The landscape of non-convex losses for statistical learning problems

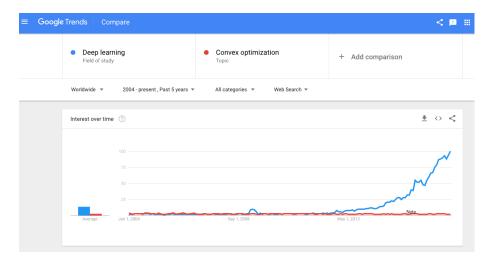
Song Mei

Stanford University

October 19, 2017

3 K K 3 K

Deep learning



Song Mei (Stanford University) The landscape of non-convex optimization October 19, 2017 2 / 32

◆ロト ◆母 ▶ ◆臣 ▶ ◆臣 ▶ ● ● ● ● ●

Deep learning

DEEP LEARNING EVERYWHERE









ancer Cell Detection Diabetic Grading Drug Discovery Video Captioning Video Search Real Time Translation Face Detection Video Surveillanc Satellite Imagen

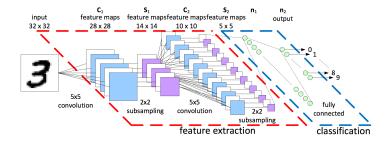
AUTONOMOUS MACHINES

Pedestrian Detection Lane Tracking Recognize Traffic Sig

イロト イポト イヨト イヨト

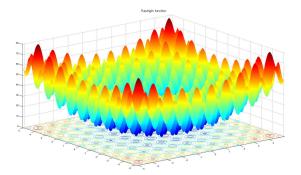
13 🥶 🚟

Convolutional Neural Network



< ロ > < 同 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ >

Non-convex optimization



< 日 > < 同 >

э

Why does non-convex neural network perform well?

Song Mei (Stanford University) The landscape of non-convex optimization October 19, 2017 5 / 32

・ロト ・ 一日 ・ ・ 日 ・ ・ 日 ・ ・ 日 ・

Why does some non-convex optimization perform well?

・ロト ・ 一日 ・ ・ 日 ・ ・ 日 ・ ・ 日

Why does some non-convex optimization perform well?

- ▶ Stochastic gradient descent escape bad local minima.
- Good initialization escape bad local minima.
- Globally there are less bad local minima.

. . . .

Non-convex optimization: analysis of global geometry

Number and locations of saddle points and local minima.

Song Mei (Stanford University) The landscape of non-convex optimization October 19, 2017 7 / 32

Let's do it!

The objective function

$$\min_{oldsymbol{W}_i}rac{1}{n}\sum_{i=1}^n \{y_i - \sigma(oldsymbol{W}_{oldsymbol{k}}\cdots\sigma(oldsymbol{W}_2\cdot\sigma(oldsymbol{W}_1x_i)))\}^2$$

イロト イヨト イヨト イヨト

Let's do it!

The objective function

$$\min_{\pmb{W}_i} rac{1}{n} \sum_{i=1}^n \{ \pmb{y}_i - \sigma(\pmb{W}_2 \cdot \sigma(\pmb{W}_1 \pmb{x}_i)) \}^2$$

イロト イヨト イヨト イヨト

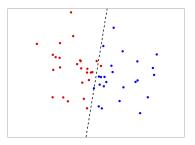
Let's do it!

The objective function

$$\min_{oldsymbol{ heta}} rac{1}{n} \sum_{i=1}^n \{y_i - \sigma(\langle oldsymbol{ heta}, x_i
angle)\}^2$$

イロト イヨト イヨト イヨト

Binary linear classification



The model

$$z_i = (x_i, y_i). \,\, x_i \in \mathbb{R}^d, \, y_i \in \{0, 1\}.$$

Convex logit loss (ℓ_c is cvx in θ)
 $\ell_c(\theta; z) = y \langle x, \theta \rangle - \log\{1 + \exp(\langle x, \theta \rangle)\}.$

▶ Non-convex loss (ℓ is not cvx in θ)

 $\ell(heta;z)=\{y-\sigma(\langle x, heta
angle)\}^2, ext{ where }\sigma(t)=1/(1+\exp(t)).$

Empirical Risk

$$\widehat{R}_n(heta) = rac{1}{n}\sum_{i=1}^n \ell(heta; z_i) = rac{1}{n}\sum_{i=1}^n \{y_i - \sigma(\langle heta, x_i
angle)\}^2.$$

Empirical risk minimizer

$$\hat{\theta}_n = rgmin_{\theta \in \mathsf{B}^d(R)} \widehat{R}_n(\theta).$$

Song Mei (Stanford University) The landscape of non-convex optimization

The model

 $z_i = (x_i, y_i). \,\, x_i \in \mathbb{R}^d, \, y_i \in \{0, 1\}.$

• Convex logit loss $(\ell_c \text{ is cvx in } \theta)$

$$\ell_c(heta;z) = y \langle x, heta
angle - \log\{1 + \exp(\langle x, heta
angle)\}.$$

► Non-convex loss (ℓ is not cvx in θ) $\ell(\theta; z) = \{y - \sigma(\langle x, \theta \rangle)\}^2$, where $\sigma(t) = 1/(1 + \exp(t))$.

Empirical Risk

$$\widehat{R}_n(heta) = rac{1}{n}\sum_{i=1}^n \ell(heta; z_i) = rac{1}{n}\sum_{i=1}^n \{y_i - \sigma(\langle heta, x_i
angle)\}^2.$$

Empirical risk minimizer

$$\hat{ heta}_n = rgmin_{ heta \in \mathrm{B}^d(R)} \widehat{R}_n(heta).$$

э.

Song Mei (Stanford University) The landscape of non-convex optimization October 19, 2017 10 / 32

The model

 $z_i = (x_i, y_i). \,\, x_i \in \mathbb{R}^d, \, y_i \in \{0, 1\}.$

• Convex logit loss $(\ell_c \text{ is cvx in } \theta)$

$$\ell_c(heta;z) = y\langle x, heta
angle - \log\{1+\exp(\langle x, heta
angle)\}.$$

► Non-convex loss (ℓ is not cvx in θ)

 $\ell(oldsymbol{ heta};z)=\{y-\sigma(\langle x,oldsymbol{ heta}
angle)\}^2, ext{ where }\sigma(t)=1/(1+\exp(t)).$

Empirical Risk

$$\widehat{R}_n(oldsymbol{ heta}) = rac{1}{n}\sum_{i=1}^n \ell(oldsymbol{ heta}; z_i) = rac{1}{n}\sum_{i=1}^n \{y_i - \sigma(\langle oldsymbol{ heta}, x_i
angle)\}^2.$$

Empirical risk minimizer

$$\hat{\theta}_n = \operatorname*{arg\,min}_{\theta \in \mathrm{B}^d(R)} \widehat{R}_n(\theta).$$

The model

 $z_i = (x_i, y_i). \; x_i \in \mathbb{R}^d, \, y_i \in \{0, 1\}.$

• Convex logit loss $(\ell_c \text{ is cvx in } \theta)$

$$\ell_c(heta;z) = y \langle x, heta
angle - \log\{1 + \exp(\langle x, heta
angle)\}.$$

▶ Non-convex loss (ℓ is not cvx in θ)

 $\ell(oldsymbol{ heta};z) = \{y - \sigma(\langle x, oldsymbol{ heta}
angle)\}^2, ext{ where } \sigma(t) = 1/(1+\exp(t)).$

Empirical Risk

$$\widehat{R}_n(oldsymbol{ heta}) = rac{1}{n}\sum_{i=1}^n \ell(oldsymbol{ heta}; z_i) = rac{1}{n}\sum_{i=1}^n \{y_i - \sigma(\langle oldsymbol{ heta}, x_i
angle)\}^2.$$

Empirical risk minimizer

$$\hat{\theta}_n = \arg\min_{\theta \in \mathsf{B}^d(R)} \widehat{R}_n(\theta).$$

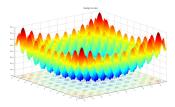
Theorem (Auer et. al. . 1996)

For the one node neural network, $\forall n, d > 0$, there exists a dataset $(x_i, y_i)_{i=1}^n$ such that the empirical risk $\widehat{R}_n(\theta)$ has $\lfloor \frac{n}{d} \rfloor^d$ distinct local minima.

Theorem (Auer et. al. . 1996)

For the one node neural network, $\forall n, d > 0$, there exists a dataset $(x_i, y_i)_{i=1}^n$ such that the empirical risk $\widehat{R}_n(\theta)$ has $\lfloor \frac{n}{d} \rfloor^d$ distinct local minima.

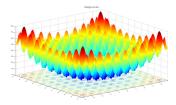
The landscape of $\widehat{R}_n(\theta)$ is very rough.



Theorem (Auer et. al. . 1996)

For the one node neural network, $\forall n, d > 0$, there exists a dataset $(x_i, y_i)_{i=1}^n$ such that the empirical risk $\widehat{R}_n(\theta)$ has $\lfloor \frac{n}{d} \rfloor^d$ distinct local minima.

The landscape of $\widehat{R}_n(\theta)$ is very rough.



Is this the end of the world of deep learning?

- The "Australian" data set from Statlog: d = 11, n = 683.
- ▶ Random initialization $\theta(0) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.
- Run gradient descent and track the path $\theta(k)$.
- Generate multiple paths with independent initializations.
- ▶ Plot standard deviation over paths $std(\theta(k))$ versus k.

э.

- The "Australian" data set from Statlog: d = 11, n = 683.
- Random initialization $\theta(0) \sim \mathcal{N}(0, I_d)$.
- Run gradient descent and track the path $\theta(k)$.
- Generate multiple paths with independent initializations.
- ▶ Plot standard deviation over paths $std(\theta(k))$ versus k.

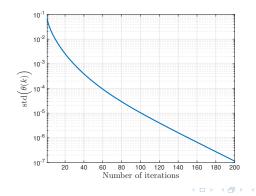
э.

- The "Australian" data set from Statlog: d = 11, n = 683.
- ▶ Random initialization $\theta(0) \sim \mathcal{N}(0, I_d)$.
- Run gradient descent and track the path $\theta(k)$.
- Generate multiple paths with independent initializations.
- ▶ Plot standard deviation over paths $std(\theta(k))$ versus k.

- The "Australian" data set from Statlog: d = 11, n = 683.
- ▶ Random initialization $\theta(0) \sim \mathcal{N}(0, I_d)$.
- Run gradient descent and track the path $\theta(k)$.
- Generate multiple paths with independent initializations.
- ▶ Plot standard deviation over paths $std(\theta(k))$ versus k.

・ロット (雪) (き) (き) (き)

- The "Australian" data set from Statlog: d = 11, n = 683.
- ► Random initialization $\theta(0) \sim \mathcal{N}(0, I_d)$.
- Run gradient descent and track the path $\theta(k)$.
- Generate multiple paths with independent initializations.
- ▶ Plot standard deviation over paths $std(\theta(k))$ versus k.



On real data, we "always" observe a unique minimum!

< 4 → </td>

э

On real data, we "always" observe a unique minimum!

Why?

< 17 ▶

On real data, we "always" observe a unique minimum!

Why?

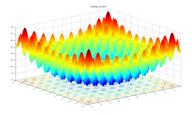
Data generated by nature is not against us!

Song Mei (Stanford University) The landscape of non-convex optimization October 19, 2017 13 / 32

Theorem (Auer et. al. . 1996)

For the one node neural network, $\forall n, d > 0$, there exists a dataset $(x_i, y_i)_{i=1}^n$ such that the empirical risk $\widehat{R}_n(\theta)$ has $\lfloor \frac{n}{d} \rfloor^d$ distinct local minima.

The landscape of $\widehat{R}_n(\theta)$ is very rough.

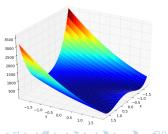


A positive result

Theorem (Mei, Bai, Montanari. 2016)

Assume Y_i are generated via $\mathbb{P}(Y_i = 1 | X_i) = \sigma(\langle X_i, \theta_0 \rangle)$ with mild assumption on X_i , as $n = \Omega(d \log d)$, with high probability: (a) $\widehat{R}_n(\theta)$ has a unique local minimizer $\widehat{\theta}_n$ in $B^d(\mathbf{0}, R)$. (b) $\widehat{\theta}_n$ satisfies $\|\widehat{\theta}_n - \theta_0\|_2 = \mathcal{O}(\sqrt{(d \log n)/n})$. (c) Gradient descent converges exponentially fast to $\widehat{\theta}_n$.

The landscape of $\widehat{R}_n(\theta)$ is actually smooth!



Why assuming a statistical model make the landscape of emprical risk smooth?

■ Assuming a statistical model $(X_i, Y_i) \stackrel{i \to a}{\sim} \mathbb{P}, i = 1, ..., n$, we can define the population risk

$$R(heta) = \mathbb{E}\left[\widehat{R}_n(heta)
ight] = \mathbb{E}_{X,Y}\left[(Y_i - \sigma(\langle heta_0, X_i))^2
ight].$$

The population risk is usually very smooth.

We can transfer the good properties of the population risk to the empirical risk using uniform convergence argument. So empirical risk will be also smooth. Why assuming a statistical model make the landscape of emprical risk smooth?

1 Assuming a statistical model $(X_i, Y_i) \stackrel{i.i.d.}{\sim} \mathbb{P}, i = 1, ..., n$, we can define the population risk

$$R(oldsymbol{ heta}) = \mathbb{E}\left[\widehat{R}_n(oldsymbol{ heta})
ight] = \mathbb{E}_{oldsymbol{X},oldsymbol{Y}}\left[(Y_i - \sigma(\langle oldsymbol{ heta}_0, X_i))^2
ight].$$

The population risk is usually very smooth.

We can transfer the good properties of the population risk to the empirical risk using uniform convergence argument. So empirical risk will be also smooth.

・ロット (雪) (き) (き) (き)

Why assuming a statistical model make the landscape of emprical risk smooth?

1 Assuming a statistical model $(X_i, Y_i) \stackrel{i.i.d.}{\sim} \mathbb{P}, i = 1, ..., n$, we can define the population risk

$$R(oldsymbol{ heta}) = \mathbb{E}\left[\widehat{R}_n(oldsymbol{ heta})
ight] = \mathbb{E}_{X,Y}\left[(Y_i - \sigma(\langle heta_0, X_i))^2
ight].$$

The population risk is usually very smooth.

We can transfer the good properties of the population risk to the empirical risk using uniform convergence argument. So empirical risk will be also smooth.

Population risk and empirical risk

The population risk has good properties under mild assumptions.

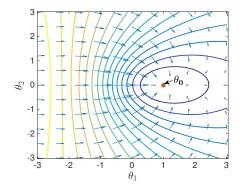


Figure: Population risk.

Population risk and empirical risk

The population risk has good properties under mild assumptions.

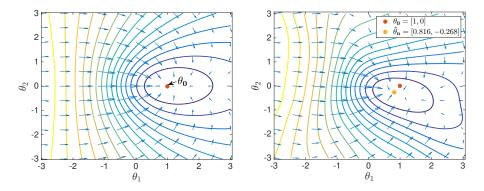


Figure: Population risk.

Figure: An instance of empirical risk.

Population risk and empirical risk

The population risk has good properties under mild assumptions.

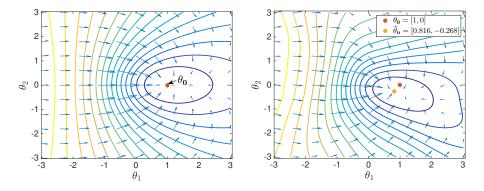


Figure: Population risk. Figure: An instance of empirical risk. How can we relate the properties of empirical risk to population risk?

Population risk and empirical risk

The population risk has good properties under mild assumptions.

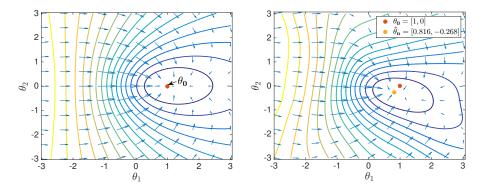


Figure: Population risk. Figure: An instance of empirical risk. How can we relate the properties of empirical risk to population risk? Uniform convergence!

Uniform convergence of gradients and Hessians.

Theorem (Uniform convergence. Informal)

ก

Under the settings, for any $\delta > 0$, there exists a positive constant C depending on (R, δ) but independent of n and d, such that as long as $n \ge Cd \log d$, we have

$$\mathbb{P}\left(\sup_{oldsymbol{ heta}\in\mathsf{B}^d(\mathbf{0},R)}\left\|
abla \widehat{R}_n(oldsymbol{ heta})-
abla R(oldsymbol{ heta})
ight\|_2\leq\sqrt{rac{Cd\log n}{n}}
ight)\geq 1-\delta.$$
 $\mathbb{P}\left(\sup_{oldsymbol{ heta}\in\mathsf{B}^d(\mathbf{0},R)}\left\|
abla^2 \widehat{R}_n(oldsymbol{ heta})-
abla^2 R(oldsymbol{ heta})
ight\|_{\mathrm{op}}\leq\sqrt{rac{Cd\log n}{n}}
ight)\geq 1-\delta.$

Proof is based on concentration inequalities and covering numbers.

Song Mei (Stanford University) The landscape of non-convex optimization October 19, 2017 18 / 32

Uniform convergence implies unique minimum of empirical risk

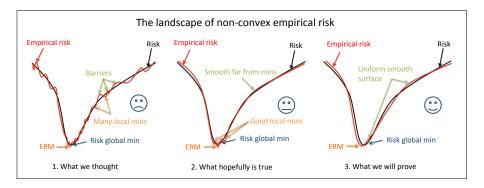


Figure: Landscape of empirical risk

・ロッ ・雪 ・ ・ ヨ ・

Numerical experiment

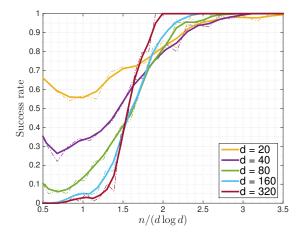


Figure: Probability to find a unique local minimum

Extensions

- ▶ Robust regression, gaussian mixture model, etc. High dimensional settings d ≫ n. [Mei et. al., 2017]
- ReLU activation. [Tian, 2017]
- Two Layers neural network. [Soltanolkotabi et. al., 2017], [Zhong et. al., 2017]
- ▶ Deep neural network. [Choromanska et. al., 2015]

・ロット (雪) (日) (日) (日)

Interlude

Before studying the complex neural network, maybe we can first study some simpler non-convex optimization problems.

(4) E > (4) E > (4)

э.

< 口 > < 同 >

MaxCut Problem



▶ G: a positively weighted graph. A_G : its adjacency matrix.

MaxCut of G: known to be NP-hard

$$\max_{x\in\{\pm1\}^n} \quad rac{1}{4}\sum_{i,j=1}^n A_{G,ij}(1-x_ix_j).$$
 (MaxCut)

 SDP relaxation: 0.878-approximate guarantee [Goemanns and Williamson, 1995]

$$\begin{array}{ll} \displaystyle \max_{X \in \mathbb{R}^{n \times n}} & \displaystyle \frac{1}{4} \sum_{i,j=1}^{n} A_{G,ij} (1-X_{ij}), \\ \text{subject to} & \displaystyle X_{ii} = 1, \\ & \displaystyle X \succeq 0. \end{array} \tag{SDPCut}$$

MaxCut Problem



- ▶ G: a positively weighted graph. A_G : its adjacency matrix.
- MaxCut of G: known to be NP-hard

$$egin{array}{lll} {\displaystyle \max_{m{x}\in\{\pm1\}^n}} & \displaystyle rac{1}{4}\sum_{i,j=1}^n A_{G,ij}(1-x_ix_j). \end{array}$$
 (MaxCut)

 SDP relaxation: 0.878-approximate guarantee [Goemanns and Williamson, 1995]

$$\begin{array}{ll} \displaystyle \max_{X \in \mathbb{R}^{n \times n}} & \displaystyle \frac{1}{4} \sum_{i,j=1}^{n} A_{G,ij} (1-X_{ij}), \\ \text{subject to} & \displaystyle X_{ii} = 1, \\ & \displaystyle X \succeq 0. \end{array} \tag{SDPCut}$$

MaxCut Problem



- ▶ G: a positively weighted graph. A_G : its adjacency matrix.
- MaxCut of G: known to be NP-hard

$$\max_{oldsymbol{x}\in\{\pm1\}^n} \quad rac{1}{4}\sum_{i,j=1}^n A_{G,ij}(1-oldsymbol{x}_ioldsymbol{x}_j). \tag{MaxCut}$$

 SDP relaxation: 0.878-approximate guarantee [Goemanns and Williamson, 1995]

$$\begin{array}{ll} \underset{X \in \mathbb{R}^{n \times n}}{\operatorname{maximize}} & \frac{1}{4} \sum_{i,j=1}^{n} A_{G,ij} (1 - X_{ij}), \\ \text{subject to} & X_{ii} = 1, \\ & X \succeq 0. \end{array}$$
(SDPCut)

The MaxCut SDP problem

• $A \in \mathbb{R}^{n \times n}$ symmetric.

$$egin{array}{lll} \max_{X\in\mathbb{R}^{n imes n}}&\langle A,X
angle\ {
m subject to}&X_{ii}=1,\ i\in[n],\ X\succeq0. \end{array}$$

< 口 > < 同 >

▶ Applications: MaxCut problem, Z₂ synchronization, Stochastic block model...

The MaxCut SDP problem

- $A \in \mathbb{R}^{n \times n}$ symmetric.
- MaxCut SDP:

$$egin{array}{lll} \displaystyle \max_{X\in \mathbb{R}^{n imes n}} & \langle A,X
angle \ {
m subject to} & X_{ii}=1, & i\in[n], \ & X\succeq 0. \end{array}$$

< □ > < /□ >

▶ Applications: MaxCut problem, Z₂ synchronization, Stochastic block model...

The MaxCut SDP problem

- $A \in \mathbb{R}^{n \times n}$ symmetric.
- MaxCut SDP:

$$egin{array}{lll} \displaystyle \max_{X\in \mathbb{R}^{n imes n}} & \langle A,X
angle \ {
m subject to} & X_{ii}=1, & i\in[n], \ & X\succeq 0. \end{array}$$

■▶ ▲ ■▶ = ● の Q @

▶ Applications: MaxCut problem, Z₂ synchronization, Stochastic block model...

Burer-Monteiro approach

▶ Convex formulation: n up to 10^3 using interior point method

$$egin{array}{lll} \max_{oldsymbol{X} \in \mathbb{R}^{n imes n}} & \langle A, oldsymbol{X}
angle \ {
m subject to} & oldsymbol{X}_{ii} = 1, & i \in [n], \ & oldsymbol{X} \succeq 0. \end{array}$$

- $\blacktriangleright \ \text{Change variable } X = \sigma \cdot \sigma^{\top}, \ \sigma \in \mathbb{R}^{n \times k}, \ k \ll n.$
- ▶ Non-convex formulation: n up to 10^5

・ロット 本語 マート キョット 日マー

Burer-Monteiro approach

▶ Convex formulation: n up to 10^3 using interior point method

$$egin{array}{lll} \max_{oldsymbol{X} \in \mathbb{R}^{n imes n}} & \langle A, oldsymbol{X}
angle \ {
m subject to} & oldsymbol{X}_{ii} = 1, & i \in [n], \ & oldsymbol{X} \succeq 0. \end{array}$$

- Change variable $X = \sigma \cdot \sigma^{\mathsf{T}}, \sigma \in \mathbb{R}^{n \times k}, k \ll n$.
- ▶ Non-convex formulation: n up to 10^5

Burer-Monteiro approach

▶ Convex formulation: n up to 10^3 using interior point method

$$egin{array}{lll} \max_{oldsymbol{X} \in \mathbb{R}^{n imes n}} & \langle A, oldsymbol{X}
angle \ {
m subject to} & X_{ii} = 1, & i \in [n], \ & X \succeq 0. \end{array}$$

- Change variable $X = \sigma \cdot \sigma^{\mathsf{T}}, \sigma \in \mathbb{R}^{n \times k}, k \ll n$.
- ▶ Non-convex formulation: n up to 10^5

 $\begin{array}{ll} \underset{\sigma \in \mathbb{R}^{n \times k}}{\operatorname{maximize}} & \langle \sigma, A \sigma \rangle \\ \text{subject to} & \sigma = [\sigma_1, \dots, \sigma_n]^\mathsf{T}, \\ & \|\sigma_i\|_2 = 1, \quad i \in [n]. \end{array} \tag{k-Nevx-SDP}$

- As k ≥ √2n, the global maxima of the Non Convex formulation coincide with the global maximizer of the Convex formulation [Pataki, 1998], [Barviok, 2001], [Burer and Monteiro, 2003].
- As k ≥ √2n, Non Convex formulation has no spurious local maxima [Boumal, et al., 2016].
- ▶ What if k remains of order 1, as n → ∞? Is there spurious local maxima?

• How is these local maxima?

- As k ≥ √2n, the global maxima of the Non Convex formulation coincide with the global maximizer of the Convex formulation [Pataki, 1998], [Barviok, 2001], [Burer and Monteiro, 2003].
- ► As k ≥ √2n, Non Convex formulation has no spurious local maxima [Boumal, et al., 2016].

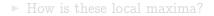
▶ What if k remains of order 1, as n → ∞? Is there spurious local maxima?

• How is these local maxima?

- As k ≥ √2n, the global maxima of the Non Convex formulation coincide with the global maximizer of the Convex formulation [Pataki, 1998], [Barviok, 2001], [Burer and Monteiro, 2003].
- As k ≥ √2n, Non Convex formulation has no spurious local maxima [Boumal, et al., 2016].
- What if k remains of order 1, as n → ∞? Is there spurious local maxima?

• How is these local maxima?

- As k ≥ √2n, the global maxima of the Non Convex formulation coincide with the global maximizer of the Convex formulation [Pataki, 1998], [Barviok, 2001], [Burer and Monteiro, 2003].
- As k ≥ √2n, Non Convex formulation has no spurious local maxima [Boumal, et al., 2016].
- What if k remains of order 1, as n → ∞? Is there spurious local maxima? Sadly, yes.



- As k ≥ √2n, the global maxima of the Non Convex formulation coincide with the global maximizer of the Convex formulation [Pataki, 1998], [Barviok, 2001], [Burer and Monteiro, 2003].
- As k ≥ √2n, Non Convex formulation has no spurious local maxima [Boumal, et al., 2016].

What if k remains of order 1, as n → ∞? Is there spurious local maxima? Sadly, yes.

How is these local maxima?

- As k ≥ √2n, the global maxima of the Non Convex formulation coincide with the global maximizer of the Convex formulation [Pataki, 1998], [Barviok, 2001], [Burer and Monteiro, 2003].
- As k ≥ √2n, Non Convex formulation has no spurious local maxima [Boumal, et al., 2016].

What if k remains of order 1, as n → ∞? Is there spurious local maxima? Sadly, yes.

▶ How is these local maxima? Empirically, they are good!

Geometry

Definition (ε -approximate concave point)

We call $\sigma \in \mathcal{M}_k$ an ε -approximate concave point of f on \mathcal{M}_k , if for any tangent vector $u \in T_{\sigma}\mathcal{M}_k$, we have

$$\langle u, \mathrm{Hess} f(\sigma)[u]
angle \leq arepsilon \langle u, u
angle.$$

Remark

A local maximizer is 0-approximate concave. An ε -approximate concave point is nearly locally optimal.

Geometry

Definition (ε -approximate concave point)

We call $\sigma \in \mathcal{M}_k$ an ε -approximate concave point of f on \mathcal{M}_k , if for any tangent vector $u \in T_{\sigma}\mathcal{M}_k$, we have

$$\langle u, \operatorname{Hess} f(\sigma)[u] \rangle \leq \varepsilon \langle u, u \rangle.$$
 (1)

Remark

A local maximizer is 0-approximate concave. An ε -approximate concave point is nearly locally optimal.

Geometry

Definition (ε -approximate concave point)

We call $\sigma \in \mathcal{M}_k$ an ε -approximate concave point of f on \mathcal{M}_k , if for any tangent vector $u \in T_{\sigma}\mathcal{M}_k$, we have

$$\langle u, \operatorname{Hess} f(\sigma)[u] \rangle \leq \varepsilon \langle u, u \rangle.$$
 (1)

Remark

A local maximizer is 0-approximate concave. An ε -approximate concave point is nearly locally optimal.

Landscape Theorem

Theorem (A Grothendieck-type inequality)

For any ε -approximate concave point $\sigma \in \mathcal{M}_k$ of the rank-k non-convex problem, we have

$$f(\sigma) \geq ext{SDP}(A) - rac{1}{k-1}(ext{SDP}(A) + ext{SDP}(-A)) - rac{n}{2}arepsilon.$$
 (2)

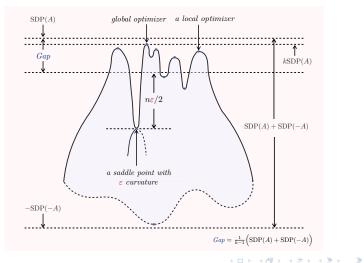
SDP(A): the maximum value of SDP with input matrix A.

Geometric iterpretation: the function value for all local maxima are within a gap of order O(1/k) within the global maximum.

・ロット 4日 マ・エット 4日 マクマク

Landscape of non-convex SDP

► $f(\sigma) \ge \text{SDP}(A) - \frac{1}{k-1}(\text{SDP}(A) + \text{SDP}(-A)) - \frac{n}{2}\varepsilon$.



Approximate MaxCut Guarantee

Theorem (Approximate MaxCut Guarantee)

For any $k \geq 3$, if σ^* is a local maximizer of corresponding rank-k non-convex problem, then we can use σ^* to find a $0.878 \times (1 - 1/(k - 1))$ -approximate MaxCut.

The global maximizer: 0.878-approximate MaxCut.

Any Local maximizers: $0.878 \times (1 - 1/(k - 1))$ -approximate MaxCut.

Approximate MaxCut Guarantee

Theorem (Approximate MaxCut Guarantee)

For any $k \geq 3$, if σ^* is a local maximizer of corresponding rank-k non-convex problem, then we can use σ^* to find a $0.878 \times (1 - 1/(k - 1))$ -approximate MaxCut.

The global maximizer: 0.878-approximate MaxCut.

Any Local maximizers: $0.878 \times (1 - 1/(k - 1))$ -approximate MaxCut.

・ロット 4日 マ・エット 4日 マクマク

Group Synchronization

SO(d) synchronization, Orthogonal Cut SDP

 $egin{array}{lll} \displaystyle \max_{X\in \mathbb{R}^{nk imes nk}} & \langle A,X
angle \ \mathrm{subject to} & X_{ii}=\mathrm{I}_k, \ & X\succeq 0. \end{array}$

(3)

Similar guarantee.

Conclusion

- Studied the global geometry of some non-convex optimization problems.
- Empirical risk minimization: uniform convergence excludes spurious local minima.
- Non-convex MaxCut SDP: all local maxima are near global maxima.

What I did not emphasize: Kac-Rice formula.