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

convex

non-convex
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!
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

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

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

- **Sparsity (1-bit CS)**: Use a number of samples \( \text{poly}(t, \log(d/\delta), 1/\epsilon) \)

**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 $\eta \leq \frac{1}{2}$, 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 $w \in \mathbb{R}^d$ is $t$-sparse
- Information-theoretic limit: $OPT + \varepsilon$
Asymmetric Noise model – Adversarial Noise

Adversarial Noise:

- The adversary can flip any $\tau$ fraction of labels of $x$.

- No result is known when $w \in \mathbb{R}^d$ 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

As 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 $h_1$
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 $\mathbb{R}^d$, for the log-concave dist. with polynomial time and probability at least $1 - \delta$:

**Theorem 1 (Bounded Noise, Learning, ABHZ'16):**
Label Complexity: $\text{poly}(d, \log(1/\delta), \log(1/\epsilon))$
Guarantee: $OPT + \epsilon$

**Theorem 2 (Bounded Noise, 1-bit CS, ABHZ'16):**
Label Complexity: $\text{poly}(t, \log(d/\delta), 1/\epsilon)$
Guarantee: $OPT + \epsilon$

**Theorem 3 (Adversarial Noise, 1-bit CS, ABHZ'16):**
Label Complexity: $O(t, \text{polylog}(d/\delta), 1/\epsilon)$
Guarantee: $OPT + O(\tau) + \epsilon$

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

Most of the errors are near the decision boundary:

\[ \Pr[\text{[ }} \ll \Pr[\text{[ ]}] \]
Intuition and Analysis

\[ err(w) = Pr[\text{red}] + Pr[\text{blue}] \]
\[ Pr[\text{red}] = Pr[\text{gray}] \times err_{band}(w) \]
\[ Pr[\text{blue}] \text{ small} \]

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 \(\text{OPT} + \varepsilon\) 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

Goal: *exactly* recover the full matrix with number of observations as small as possible.
Non-Convex Form of Matrix Completion

What we have right now?

- the observed entries
- \( \text{rank}(X) \leq r \)

What if we solve the non-convex problem?

\[
\min_{X,U,V} \|X\|_F, \text{ s.t. } P_{\Omega}(X) = P_{\Omega}(X^*), \quad X = UV. \tag{1}
\]
Worst Case of Matrix Completion

What if we solve the non-convex problem?

\[
\min_{X,U,V} \|X\|_F, \text{ s.t. } P_\Omega(X) = P_\Omega(X^*),
X = UV.
\]
Information-Theoretic Upper Bound

What if we solve the non-convex problem?

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

$$X = UV.$$  \hfill (1)

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

[BLWZ] Matrix Completion and Related Problems via Strong Duality, ITCS’18
Proof Idea

What if we solve the non-convex problem?

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

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

consistent with obs

low-rank matrices

X*
Challenges
Our Methodology --- Strong Duality

common global optimality
Our Methodology --- Strong Duality

We consider the optimization problem of matrix completion, which is non-convex, and prove that strong duality holds under an incoherence condition.

**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

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

(non-convex)

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

(convex)

---

[BLWZ] Matrix Completion and Related Problems via Strong Duality, ITCS’18
Proof Outline

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

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

Reduction to PCA:

\[
F(U, V) = \frac{1}{2} \|UV\|_F^2 + H(UV)
\]

\[
= \frac{1}{2} \|UV\|_F^2 + H^\ast(UV)
\]

\[
= \max_{\Lambda} \frac{1}{2} \|UV\|_F^2 + \langle \Lambda, UV \rangle - H^\ast(\Lambda)
\]

\[
= \max_{\Lambda} \frac{1}{2} \|-\Lambda - UV\|_F^2 - \frac{1}{2} \|\Lambda\|_F^2 - H^\ast(\Lambda)
\]

\(H(\cdot)\) is convex

Def. of bi-conjugate

PCA for fixed \(\Lambda\)!!!

Find \(\widetilde{\Lambda}\) by dual certificate:

(1) \(\widetilde{\Lambda} \in \partial H(X^*) = \Omega\)

(2) \(P_T(\widetilde{\Lambda}) = X^*\)

(3) \(\|P_{T^\perp} \widetilde{\Lambda}\| < \sigma_r(X^*)\)

\(\theta\) is small as the samples grow
Hardness results

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

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

\[
\min_{U,V} F(U, V) = \frac{1}{2} \|UV\|_F^2 + H(UV),
\]

\(H\) is convex function.

[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