The Landscape of Non-convex Empirical Risk with Degenerate Population Risk

Shuang Li, Gongguo Tang, and Michael B. Wakin
Electrical Engineering, Colorado School of Mines

Motivation

What if the Hessian is degenerate?

• S. Mei, et al., "The landscape of empirical risk for non-convex losses", 2016.

Assumptions

1. In $\overline{\mathcal{D}} \triangleq \{\boldsymbol{x}\in\mathcal{B}(l): \|\text{grad } g(\boldsymbol{x})\|_2 \le \epsilon\}$:

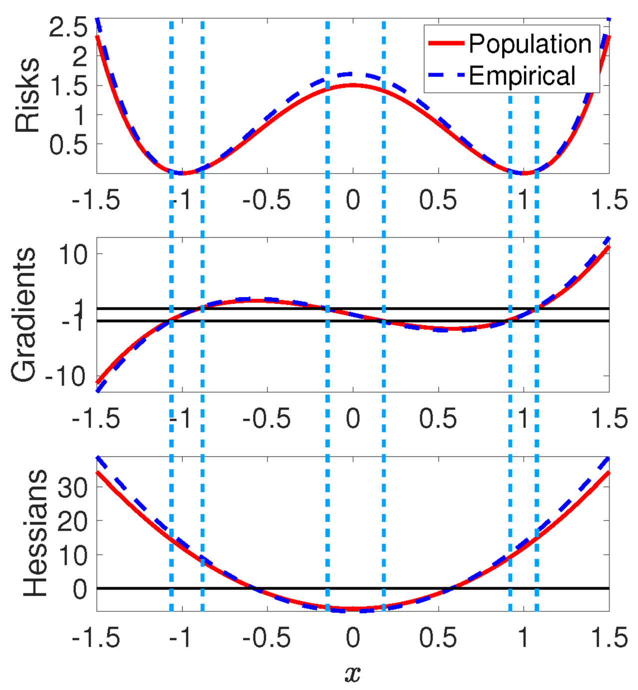
$$|\lambda_{\min}(\text{hess } g(\boldsymbol{x}))| \ge \eta$$

2. Gradient proximity:

$$\sup_{\boldsymbol{x}\in\mathcal{B}(l)} \|\text{grad } f(\boldsymbol{x}) - \text{grad } g(\boldsymbol{x})\|_2 \le \frac{\epsilon}{2}$$

3. Hessian proximity:

$$\sup_{\boldsymbol{x}\in\mathcal{B}(l)} \|\text{hess } f(\boldsymbol{x}) - \text{hess } g(\boldsymbol{x})\|_2 \le \frac{\eta}{2}$$



Phase retrieval with $N=1$, $\boldsymbol{x}^\star=1$, and $M=30$. $g(x)=\frac{3}{2}(x^2-1)^2$. $f(x)=\frac{1}{2M}\sum_{m=1}^M a_m^4(x^2-1)^2$.

Main Theorem

Theorem

Denote f and g as the empirical risk and the population risk. Let $\mathcal{D}$ be a maximal connected and compact subset of $\overline{\mathcal{D}}$. With the above assumptions, we have

(a) $\mathcal{D}$ contains at most one local minimum of g. If g has $K (K=0,1)$ local minima in $\mathcal{D}$, then f also has K local minima in $\mathcal{D}$.

(b) If g has strict saddles in $\mathcal{D}$, then if f has any critical points in $\mathcal{D}$, they must be strict saddle points.

Local Minima Distance

Corollary

$\{\widehat{\boldsymbol{x}}_k\}_{k=1}^K, \{\boldsymbol{x}_k\}_{k=1}^K$: local minima of f and g. $\mathcal{D}_k$: maximal connected and compact subset of $\overline{\mathcal{D}}$ containing $\boldsymbol{x}_k$ and $\widehat{\boldsymbol{x}}_k$. ρ: *injectivity radius* of $\mathcal{M}$. Suppose the pre-image of $\mathcal{D}_k$ under the exponential mapping $\text{Exp}_{\boldsymbol{x}_k}(\cdot)$ is contained in the ball at the origin of $\mathcal{T}_{\boldsymbol{x}_k}\mathcal{M}$ with radius ρ. Suppose the pullback of g onto $\mathcal{T}_{\boldsymbol{x}_k}\mathcal{M}$ has Lipschitz Hessian with constant L_H at the origin. Then as long as $\epsilon \le \frac{\eta^2}{2\sigma L_H}$, we have

$$\text{dist}(\widehat{\boldsymbol{x}}_k, \boldsymbol{x}_k) \le 2\sigma\epsilon/\eta, \quad 1\le k\le K.$$

Matrix Sensing

▸ Empirical risk: ($\mathbf{U}\in\mathbb{R}^{N\times k}$, $\mathbf{X}\in\mathbb{R}^{N\times N}$ is PSD with rank r)

$$f(\mathbf{U}) = \frac{1}{4}\|\mathcal{A}(\mathbf{U}\mathbf{U}^\top - \mathbf{X})\|_2^2$$

▸ Population risk:

$$g(\mathbf{U}) = \mathbb{E}f(\mathbf{U}) = \frac{1}{4}\|\mathbf{U}\mathbf{U}^\top - \mathbf{X}\|_F^2$$

Lemma 1

▸ Assumption 1 is true by setting

$$\epsilon \le \min\{1/80, 1/60\kappa^{-1}\}\lambda_k^{\frac{3}{2}}, \ \eta = 0.06\lambda_k$$

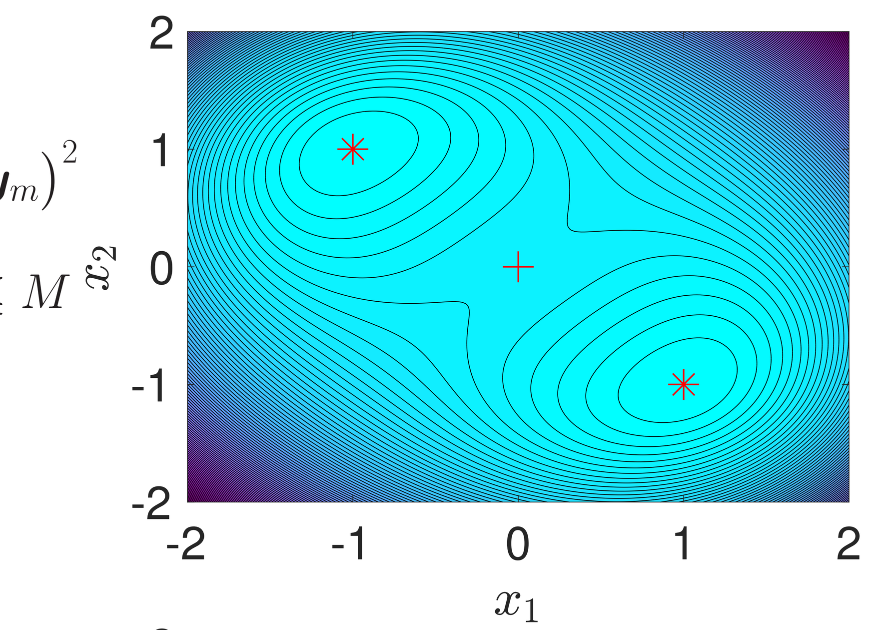
▸ Assumptions 2&3 are true if the constructed symmetric operator $\mathcal{A}$ from $\mathcal{B}$ ($\mathbf{A}_m = \frac{1}{2}(\mathbf{B}_m + \mathbf{B}_m^\top)$) satisfies the RIP:

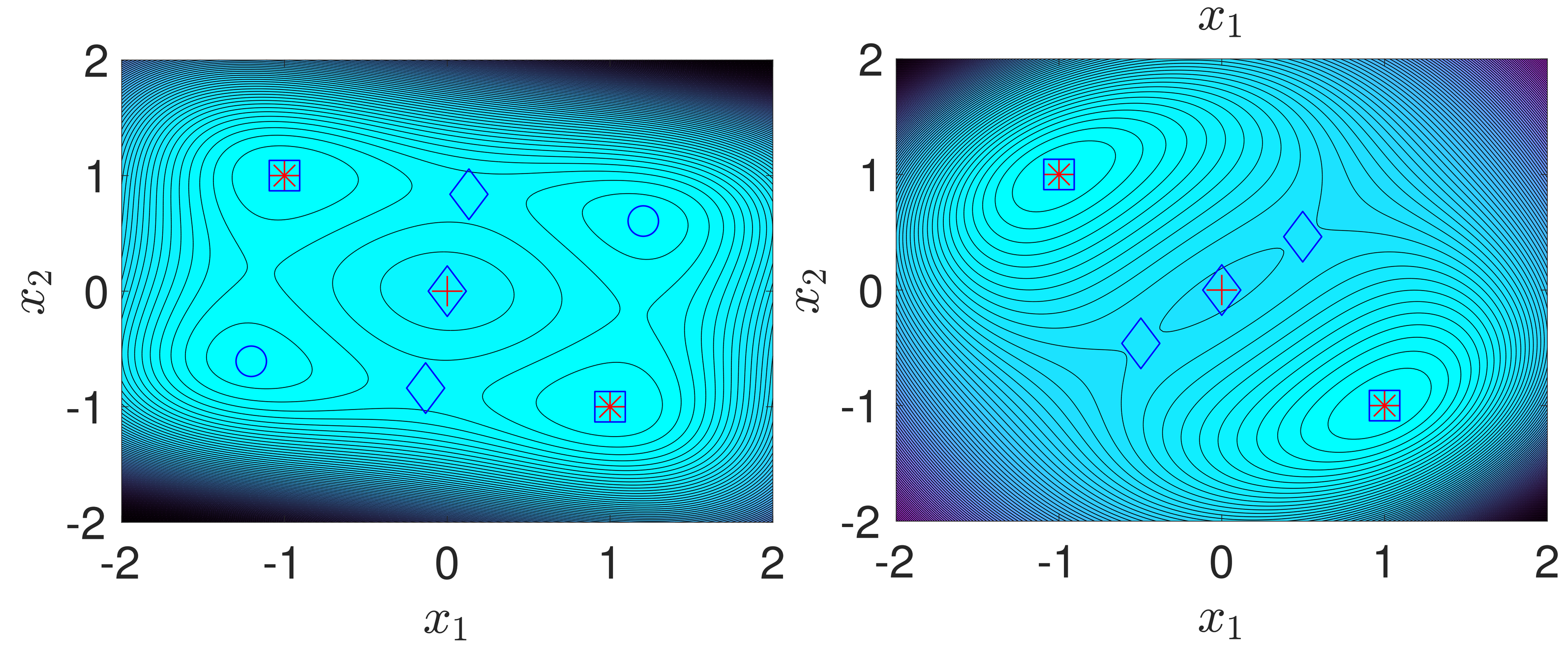
$$\delta_{r+k} \le \min\left\{\frac{\epsilon}{2\sqrt{\frac{8}{7}}k^{\frac{1}{4}}(\frac{8}{7}\|\mathbf{U}^\star\mathbf{U}^{\star\top}\|_F + \|\mathbf{X}\|_F)\|\mathbf{U}^\star\mathbf{U}^{\star\top}\|_F^{\frac{1}{2}}}, \frac{1}{36}, \frac{\eta}{2(\frac{16}{7}\sqrt{k}\|\mathbf{U}^\star\mathbf{U}^{\star\top}\|_F + \frac{8}{7}\|\mathbf{U}^\star\mathbf{U}^{\star\top}\|_F + \|\mathbf{X}\|_F)}\right\}$$

Note that the RIP holds w.h.p if $M \ge C(r+k)N/\delta_{r+k}^2$.

Rank-1 matrix sensing:

- $f(\boldsymbol{x}) = \frac{1}{4M}\sum_{m=1}^M (\langle \mathbf{A}_m, \boldsymbol{x}\boldsymbol{x}^\top\rangle - \boldsymbol{y}_m)^2$
- $\boldsymbol{y}_m = \langle \mathbf{A}_m, \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\rangle, \quad 1\le m\le M$
- $N=2, \ \boldsymbol{x}^\star = [1 \ -1]^\top$

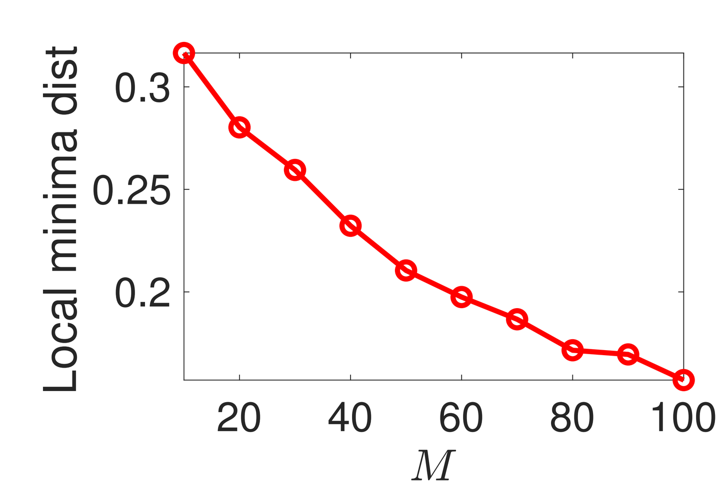




$M=3$ $M=10$

Rank-2 matrix sensing:

- $k=2, \ r=3, \ N=8$
- $\mathbf{X} = \mathbf{U}^\star\mathbf{U}^{\star\top}, \ \mathbf{U}^\star = \mathbf{I}_{N\times r}$
- Average results over 100 trials



Phase Retrieval

▸ Empirical risk: ($\boldsymbol{y}_m = |\langle \boldsymbol{a}_m, \boldsymbol{x}^\star\rangle|^2, \ 1\le m\le M, \ \boldsymbol{x}^\star\in\mathbb{R}^N$)

$$f(\boldsymbol{x}) = \frac{1}{2M}\sum_{m=1}^M \left(|\langle \boldsymbol{a}_m, \boldsymbol{x}\rangle|^2 - \boldsymbol{y}_m\right)^2$$

▸ Population risk:

$$g(\boldsymbol{x}) = \mathbb{E}f(\boldsymbol{x}) = \|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\|_F^2 + \frac{1}{2}(\|\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}^\star\|_2^2)^2$$

Lemma 2

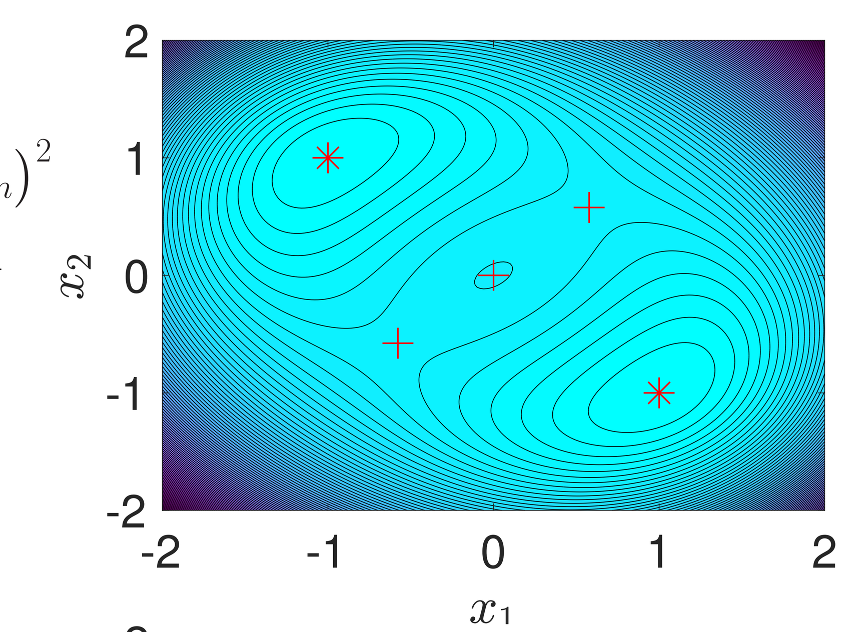
▸ Assumption 1 is true by setting

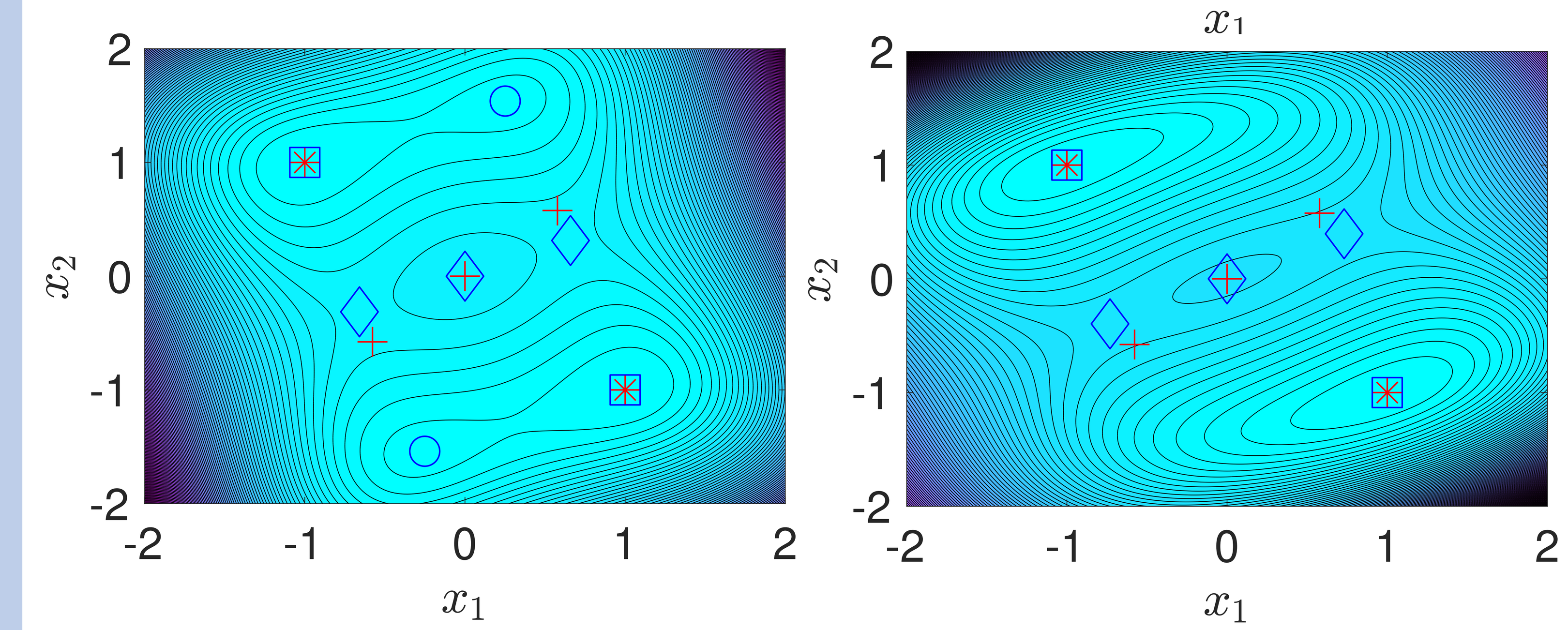
$$\epsilon \le 0.3963\|\boldsymbol{x}^\star\|_2^3, \ \eta = 0.22\|\boldsymbol{x}^\star\|_2^2$$

▸ Assumptions 2&3 hold w.h.p if $M \ge CN^2$ and $\boldsymbol{a}_m\in\mathbb{R}^N$ is a Gaussian random vector with entries following $\mathcal{N}(0,1)$.

Phase retrieval:

- $f(\boldsymbol{x}) = \frac{1}{2M}\sum_{m=1}^M \left(|\langle \boldsymbol{a}_m, \boldsymbol{x}\rangle|^2 - \boldsymbol{y}_m\right)^2$
- $\boldsymbol{y}_m = |\langle \boldsymbol{a}_m, \boldsymbol{x}^\star\rangle|^2, \quad 1\le m\le M$
- $N=2, \ \boldsymbol{x}^\star = [1 \ -1]^\top$





$M=3$ $M=10$

Conclusions

We established a correspondence between the critical points of the empirical risk and its population risk **without** the strongly Morse assumption.

Acknowledgement

This work was supported by NSF grant CCF-1704204, and the DARPA Lagrange Program under ONR/SPAWAR contract N660011824020.

email: shuangli,gtang,mwakin@mines.edu