The landscape of empirical risk for non-convex losses

Song Mei

ICME, Stanford

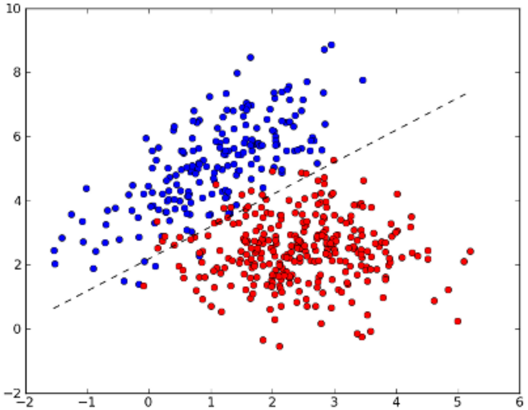
December 3, 2016

Joint work with Yu Bai and Andrea Montanari

Binary linear classification

The model

$\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$. $\boldsymbol{X}_i \in \mathbb{R}^d$, $Y_i \in \{0, 1\}$, $\mathrm{i} = 1, \ldots, \mathrm{n}$.



Non-convex formulation of binary classification

The model

$\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$. $\boldsymbol{X}_i \in \mathbb{R}^d$, $Y_i \in \{0, 1\}$, i = 1,. . . , n.

- Convex logit loss (ℓ_c is cvx in $\boldsymbol{\theta}$)

$$\ell_c(\boldsymbol{\theta}; \boldsymbol{Z}) = Y\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle - \log\Big(1 + \exp(\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle)\Big).$$

- Non-convex loss (ℓ is not cvx in $\boldsymbol{\theta}$)

$$\ell(\boldsymbol{\theta}; \boldsymbol{Z}) = \Big(Y - \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle)\Big)^2, \text{ where } \sigma(t) = 1/(1 + \exp(t)).$$

- Empirical Risk

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{Z}_i).$$

- Empirical risk minimizer

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

Non-convex formulation of binary classification

The model

$\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$. $\boldsymbol{X}_i \in \mathbb{R}^d$, $Y_i \in \{0, 1\}$, i = 1,. . . , n.

- Convex logit loss (ℓ_c is cvx in $\boldsymbol{\theta}$)

$$\ell_c(\boldsymbol{\theta}; \boldsymbol{Z}) = Y\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle - \log\Big(1 + \exp(\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle)\Big).$$

- Non-convex loss (ℓ is not cvx in $\boldsymbol{\theta}$)

$$\ell(\boldsymbol{\theta}; \boldsymbol{Z}) = \Big(Y - \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle)\Big)^2, \text{ where } \sigma(t) = 1/(1 + \exp(t)).$$

- Empirical Risk

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{Z}_i).$$

- Empirical risk minimizer

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

Non-convex formulation of binary classification

The model

$\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$. $\boldsymbol{X}_i \in \mathbb{R}^d$, $Y_i \in \{0, 1\}$, i = 1,. . . , n.

- Convex logit loss (ℓ_c is cvx in $\boldsymbol{\theta}$)

$$\ell_c(\boldsymbol{\theta}; \boldsymbol{Z}) = Y\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle - \log\Big(1 + \exp(\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle)\Big).$$

- Non-convex loss (ℓ is not cvx in $\boldsymbol{\theta}$)

$$\ell(\boldsymbol{\theta}; \boldsymbol{Z}) = \Big(Y - \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle)\Big)^2, \text{ where } \sigma(t) = 1/(1 + \exp(t)).$$

- Empirical Risk

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{Z}_i).$$

- Empirical risk minimizer

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

Non-convex formulation of binary classification

The model

$Z_i = (X_i, Y_i)$. $X_i \in \mathbb{R}^d$, $Y_i \in \{0, 1\}$, i = 1,. . . , n.

- Convex logit loss (ℓ_c is cvx in θ)

$$\ell_c(\theta; Z) = Y\langle X, \theta\rangle - \log\left(1 + \exp(\langle X, \theta\rangle)\right).$$

- Non-convex loss (ℓ is not cvx in θ)

$$\ell(\theta; Z) = \left(Y - \sigma(\langle X, \theta\rangle)\right)^2, \text{ where } \sigma(t) = 1/(1 + \exp(t)).$$

- Empirical Risk

$$\widehat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i).$$

- Empirical risk minimizer

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

Why use non-convex loss?

- Comparing to logistic regression, non-convex formulation is robust to ourliers.

- This model is the same as neural network with a single layer and a single node.

Why use non-convex loss?

- Comparing to logistic regression, non-convex formulation is robust to ourliers.
- This model is the same as neural network with a single layer and a single node.

A negative theoretical result

Theorem (Auer *et. al.* 1996 [AHW+96])

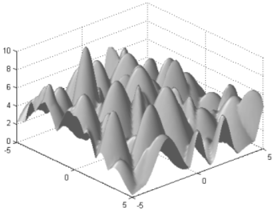
For the non-convex binary classification problem, for any n and d, there exists a dataset $(x_i, y_i)_{i=1}^n$ such that the empirical risk $\widehat{R}_n(\boldsymbol{\theta})$ has $\lfloor \frac{n}{d} \rfloor^d$ distinct local minima.

A negative theoretical result

Theorem (Auer *et. al.* 1996 [AHW+96])

For the non-convex binary classification problem, for any n and d, 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.

Seems to imply the landscape of the non-convex empirical risk $\widehat{R}_n(\theta)$ is very rough.

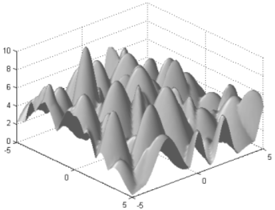


A negative theoretical result

Theorem (Auer *et. al.* 1996 [AHW+96])

For the non-convex binary classification problem, for any n and d, 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.

Seems to imply the landscape of the non-convex empirical risk $\widehat{R}_n(\theta)$ is very rough.



Is this the end of the world of non-convex binary classification?

Non-convex formulation of binary classification

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

Non-convex formulation of binary classification

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

Why?

Non-convex formulation of binary classification

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

Why?

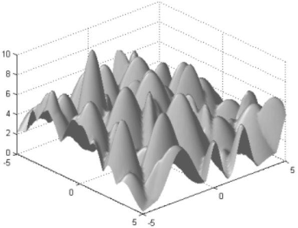
Data generated by nature is not against us!

A negative theoretical result

Theorem (Auer *et. al.* . 1996 [AHW+96])

For the non-convex binary classification problem, for all $n > 0$ there exists a dataset $(x_i, y_i)_{i=1}^n$ such that the empirical risk $\widehat{R}_n(\boldsymbol{\theta})$ has $\lfloor \frac{n}{d} \rfloor^d$ distinct local minima.

Seems to imply the landscape of the non-convex empirical risk $\widehat{R}_n(\boldsymbol{\theta})$ is very rough.



Our main positive result

Theorem (Mei, Bai, Montanari. 2016 [MBM16])

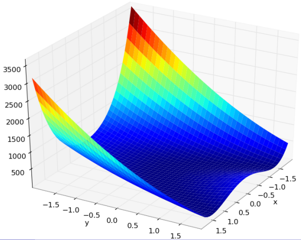
Assume $\boldsymbol{X}_i$ are i.i.d. sub-Gaussian random vectors, and Y_i are generated via $\mathbb{P}(Y_i = 1|\boldsymbol{X}_i) = \sigma(\langle \boldsymbol{X}_i, \boldsymbol{\theta}_0 \rangle)$. Then there exists a constant C depending on δ, such that as long as $n \geq C d \log d$, the following happens with probability at least $1 - \delta$:

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

(b) $\hat{\boldsymbol{\theta}}_n$ *satisfies* $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 \leq C\sqrt{(d \log n)/n}$.

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

The landscape of the non-convex empirical risk $\widehat{R}_n(\boldsymbol{\theta})$ is actually smooth!



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

1. Assuming a statistical model $Z_i \stackrel{i.i.d.}{\sim} \mathbb{P}_Z$, $i = 1, \ldots, n$, we can define the population risk

$$R(\theta) = \mathbb{E}_Z\left[\widehat{R}_n(\theta)\right] = \mathbb{E}_Z\left[\frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i)\right].$$

   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 $Z_i \overset{i.i.d.}{\sim} \mathbb{P}_Z$, $i = 1, \ldots, n$, we can define the population risk

$$R(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Z}}\left[\widehat{R}_n(\boldsymbol{\theta})\right] = \mathbb{E}_{\boldsymbol{Z}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}; \boldsymbol{Z}_i)\right].$$

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 $Z_i \overset{i.i.d.}{\sim} \mathbb{P}_Z$, $i = 1, \ldots, n$, we can define the population risk

$$R(\theta) = \mathbb{E}_Z\left[\widehat{R}_n(\theta)\right] = \mathbb{E}_Z\left[\frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i)\right].$$

   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.

Recap: Non-convex binary linear classification

The model

$\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$. $\boldsymbol{X}_i \in \mathbb{R}^d$, $Y_i \in \{0, 1\}$, i = 1,. . . , n.

- Non-convex square loss (ℓ is not cvx in $\boldsymbol{\theta}$)

$$\ell(\boldsymbol{\theta}; \boldsymbol{Z}) = \Big(Y - \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta} \rangle)\Big)^2, \text{ where } \sigma(t) = 1/(1 + \exp(t))$$

- Empirical Risk

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{Z}_i).$$

Recap: Non-convex binary linear classification

The model

$\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$. $\boldsymbol{X}_i \in \mathbb{R}^d$, $Y_i \in \{0, 1\}$, i = 1,. . . , n.

- Non-convex square loss (ℓ is not cvx in $\boldsymbol{\theta}$)

$$\ell(\boldsymbol{\theta}; \boldsymbol{Z}) = \Big(Y - \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta} \rangle)\Big)^2, \text{ where } \sigma(t) = 1/(1 + \exp(t))$$

- Empirical Risk

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{Z}_i).$$

- Assume $\boldsymbol{X}_i \overset{i.i.d.}{\sim} \mathbb{P}_{\boldsymbol{X}}$, ($\mathbb{P}_{\boldsymbol{X}}$ is sub-Gaussian), and $\mathbb{P}(Y_i = 1 | \boldsymbol{X}_i) = \sigma(\langle \boldsymbol{X}_i, \boldsymbol{\theta}_0 \rangle)$ with $\sigma(t) = 1/(1 + \exp(t))$, $\boldsymbol{\theta}_0 \in \mathbb{R}^d$.

Recap: Non-convex binary linear classification

The model

$\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$. $\boldsymbol{X}_i \in \mathbb{R}^d$, $Y_i \in \{0,1\}$, i = 1,. . . , n.

- Non-convex square loss (ℓ is not cvx in $\boldsymbol{\theta}$)

$$\ell(\boldsymbol{\theta}; \boldsymbol{Z}) = \Big(Y - \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta} \rangle)\Big)^2, \text{ where } \sigma(t) = 1/(1+\exp(t))$$

- Empirical Risk

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}; \boldsymbol{Z}_i).$$

- Assume $\boldsymbol{X}_i \overset{i.i.d.}{\sim} \mathbb{P}_{\boldsymbol{X}}$, ($\mathbb{P}_{\boldsymbol{X}}$ is sub-Gaussian), and $\mathbb{P}(Y_i = 1 | \boldsymbol{X}_i) = \sigma(\langle \boldsymbol{X}_i, \boldsymbol{\theta}_0 \rangle)$ with $\sigma(t) = 1/(1+\exp(t))$, $\boldsymbol{\theta}_0 \in \mathbb{R}^d$.
- Population risk:

$$R(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Z}}[(Y - \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta} \rangle))^2] = \mathbb{E}_{\boldsymbol{Z}}[(\sigma(\langle \boldsymbol{X}, \boldsymbol{\theta}_0 \rangle) - \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta} \rangle))^2] + c.$$

Recap: Non-convex binary linear classification

The model

$\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$. $\boldsymbol{X}_i \in \mathbb{R}^d$, $Y_i \in \{0, 1\}$, i = 1,. . . , n.

- Non-convex square loss (ℓ is not cvx in $\boldsymbol{\theta}$)

$$\ell(\boldsymbol{\theta}; \boldsymbol{Z}) = \Big(Y - \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta} \rangle)\Big)^2, \text{ where } \sigma(t) = 1/(1 + \exp(t))$$

- Empirical Risk

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{Z}_i).$$

- Assume $\boldsymbol{X}_i \overset{i.i.d.}{\sim} \mathbb{P}_{\boldsymbol{X}}$, ($\mathbb{P}_{\boldsymbol{X}}$ is sub-Gaussian), and $\mathbb{P}(Y_i = 1 | \boldsymbol{X}_i) = \sigma(\langle \boldsymbol{X}_i, \boldsymbol{\theta}_0 \rangle)$ with $\sigma(t) = 1/(1 + \exp(t))$, $\boldsymbol{\theta}_0 \in \mathbb{R}^d$.
- Population risk:

$$R(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Z}}[(Y - \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta} \rangle))^2] = \mathbb{E}_{\boldsymbol{Z}}[(\sigma(\langle \boldsymbol{X}, \boldsymbol{\theta}_0 \rangle) - \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta} \rangle))^2] + c.$$

- $R(\boldsymbol{\theta})$ has a unique minimum which is $\boldsymbol{\theta}_0$.

Population risk and empirical risk

The population risk has good properties under mild assumptions.

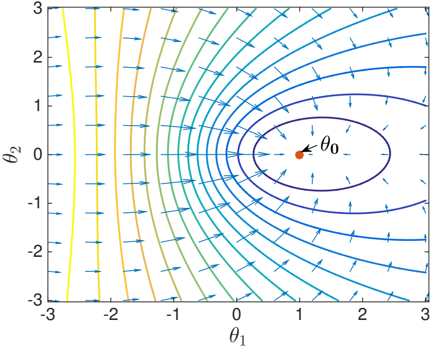


Figure: Population risk.

Population risk and empirical risk

The population risk has good properties under mild assumptions.

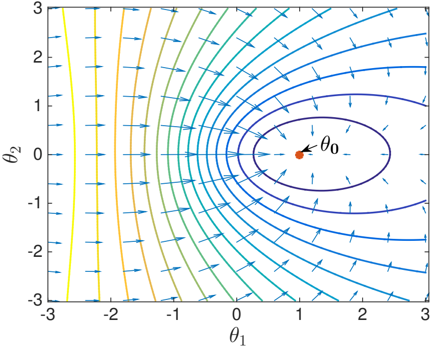


Figure: Population risk.

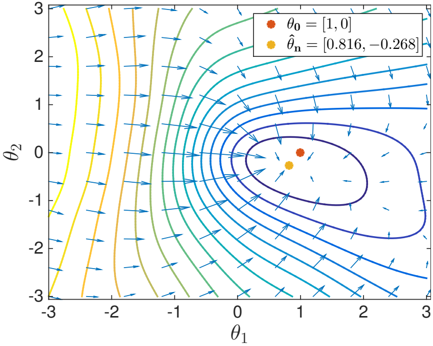


Figure: An instance of empirical risk.

Population risk and empirical risk

The population risk has good properties under mild assumptions.

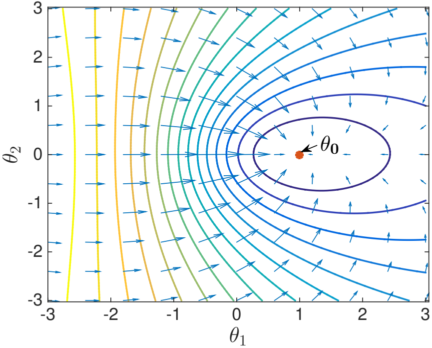


Figure: Population risk.

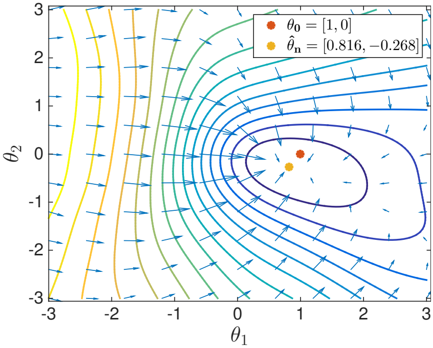


Figure: An instance of empirical risk.

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.

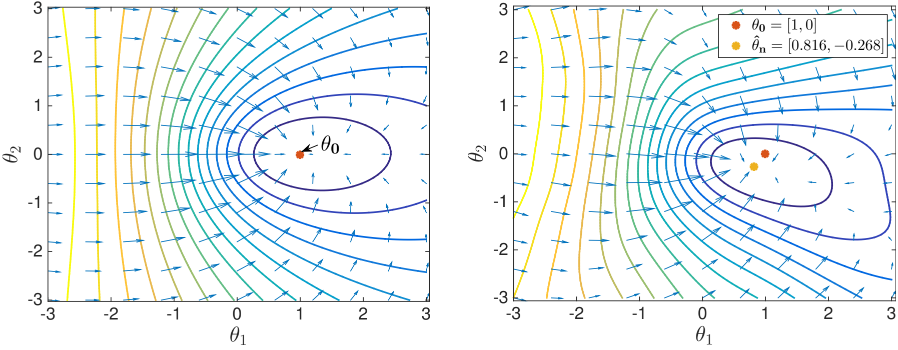


Figure: Population risk. Figure: An instance of empirical risk.

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

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

Proof is based on concentration inequalities and covering numbers.

Uniform convergence implies unique minimum of empirical risk

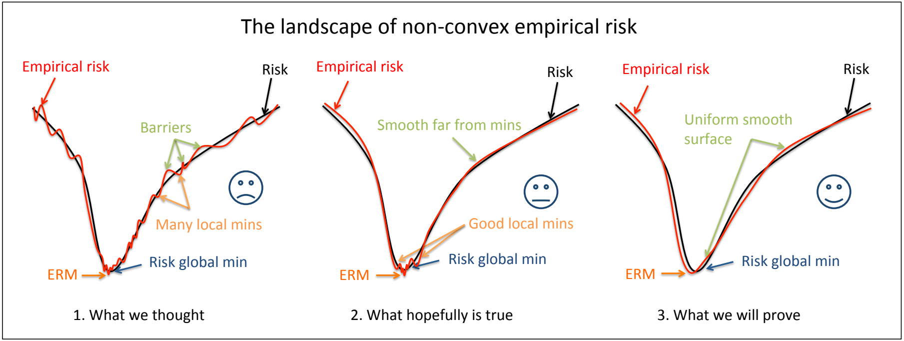


Figure: Landscape of empirical risk

Numerical experiment

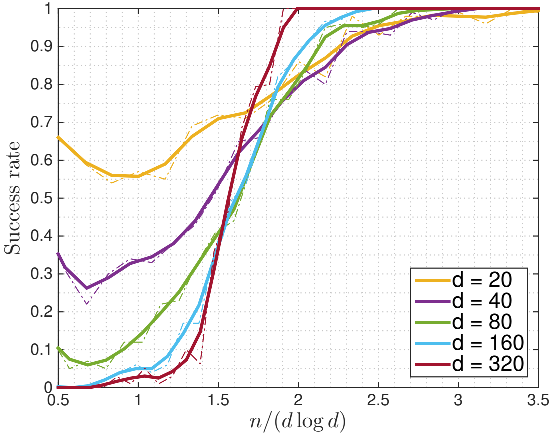


Figure: Probability to find a unique local minimum

Extension to other models

- Robust regression.
  Linear regression with bounded loss. Robust to outliers.
- Gaussian mixture model with two equal-proportion Gaussians.
  Two local minimum connected with a saddle point.
- Very high dimensional regime. $d \gg n$. Sparse θ_0.
  Uniform convergence of gradient in the sense of l_1 norm.

Conclusion

1. For non-convex empirical risk minimization problem, in the worst case, there could be exponentially many local minimum.
2. If there are enough data generated by a statistical model, the landscape of empirical risk is smooth.
3. The uniform convergence of gradients and Hessians is a powerful tool and can supplement the classical empirical risk minimization theory.

Bibliography

- Peter Auer, Mark Herbster, Manfred K Warmuth, et al., *Exponentially many local minima for single neurons*, Advances in neural information processing systems (1996), 316–322.
- Song Mei, Yu Bai, and Andrea Montanari, *The landscape of empirical risk for non-convex losses*, arXiv preprint arXiv:1607.06534 (2016).