New Paradigms and Global Optimality in Non-Convex Optimization

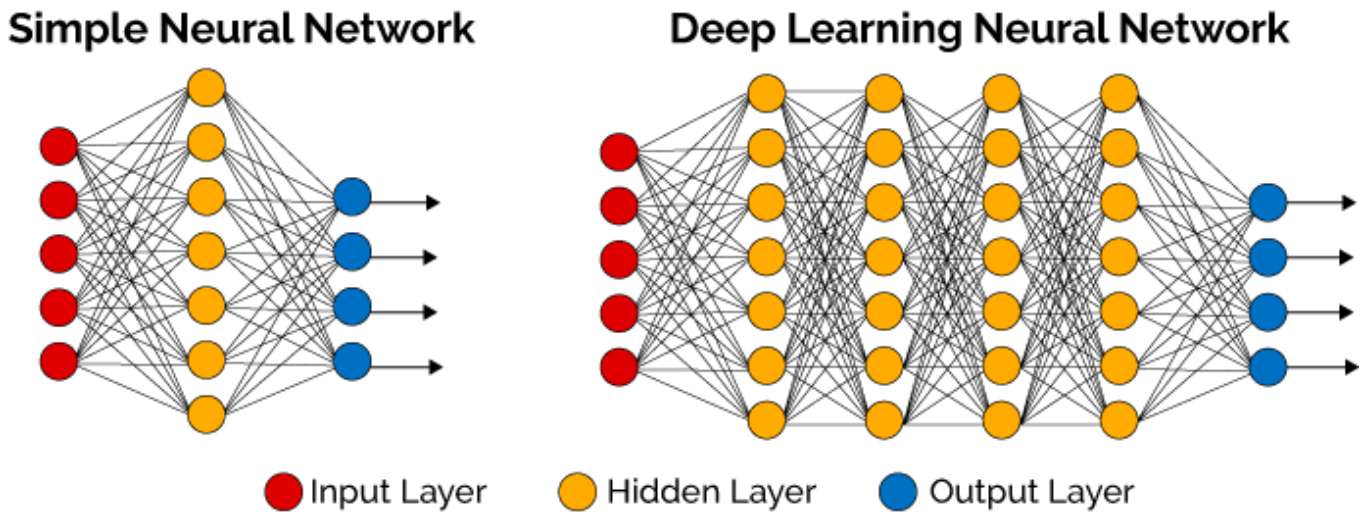
Hongyang Zhang, Machine Learning Dept., CMU
Joint work with Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab
Yingyu Liang, David P. Woodruff

CMU Theory Lunch

Optimizations are Everywhere

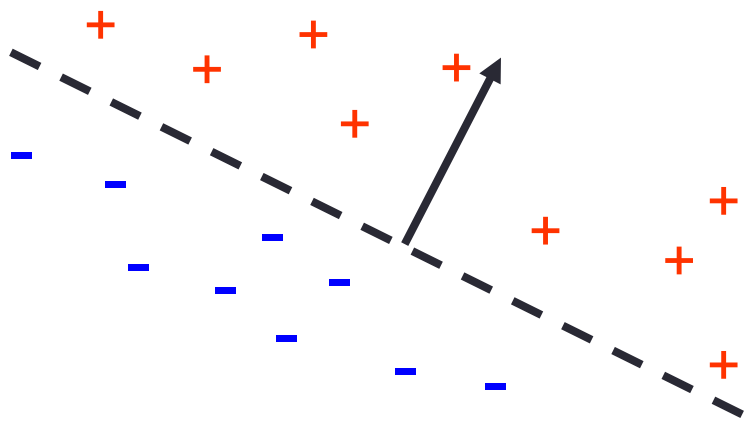
Examples of Non-Convex Optimizations

- Non-linear mapping (deep neural networks)



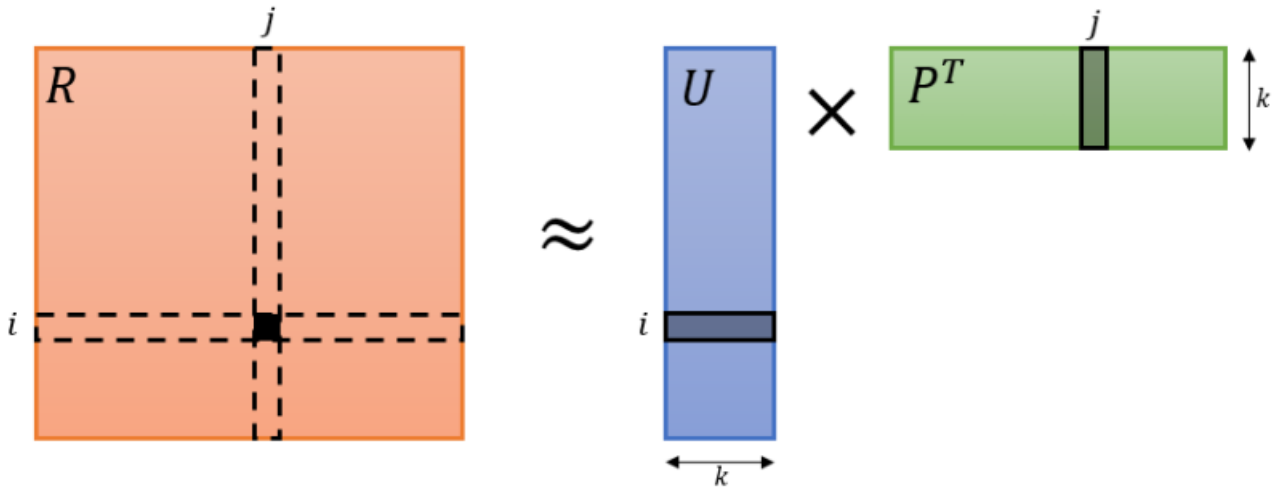
Examples of Non-Convex Optimizations

- Discrete loss (learning halfspace)

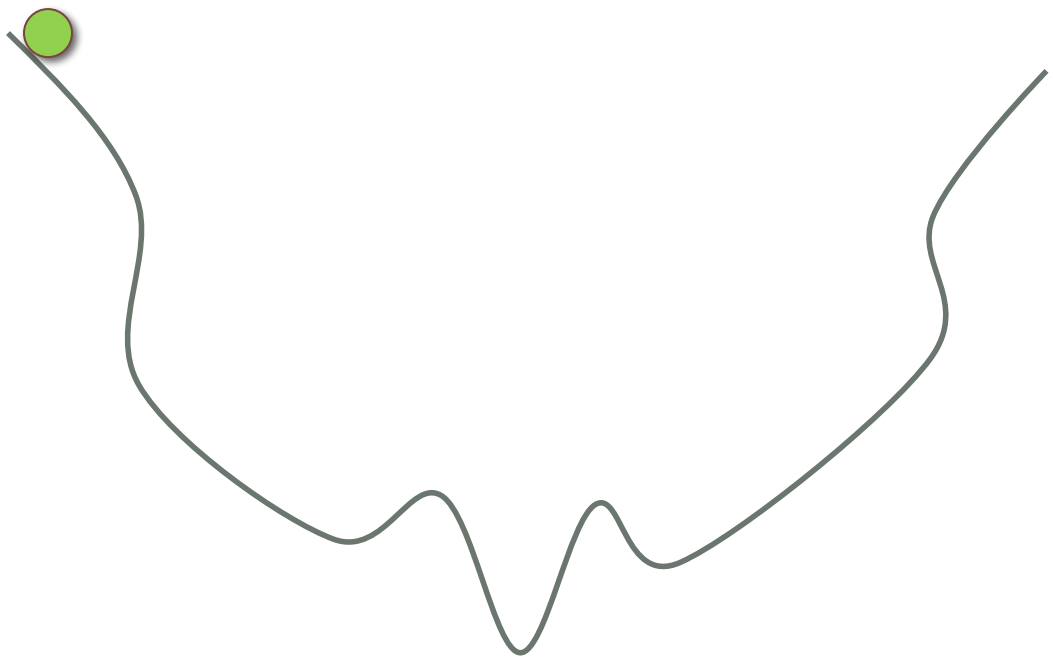


Examples of Non-Convex Optimizations

- Coupling of variables (matrix factorization)



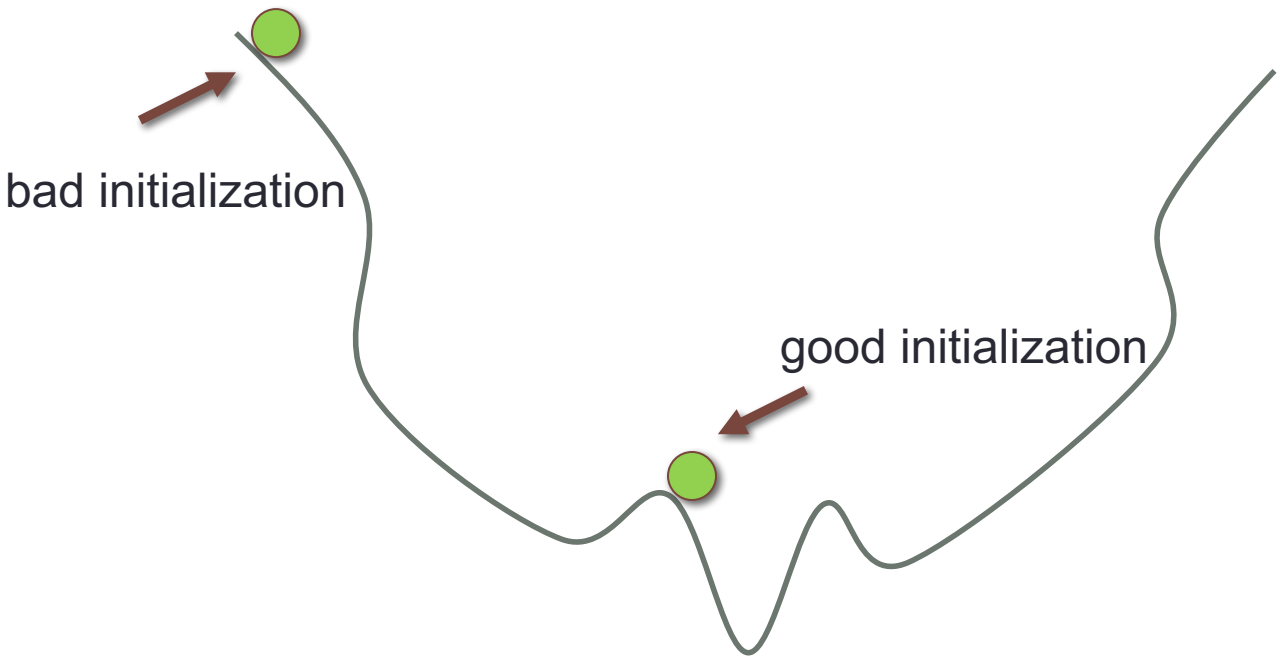
Challenges in Non-Convex Problems



NP-hard in general!

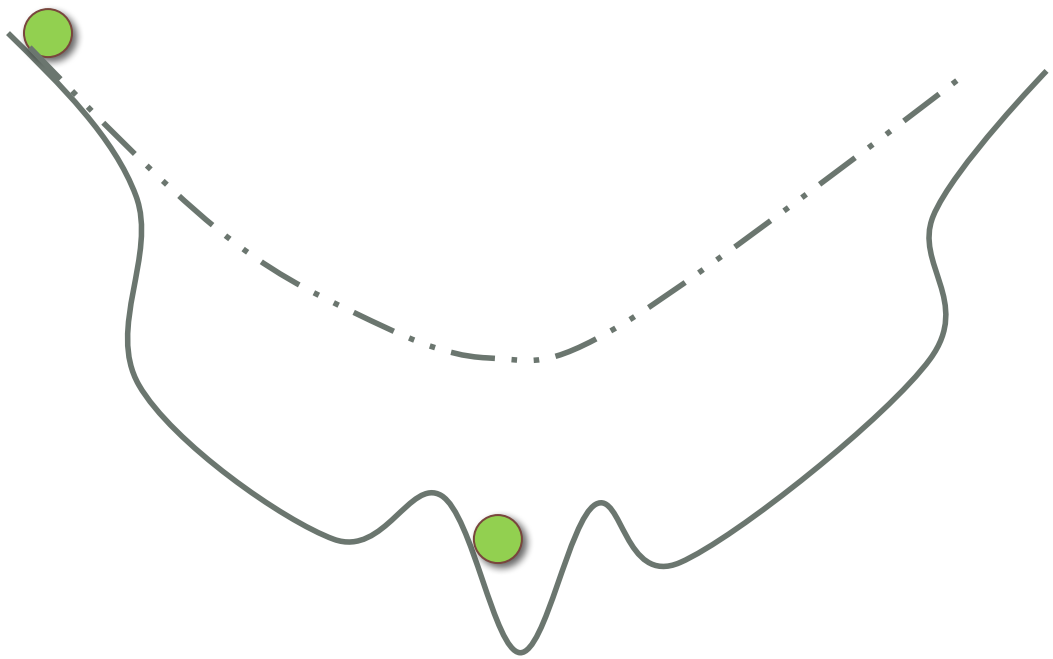
Tame Non-Convex Problems

1. By good initialization



Tame Non-Convex Problems

1. By good initialization 2. By sequential convex probs.

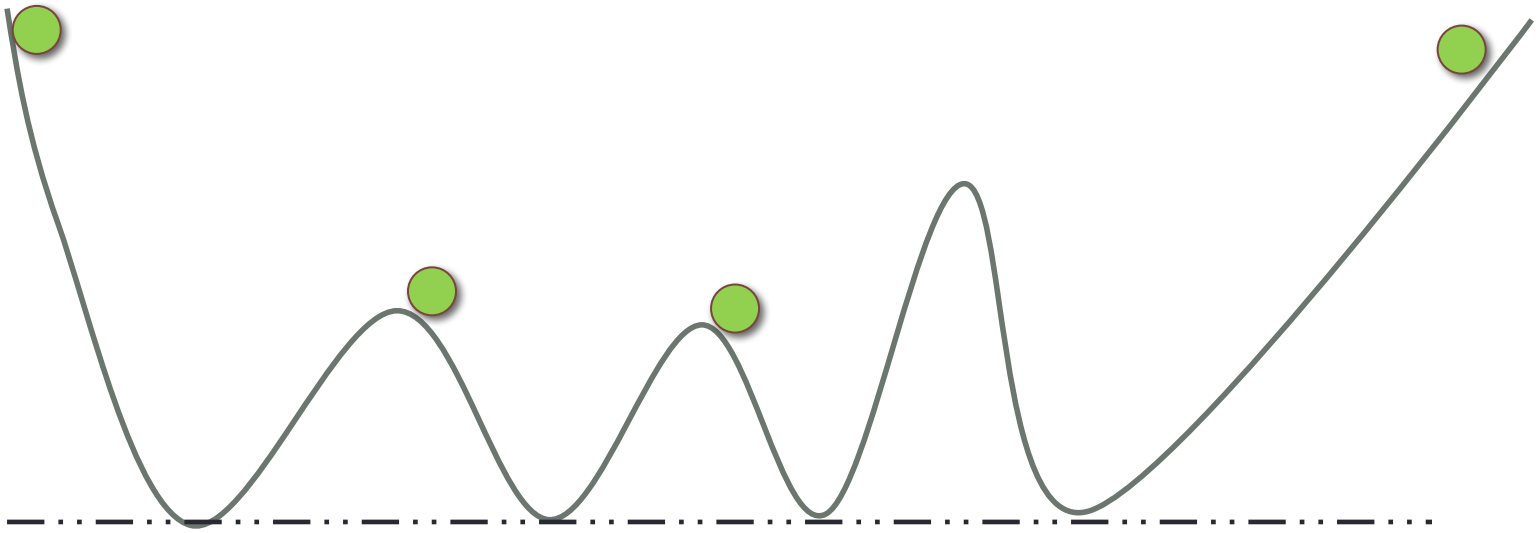


Tame Non-Convex Problems

1. By good initialization 2. By sequential convex probs. 3. By landscape

Part I Part II

the focus of this talk



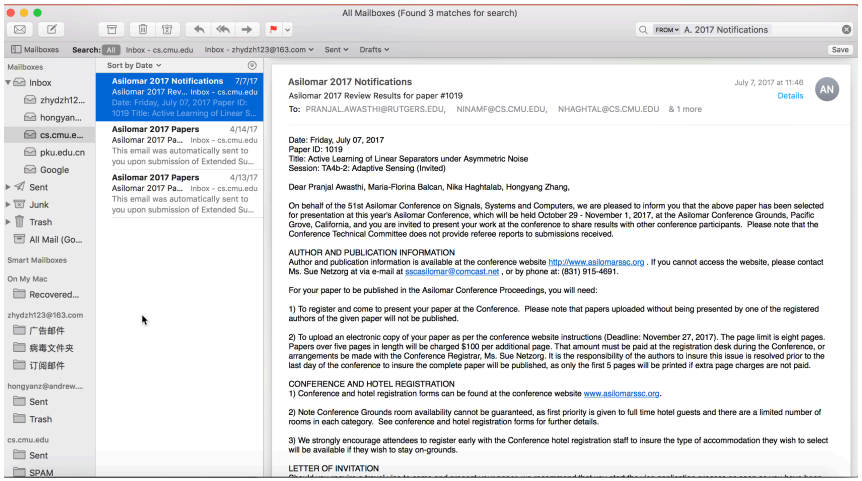
Part I

Learning of Halfspaces and 1-Bit Compressed Sensing

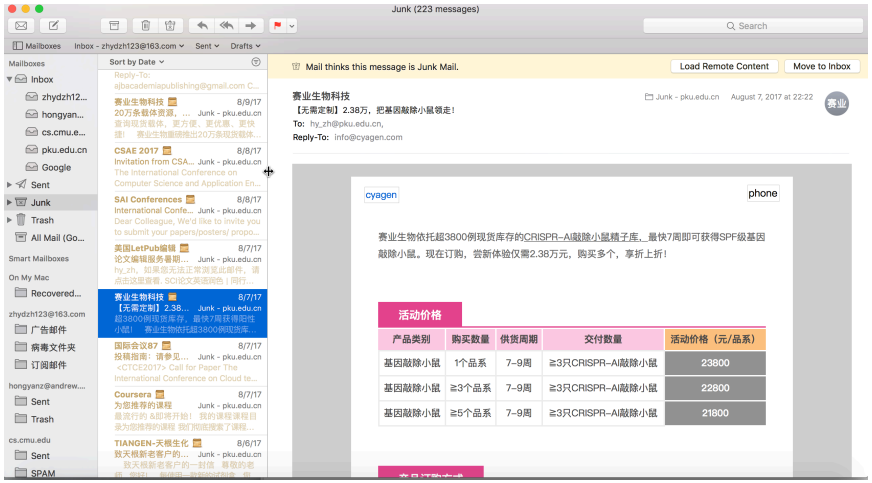
(by sequential convex probs.)

Learning of Halfspaces and 1-bit CS

Not Spam

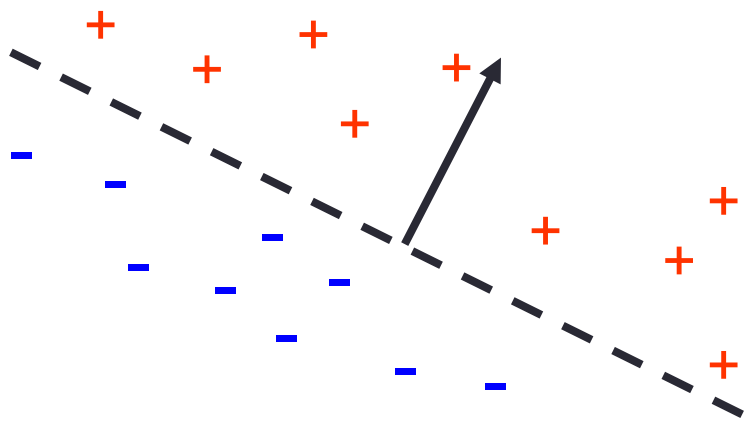


Spam



Goal: use emails seen so far to produce good prediction rule for future data.

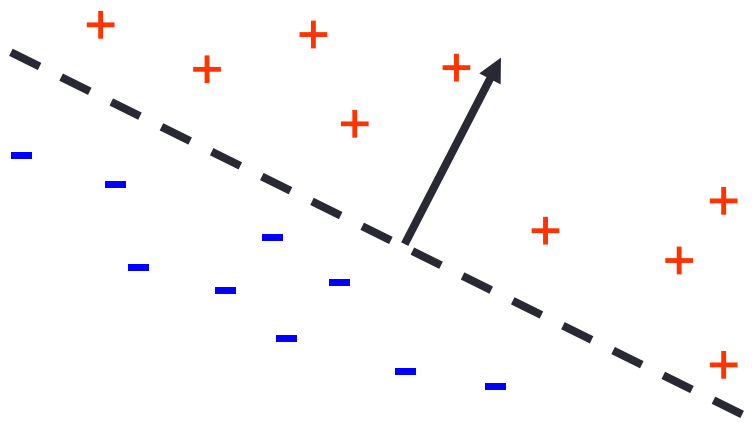
Learning of Halfspaces



What if we know the classifier is sparse? Is it possible that we require fewer samples?

[ABL] The Power of Localization for Efficiently Learning Linear Separators with Noise, JACM'17
[KLS] Learning halfspaces with malicious noise, JMLR'09
[KKMS] Agnostically learning halfspaces, FOCS'05

1-Bit Compressed Sensing



What if we know the classifier is sparse? Is it possible that we require fewer samples?

Difference with learning: Impose additional sparsity constraint

[PV] Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach, IEEE TIT'13

Optimization Formulation

- No Noise: Easy – solve ERM via a linear program

Find w such that $\forall i,\ y_i(w \cdot x_i) \geq 0$

- With Noise: Solve a non-convex problem

$\min_w \Pr_{(x,y)\sim\tilde{D}}[\mathrm{sgn}(w \cdot x) \neq y],\ (\text{s.t. } \|w\|_0 \leq t \text{ for 1-bit CS})$ (log-concave dist.)

- Sparsity (1-bit CS): Use a number of samples poly*(t,* log*(d/δ), 1/ε)*

Can we minimize the objective function to the accuracy of the information-theoretic limit under asymmetric noise model, although its formulation is non-convex?

The answer is affirmative!

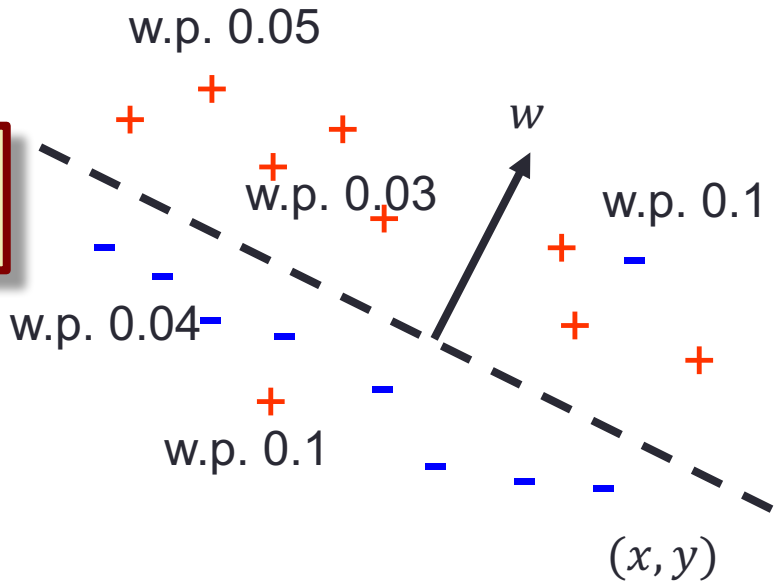
Part I Outline

- Motivation and examples
- **Our settings**
- Our algorithms
- Our hardness results
- Conclusions

Asymmetric Noise model – Bounded Noise

Bounded Noise (a.k.a. Massart Noise):

For a fixed $\eta \leq ½$, for each x, the adversary flips the label of x with probability $\eta(x) \leq \eta$.



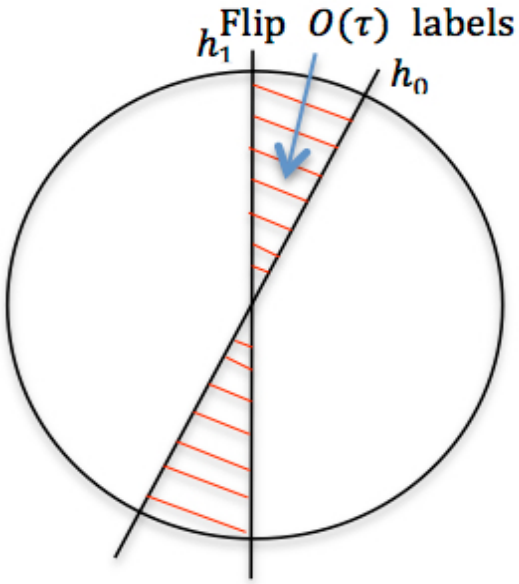
- Generalization of the RCN model
- Prior Result: $\eta \approx 10^{-6}$
- No result is known when $w \in \Re^d$ is *t*-sparse
- Information-theoretic limit: $OPT + \varepsilon$

Asymmetric Noise model – Adversarial Noise

Adversarial Noise:

The adversary can flip any τ fraction of labels of x.

- No result is known when $w \in \Re^d$ is *t*-sparse
- Information-theoretic limit: $OPT + \tau + \varepsilon$

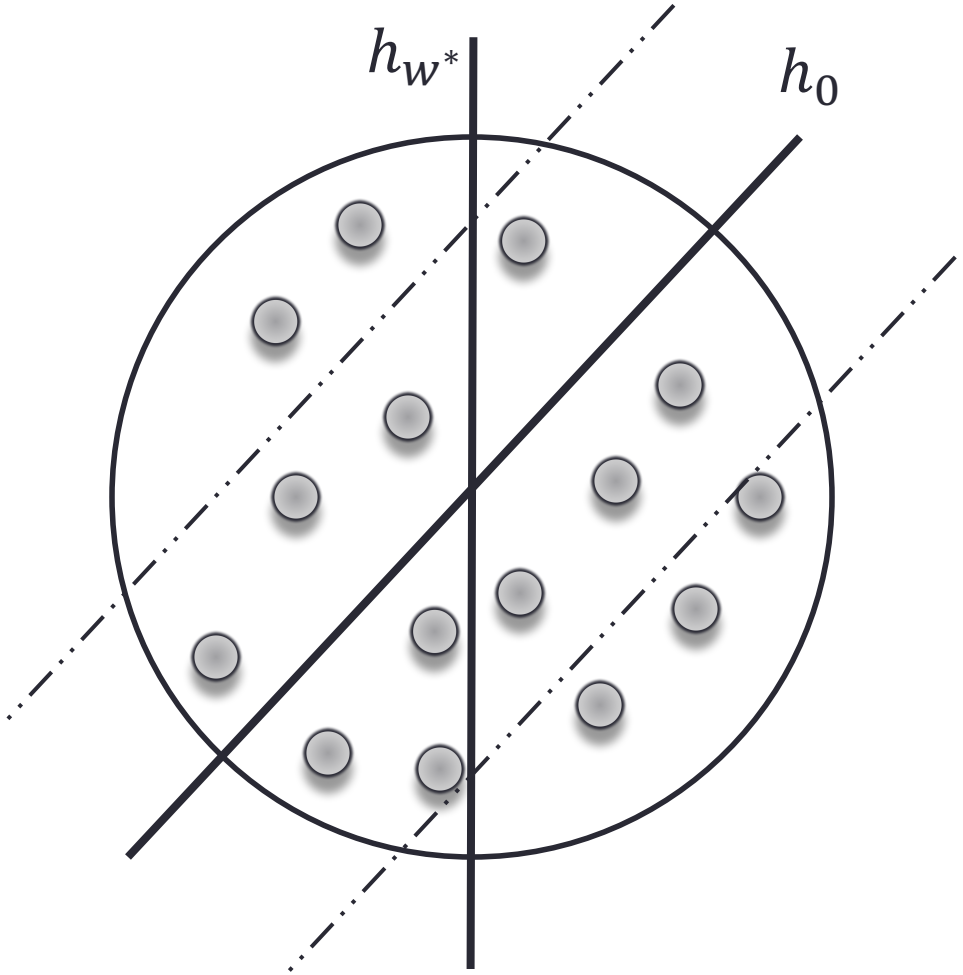


Part I Outline

- Motivation and examples
- Our settings
- **Our algorithms**
- Our hardness results
- Conclusions

Algorithm

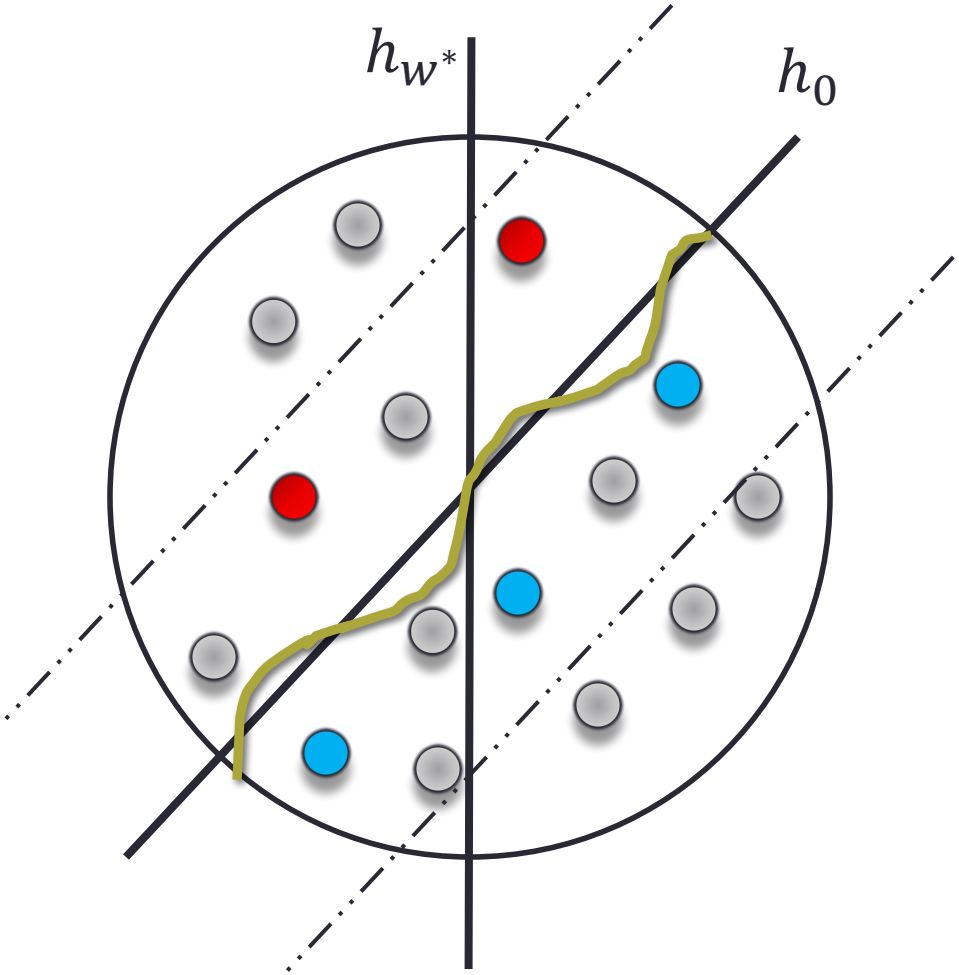
- Idea: Adaptively solve a sequence of convex programs



Sample unlabeled data
and have an initial guess

Algorithm

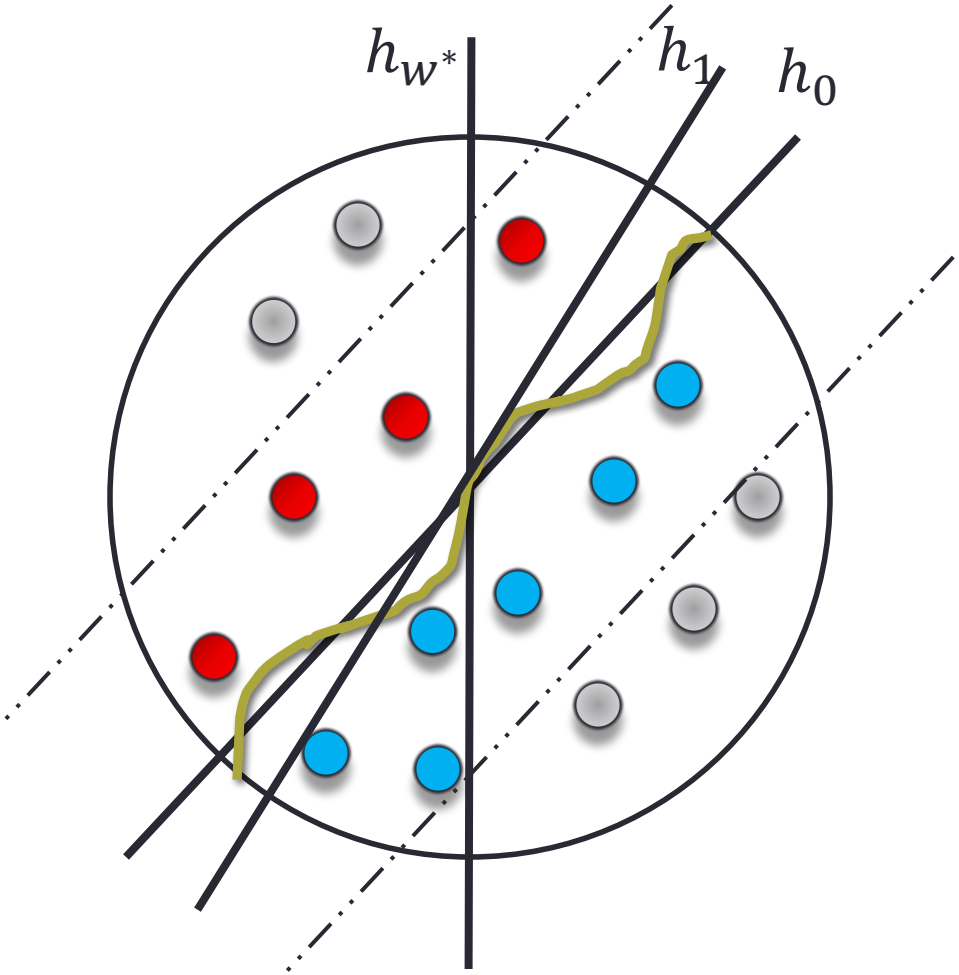
- Idea: Adaptively solve a sequence of convex programs



Ask some of labels in the band,
fit a polynomial to constant error
(require exp. time on 1/error)

Algorithm

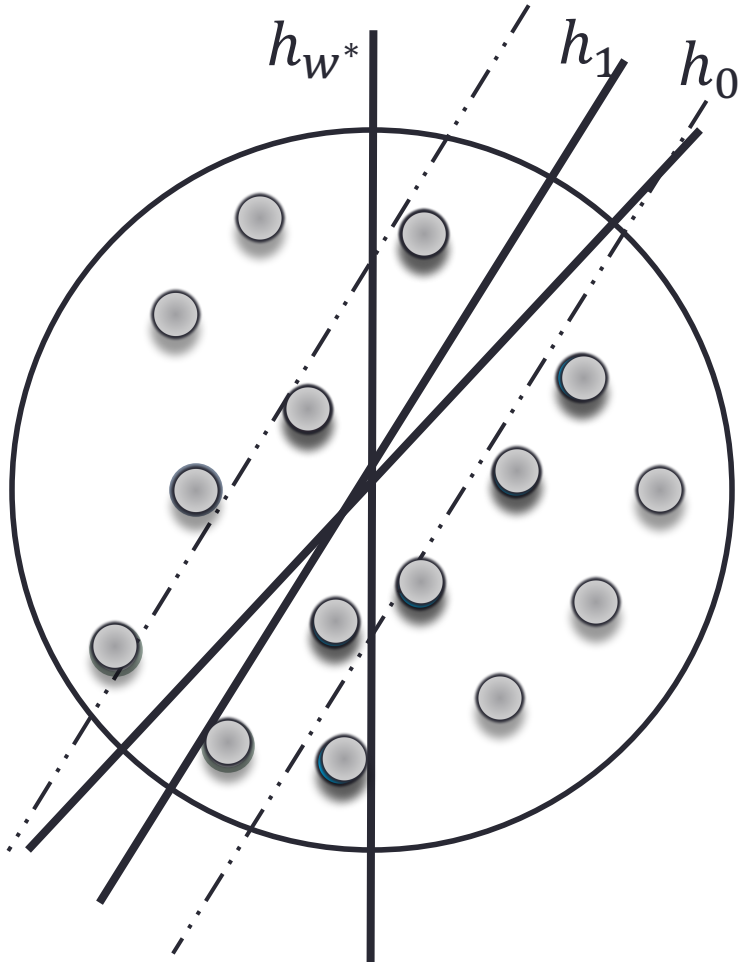
- Idea: Adaptively solve a sequence of convex programs



Label points in band by the polynomial, do hinge loss minimization to constant error, and obtain h_1

Algorithm

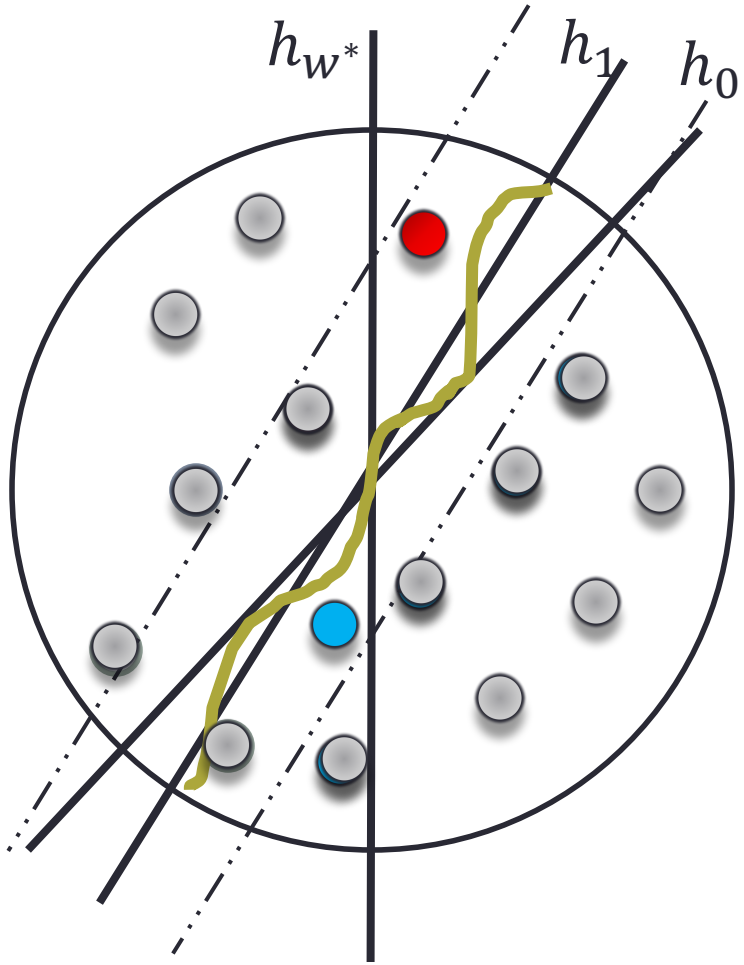
- Idea: Adaptively solve a sequence of convex programs



Halve the bandwidth around h_1, ask labels in the band, fit polynomial

Algorithm

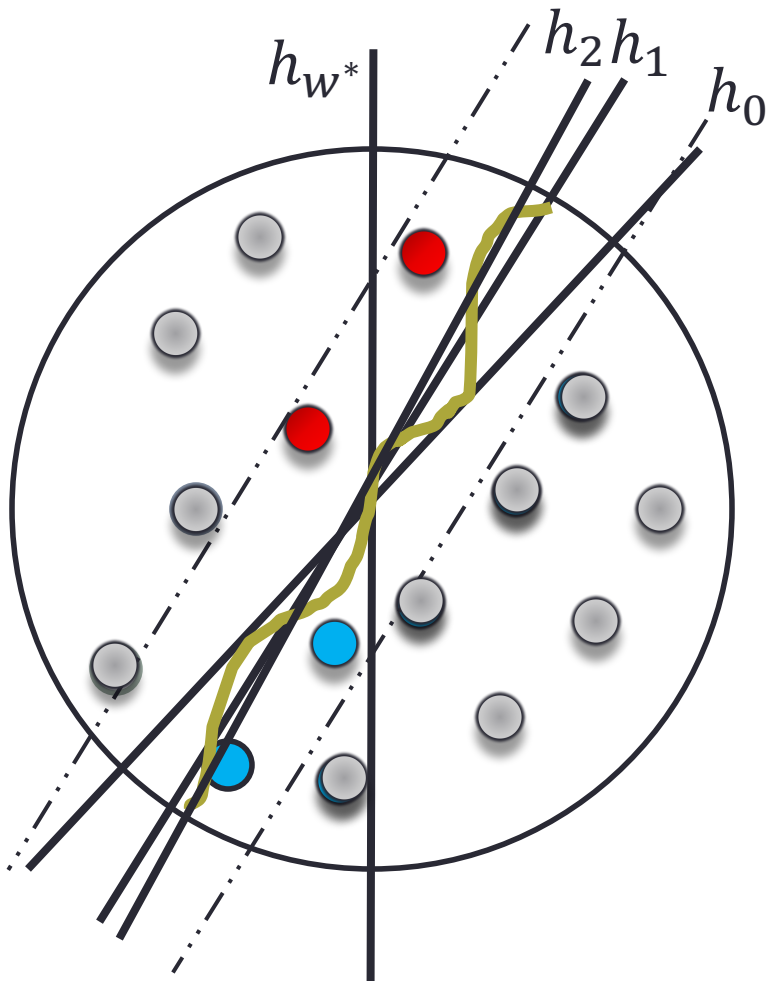
- Idea: Adaptively solve a sequence of convex programs



Halve the bandwidth around h_1, ask labels in the band, fit polynomial

Algorithm

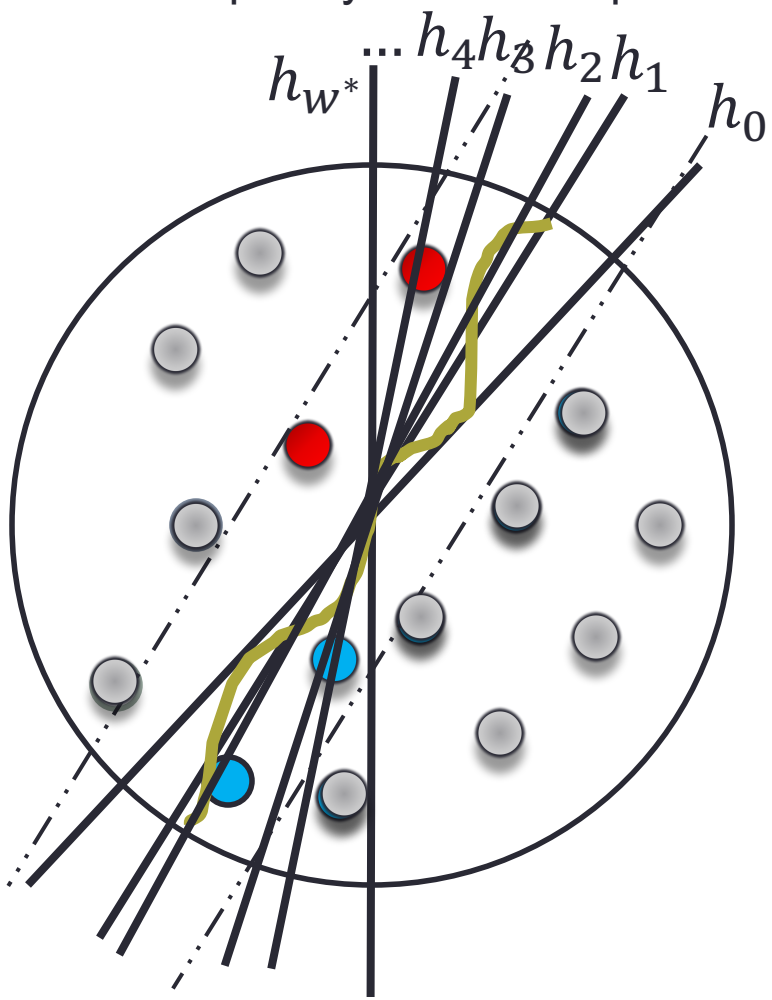
- Idea: Adaptively solve a sequence of convex programs



Label points in band by polynomial, do hinge loss minimization to constant error, and obtain h_2

Algorithm

- Idea: Adaptively solve a sequence of convex programs



Repeat $\log(1/\varepsilon)$ rounds

Main Results

In $\Re^d$, for the log-concave dist. with polynomial time and probability at least 1- δ:

Theorem 1 (Bounded Noise, Learning, ABHZ'16):

Label Complexity: poly*(d,* log*(1/δ),* log*(1/ε))*
Guarantee: $OPT + \varepsilon$

Theorem 2 (Bounded Noise, 1-bit CS, ABHZ'16):

Label Complexity: poly*(t,* log*(d/δ), 1/ε)*
Guarantee: $OPT + \varepsilon$

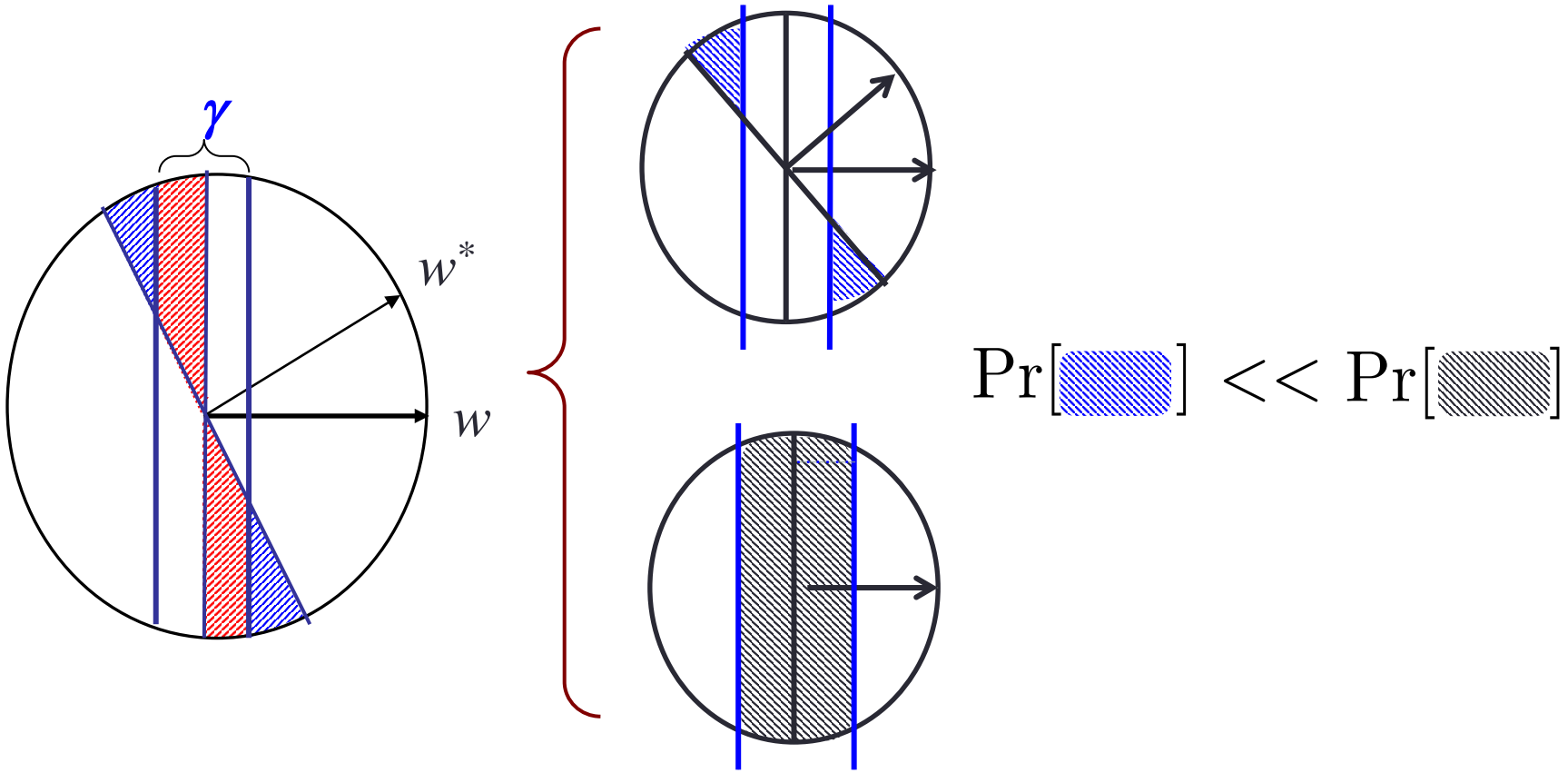
Theorem 3 (Adversarial Noise, 1-bit CS, ABHZ'16):

Label Complexity: *O (t,* polylog*(d/δ), 1/ε)*
Guarantee: $OPT + O(\tau) + \varepsilon$

[ABHZ] Learning and 1-bit Compressed Sensing under Asymmetric Noise, COLT'16

Intuition and Analysis

Most of the errors are near the decision boundary:



Intuition and Analysis

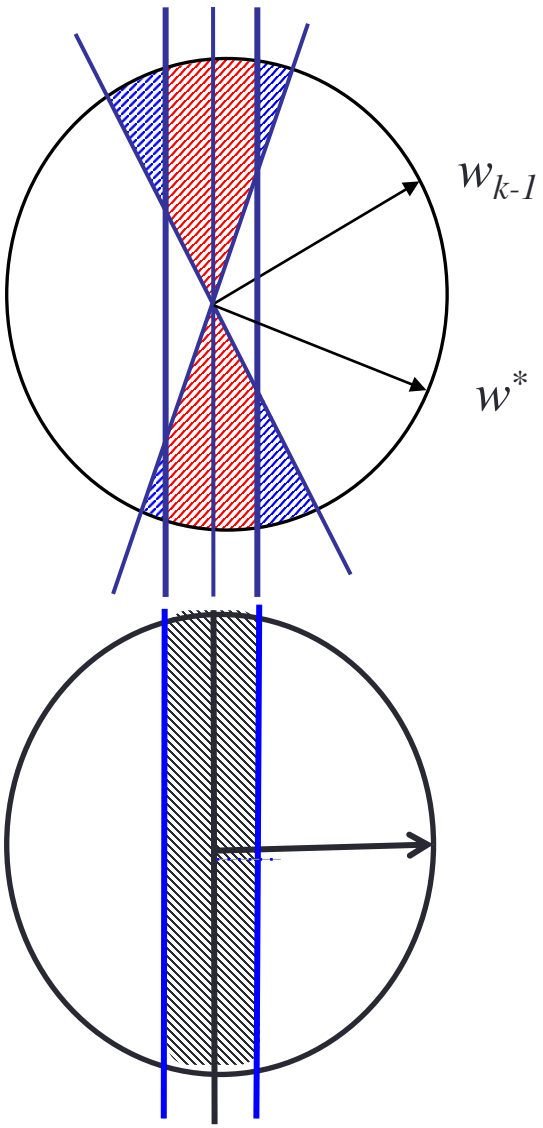
$err(w) = \Pr[\,] + \Pr[\,]$

$\Pr[\,] = \Pr[\,] \times \mathrm{err}_{band}(w)$

$\Pr[\,]$ small

How to find w?

- Hinge loss minimization
- Works only when $\eta \approx 10^{-6}$
- Poly Regression [Kalai et al.] with constant error
- Return a poly, rather than a halfspace
- Combine two together



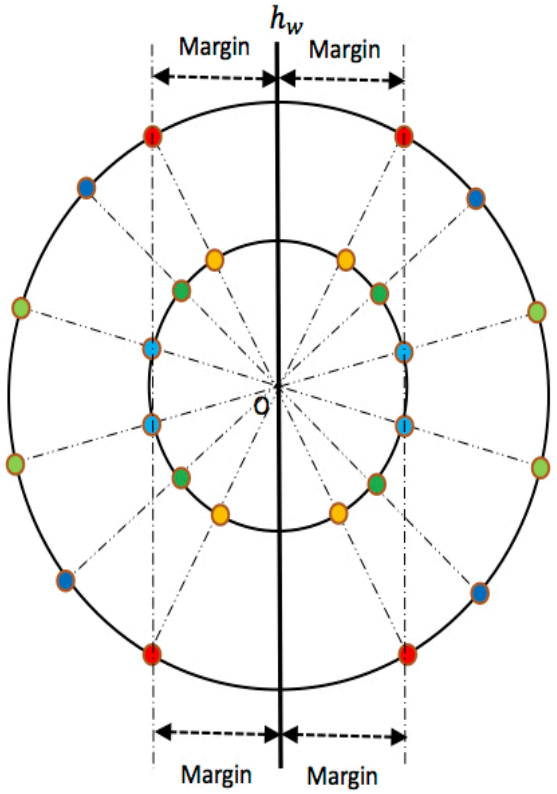
Part I Outline

- Motivation and examples
- Our settings
- Our algorithms
- **Our hardness results**
- Conclusions

Hardness – One shot minimization

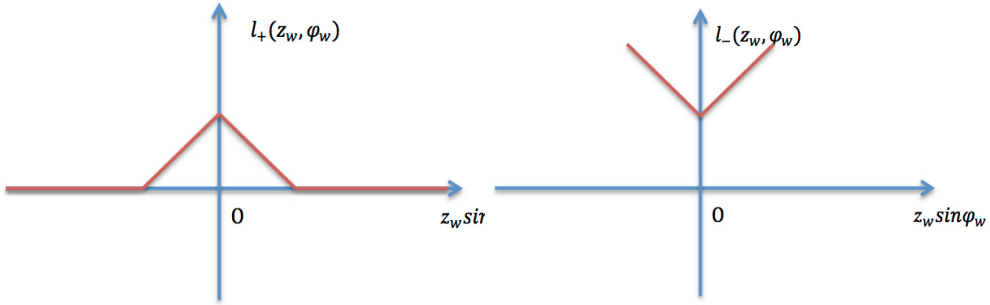
Continuous loss function on h_w satisfies:

- Symmetric w.r.t. h_w
- The loss is larger if h_w is inconsistent with the true label



A couple of examples:

- Hinge, Logistic, Square, Exponential Loss, etc.
  E.g. Hinge:



Hardness – One shot minimization

Theorem 4 (bounded noise, ABHZ'16):

Any one-shot minimization of function satisfying above properties cannot achieve $OPT + \varepsilon$ error under log-concave distribution with bounded noise.

[ABHZ] Learning and 1-bit Compressed Sensing under Asymmetric Noise, COLT'16

Part I Outline

- Motivation and examples
- Our settings
- Our algorithms
- Our hardness results
- **Conclusions**

Part I Conclusions

- Learning of halfspaces and 1-bit CS
  - Polynomial-time algorithm
  - Noise-tolerant for bounded and adversarial noise models
  - Achieve information-theoretic limits
  - Solve a non-convex problem via a sequence of convex programs
- Hardness results
  - One-shot minimization does not work
- Future work
  - Explore the localization technique to the other applications

Part II

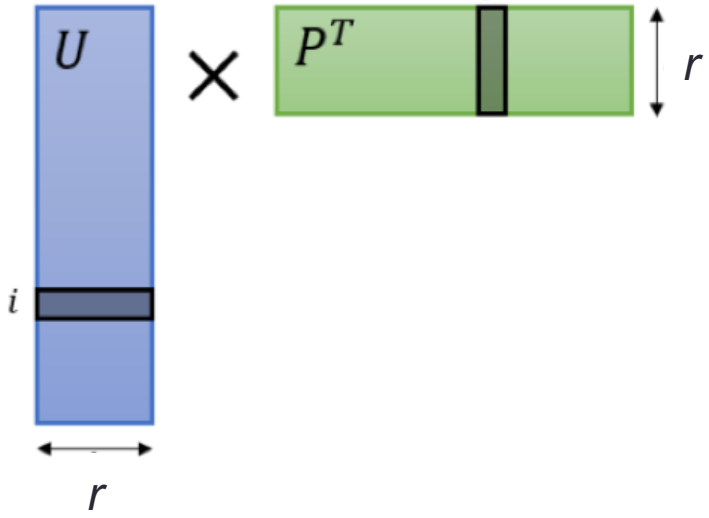
Matrix Completion and Related Problems

(by nice landscape)

Matrix Completion

8	8	1	1
8	8	1	1
4	4	2	2
4	4	2	2
1	1	4	4
1	1	4	4
0	0	8	8
0	0	8	8

$\approx$



Goal: exactly recover the full matrix with #observations as small as possible

Non-Convex Form of Matrix Completion

What we have right now?

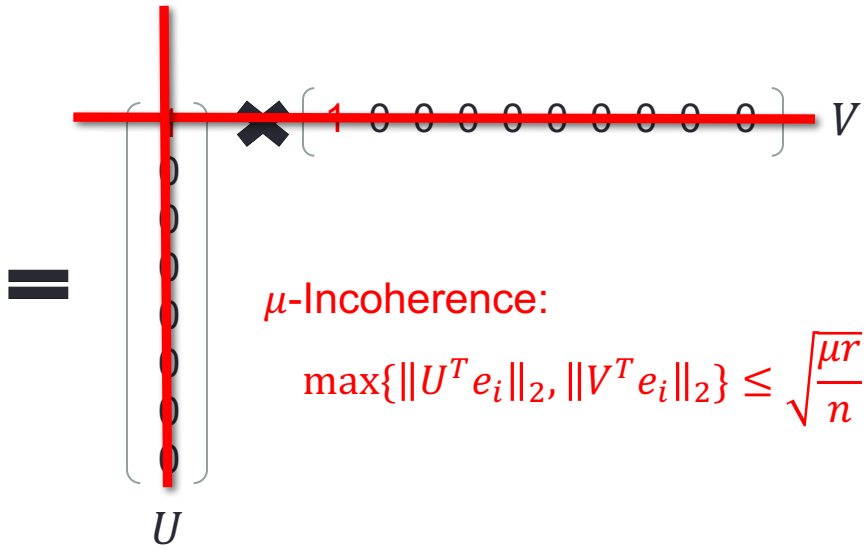
- the observed entries
- $rank(X) \leq r$

What if we solve the non-convex problem?

$$\min_{X,U,V} \|X\|_F, s.t. P_\Omega(X) = P_\Omega(X^*), \quad X = UV. \tag{1}$$

Worst Case of Matrix Completion

=



μ-Incoherence:

$$\max\{\|U^T e_i\|_2, \|V^T e_i\|_2\} \leq \sqrt{\frac{\mu r}{n}}$$

What we have right now?

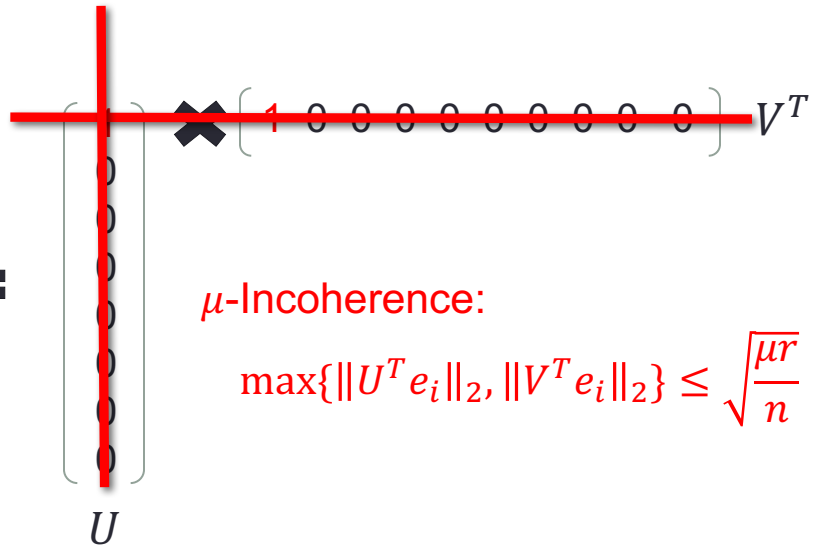
- the observed entries
- $rank(X) \leq r$

What if we solve the non-convex problem?

$$\min_{X,U,V} \|X\|_F, s.t. P_\Omega(X) = P_\Omega(X^*), \quad X = UV. \tag{1}$$

Information-Theoretic Upper Bound

=



μ-Incoherence:

$$\max\{\|U^T e_i\|_2, \|V^T e_i\|_2\} \le \sqrt{\frac{\mu r}{n}}$$

Theorem 1 (BLWZ'18):

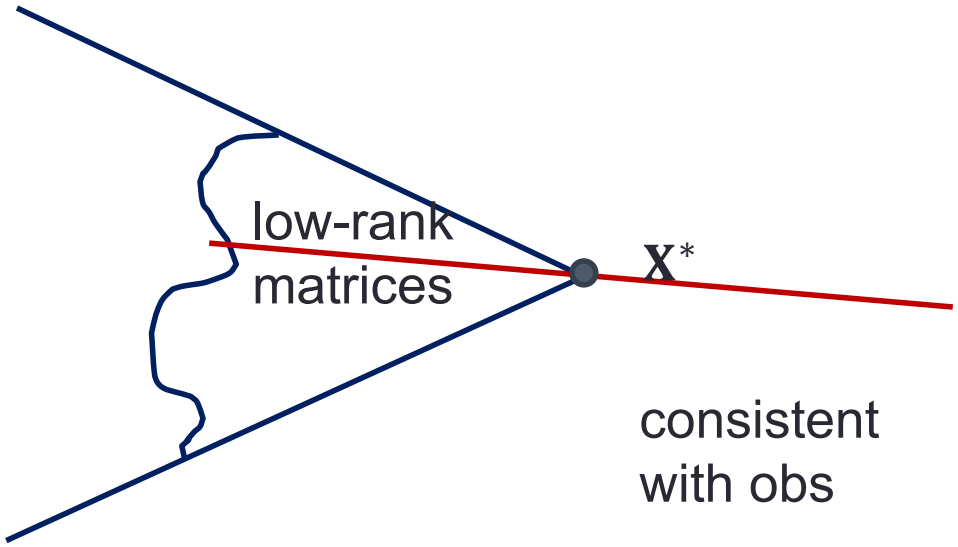
Sample Complexity: $O(\mu n r \log n)$ exactly match lower bound $\Omega(\mu n r \log n)$!
Guarantee: Exact recovery by (1), under incoherence condition

What if we solve the non-convex problem?

$$\min_{X,U,V} \|X\|_F, s.t. P_\Omega(X) = P_\Omega(X^*), \quad X = UV. \qquad (1)$$

[BLWZ] Matrix Completion and Related Problems via Strong Duality, ITCS'18

Proof Idea



Theorem 1 (BLWZ'18):

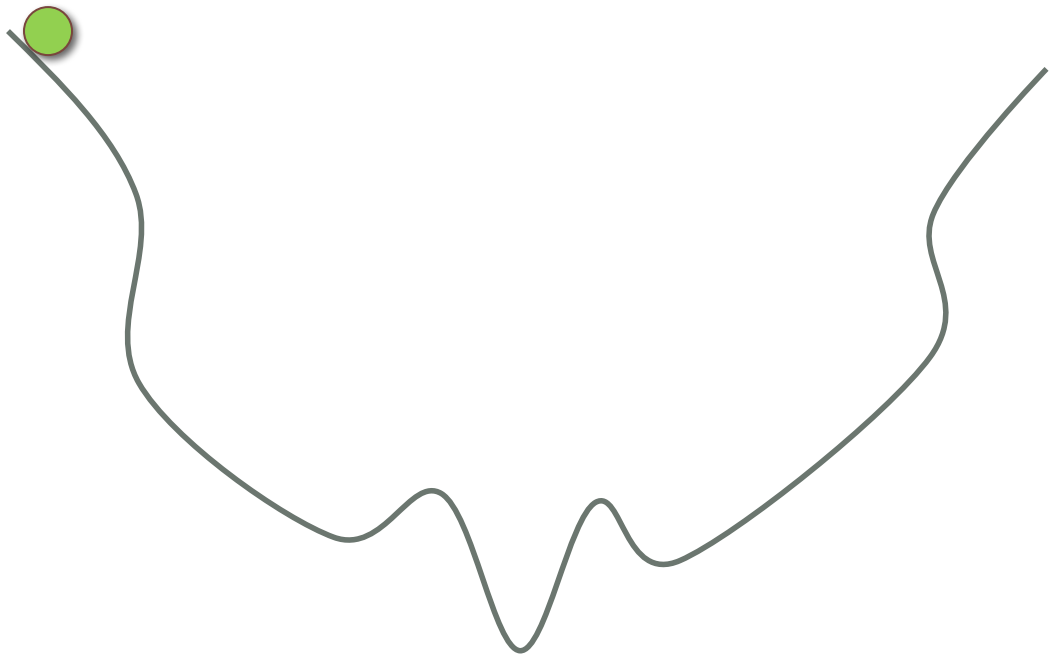
Sample Complexity: $O(\mu nr \log n)$ exactly match lower bound $\Omega(\mu nr \log n)!$
Guarantee: Exact recovery by (1), under incoherence condition

What if we solve the non-convex problem?

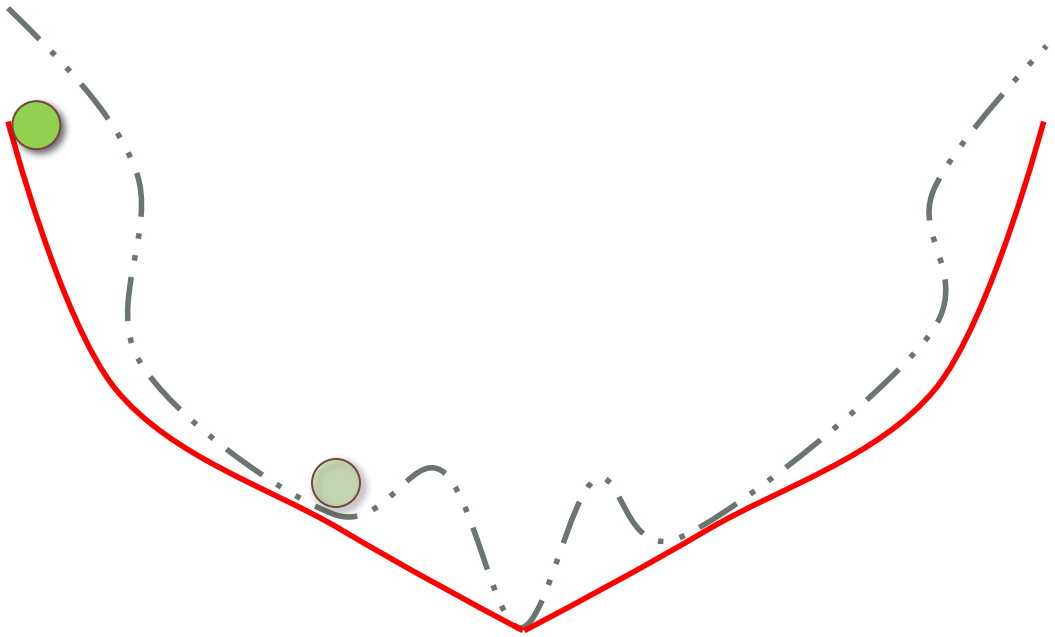
$$\min_{X,U,V} \|X\|_F, s.t. P_\Omega(X) = P_\Omega(X^*), \quad X = UV. \qquad (1)$$

How to solve it efficiently?

Challenges

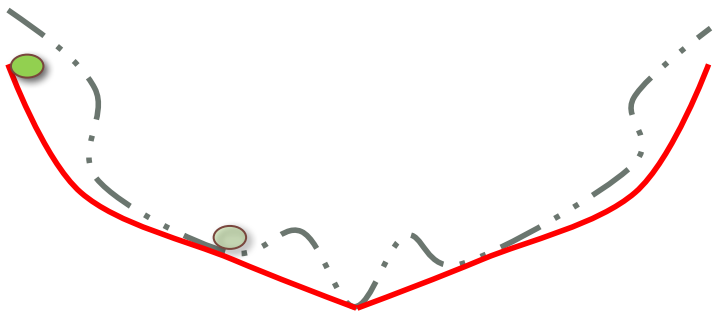


Our Methodology --- Strong Duality



common global optimality

Our Methodology --- Strong Duality



common global optimality

Work	Sample Complexity
[JNS13]	$\mathcal{O}\left(\kappa^4\mu^2 r^{4.5} n_{(1)} \log n_{(1)} \log\left(\frac{r\|\mathbf{X}^*\|_F}{\epsilon}\right)\right)$
[Har14]	$\mathcal{O}\left(\mu r n_{(1)}(r + \log\left(\frac{n_{(1)}\|\mathbf{X}^*\|_F}{\epsilon}\right)\frac{\|\mathbf{X}^*\|_F^2}{\sigma_r^2}\right)$
[GLM16]	$\mathcal{O}(\max\{\mu^6\kappa^{16}r^4, \mu^4\kappa^4 r^6\} n_{(1)} \log^2 n_{(1)})$
[SL15]	$\mathcal{O}(r n_{(1)}\kappa^2 \max\left\{\mu \log n_{(2)}, \sqrt{\frac{n_{(1)}}{n_{(2)}}}\mu^2 r^6 \kappa^4\right\}$
[ZL16]	$\mathcal{O}(\mu r^2 n_{(1)}\kappa^2 \max(\mu, \log n_{(1)}))$
[GLZ17]	$\mathcal{O}\left(\left(\mu^2 r^4\kappa^2 + \mu r \log\left(\frac{\|\mathbf{X}^*\|_F}{\epsilon}\right)\right) n_{(1)} \log\left(\frac{\|\mathbf{X}^*\|_F}{\epsilon}\right)\right)$
[ZWL15]	$\mathcal{O}\left(\mu r^3 n_{(1)} \log n_{(1)} \log\left(\frac{1}{\epsilon}\right)\right)$
[KMO10a]	$\mathcal{O}\left(n_{(2)} r\sqrt{\frac{n_{(1)}}{n_{(2)}}}\kappa^2 \max\left\{\mu \log n_{(2)}, \mu^2 r\sqrt{\frac{n_{(1)}}{n_{(2)}}}\kappa^4\right\}\right)$
[Gro11]	$\mathcal{O}(\mu r n_{(1)} \log^2 n_{(1)})$
[Che15]	$\mathcal{O}(\mu r n_{(1)} \log^2 n_{(1)})$
Ours	$\mathcal{O}(\mu r n_{(1)} \log n_{(1)})$ $\mathcal{O}(\kappa^2 \mu r n_{(1)} \log(n_{(1)}) \log_{2\kappa}(n_{(1)}))$
Lower Bound[1] [CT10]	$\Omega(\mu r n_{(1)} \log n_{(1)})$

Theorem 2 (BLWZ'18):

Sample Complexity: $O(\kappa^2 \mu n r \log n \log_{2\kappa} n)$

Guarantee: Strong duality holds under incoherence condition

$$\min_{X,U,V} \frac{1}{2}\|X\|_F^2, \quad \text{s.t. } P_\Omega(X) = P_\Omega(X^*), \quad X = UV.$$

(non-convex)

share common global optimality

$$\min_X \|X\|_{r*}, \quad \text{s.t. } P_\Omega(X) = P_\Omega(X^*).$$

(convex)

[BLWZ] Matrix Completion and Related Problems via Strong Duality, ITCS'18

Proof Outline

$$\min_{X,U,V} \frac{1}{2}\|X\|_F^2, \text{s.t.}\, P_\Omega(X) = P_\Omega(X^*), X = UV.$$

$$\Updownarrow$$

$$\min_{U,V} \frac{1}{2}\|UV\|_F^2 + H(UV), \text{ where } H \text{ is the indicator function of } P_\Omega(X) = P_\Omega(X^*)$$

Reduction to PCA:

$$\begin{aligned} F(U,V) &= \frac{1}{2}\|UV\|_F^2 + H(UV) \\ &= \frac{1}{2}\|UV\|_F^2 + H^{**}(UV) \\ &= \max_{\Lambda} \frac{1}{2}\|UV\|_F^2 + \langle \Lambda, UV\rangle - H^*(\Lambda) \\ &= \max_{\Lambda} \frac{1}{2}\|-\Lambda - UV\|_F^2 - \frac{1}{2}\|\Lambda\|_F^2 - H^*(\Lambda) \end{aligned}$$

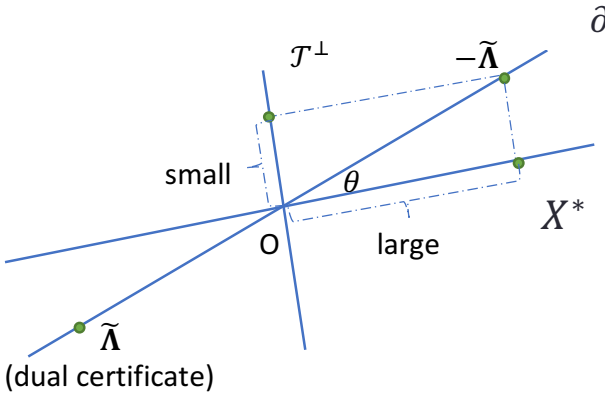
$H(\cdot)$ is convex

Def. of bi-conjugate

PCA for fixed Λ!!!

Find $\widetilde{\Lambda}$ by dual certificate:

(1) $\widetilde{\Lambda} \in \partial H(X^*) = \Omega$
(2) $P_T(-\widetilde{\Lambda}) = X^*$
(3) $\|P_{T^\perp}\widetilde{\Lambda}\| < \sigma_r(X^*)$



θ is small as the samples grow

Hardness results

$$\min_{U,V} F(U,V) = \frac{1}{2}\|UV\|_F^2 + H(UV),$$
H is convex function

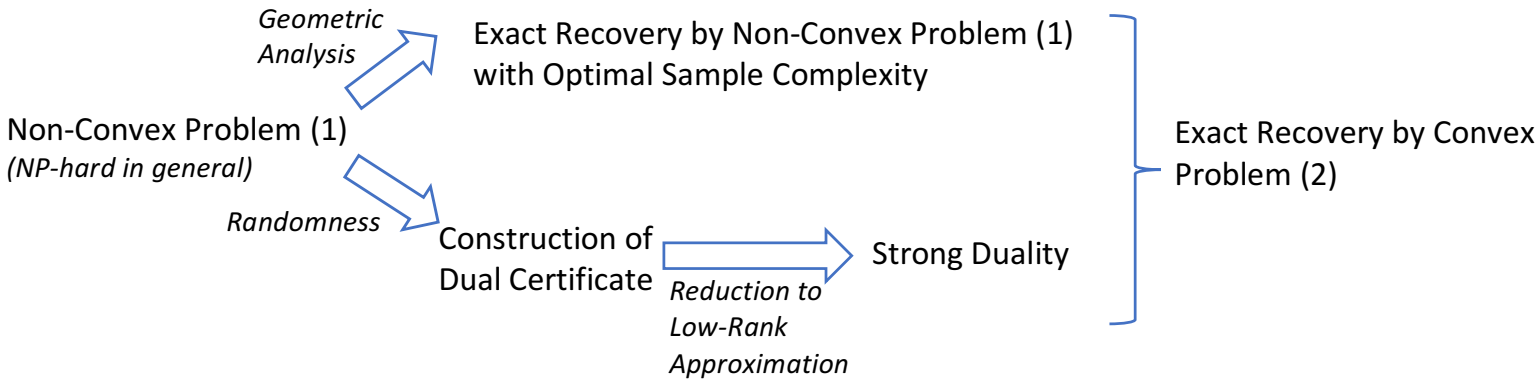
Theorem 3 (Hardness of matrix factorization, BLWZ'18):

Assume the hardness of 4-SAT. Any deterministic algorithm achieving $(1+\varepsilon)OPT$ requires $2^{\Omega(n)}$ time.

[BLWZ] Matrix Completion and Related Problems via Strong Duality, ITCS'18

Part II Conclusions

- Matrix Completion
  - Information-theoretic upper bound
  - A computationally efficient algorithm by strong duality
- Hardness results
  - Generic matrix factorization requires $2^{\Omega(n)}$ time to get $(1+\varepsilon)OPT$
- Future work
  - Explore the strong duality of other problems, e.g., dictionary learning



Thank You