 \neq \mathbf{0}$

$\forall \epsilon > 0$, $C(L, \theta, \epsilon)$ has infinite volume.

Note:

- $T_\alpha : (\theta_1, \theta_2) \mapsto (\alpha\theta_1, \alpha^{-1}\theta_2)$ has Jacobian determinant $\alpha^{n_1 - n_2}$.
- $T_\alpha \circ T_\beta = T_{\alpha\beta}$.

ϵ-flatness

proof

Case 1: $n_1 \neq n_2$

- goal: show that $\forall \epsilon > 0$ there exists a region of approximately constant loss with infinite volume.
  1. $\exists r > 0 \quad s.t. \quad B_\infty(\epsilon, \theta) \subseteq C(L, \theta, \epsilon)$.
     In fact, L continuous $\implies$ $L^{-1}(B_2(\epsilon, L(\theta))$ is open in Θ, so $r > 0$.
  2. since $\theta_1 \neq \mathbf{0}$ and $\theta_2 \neq \mathbf{0}$, $B_\infty(\epsilon, \theta)$ has volume $v = 2r^{n_1+n_2} > 0$.
  3. the volume of $T_\alpha(B_\infty(\epsilon, \theta))$ is $v\alpha^{n_1-n_2}$.
  4. hence, by picking α arbitrarily large, the volume of $C(L, \theta, \epsilon)$ can be controlled.

ϵ-flatness

proof

Case 2: $n_1 = n_2$

- goal: show that $\forall \epsilon > 0$ there exists a region of approximately constant loss with infinite volume.
  1. let $C' = \bigcup_{\alpha' > 0} T_{\alpha'}(B_\infty(r, \theta))$
     $T_{\alpha'}(B_\infty(r, \theta))$ has volume v, $\quad \forall \alpha' > 0$
  2. C' is a connected region with approximately constant volume.

  Goal: lowerbound C' with a region of infinite volume.

ϵ-flatness

proof

Case 2: $n_1 = n_2$

- goal: lowerbound C' with a region of infinite volume.
  3. $B_\infty(r, \theta) = B_\infty(r, \theta_1) \times B_\infty(r, \theta_2)$.
  4. then $T_\alpha(B_\infty(r, \theta)) = B_\infty(\alpha r, \alpha\theta_1) \times B_\infty(\alpha^{-1} r, \alpha^{-1}\theta_2)$ for $\alpha = 2\frac{||\theta_1||_\infty + r}{||\theta_2||_\infty - r}$
  5. and $T_\alpha(B_\infty(r, \theta)) \cap B_\infty(r, \theta) = 0$.
  6. similarly $\{T_\alpha^k(B_\infty(r, \theta))\}_{k \geq 0}$ are disjoint sets, each of volume v
  7. then, volume $C' > v + v + v + \ldots$.

ϵ-sharpness

proof

One-layer rectifier network:

$$y = \phi_{relu}(x \cdot \theta_1) \cdot \theta_2$$

If $\theta \in \Theta$ minimum s.t. $\theta_1 \neq \mathbf{0}, \theta_2 \neq \mathbf{0}$

$\forall \epsilon > 0, \exists \theta' \in \Theta$ with higher ϵ-sharpness than θ.

Note:

- For $(\theta_1, \theta_2) = (\mathbf{0}, \theta_2)$ the prediction function degenerates to $y \equiv 0, \quad \forall x \in \mathcal{X}$.

ϵ-sharpness

proof

- goal: show that $\forall \epsilon > 0, \quad \exists \theta' \in \Theta \quad$ s.t.

$$\{(\theta_1, \theta_2) \in \Theta, \theta_1 = 0\} \cap B_2(\epsilon, \theta') \neq \emptyset$$

1. define $\alpha = \frac{\epsilon}{||\theta_1||_2}$
2. let $T_\alpha(\theta_1, \theta_2) = (\epsilon \frac{\theta_1}{||\theta_1||_2}, \alpha^{-1}\theta_2)$.
3. since $||\theta_1||_2 \leq ||\theta||_2$ then $(0, \alpha^{-1}\theta_2) \in B_2(\epsilon, T_\alpha(\theta))$.

Spectral norm

proof

One-layer rectifier network:

$$y = \phi_{relu}(x \cdot \theta_1) \cdot \theta_2$$

If $\theta \in \Theta$ minimum s.t. $(\nabla^2 L)(\theta) \neq \mathbf{0}$

$\forall M > 0, \exists \alpha > 0$ s.t. $||(\nabla^2 L)(T_\alpha(\theta))||_2 > M.$

Spectral norm

proof

- goal: lower bound $||(\nabla^2(L)(T_\alpha(\theta))||_2$ with arbitrary $M > 0$
  1. if hessian is non-zero, since θ is minimum, there exists a positive eigenvalue $\gamma > 0$, e.g. for θ_1.
  2. hence, the Frobenius norm $||(\nabla^2(L)(T_\alpha(\theta))||_F$ is at least $\alpha^{-2}\gamma$.
  3. by the equivalence of norms in finite spaces, $\exists r > 0$ s.t. $r||A||_F \leq ||A||_2$ for any symmetric matrix A.
  4. by picking $\alpha < \sqrt{\frac{r\gamma}{M}}$, the thesis follows.

Hessian in many dimensions

K-layer rectifier network:

$$y = \phi_{relu}(\phi_{relu}(\ldots \phi_{relu}(x \cdot \theta_1)\ldots)\theta_{K-1})\theta_K$$

If $\theta = (\theta_1, \ldots, \theta_K) \in \Theta$ minimum s.t. hessian in θ has rank r

$\forall M > 0, \exists \alpha > 0$ so that $r - \min_{k \leq K}(n_k)$ eigenvalues are greater than M.

Note:

- $(\nabla^2 L)(T_\alpha(\theta)) = D_\alpha(\nabla^2 L)(\theta)D_\alpha$

Hessian in many dimensions

- goal: sort eigenvalues of the hessian and lower bound $r - \min_{k \leq K} n_k$ of them with an arbitrary $M > 0$
  1. the hessian of L is assumed to be positive semidefinite and symmetric in a neighbourhood of θ.
- idea: compute the singular values of $D_\alpha(\nabla^2 L)(\theta)D_\alpha$ and apply Horn's inequalities
- Horn's inequalities: given the singular values of A and B, relationship on the singular values of AB

Hessian in many dimensions

- goal: sort eigenvalues of the hessian and lower bound $r - \min_{k \leq K} n_k$ of them with an arbitrary $M > 0$

2. To apply the inequalities, we work on the singular values $\sqrt{(\nabla^2 L)(\theta) D_\alpha^2}$, which are the square root of the eigenvalues of the $D_\alpha (\nabla^2 L)(\theta) D_\alpha$
3. for $k \leq K$, α_k is chosen as: $\alpha_k = \beta^{-1}$ and $\alpha_K = \beta^{K-1}$.
4. Horn's inequalities: $\forall i \leq n, \quad j \leq (n - n_K)$:

$$\lambda_{i+j-n}((\nabla^2 L)(\theta) D_\alpha^2) \geq \lambda_i((\nabla^2 L)(\theta))\beta^2, \quad \text{for any } \beta > 0$$

Hessian in many dimensions

- goal: sort eigenvalues of the hessian and lower bound $r - \min_{k \leq K} n_k$ of them with an arbitrary $M > 0$

5. for $\beta > \sqrt{\frac{M}{\lambda_r(\nabla^2 L)(\theta)}}$
6. $\forall i \leq (r - n_K)$

$$\lambda_i((\nabla^2 L)(\theta) D_\alpha^2) \geq \lambda_{i+n_k}((\nabla^2 L)(\theta))\beta^2 \geq \lambda_r((\nabla^2 L)(\theta))\beta^2 > M$$

Weight normalisation

Weight normalisation allows to define isotropic rescalings:

- non-zero weight $\mathbf{v}$
- normalised as $\mathbf{w} \leftarrow s\frac{\mathbf{v}}{||\mathbf{v}||_2}, \quad s$ scale
- since $\mathbf{w}$ is invariant to rescaling of $\mathbf{v}$
- define $T_\alpha = v \mapsto \alpha v, \quad \alpha \neq 0.$

Weight normalisation

Implications

- every minimum has infinite volume ϵ-flatness
- every minimum is obs. equivalent to an infinitely sharp minimum and to an infinitely flat minimum when considering the eigenvalues of the hessian
- every minimum is obs. equivalent to a minimum with arbitrarily low full-space ϵ-sharpness and a minimum with high full-space ϵ-sharpness.