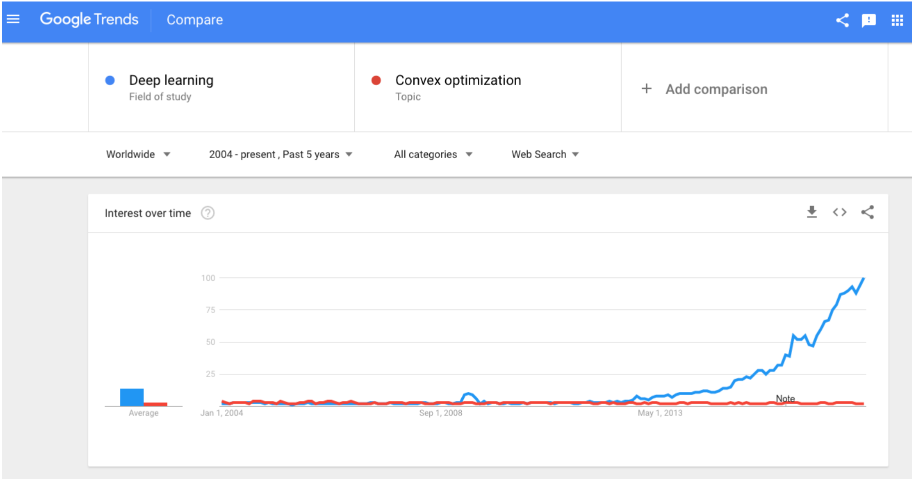
The landscape of non-convex losses for statistical learning problems

Song Mei

Stanford University

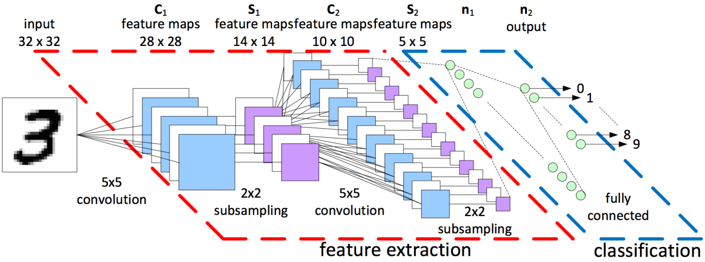
October 19, 2017

Deep learning

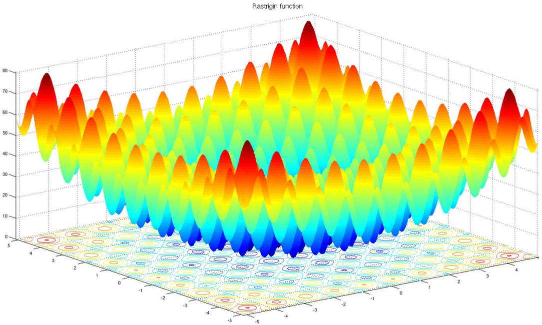


Deep learning

Convolutional Neural Network



Non-convex optimization



Why does non-convex neural network perform well?

Why does some non-convex optimization perform well?

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.

Let's do it!

The objective function

$$\min_{W_i} \frac{1}{n}\sum_{i=1}^{n}\{y_i - \sigma(W_k \cdots \sigma(W_2 \cdot \sigma(W_1 x_i)))\}^2$$

Let's do it!

The objective function

$$\min_{W_i} \frac{1}{n} \sum_{i=1}^{n} \{y_i - \sigma(W_2 \cdot \sigma(W_1 x_i))\}^2$$

Let's do it!

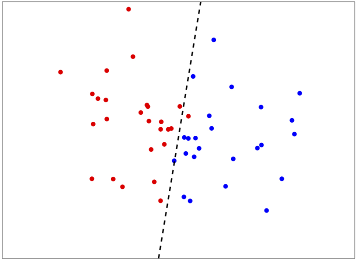
The objective function

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \{y_i - \sigma(\langle \theta, x_i \rangle)\}^2$$

Binary linear classification

The model

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



One node neural network

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)

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

- Empirical Risk

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

- Empirical risk minimizer

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

One node neural network

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)

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

- Empirical Risk

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

- Empirical risk minimizer

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

One node neural network

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)

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

- Empirical Risk

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

- Empirical risk minimizer

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

One node neural network

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)

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

- Empirical Risk

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

- Empirical risk minimizer

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

A negative theoretical result

Theorem (Auer *et. al.* . 1996)

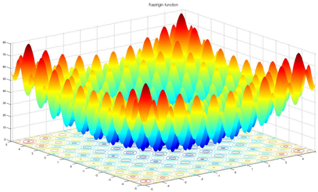
For the one node neural network, $\forall n, d > 0$, there exists a dataset $(x_i, y_i)_{i=1}^n$ such that the empirical risk $\widehat{R}_n(\theta)$ has $\lfloor \frac{n}{d} \rfloor^d$ distinct local minima.

A negative theoretical result

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.

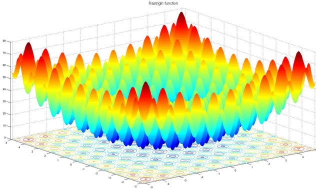


A negative theoretical result

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?

Real data experiment

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

Real data experiment

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

Real data experiment

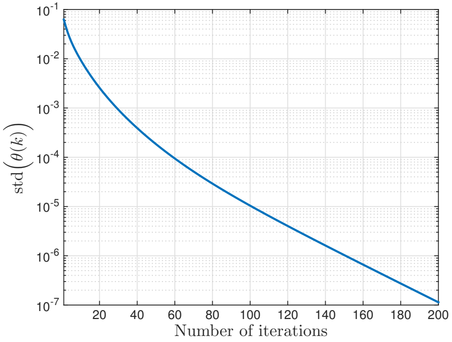
- The "Australian" data set from Statlog: $d = 11$, $n = 683$.
- Random initialization $\theta(0) \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$.
- Run gradient descent and track the path $\theta(k)$.
- Generate multiple paths with independent initializations.
- Plot standard deviation over paths $\mathrm{std}(\theta(k))$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

Real data experiment

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

Real data experiment

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



One node neural network

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

One node neural network

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

Why?

One node neural network

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

Why?

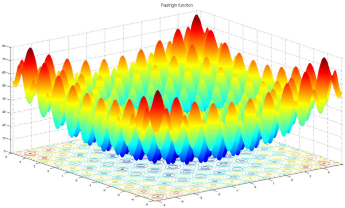
Data generated by nature is not against us!

A negative theoretical result

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)

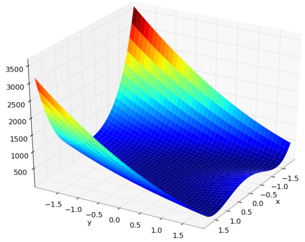
Assume Y_i are generated via $\mathbb{P}(Y_i = 1|X_i) = \sigma(\langle X_i, \theta_0 \rangle)$ with mild assumption on X_i, as $n = \Omega(d \log d)$, with high probability:

(a) $\widehat{R}_n(\theta)$ *has a unique local minimizer* $\hat{\theta}_n$ *in* $\mathsf{B}^d(\mathbf{0}, R)$.

(b) $\hat{\theta}_n$ *satisfies* $\|\hat{\theta}_n - \theta_0\|_2 = \mathcal{O}(\sqrt{(d \log n)/n})$.

(c) *Gradient descent converges exponentially fast to* $\hat{\theta}_n$.

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



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

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

$$R(\theta) = \mathbb{E}\left[\widehat{R}_n(\theta)\right] = \mathbb{E}_{X,Y}\left[(Y_i - \sigma(\langle \theta_0, X_i \rangle))^2\right].$$

   The population risk is usually very smooth.

2. 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 $(X_i, Y_i) \overset{i.i.d.}{\sim} \mathbb{P}$, $i = 1, \ldots, n$, we can define the population risk

$$R(\theta) = \mathbb{E}\left[\widehat{R}_n(\theta)\right] = \mathbb{E}_{X,Y}\left[(Y_i - \sigma(\langle \theta_0, X_i \rangle))^2\right].$$

   The population risk is usually very smooth.

2. 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 $(X_i, Y_i) \overset{i.i.d.}{\sim} \mathbb{P}$, $i = 1, \ldots, n$, we can define the population risk

$$R(\theta) = \mathbb{E}\left[\widehat{R}_n(\theta)\right] = \mathbb{E}_{X,Y}\left[(Y_i - \sigma(\langle \theta_0, X_i \rangle))^2\right].$$

   The population risk is usually very smooth.

2. 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.

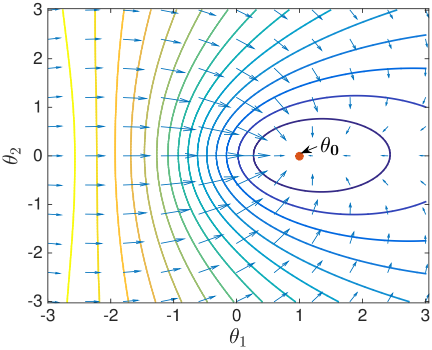


Figure: Population risk.

Population risk and empirical risk

The population risk has good properties under mild assumptions.

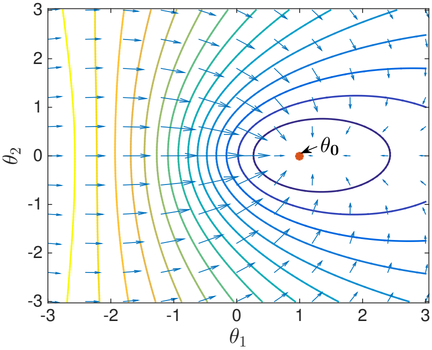


Figure: Population risk.

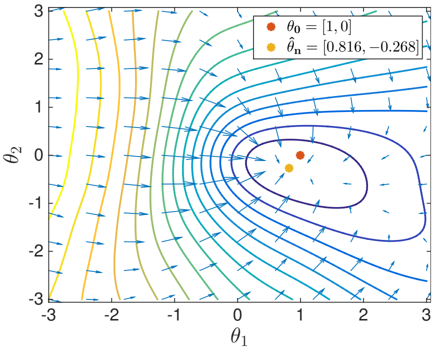


Figure: An instance of empirical risk.

Population risk and empirical risk

The population risk has good properties under mild assumptions.

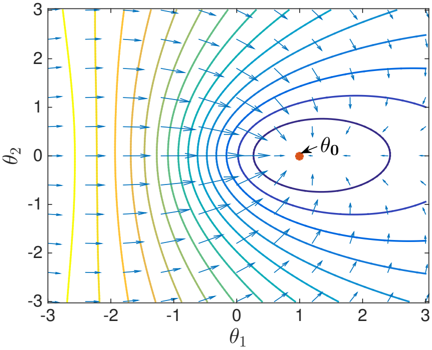


Figure: Population risk.

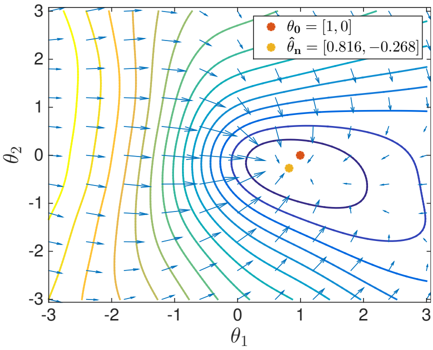


Figure: An instance of empirical risk.

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.

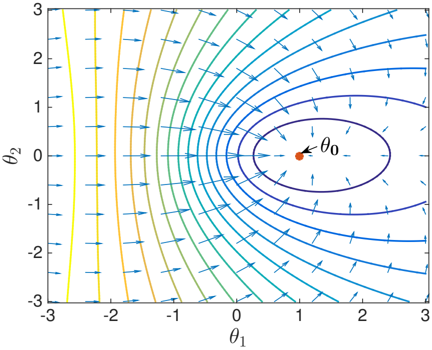


Figure: Population risk.

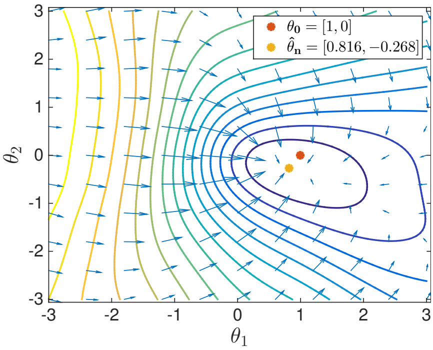


Figure: An instance of empirical risk.

How can we relate the properties of empirical risk to population risk?
Uniform convergence!

Uniform convergence of gradients and Hessians.

Theorem (Uniform convergence. Informal)

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 $n \geq C d \log d$, we have

1. $$\mathbb{P}\left(\sup_{\theta \in \mathsf{B}^d(\mathbf{0}, R)} \left\| \nabla \widehat{R}_n(\theta) - \nabla R(\theta) \right\|_2 \leq \sqrt{\frac{C d \log n}{n}} \right) \geq 1 - \delta.$$

2. $$\mathbb{P}\left(\sup_{\theta \in \mathsf{B}^d(\mathbf{0}, R)} \left\| \nabla^2 \widehat{R}_n(\theta) - \nabla^2 R(\theta) \right\|_{\mathrm{op}} \leq \sqrt{\frac{C d \log n}{n}} \right) \geq 1 - \delta.$$

Proof is based on concentration inequalities and covering numbers.

Uniform convergence implies unique minimum of empirical risk

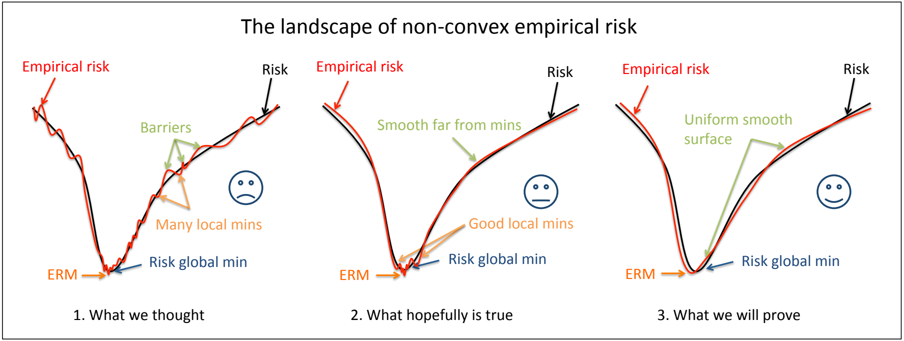


Figure: Landscape of empirical risk

Numerical experiment

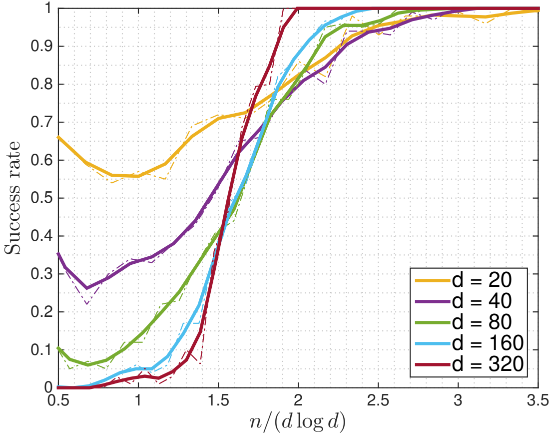


Figure: Probability to find a unique local minimum

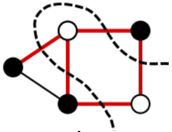
Extensions

- Robust regression, gaussian mixture model, etc. High dimensional settings $d \gg n$. [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]

Interlude

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

MaxCut Problem



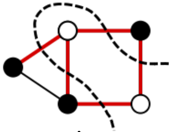
- G: a positively weighted graph. A_G: its adjacency matrix.
- MaxCut of G: known to be NP-hard

$$\underset{x\in\{\pm 1\}^n}{\text{maximize}} \quad \frac{1}{4}\sum_{i,j=1}^{n} A_{G,ij}(1 - x_i x_j). \qquad \text{(MaxCut)}$$

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

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

MaxCut Problem



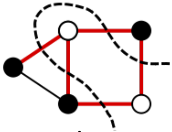
- G: a positively weighted graph. A_G: its adjacency matrix.
- MaxCut of G: known to be NP-hard

$$\underset{x \in \{\pm 1\}^n}{\text{maximize}} \quad \frac{1}{4} \sum_{i,j=1}^{n} A_{G,ij}(1 - x_i x_j). \qquad \text{(MaxCut)}$$

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

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

MaxCut Problem



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

$$\underset{x \in \{\pm 1\}^n}{\text{maximize}} \quad \frac{1}{4} \sum_{i,j=1}^{n} A_{G,ij}(1 - x_i x_j). \qquad \text{(MaxCut)}$$

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

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

The MaxCut SDP problem

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

$$\begin{aligned} \underset{X \in \mathbb{R}^{n\times n}}{\text{maximize}} \quad & \langle A, X \rangle \\ \text{subject to} \quad & X_{ii} = 1, \quad i \in [n], \\ & X \succeq 0. \end{aligned} \tag{SDP}$$

- Applications: MaxCut problem, $\mathbb{Z}_2$ synchronization, Stochastic block model...

The MaxCut SDP problem

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

$$\begin{aligned} \underset{X \in \mathbb{R}^{n \times n}}{\text{maximize}} \quad & \langle A, X \rangle \\ \text{subject to} \quad & X_{ii} = 1, \quad i \in [n], \\ & X \succeq 0. \end{aligned} \tag{SDP}$$

- Applications: MaxCut problem, $\mathbb{Z}_2$ synchronization, Stochastic block model...

The MaxCut SDP problem

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

$$\begin{aligned} \underset{X \in \mathbb{R}^{n \times n}}{\text{maximize}} \quad & \langle A, X \rangle \\ \text{subject to} \quad & X_{ii} = 1, \quad i \in [n], \\ & X \succeq 0. \end{aligned} \tag{SDP}$$

- Applications: MaxCut problem, $\mathbb{Z}_2$ synchronization, Stochastic block model...

Burer-Monteiro approach

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

$$\begin{aligned} \underset{X \in \mathbb{R}^{n \times n}}{\text{maximize}} \quad & \langle A, X \rangle \\ \text{subject to} \quad & X_{ii} = 1, \quad i \in [n], \\ & X \succeq 0. \end{aligned} \tag{SDP}$$

- 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

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

Burer-Monteiro approach

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

$$\begin{aligned} \underset{X \in \mathbb{R}^{n \times n}}{\text{maximize}} \quad & \langle A, X \rangle \\ \text{subject to} \quad & X_{ii} = 1, \quad i \in [n], \\ & X \succeq 0. \end{aligned} \tag{SDP}$$

- 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

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

Burer-Monteiro approach

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

$$\begin{aligned} \underset{X \in \mathbb{R}^{n \times n}}{\text{maximize}} \quad & \langle A, X \rangle \\ \text{subject to} \quad & X_{ii} = 1, \quad i \in [n], \\ & X \succeq 0. \end{aligned} \tag{SDP}$$

- 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

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

Related literatures

- As $k \geq \sqrt{2n}$, 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 $k \geq \sqrt{2n}$, Non Convex formulation has no spurious local maxima [Boumal, *et al.*, 2016].

- What if k remains of order 1, as $n \to \infty$? Is there spurious local maxima?

- How is these local maxima?

Related literatures

- As $k \geq \sqrt{2n}$, 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 $k \geq \sqrt{2n}$, Non Convex formulation has no spurious local maxima [Boumal, *et al.*, 2016].

- What if k remains of order 1, as $n \to \infty$? Is there spurious local maxima?

- How is these local maxima?

Related literatures

- As $k \geq \sqrt{2n}$, 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 $k \geq \sqrt{2n}$, Non Convex formulation has no spurious local maxima [Boumal, *et al.*, 2016].

- What if k remains of order 1, as $n \to \infty$? Is there spurious local maxima?

- How is these local maxima?

Related literatures

- As $k \geq \sqrt{2n}$, 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 $k \geq \sqrt{2n}$, Non Convex formulation has no spurious local maxima [Boumal, *et al.*, 2016].

- What if k remains of order 1, as $n \to \infty$? Is there spurious local maxima? Sadly, yes.

- How is these local maxima?

Related literatures

- As $k \geq \sqrt{2n}$, 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 $k \geq \sqrt{2n}$, Non Convex formulation has no spurious local maxima [Boumal, *et al.*, 2016].

- What if k remains of order 1, as $n \to \infty$? Is there spurious local maxima? Sadly, yes.

- How is these local maxima?

Related literatures

- As $k \geq \sqrt{2n}$, 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 $k \geq \sqrt{2n}$, Non Convex formulation has no spurious local maxima [Boumal, *et al.*, 2016].

- What if k remains of order 1, as $n \to \infty$? Is there spurious local maxima? Sadly, yes.

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

Geometry

$$\begin{array}{lll} \underset{\sigma \in \mathbb{R}^{n \times k}}{\text{maximize}} & \langle \sigma, A\sigma \rangle & := f(\sigma) \\ \text{subject to} & \|\sigma_i\|_2 = 1. & \end{array} \Bigg\} \ \mathcal{M}_k = \{\sigma \in \mathbb{R}^{n \times k} : \|\sigma_i\|_2 = 1\}.$$

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

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

Remark

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

Geometry

$$
\left.\begin{aligned}
&\underset{\sigma\in\mathbb{R}^{n\times k}}{\text{maximize}} && \langle \sigma, A\sigma\rangle && := f(\sigma)\\
&\text{subject to} && \|\sigma_i\|_2 = 1.
\end{aligned}\right\} \mathcal{M}_k = \{\sigma \in \mathbb{R}^{n\times k} : \|\sigma_i\|_2 = 1\}.
$$

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

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

Remark

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

Geometry

$$\left.\begin{array}{lll} \underset{\sigma\in\mathbb{R}^{n\times k}}{\text{maximize}} & \langle\sigma, A\sigma\rangle & := f(\sigma) \\ \text{subject to} & \|\sigma_i\|_2 = 1. \end{array}\right\} \mathcal{M}_k = \{\sigma \in \mathbb{R}^{n\times k} : \|\sigma_i\|_2 = 1\}.$$

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

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

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

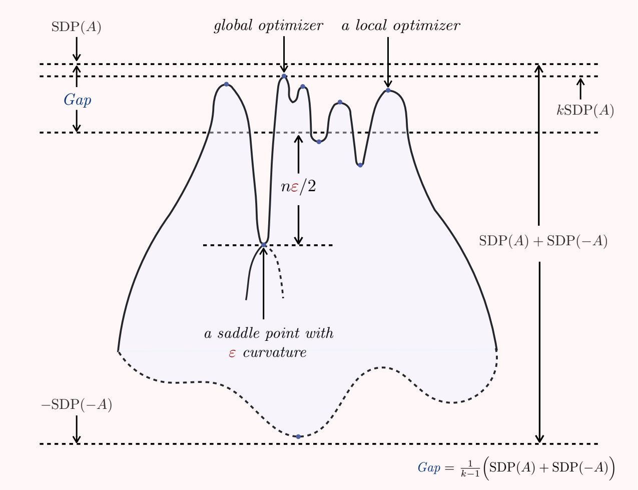
$$f(\sigma) \geq \mathrm{SDP}(A) - \frac{1}{k-1}(\mathrm{SDP}(A) + \mathrm{SDP}(-A)) - \frac{n}{2}\varepsilon. \tag{2}$$

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.

Landscape of non-convex SDP

- $f(\sigma) \geq \mathrm{SDP}(A) - \frac{1}{k-1}(\mathrm{SDP}(A) + \mathrm{SDP}(-A)) - \frac{n}{2}\varepsilon.$



Approximate MaxCut Guarantee

Theorem (Approximate MaxCut Guarantee)

For any $k \geq 3$, if $\sigma^\star$ is a local maximizer of corresponding rank-k non-convex problem, then we can use $\sigma^\star$ 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 $\sigma^\star$ is a local maximizer of corresponding rank-k non-convex problem, then we can use $\sigma^\star$ 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.

Group Synchronization

- SO(d) synchronization, Orthogonal Cut SDP

$$
\begin{aligned}
\underset{X \in \mathbb{R}^{nk \times nk}}{\text{maximize}} \quad & \langle A, X \rangle \\
\text{subject to} \quad & X_{ii} = \mathrm{I}_k, \\
& X \succeq 0.
\end{aligned} \tag{3}
$$

- 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.