Friedrich-Alexander-Universität
Naturwissenschaftliche Fakultät

Friedrich-Alexander-Universität
DYNAMICS, CONTROL, MACHINE LEARNING AND NUMERICS

ON THE GEOMETRY OF SHARP MINIMIZERS

Alberto Domínguez Corella

Chair in Dynamics, Control, and Numerics, FAU Erlangen Nuernberg.
Institut de Mathématiques de Jussieu, Sorbonne Université.

Introduction

A sharp minimizer is a point where a function not only reaches its minimum but also has a steep slope away from that point. This means that the function increases quickly as you move away from the minimizer, making the minimum "sharp". The sharpness implies that, for the minimizer. small changes or perturbations in the function will not shift the location of the minimizer.

The central questions addressed are: Can we always fit a cone below a function if small perturbations do not change the minimizer? Moreover, how can we characterize the behavior of sharp minimizers under both linear and nonlinear perturbations? We answer these questions by introducing key assumptions and results that extend the classical understanding of sharp minimizers.

Additionally, we draw parallels with deep learning, where similar concepts arise. Despite their expressive power, neural networks often generalize well, avoiding bad minima even in high-dimensional spaces. We discuss how the volume disparity between flat and sharp minima helps explain why stochastic optimizers tend to favor solutions that generalize to unseen data.

Sharp minimizers and infinitesimal tilting

Let $(X, \|\cdot\|)$ be a real normed space and $f : X \to \mathbb{R} \cup \{+\infty\}$ a proper function. We fix a reference point $\bar{x} \in X$.
We say that $\bar{x} \in X$ of f is a sharp minimizer if there exists $\alpha > 0$ such that

$$f(x) \geq f(\bar{x}) + \alpha\|x - \bar{x}\| \quad \forall x \in X. \tag{1}$$

The best constant in (1) is called the sharpness modulus. Intuitively, we can think of the function as lying above a cone with vertex at the minimizer.
Geometrically, if a function is bounded below by a cone, infinitesimal tilting will not affect the minimizer. The following proposition formalizes this intuition. Although its proof is not complicated, we will provide it in the next section.

Proposition 1. Let $\bar{x} \in X$ be a sharp minimizer with modulus $\alpha > 0$. Then, for any $\xi \in X^*$ with $\|\xi\| < \alpha$,

$$x \in \operatorname*{arg\,min}_{y\in X}\big\{f(y) - \langle \xi, y\rangle\big\} \quad \Longrightarrow \quad x = \bar{x}.$$

An interesting question arises, if infinitesimal tilting holds, i.e., if infinitesimally small perturbations do not change the minimizer, can we fit a cone below the function?

If there are no perturbations yielding minimizers, e.g., $f(x) = \|x\|^{\frac{1}{2}}$, then the answer is negative. We can circumvent this issue by assuming that there are enough perturbations yielding minimizers.

Assumption 1. Assume that there exists a dense set $\Theta \subset X^*$ such that

$$\operatorname*{arg\,min}_{y\in X}\big\{f(y) - \langle \xi, y\rangle\big\} \neq \emptyset \quad \forall \xi \in \Theta.$$

Sharp minimizers and infinitesimal tilting

Proposition 2. Let Assumption 1 be fulfilled. Suppose there exists $\alpha > 0$ such that, for any $\xi \in X^*$ with $\|\xi\| < \alpha$, it holds

$$x \in \operatorname*{arg\,min}_{y\in X}\big\{f(y) - \langle \xi, y\rangle\big\} \quad \Longrightarrow \quad x = \bar{x}.$$

Then, $\bar{x}$ is a sharp minimizer with modulus α.

We can strengthen this result and prove that, under Assumption 1, a sharp minimizer is characterized by the property that all sufficiently small perturbations yield a minimizer, and this minimizer coincides with the original one.

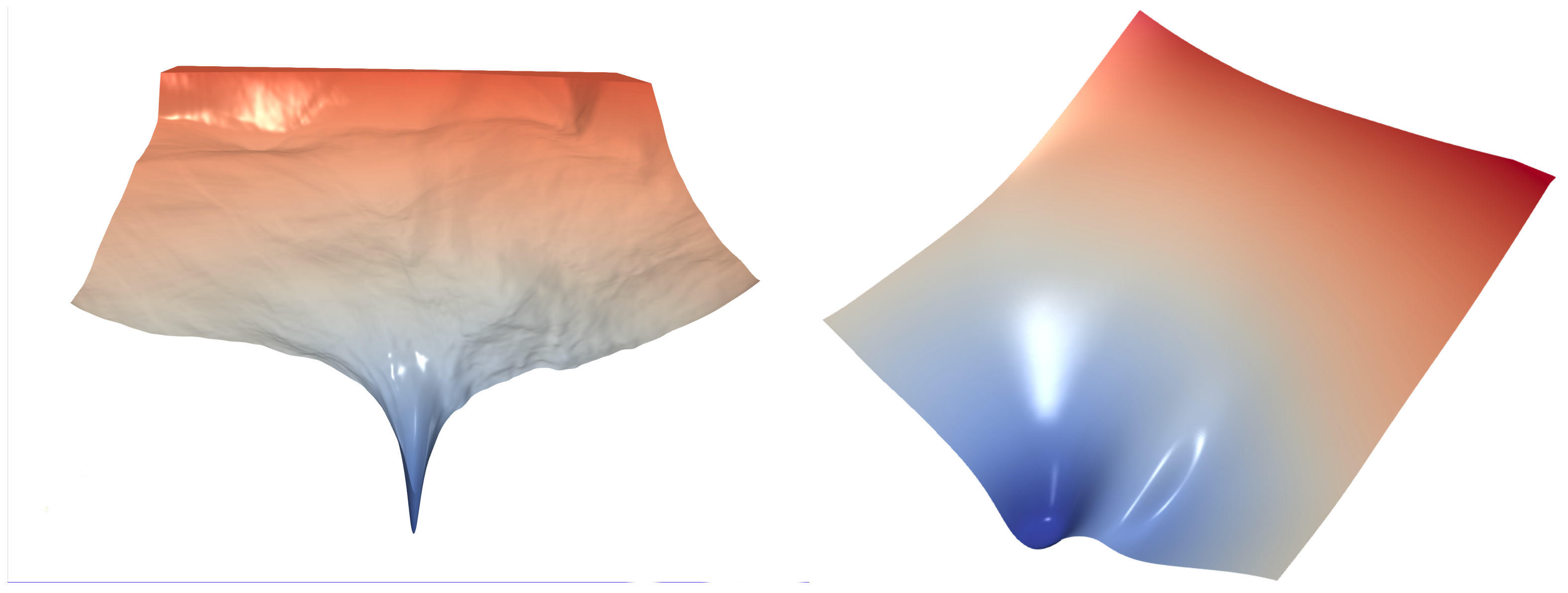
Theorem 1. Let Assumption 1 be fulfilled. Let $\alpha > 0$. The following statements are equivalent.
(i) $\bar{x}$ is a sharp minimizer with modulus α.
(ii) $\operatorname{arg\,min}_{y\in X}\big\{f(y) - \langle \xi, y\rangle\big\} = \{\bar{x}\}$ for every $\xi \in X^*$ satisfying $\|\xi\| < \alpha$.

One can then think about other type of perturbations besides linear ones. It turns out that it is enough to study linear ones to understand nonlinear ones.

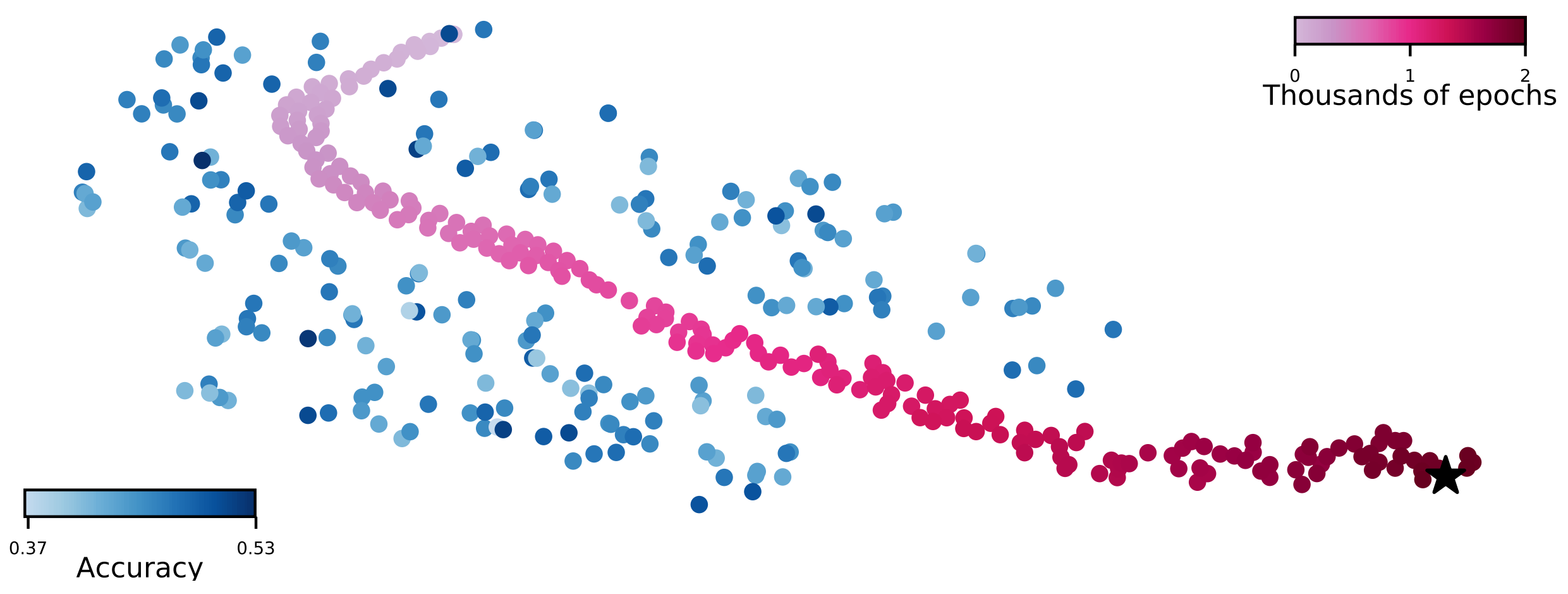
Theorem 2. Let Assumption 1 be fulfilled. The following statements are equivalent.
(i) $\operatorname*{arg\,min}_{y\in X}\big\{f(y) - \langle \xi, y\rangle\big\} = \{\bar{x}\}$ for every $\xi \in X^*$ satisfying $\|\xi\| < \alpha$.
(ii) $\operatorname*{arg\,min}_{y\in X}\big\{f(y) + \zeta(y)\big\} = \{\bar{x}\}$ for every Lipschitz function $\zeta : X \to \mathbb{R}$ with $\operatorname{Lip}\zeta < \alpha$.

A few word on Deep Learning

The strength of neural networks lies in their ability to generalize to unseen data, yet the reasons behind this remain elusive. Despite numerous rigorous efforts to explain generalization, existing bounds are often loose, and the analysis doesn't always provide a deep understanding.



Why is generalization a mystery? The generalization ability of neural networks seems paradoxical given their expressiveness. Training algorithms minimize a loss function based solely on the performance over training data. Due to their flexibility, neural networks can find parameter configurations that perfectly fit the training data but perform poorly on test data. Yet, remarkably, commonly used optimizers consistently avoid such "bad" minima and instead find "good" minima that generalize well.



The volume disparity between flat and sharp minima promotes generalization. The bias of stochastic optimizers toward good minima can be attributed to the volume disparity between the attraction basins of good and bad minima. Flat minima that generalize well have wide attraction basins, occupying a large portion of the parameter space. In contrast, sharp minima have narrow basins with comparatively small volumes. Consequently, a random search is more likely to land in the basin of a good minimizer than a bad one.

This volume disparity between good and bad minima is significantly amplified by the curse (or perhaps the blessing) of dimensionality. While the difference in width between good and bad basins may seem modest in 2D visualizations, in high-dimensional spaces, this difference translates into exponentially large disparities in volume. As a result, the likelihood of encountering a good minimizer during a random search grows dramatically in high-dimensional parameter spaces, where small differences in sharpness lead to vast differences in the basin volumes.

Related publications

[1] Domínguez Corella, A. **On the geometry of sharp minimizers.** To appear in Proceedings of the American Mathematical Society.

[2] Domínguez Corella, A and Le, T. **On the growth of nonconvex functionals at strict local minimizers.** To appear in SIAM Journal on Optimization.

[3] Huang, W. et al. **Understanding generalization through visualizations.** Proceedings of Machine Learning Research.

https://dcn.nat.fau.eu/ 09/2024

Department MATHEMATIK

Unterstützt von / Supported by
Alexander von Humboldt Stiftung/Foundation