Sharp Minima Can Generalize For Deep Nets
Supplementary Material

Laurent Dinh [1] **Razvan Pascanu** [2] **Samy Bengio** [3] **Yoshua Bengio** [1 4]

A Radial transformations

We show an elementary transformation to locally perturb the geometry of a finite-dimensional vector space and therefore affect the relative flatness between a finite number minima, at least in terms of spectral norm of the Hessian. We define the function:

$$\begin{aligned}
&\forall \delta > 0, \forall \rho \in]0, \delta[, \forall (r, \hat{r}) \in \mathbb{R}_+ \times]0, \delta[, \\
&\quad \psi(r, \hat{r}, \delta, \rho) = \mathbb{1}\big(r \notin [0, \delta]\big)\, r + \mathbb{1}\big(r \in [0, \hat{r}]\big)\, \rho\, \frac{r}{\hat{r}} \\
&\qquad\qquad + \mathbb{1}\big(r \in]\hat{r}, \delta]\big)\, \Big((\rho - \delta)\, \frac{r - \delta}{\hat{r} - \delta} + \delta\Big) \\
&\quad \psi'(r, \hat{r}, \delta, \rho) = \mathbb{1}\big(r \notin [0, \delta]\big) + \mathbb{1}\big(r \in [0, \hat{r}]\big)\, \frac{\rho}{\hat{r}} \\
&\qquad\qquad + \mathbb{1}\big(r \in]\hat{r}, \delta]\big)\, \frac{\rho - \delta}{\hat{r} - \delta}
\end{aligned}$$

For a parameter $\hat{\theta} \in \Theta$ and $\delta > 0, \rho \in]0, \delta[, \hat{r} \in]0, \delta[$, inspired by the *radial flows* (Rezende & Mohamed, 2015) in we can define the *radial transformations*

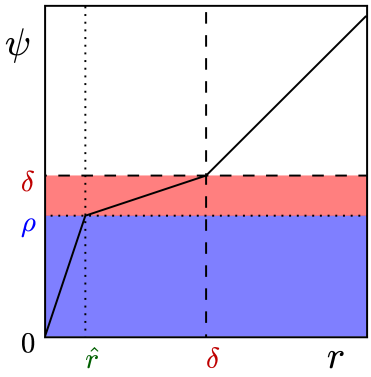
$$\forall \theta \in \Theta,\ g^{-1}(\theta) = \frac{\psi\Big(\|\theta - \hat{\theta}\|_2, \hat{r}, \delta, \rho\Big)}{\|\theta - \hat{\theta}\|_2} \big(\theta - \hat{\theta}\big) + \hat{\theta}$$

with Jacobian

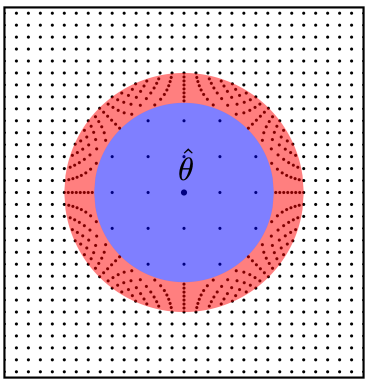
$$\begin{aligned}
\forall \theta \in \Theta,\ (\nabla g^{-1})(\theta) = &\ \psi'(r, \hat{r}, \delta, \rho)\, \mathbb{I}_n \\
&- \mathbb{1}\big(r \in]\hat{r}, \delta]\big) \frac{\delta(\hat{r} - \rho)}{r^3(\hat{r} - \delta)} (\theta - \hat{\theta})^T (\theta - \hat{\theta}) \\
&+ \mathbb{1}\big(r \in]\hat{r}, \delta]\big) \frac{\delta(\hat{r} - \rho)}{r(\hat{r} - \delta)}\, \mathbb{I}_n,
\end{aligned}$$

with $r = \|\theta - \hat{\theta}\|_2$.

First, we can observe in Figure 1 that these transformations are purely local: they only have an effect inside the ball $B_2(\hat{\theta}, \delta)$. Through these transformations, you can arbitrarily perturb the ranking between several minima in terms of flatness as described in subsection 5.1.



(a) $\psi(r, \hat{r}, \delta, \rho)$



(b) $g^{-1}(\theta)$

Figure 1: An example of a radial transformation on a 2-dimensional space. We can see that only the area in blue and red, i.e. inside $B_2(\hat{\theta}, \delta)$, are affected. Best seen with colors.

[1]Université of Montréal, Montréal, Canada [2]DeepMind, London, United Kingdom [3]Google Brain, Mountain View, United States [4]CIFAR Senior Fellow. Correspondence to: Laurent Dinh <laurent.dinh@umontreal.ca>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

B Considering the bias parameter

When we consider the bias parameter for a one (hidden) layer neural network, the non-negative homogeneity property translates into

$$\begin{aligned} y &= \phi_{rect}(x \cdot \theta_1 + b_1) \cdot \theta_2 + b_2 \\ &= \phi_{rect}(x \cdot \alpha\theta_1 + \alpha b_1) \cdot \alpha^{-1}\theta_2 + b_2, \end{aligned}$$

which results in conclusions similar to section 4.

For a deeper rectified neural network, this property results in

$$\begin{aligned} y &= \phi_{rect}\Big(\phi_{rect}\big(\cdots \phi_{rect}(x \cdot \theta_1 + b_1)\cdots\big) \cdot \theta_{K-1} + b_{K-1}\Big) \\ &\quad \cdot \theta_K + b_K \\ &= \phi_{rect}\Big(\phi_{rect}\big(\cdots \phi_{rect}(x \cdot \alpha_1\theta_1 + \alpha_1 b_1)\cdots\big) \\ &\quad \cdot \alpha_{K-1}\theta_{K-1} + \prod_{k=1}^{K-1} \alpha_k b_{K-1}\Big) \cdot \alpha_K\theta_K + b_K \end{aligned}$$

for $\prod_{k=1}^{K} \alpha_k = 1$. This can decrease the amount of eigenvalues of the Hessian that can be arbitrarily influenced.

C Rectified neural network and Lipschitz continuity

Relative to recent works (Hardt et al., 2016; Gonen & Shalev-Shwartz, 2017) assuming *Lipschitz continuity* of the loss function to derive uniform stability bound, we make the following observation:

Theorem 1. *For a one-hidden layer rectified neural network of the form*

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

if L is not constant, then it is not Lipschitz continuous.

Proof. Since a Lipschitz function is necessarily absolutely continuous, we will consider the cases where L is absolutely continuous. First, if L has zero gradient almost everywhere, then L is constant.

Now, if there is a point θ with non-zero gradient, then by writing

$$\begin{aligned} (\nabla L)(\theta_1, \theta_2) = [&(\nabla_{\theta_1} L)(\theta_1, \theta_2) \\ &(\nabla_{\theta_2} L)(\theta_1, \theta_2)], \end{aligned}$$

we have

$$\begin{aligned} (\nabla L)(\alpha\theta_1, \alpha^{-1}\theta_2) = [&\alpha^{-1}(\nabla_{\theta_1} L)(\theta_1, \theta_2) \\ &\alpha(\nabla_{\theta_2} L)(\theta_1, \theta_2)]. \end{aligned}$$

Without loss of generality, we consider $(\nabla_{\theta_1} L)(\theta_1, \theta_2) \neq 0$. Then the limit of the norm

$$\begin{aligned} \|(\nabla L)(\alpha\theta_1, \alpha^{-1}\theta_2)\|_2^2 = \; &\alpha^{-2}\|(\nabla_{\theta_1} L)(\theta_1, \theta_2)\|_2^2 \\ &+ \alpha^2\|(\nabla_{\theta_2} L)(\theta_1, \theta_2)\|_2^2 \end{aligned}$$

of the gradient goes to $+\infty$ as α goes to 0. Therefore, L is not Lipschitz continuous. □

This result can be generalized to several other models containing a one-hidden layer rectified neural network, including deeper rectified networks.

D Euclidean distance and input representation

A natural consequence of subsectio 5.2 is that metrics relying on Euclidean metric like *mean square error* or *Earth-mover distance* will rank very differently models depending on the input representation chosen. Therefore, the choice of input representation is critical when ranking different models based on these metrics. Indeed, bijective transformations as simple as *feature standardization* or *whitening* can change the metric significantly.

On the contrary, ranking resulting from metrics like *f-divergence* and *log-likelihood* are not perturbed by bijective transformations because of the *change of variables formula*.

E Eigenspectrum of Hessian

In section 4.2, we show how to manipulate the spectral radius and trace of the Hessian as a notion of sharpness. In However, some notion of sharpness might take into account the entire eigenspectrum of the Hessian as opposed to its largest eigenvalue, for instance, Chaudhari et al. (2017) describe the notion of *wide valleys*, allowing the presence of very few large eigenvalues. We can generalize the transformations between observationally equivalent parameters to deeper neural networks with $K-1$ hidden layers: for $\alpha_k > 0, T_\alpha : (\theta_k)_{k\leq K} \mapsto (\alpha_k\theta_k)_{k\in K}$ with $\prod_{k=1}^{K} \alpha_k = 1$. If we define

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

then the first and second derivatives at $T_\alpha(\theta)$ will be

$$\begin{aligned} (\nabla L)\big(T_\alpha(\theta)\big) &= (\nabla L)(\theta) D_\alpha \\ (\nabla^2 L)\big(T_\alpha(\theta)\big) &= D_\alpha (\nabla^2 L)(\theta) D_\alpha. \end{aligned}$$

We will show to which extent you can increase several eigenvalues of $(\nabla^2 L)\big(T_\alpha(\theta)\big)$ by varying α.

Definition 1. *For each $n \times n$ matrix A, we define the vector $\lambda(A)$ of sorted singular values of A with their multiplicity $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$.*

If A is symmetric positive semi-definite, $\lambda(A)$ is also the vector of its sorted eigenvalues.

Theorem 2. *For a $(K-1)$-hidden layer rectified neural network of the form*

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

and critical point $\theta = (\theta_k)_{k \leq K}$ being a minimum for L, such that $(\nabla^2 L)(\theta)$ has rank $r = rank\big((\nabla^2 L)(\theta)\big)$, $\forall M > 0, \exists \alpha > 0$ such that $\Big(r - \min_{k \leq K}(n_k)\Big)$ eigenvalues are greater than M.

Proof. For simplicity, we will note $\sqrt{M}$ the *principal square root* of a symmetric positive-semidefinite matrix M. The eigenvalues of $\sqrt{M}$ are the square root of the eigenvalues of M and are its *singular values*. By definition, the *singular values* of $\sqrt{(\nabla^2 L)(\theta)} D_\alpha$ are the square root of the eigenvalues of $D_\alpha (\nabla^2 L)(\theta) D_\alpha$. Without loss of generality, we consider $\min_{k \leq K}(n_k) = n_K$ and choose $\forall k < K, \alpha_k = \beta^{-1}$ and $\alpha_K = \beta^{K-1}$. Since D_α and $\sqrt{(\nabla^2 L)(\theta)}$ are positive symmetric semi-definite matrices, we can apply the multiplicative Horn inequalities (Klyachko, 2000) on singular values of the product $\sqrt{(\nabla^2 L)(\theta)} D_\alpha$:

$$\forall i \leq n, j \leq (n - n_K),$$
$$\lambda_{i+j-n}\big((\nabla^2 L)(\theta) D_\alpha^2\big) \geq \lambda_i\big((\nabla^2 L)(\theta)\big)\beta^2.$$

By choosing $\beta > \sqrt{\frac{M}{\lambda_r\big((\nabla^2 L)(\theta)\big)}}$, since we have $\forall i \leq r, \lambda_i\big((\nabla^2 L)(\theta)\big) \geq \lambda_r\big((\nabla^2 L)(\theta)\big) > 0$ we can conclude that

$$\begin{aligned} &\forall i \leq (r - n_K), \\ &\lambda_i\big((\nabla^2 L)(\theta) D_\alpha^2\big) \geq \lambda_{i+n_k}\big((\nabla^2 L)(\theta)\big)\beta^2 \\ &\qquad\qquad \geq \lambda_r\big((\nabla^2 L)(\theta)\big)\beta^2 > M. \end{aligned}$$

□

It means that there exists an observationally equivalent parameter with at least $\Big(r - \min_{k \leq K}(n_k)\Big)$ arbitrarily large eigenvalues. Since Sagun et al. (2016) seems to suggests that rank deficiency in the Hessian is due to over-parametrization of the model, one could conjecture that $\Big(r - \min_{k \leq K}(n_k)\Big)$ can be high for thin and deep neural networks, resulting in a majority of large eigenvalues. Therefore, it would still be possible to obtain an equivalent parameter with large Hessian eigenvalues, i.e. sharp in multiple directions.

References

Chaudhari, Pratik, Choromanska, Anna, Soatto, Stefano, LeCun, Yann, Baldassi, Carlo, Borgs, Christian, Chayes, Jennifer, Sagun, Levent, and Zecchina, Riccardo. Entropy-sgd: Biasing gradient descent into wide valleys. In *ICLR'2017, arXiv:1611.01838*, 2017.

Gonen, Alon and Shalev-Shwartz, Shai. Fast rates for empirical risk minimization of strict saddle problems. *arXiv preprint arXiv:1701.04271*, 2017.

Hardt, Moritz, Recht, Ben, and Singer, Yoram. Train faster, generalize better: Stability of stochastic gradient descent. In Balcan, Maria-Florina and Weinberger, Kilian Q. (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1225–1234. JMLR.org, 2016. URL http://jmlr.org/proceedings/papers/v48/hardt16.html.

Klyachko, Alexander A. Random walks on symmetric spaces and inequalities for matrix spectra. *Linear Algebra and its Applications*, 319(1-3):37–59, 2000.

Rezende, Danilo Jimenez and Mohamed, Shakir. Variational inference with normalizing flows. In Bach, Francis R. and Blei, David M. (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1530–1538. JMLR.org, 2015. URL http://jmlr.org/proceedings/papers/v37/rezende15.html.

Sagun, Levent, Bottou, Léon, and LeCun, Yann. Singularity of the hessian in deep learning. *arXiv preprint arXiv:1611.07476*, 2016.