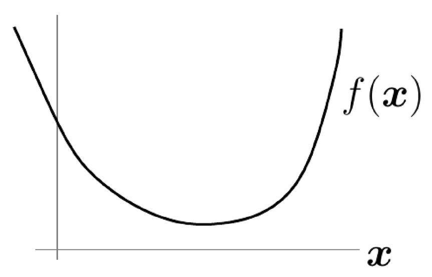
When Are Nonconvex Optimization Problems Not Scary?

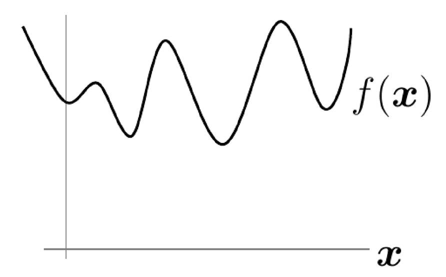
Ju Sun, Qing Qu, and John Wright, Department of Electrical Engineering, Columbia University

Nonconvex Optimization at Large

$\min f(\boldsymbol{x})$
s.t. $\boldsymbol{x} \in \mathcal{D}$



"Easy"



"Hard"

Nonconvex Optimization Problems

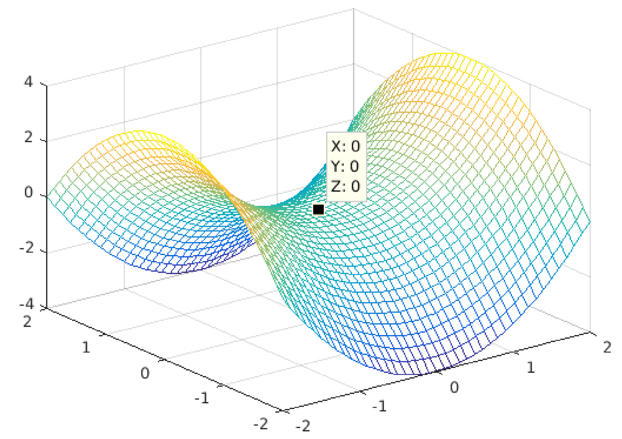
In theory: NP-hard even to compute a local minimizer
In practice: Heuristics such as gradient descent and alternating minimization often highly successful

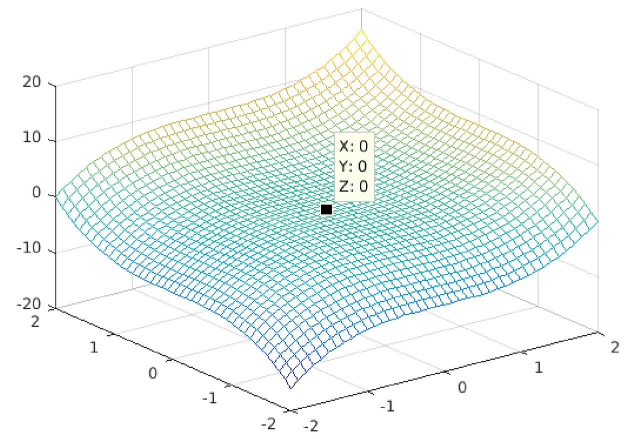
Which Nonconvex Optimization Problems are "Easy"?

We will focus on the family that **all local minimizers are equally "nice"** – focus on escaping local maximizers and saddle points and finding a local minimizer

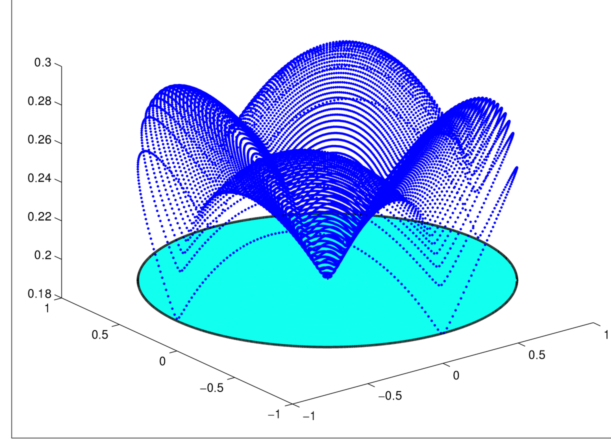
- Any local maximizer is strict at least in one direction – easy to escape
- **Any saddle points are second-order (strict or ridable saddle) – second-order info can be used for escaping**

Not all saddles are second-order (ridable) – i.e., with indefinite Hessian!





One example from the family...



Quantitative Characterization: Ridable (Strict-Saddle) Functions

$$\operatorname{minimize} f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{x} \in \mathcal{M}.$$

A function f over manifold $\mathcal{M}$ is $(\alpha, \beta, \gamma, \delta)$-ridable ($\alpha, \beta, \gamma, \delta$ strictly positive) if any point $\boldsymbol{x} \in \mathcal{M}$ obeys at least one of the following: ($T_{\boldsymbol{x}}\mathcal{M}$ is the tangent space of $\mathcal{M}$ at point $\boldsymbol{x}$)
1) [**Negative curvature**] There exists $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{M}$ with $\|\boldsymbol{v}\| = 1$ such that $\langle \operatorname{Hess} f(\boldsymbol{x})[\boldsymbol{v}], \boldsymbol{v}\rangle \le -\alpha$;
2) [**Strong gradient**] $\|\operatorname{grad} f(\boldsymbol{x})\| \ge \beta$;
3) [**Strong convexity around minimizers**] There exists a local minimizer $\boldsymbol{x}_\star$ such that $\|\boldsymbol{x} - \boldsymbol{x}_\star\| \le \delta$, and for all $\boldsymbol{y} \in \mathcal{M}$ that is in 2δ neighborhood of $\boldsymbol{x}_\star$, $\langle \operatorname{Hess} f(\boldsymbol{y})[\boldsymbol{v}], \boldsymbol{v}\rangle \ge \gamma$ for any $\boldsymbol{v} \in T_{\boldsymbol{y}}\mathcal{M}$ with $\|\boldsymbol{v}\| = 1$, i.e., the function f is γ-strongly convex in 2δ neighborhood of $\boldsymbol{x}_\star$.

Example I: Eigenvector Problem [Classic]

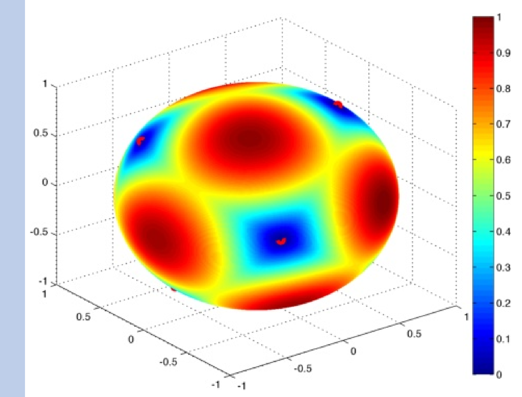
For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$,

$$\operatorname{maximize}_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \quad \text{subject to} \quad \|\boldsymbol{x}\| = 1.$$

- Critical points: $\{\pm \boldsymbol{v}_i\}$
- Suppose $\lambda_1 > \lambda_2 \ge \dots \lambda_{n-1} > \lambda_n$. The only global minimizers are $\pm\boldsymbol{v}_n$, and all $\{\pm\boldsymbol{v}_i\}$ for $2 \le i \le n-1$ are second-order saddle points.
- Quantitatively, the function is $(c(\lambda_{n-1}-\lambda_n), c(\lambda_{n-1}-\lambda_n)/\lambda_1, c(\lambda_{n-1}-\lambda_n), 2c(\lambda_{n-1}-\lambda_n)/\lambda_1)$-ridable over $\mathbb{S}^{n-1}$.

Example II: Sparse (Complete) Dictionary Learning [Sun et al'15]

Given $\boldsymbol{Y}$, find $(\boldsymbol{A}, \boldsymbol{X})$ such that $\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{X}$, with $\boldsymbol{X}$ as sparse as possible.
When $\boldsymbol{A}$ square, invertible, $\operatorname{row}(\boldsymbol{Y}) = \operatorname{row}(\boldsymbol{X}) \Longrightarrow$ Finding sparse vectors in $\operatorname{row}(\boldsymbol{Y})$

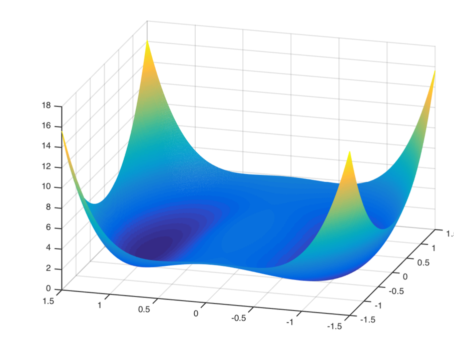


$$\operatorname{minimize} f(\boldsymbol{q}) \doteq \frac{1}{p}\sum_{k=1}^{p} h(\boldsymbol{q}^\top \overline{\boldsymbol{y}}_k) \quad \text{subject to} \quad \|\boldsymbol{q}\|_2 = 1. \quad (h \text{ promotes sparsity})$$

- Trivial equivalence due to sign and permutation
- Under probability model on the true $\boldsymbol{X}$, f is $(c\theta, c\theta, c\theta/\mu, \sqrt{2}\mu/7)$-ridable over $\mathbb{S}^{n-1}$

Example III: Generalized Phase Retrieval [Sun et al'15]

Recover $\boldsymbol{x}$ from nonlinear measurements $|\boldsymbol{a}_i^*\boldsymbol{x}|^2$, $i = 1, \dots, m$, encountered in many optic imaging systems



$$\operatorname{minimize}_{\boldsymbol{z}\in\mathbb{C}^n} f(\boldsymbol{z}) \doteq \frac{1}{4m}\sum_{k=1}^{m}(y_k - |\boldsymbol{a}_k^*\boldsymbol{z}|^2)^2.$$

- Trivial equivalence by global phase offset
- The function f is $(c, c/(n\log m), c, c/(n\log m))$-ridable

Example IV: Tensor Decomposition and ICA [Ge et al' 15]

Given orthogonal decomposable d-th order tensors $\mathcal{T}$, i.e., $\mathcal{T} = \sum_{i=1}^{r} \boldsymbol{a}_i^{\otimes d}$, $\boldsymbol{a}_i^\top \boldsymbol{a}_j = \delta_{ij}\ \forall\, i, j$, ($\boldsymbol{a}_i \in \mathbb{R}^n\ \forall$, find $\boldsymbol{a}_i$'s up to sign and permutation.

$$\operatorname{minimize} g(\boldsymbol{u}_1, \dots, \boldsymbol{u}_r) \doteq \sum_{i\ne j} \mathcal{T}(\boldsymbol{u}_i, \boldsymbol{u}_i, \boldsymbol{u}_j, \boldsymbol{u}_j) = \sum_{i\ne j}\sum_{k=1}^{r} (\boldsymbol{a}_k^\top \boldsymbol{u}_i)^2 (\boldsymbol{a}_k^\top \boldsymbol{u}_j)^2, \quad \text{subject to} \quad \|\boldsymbol{u}_i\| = 1\ \forall i \in [r].$$

All local minimizers of g are equivalent (i.e., signed permuted) copies of $[\boldsymbol{a}_1, \dots, \boldsymbol{a}_r]$. Moreover, g is $(1/\operatorname{poly}(r), 1/\operatorname{poly}(r), 1, 1/\operatorname{poly}(r))$-ridable.

Algorithm: Second-order Trust-region Method

Iteratively construct approximation of the form

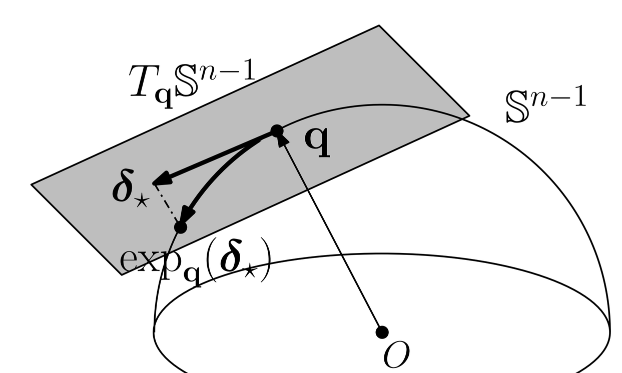
$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}^{(k)}) \doteq f(\boldsymbol{x}^{(k)}) + \left\langle \boldsymbol{\delta}, \operatorname{grad} f(\boldsymbol{x}^{(k)})\right\rangle + \frac{1}{2}\left\langle \operatorname{Hess} f(\boldsymbol{x}^{(k)})[\boldsymbol{\delta}], \boldsymbol{\delta}\right\rangle$$

The next iterate determined by minimizing the quadratic approximation within a small radius Δ:

$$\boldsymbol{\delta}^{(k+1)} \doteq \operatorname*{arg\,min}_{\boldsymbol{\delta}\in T_{\boldsymbol{x}^{(k)}}\mathcal{M}, \|\boldsymbol{\delta}\|_2 \le \Delta} \widehat{f}\left(\boldsymbol{\delta}; \boldsymbol{x}^{(k)}\right).$$

The next iterate is obtained by retracting the resulting vector back to the manifold:

$$\boldsymbol{x}^{(k+1)} = R_{\boldsymbol{x}^{(k)}}(\boldsymbol{x}^{(k)} + \boldsymbol{\delta}^{(k+1)}).$$



References

Sun et al. When Are Nonconvex Problems Not Scary?. arXiv:1510.06096
Sun et al. Complete Dictionary Learning over the Sphere.. arXiv:1504.06785
Sun et al. A Geometric Analysis of Phase Retrieval.
Ge et al. Escaping From Saddle Points – Online Stochastic Gradient for Tensor Decomposition.
arXiv:1503.02101

Email: <js4038, qq2105, jw2966>@columbia.edu