Loss Landscape Characterization of Neural Networks without Over-Parametrziation

Rustem Islamov[1] Niccoló Ajroldi[2] Antonio Orvieto[2,3,4] Aurelien Lucchi[1]

[1]University of Basel [2]Max Planck Institute for Intelligent Systems [3]ELLIS Institute Tübingen [4]Tübingen AI Center

Problem Formulation

We want to solve the finite-sum optimization problem

$$\min_{x\in\mathbb{R}^d}\left\{f(x)\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}f_i(x)\right\}\quad(1)$$

$f^*=\min_x f(x)$ non-convex; # of data points; $f_i^*=\min_x f_i(x)$; # model parameters; Empirical risk/loss; Loss associated with one data point; $\mathcal{S}$ is the set of global minimizers

Limitations of Existing Conditions

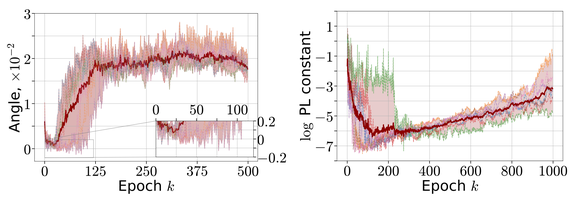


Figure 1:Training of 3 layer LSTM model that shows Aiming condition does not always hold since the angle $\angle(\nabla f(x^k), x^k-x^K)$ can be negative. The right figure demonstrates that the possible constant μ in PL condition should be small.

- **Necessity of Over-parameterization.** The theoretical justification of conditions such as Aiming [2] and PL [3] require a significant amount of overparameterization.
- **Necessity of Invexity.** The conditions imply that any stationary point is a global minimum (i.e., exclusion of saddle points and local minima).
- **Lack of Theory.** Several works have studied the empirical properties of the loss landscape of neural networks but fall short of providing theoretical explanations for this observed phenomenon.
- **Lack of Empirical Evidence.** Several theoretical works prove results on the loss landscape without supporting their claims using experimental validation on deep learning benchmarks.

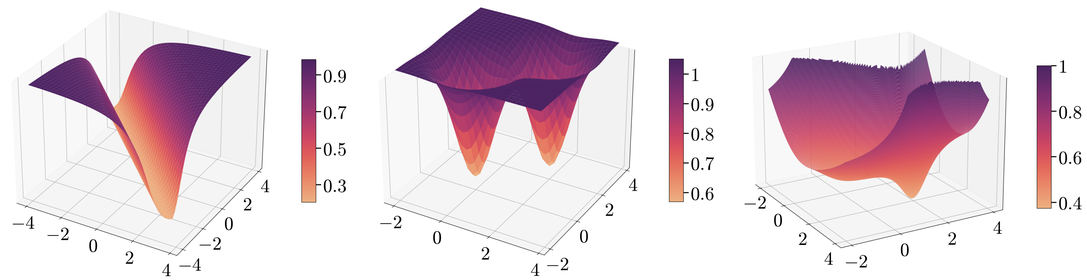
Main Contributions

- We introduce the α-β-condition and theoretically demonstrate its applicability to a wide range of complex functions, notably those that include local saddle points and local minima.
- We empirically validate that the α-β-condition is a meaningful assumption that captures a wide range of practical functions, including matrix factorization and neural networks (ResNet, LSTM, GNN, Transformer, and other architectures).
- We analyze the theoretical convergence of several optimizers under α-β-condition, including vanilla SGD, $\text{SPS}_{\max}$, and NGN.
- We provide empirical and theoretical counter-examples where the weakest assumptions, such as the PL and Aiming conditions, do not hold, but the α-β-condition does.

Table 1:Summary of existing assumptions on the optimization problem and their limitations. Here $\mathcal{S}$ denotes the set of minimizers of f and $f_i^*:=\text{argmin}_x f_i(x)$.

Condition	Definition	Comments
QCvx [1]	$\langle\nabla f(x), x-x^*\rangle\geq\theta(f(x)-f(x^*))$ for some fixed $x^*\in\mathcal{S}$	- excludes saddle points
Aiming [2]	$\langle\nabla f(x), x-\text{Proj}(x,\mathcal{S})\rangle\geq\theta f(x)$	- excludes saddle points - in theory requires over-parameterization [2] - does not always hold in practice [Fig. 1 a-b]
PL [3]	$\|\nabla f(x)\|^2\geq 2\mu(f(x)-f^*)$	- excludes saddle points - in theory requires over-parameterization [4] - does not always hold in practice [Fig. 1 c-d]
α-β-condition **[This work]**	$\langle\nabla f_i(x), x-\text{Proj}(x,\mathcal{S})\rangle\geq\alpha(f_i(x)-f_i(\text{Proj}(x,\mathcal{S})))-\beta(f_i(x)-f_i^*)$	- might have saddles and local minima [Fig. 2 (b-c)] - in practice does not require over-parameterization [2 layer NN ex.]

The Proposed Condition and Examples



(a) $f_1(x)=\frac{(x+y)^2}{1+(x+y)^2}$, $f_2(x)=\frac{(x+y+1)^2}{1+(x+y+1)^2}$ (b) $f_1(x)=1-e^{-x^2-y^2}$, $f_2(x)=1-e^{-(x-2)^2-(y-2)^2}$ (c) $f_1(x)=\frac{1+x^2+y^2}{4+x^2+y^2}$, $f_2(x)=\frac{(x-2.5)^2+(y-2.5)^2}{4+(x-2.5)^2+(y-2.5)^2}$

Figure 2:Loss landscape of f that satisfy α-β-condition. These examples demonstrate that the problem (1) that satisfies α-β-condition might have an unbounded set of minimizers $\mathcal{S}$ (**left**), a saddle point (**center**), and local minima (**right**) in contrast to the PL and Aiming conditions.

Definition of α-β-condition

Let $\mathcal{X}\subseteq\mathbb{R}^d$ be a set and consider a function $f\colon\mathcal{X}\to\mathbb{R}$ as defined in (1). Then f satisfies the α-β-condition with positive parameters α and β such that $\alpha>\beta$ if for any $x\in\mathcal{X}$ there exists $x_p\in\text{Proj}(x,\mathcal{S})$ such that for all $i\in[n]$

$$\langle\nabla f_i(x), x-x_p\rangle\geq\alpha(f_i(x)-f_i(x_p))-\beta(f_i(x)-f_i^*).$$

Matrix Factorization. Let f, f_{ij} be such that

$$f(W,S)=\frac{1}{2nm}\|X-W^\top S\|_{\text{F}}^2=\frac{1}{2nm}\sum_{i,j}(X_{ij}-w_i^\top s_j)^2,$$

$$f_{ij}(W,S)=\frac{1}{2}(X_{ij}-w_i^\top s_j)^2,$$

where $X\in\mathbb{R}^{n\times m}, W=(w_i)_{i=1}^n\in\mathbb{R}^{k\times n}, S=(s_j)_{j=1}^m\in\mathbb{R}^{k\times m}$, and $\text{rank}(X)=r\geq k$. We assume that X is generated using matrices W^* and S^* with non-zero additive noise that minimize empirical loss, namely, $X=(W^*)^\top S^*+(\varepsilon_{ij})_{i\in[n],j\in[m]}$ where $W^*,S^*=\text{argmin}_{W,S}f(W,S)$. Let $\mathcal{X}$ be any bounded set that contains $\mathcal{S}$. Then α-β-condition is satisfied with $\alpha=\beta+1$ and some $\beta>0$.

Two Layer Neural Network. Consider training a two-layer neural network with a logistic loss

$$f(W,v)=\frac{1}{n}\sum_{i=1}^n f_i(W,v),\quad f_i(W,v)=\phi(y_i\cdot v^\top\sigma(Wx_i))$$

for a classification problem where $\phi(t):=\log(1+\exp(-t))$, $W\in\mathbb{R}^{k\times d}, v\in\mathbb{R}^k$, σ is a ReLU function applied coordinate-wise, $y_i\in\{-1,+1\}$ is a label and $x_i\in\mathbb{R}^d$ is a feature vector. Let $\mathcal{X}$ be any bounded set that contains $\mathcal{S}$. Then the α-β-condition holds in $\mathcal{X}$ for some $\alpha\geq1$ and $\beta=\alpha-1$.

Convergence under α-β-condition

Theorem. Assume that each f_i is L-smooth and the interpolation error $\sigma_{\text{int}}^2:=\mathbb{E}[f^*-f_i^*]$ is bounded. Then the iterates of SGD with stepsize $\gamma\leq\frac{\alpha-\beta}{2L}$ satisfy

$$\min_{0\leq k<K}\mathbb{E}[f(x^k)-f^*]\leq\frac{\mathbb{E}[\text{dist}(x^0,\mathcal{S})^2]}{K}\frac{1}{\gamma(\alpha-\beta)}+\frac{2L\gamma}{\alpha-\beta}\sigma_{\text{int}}^2+\frac{2\beta}{\alpha-\beta}\sigma_{\text{int}}^2.$$

Empirical Verification

	Model's width ↗	Model's depth ↗	Batch size ↗
Change in $\beta\sigma_{\text{int}}^2$	↘	↘	↘

Table 2:Summary of how the non-vanishing term $\beta\sigma_{\text{int}}^2$ changes as a function of specific quantities of interest.

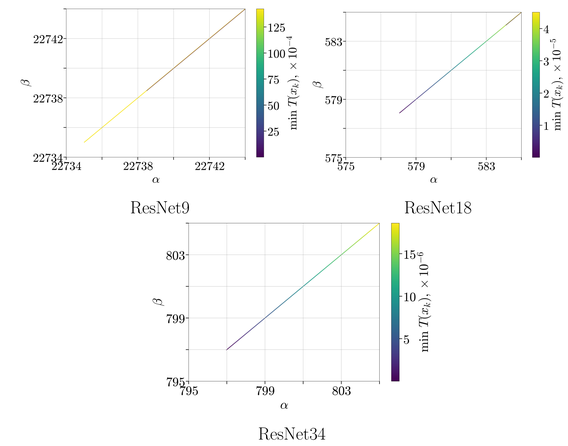
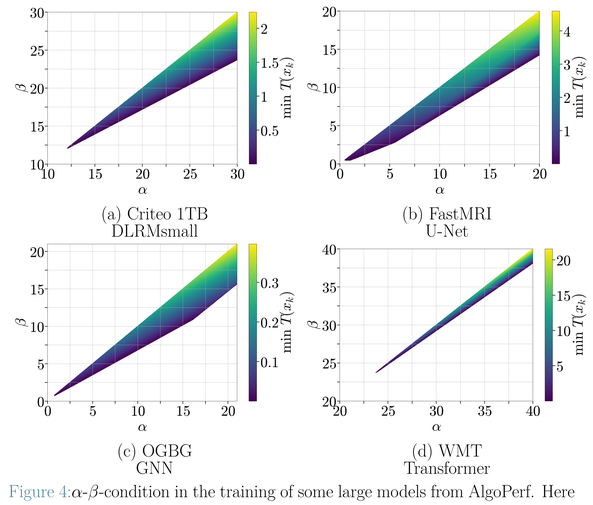


Figure 3:Training of ResNet models on CIFAR100 dataset. Here $T(x_k)=\langle\nabla f_{i_k}(x^k), x^k-x^K\rangle-\alpha(f_{i_k}(x^k)-f_{i_k}(x^K))-\beta f_{i_k}(x^k)$ assuming that $f_i^*=0$. Minimum is taken across all runs and iterations for a given pair of (α,β).



(a) Criteo 1TB DLRMsmall (b) FastMRI U-Net (c) OGBG GNN (d) WMT Transformer

Figure 4:α-β-condition in the training of some large models from AlgoPerf. Here $T(x_k)=\langle\nabla f_{i_k}(x^k), x^k-x^K\rangle-\alpha(f_{i_k}(x^k)-f_{i_k}(x^K))-\beta f_{i_k}(x^k)$ assuming that $f_i^*=0$. Minimum is taken across all runs and iterations for a given pair of (α,β).

References

[1] Hardt et al., Gradient descent learns linear dynamical systems. JMLR, 2018.

[2] Liu et al., Aiming towards the minimizers: fast convergence of SGD for overparametrized problems. NeurIPS, 2023.

[3] Polyak. Gradient methods for the minimisation of functionals. USSR Comp. Math. and Math. Phys., 1963.

[4] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in overparameterized non-linear systems and neural networks. Appl. and Comp. Harm. Analysis, 2022.