Recent Progress on Algorithmic Phase Retrieval

Yuejie Chi

Department of Electrical and Computer Engineering

June 2017
AFRL ATR Summer Program

Acknowledgements

- This talk aims to provide an introduction to recent advances in algorithmic phase retrieval. For the purpose of keeping the flow of the exposition, we centered around fast algorithms using a particular loss function, leaving some relevant recent work not covered (such as convex methods and methods using other loss functions).

- Developments of the talk materials are supported by AFOSR Young Investigator Program Award FA9550-15-1-0205, and NSF I/UCRC Center for Surveillance Research.

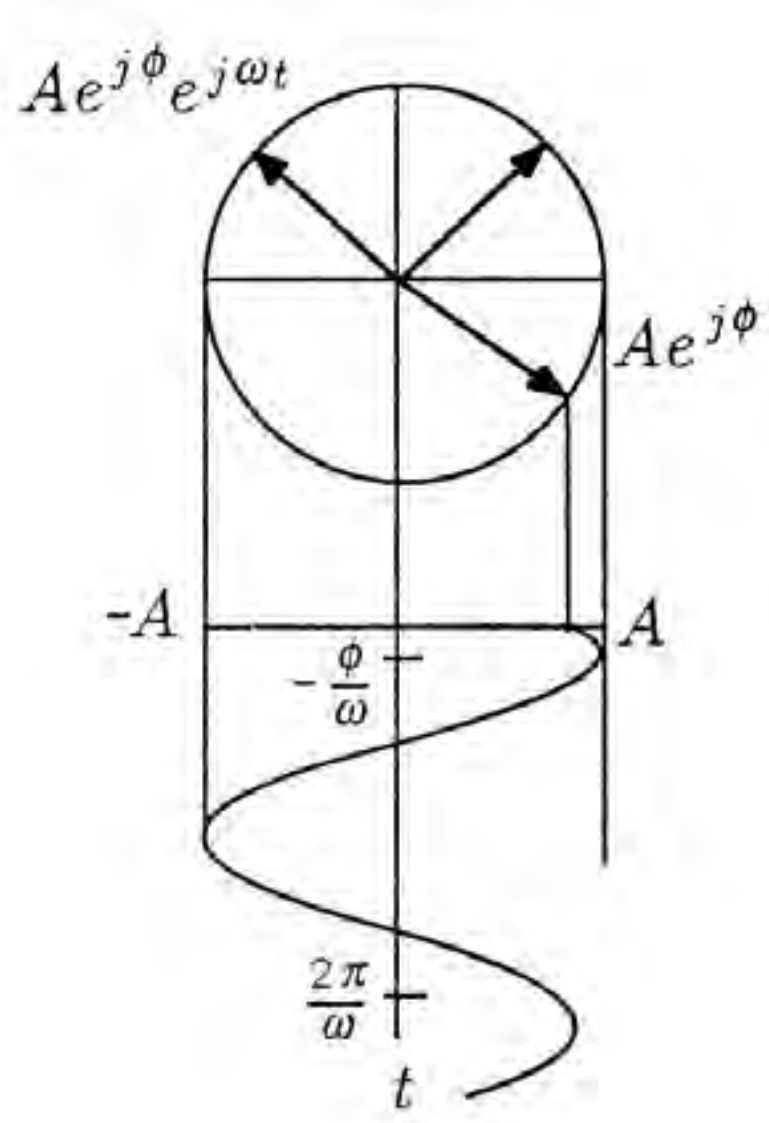
EE 101: Phasors!

A **phasor** is a complex number used to represent a sinusoid.

$$x(t) = A\cos(\omega t + \phi),$$

$$\Updownarrow$$

$$A\angle\phi = Ae^{j\phi}$$

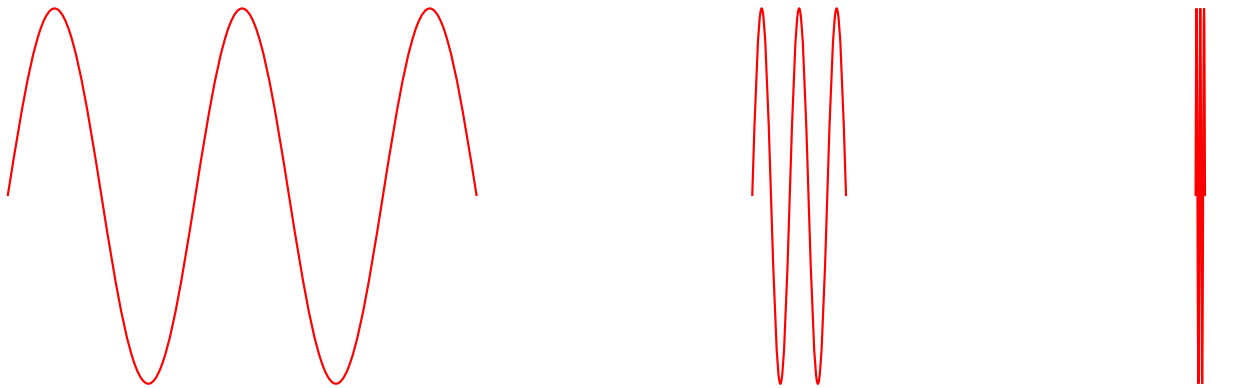


Phasors are convenient tools to represent and manipulate sinusoidal signals (e.g. electromagnetic waves).

- A is the magnitude;
- ϕ is the phase;

Phase Retrieval: The Missing Phase Problem

- In high-frequency (e.g. optical) applications, the (optical) detection devices [e.g., CCD cameras, photosensitive films, and the human eye] **cannot** measure the phase of a light wave.

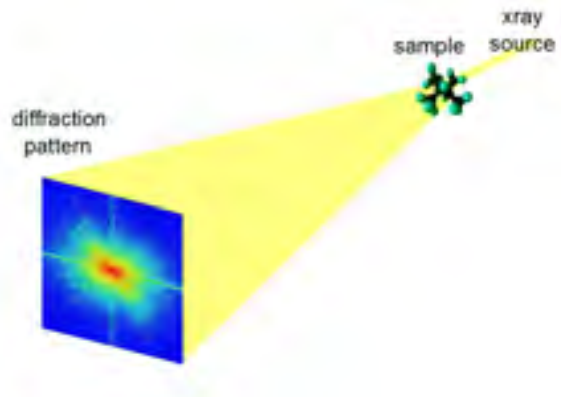


- Optical devices measure the *photon flux* (no. of photons per second per unit area), which is proportional to the magnitude.

- This leads to the so-called *phase retrieval* problem — inference with only intensity measurements.

Coherent Diffraction Imaging

- Given an object illuminated by coherent light, in the far field we obtain the intensity of its Fourier transform.



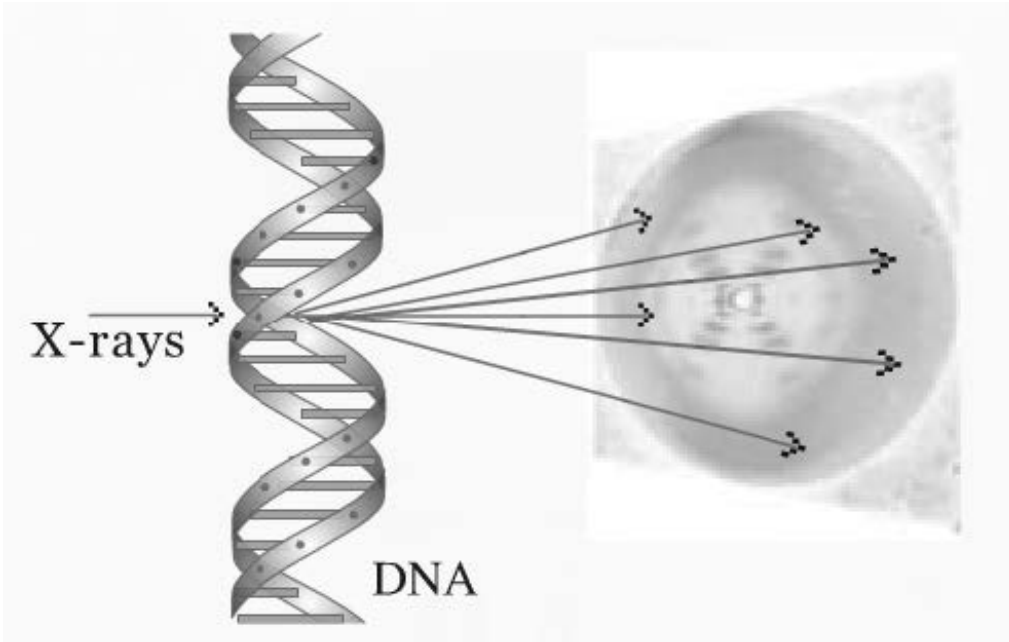
- Mathematically, consider 2-D signal $x(t_1, t_2)$, and its Fourier transform:

$$\hat{X}(\omega_1, \omega_2) = \iint x(t_1, t_2) e^{-j2\pi(t_1\omega_1 + t_2\omega_2)} dt_1 dt_2$$

- We measure $|\hat{X}(\omega_1, \omega_2)|^2$, and want to recover $\hat{X}(\omega_1, \omega_2)$, or equivalently $x(t_1, t_2)$.

X-ray Crystallography and DNA structures

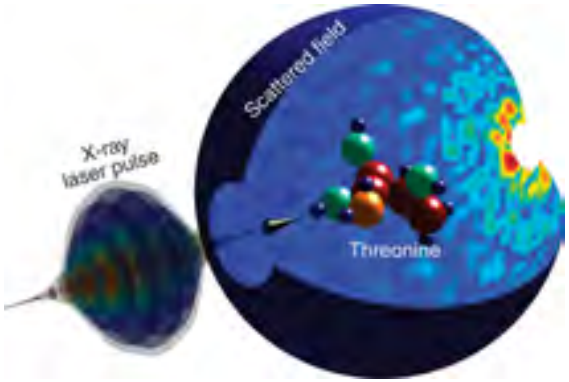
Aided the discovery of the double helix structure of the DNA with X-ray crystallography in 1951.



Nobel Prize for Watson, Crick, and Wilkins in 1962.

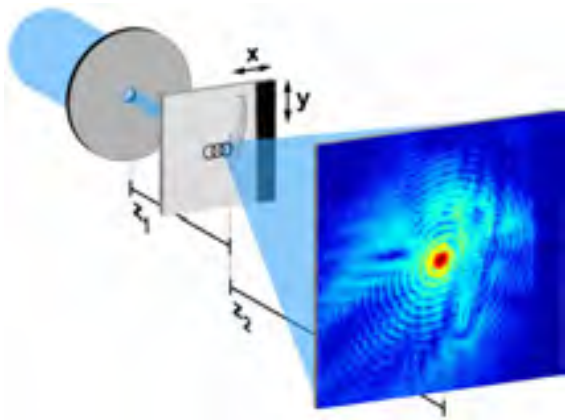
Computational Imaging

Phase retrieval is the foundation for modern computational imaging.

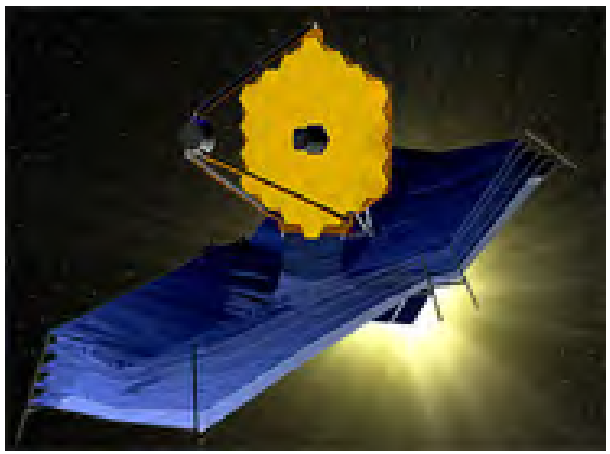


Ankylography

Terahertz Imaging



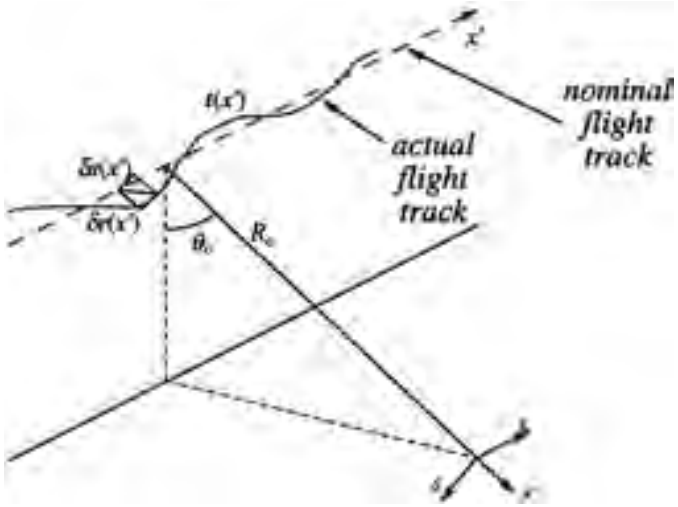
Ptychography



Space Telescope

Phase Retrieval for SAR imaging

- The platform motion instability and electromagnetic propagation in turbulent media affect the phase of the SAR received signal.
- Instead of receiving the nominal signal:



$$h(x', r') = \iint_S \gamma(x, r) g(x' - x, r' - r; x, r) dx dr$$

where (x', r') are the azimuth and range coordinates, γ is the ground reflectivity function, g is the SAR space-dependent unit response, we receive its phase-corrupted version:

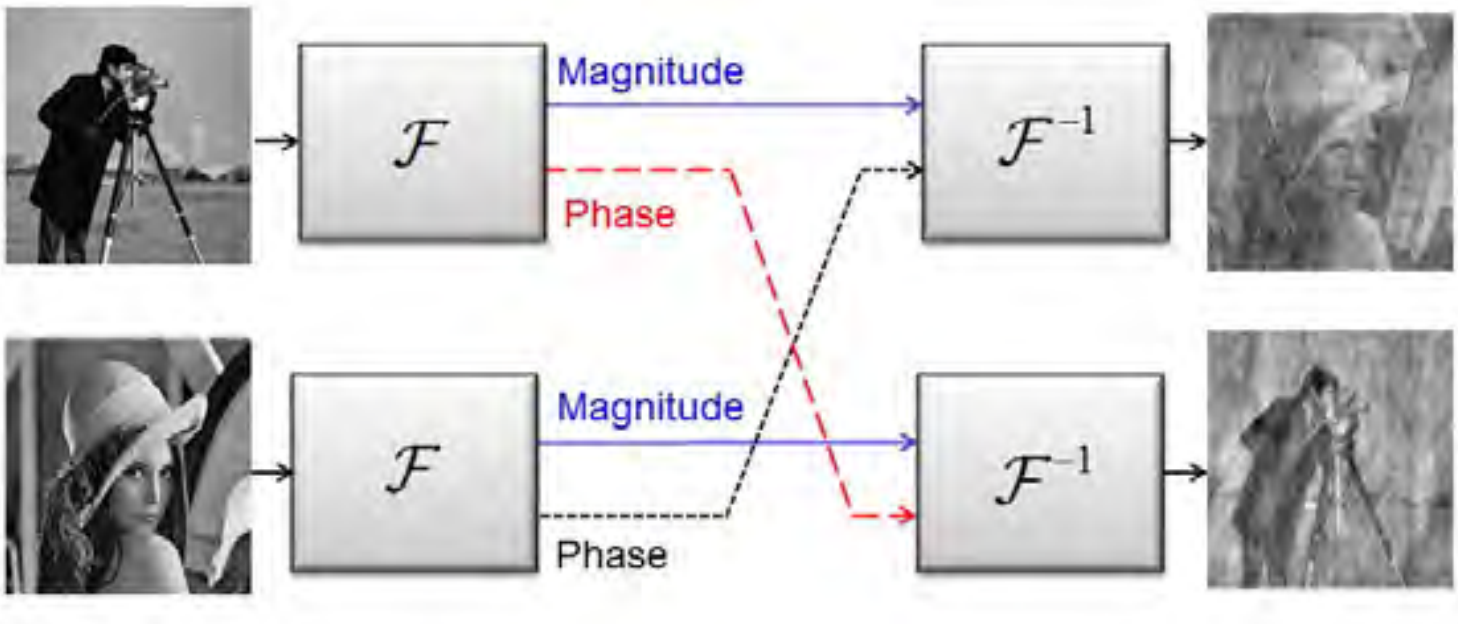
$$\tilde{h}(x', r') = |h(x', r')| e^{j\theta(x', r')}.$$

where $\theta(x', r')$ is the phase error.

Isernia, T., et al. "Image reconstruction from Fourier transform magnitude with applications to synthetic aperture radar imaging." JOSA A 13.5 (1996): 922-934.

Phase information is critical

What happens if we swap the phase of two images in the Fourier domain?



The phase contains much information about the image content.

Figure credit: Shechtman et al. "Phase retrieval with application to optical imaging: a contemporary overview." IEEE Signal Processing Magazine 32.3 (2015): 87-109.

Mathematical Setup

- **Phase retrieval:** estimate $\boldsymbol{x}^\star \in \mathbb{R}^n/\mathbb{C}^n$ from m phaseless measurements:

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}^\star \rangle|, \quad i = 1, \dots, m$$

  where $\boldsymbol{a}_i$ corresponds to the ith measurement vector.
  - $\boldsymbol{a}_i$'s are (coded or oversampled) Fourier transform vectors;
  - $\boldsymbol{a}_i$'s are short-time Fourier transform vectors;
  - $\boldsymbol{a}_i$'s are "generic" vectors such as random Gaussian vectors.

- In a vectorized notation, we write

$$\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}^\star| \in \mathbb{R}^n_+, \quad \text{where} \quad \boldsymbol{A} = \begin{bmatrix} -\boldsymbol{a}_1^*- \\ -\boldsymbol{a}_2^*- \\ \vdots \\ -\boldsymbol{a}_m^*- \end{bmatrix} \in \mathbb{R}/\mathbb{C}^{m\times n}.$$

- Phase retrieval solves a quadratic nonlinear system since:

$$y_i^2 = |\langle \boldsymbol{a}_i, \boldsymbol{x}^\star \rangle|^2 = (\boldsymbol{x}^\star)^* \boldsymbol{a}_i \boldsymbol{a}_i^* \boldsymbol{x}^\star, \quad i = 1, \dots, m,$$

Identifiability

- **Identifiability/Uniqueness:** For any $\phi \in [0, 2\pi)$, $\boldsymbol{x}^\star$ and $e^{j\phi}\boldsymbol{x}^\star$ produce the same measurements:

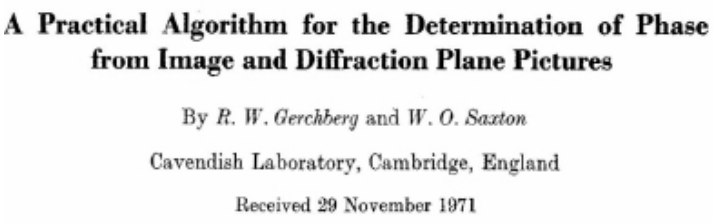
$$|\langle \boldsymbol{a}_i, \boldsymbol{x}^\star e^{j\phi} \rangle| = |\langle \boldsymbol{a}_i, \boldsymbol{x}^\star \rangle|.$$

  therefore, we can only hope to recover/identify $\boldsymbol{x}^\star$ up to a global phase difference.
- Often requires $m > n$ (oversampling!) for identifiability.
- The rule-of-thumb:
  – real-valued $\boldsymbol{x}^\star$: $m \gtrsim 2n$
  – complex-valued $\boldsymbol{x}^\star$: $m \gtrsim 4n$
- We can further reduce the sample complexity if more priors of $\boldsymbol{x}^\star$ can be exploited (such as sparsity and nonnegativity).

Shechtman et al. "Phase retrieval with application to optical imaging: a contemporary overview." IEEE Signal Processing Magazine 32.3 (2015): 87-109.

Algorithms for Phase Retrieval

- The classical algorithms, which started in the 1970s, were proposed by Gerchberg and Saxton (Error Reduction), and later refined by Fienup (Hybrid Input-Output).



A Practical Algorithm for the Determination of Phase from Image and Diffraction Plane Pictures

By *R. W. Gerchberg* and *W. O. Saxton*

Cavendish Laboratory, Cambridge, England

Received 29 November 1971

Reconstruction of an object from the modulus of its Fourier transform

J. R. Fienup

Environmental Research Institute of Michigan, P.O. Box 8618, Ann Arbor, Michigan 48107

Received February 23, 1978

- A lot of recent interest because of
  - modern applications in *computational imaging*: algorithm and sensing co-design;
  - connections with *machine learning*: understanding when nonconvex problems can be solved in a provable manner using simple algorithms.

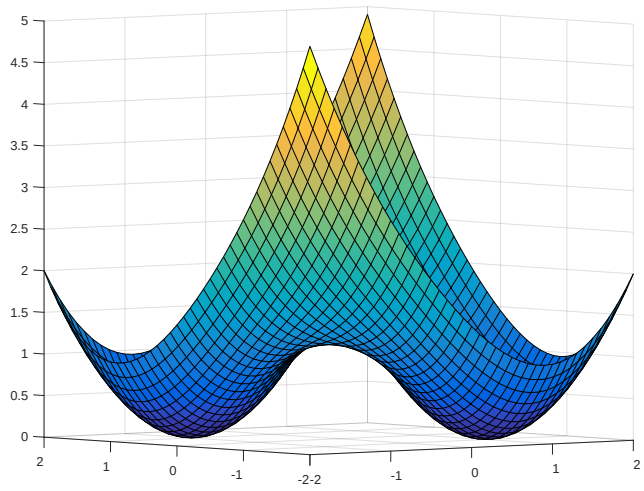
- This talk will focus on iterative algorithms: alternating minimization and gradient descent.

Quadratic Loss of Amplitudes

One can directly recover $\boldsymbol{x}$ by attempting to minimize the quadratic loss of amplitude measurements:

$$\begin{aligned}\ell(\boldsymbol{x}) &:= \frac{1}{m}\left\|\boldsymbol{y} - |\boldsymbol{A}\boldsymbol{x}|\right\|_2^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}\ell(y_i;\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|\right)^2,\end{aligned}$$

which is **nonconvex** and **nonsmooth**.



The expected loss surface when $\boldsymbol{a}_i$'s are Gaussian.

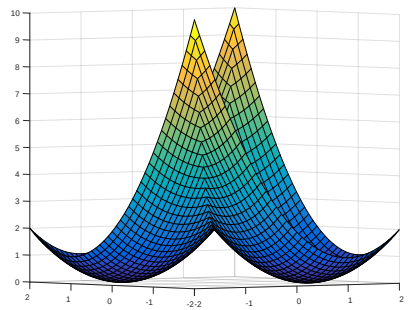
Other choices of loss functions are also possible such as a Poisson loss. The amplitude loss has been observed to have performance advantages in practice, and has been selected in this presentation to maintain a focused exposition.

The Choice of Loss Function is Important

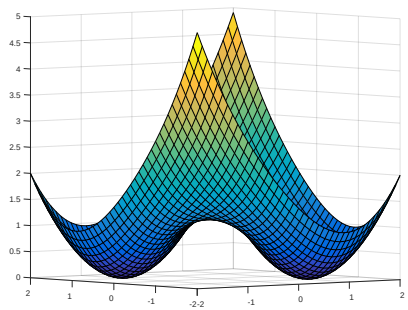
Compare with the intensity-based loss surface:

$$\ell_{WF}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\left(y_i^2 - |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2\right)^2,$$

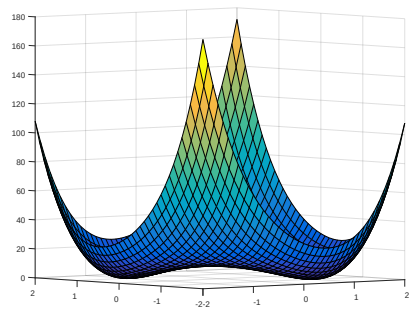
the amplitude-based one has much better curvature.



(a) Expected loss of LS



(b) Amplitude-based



(c) Intensity-based

Figure: Surface of the expected loss function of (a) least-squares (mirrored symmetrically), (b) quadratic loss of amplitudes, and (c) quadratic loss of intensity when $\boldsymbol{x} = [1, -1]^T$.

Phase Retrieval via Alternating Minimization

Error Reduction (ER), proposed by Gerchberg and Saxton in 1972 is based on alternating minimization.

- Define the unit-modulo phase vector $\boldsymbol{b}^\star \in \mathbb{C}^m$ as

$$\boldsymbol{b}^\star = \mathsf{sign}(\boldsymbol{A}\boldsymbol{x}^\star), \quad \text{with} \quad b_i = e^{j\angle\langle \boldsymbol{a}_i, \boldsymbol{x}^\star\rangle}$$

- The magnitude measurements can be written as

$$\mathsf{diag}(\boldsymbol{b}^\star)\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star.$$

- Notice that the loss function $\ell(\boldsymbol{x})$ can be equivalently written as

$$\ell(\boldsymbol{x}) = \min_{|b_i|=1} \|\mathsf{diag}(\boldsymbol{b})\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

One may solve for $(\boldsymbol{x}^\star, \boldsymbol{b}^\star)$ by alternating minimization (AltMin).

Error Reduction

Start with an initialization $\boldsymbol{x}_0$. At iteration $t = 0, 1, \ldots$

1. update the phase as

$$\boldsymbol{b}_{t+1} = \underset{|b_i|=1}{\operatorname{argmin}} \|\mathsf{diag}(\boldsymbol{b})\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_t\|_2^2 = \mathsf{sign}(\boldsymbol{A}\boldsymbol{x}_t),$$

2. update the signal as

$$\begin{aligned}\boldsymbol{x}_{t+1} &= \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\mathsf{diag}(\boldsymbol{b}_{t+1})\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 = \boldsymbol{A}^\dagger \mathsf{diag}(\boldsymbol{b}_{t+1})\boldsymbol{y} \\ &= (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^* \mathsf{diag}(\boldsymbol{y})\mathsf{sign}(\boldsymbol{A}\boldsymbol{x}_t)\end{aligned}$$

The algorithm is guaranteed to not increasing the loss function:

$$\ell(\boldsymbol{x}_{t+1}) \leq \ell(\boldsymbol{x}_t)$$

- ER converges to a **stationary point** of $\ell(x)$, but does not provide guarantees on global convergence or convergence rates.
- In practice a **random initialization** is typically used, and the performance is sensitive to the initialization.

Phase Retrieval via Gradient Descent

- The generalized gradient of $\ell(\boldsymbol{x})$ can be calculated as

$$\nabla \ell(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} \left(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - y_i \cdot \mathsf{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle) \right) \boldsymbol{a}_i$$

- Start with an initialization $\boldsymbol{x}_0$. At iteration $t = 0, 1, \ldots$

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{x}_t - \mu \nabla \ell(\boldsymbol{x}_t) \\ &= \left(\boldsymbol{I} - \frac{\mu}{m} \boldsymbol{A}^* \boldsymbol{A} \right) \boldsymbol{x}_t + \frac{\mu}{m} \boldsymbol{A}^* \mathsf{diag}(\boldsymbol{y}) \mathsf{sign}(\boldsymbol{A} \boldsymbol{x}_t), \end{aligned}$$

  where μ is the step size.

- Referred to as the **Reshaped Wirtinger Flow (RWF)**.
- Side-by-side comparison with the AltMin update:

$$\boldsymbol{x}_{t+1} = (\boldsymbol{A}^* \boldsymbol{A})^{-1} \boldsymbol{A}^* \mathsf{diag}(\boldsymbol{y}) \mathsf{sign}(\boldsymbol{A} \boldsymbol{x}_t)$$

Statistical Measurement Model

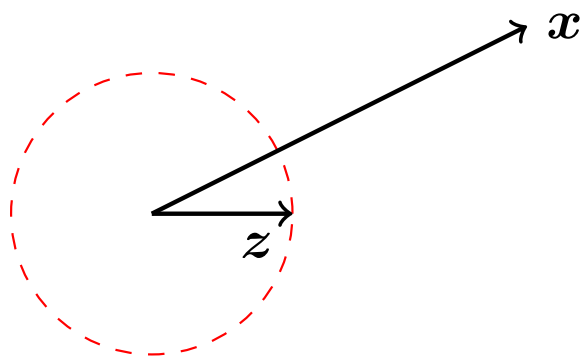
Strong performance guarantees are possible by leverage statistical properties of the measurement ensemble.

- **Gaussian measurement model:**

$$\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \quad \text{i.i.d.} \quad \text{if real-valued,}$$
$$\boldsymbol{a}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}) \quad \text{i.i.d.} \quad \text{if complex-valued,}$$

- Distance measure:

$$\text{dist}(\boldsymbol{x}, \boldsymbol{z}) = \min_{\phi \in [0, 2\pi)} \|\boldsymbol{x} - e^{j\phi}\boldsymbol{z}\|.$$



Local Linear Convergence of AltMin (ER)

Theorem (Waldspurger 2016)

Assume the random Gaussian measurement model. There exist universal constants C, c_1, c_2 such as long as $m \geq Cn$, provided that we initialize in the neighborhood of the ground truth $\boldsymbol{x}^\star$, i.e.

$$dist(\boldsymbol{x}_0, \boldsymbol{x}^\star) \leq \frac{1}{10}\|\boldsymbol{x}^\star\|,$$

then with probability at least $1 - c_1 \exp(-c_2 m)$, the iterates of ER or AltMin algorithm satisfies for some $0 < \rho < 1$:

$$dist(\boldsymbol{x}_t, \boldsymbol{x}^\star) \leq (1-\rho)^t \|\boldsymbol{x}^\star\|, \quad \forall t \in \mathbb{N}_+.$$

- **Sample complexity:** only $m = O(n)$ samples to guarantee local convergence;
- **Linear rate of convergence:** only $\log(1/\epsilon)$ iterations to reach an accuracy $\text{dist}(\boldsymbol{x}_t, \boldsymbol{x}^\star)/\|\boldsymbol{x}^\star\| \leq \epsilon$.

Waldspurger, "Phase retrieval with random Gaussian sensing vectors by alternating projections", arXiv:1609.0308.

Local Linear Convergence of Gradient Descent

Theorem (Zhang, Zhou, Liang, C., 2016)

Assume the random Gaussian measurement model. There exist universal constants C, c_1, c_2 such as long as $m \geq Cn$, provided that we initialize in the neighborhood of the ground truth $\boldsymbol{x}^\star$, i.e.

$$\mathit{dist}(\boldsymbol{x}_0, \boldsymbol{x}^\star) \leq \frac{1}{10}\|\boldsymbol{x}^\star\|,$$

then with probability at least $1 - c_1 \exp(-c_2 m)$, the iterates of RWF satisfies for some $0 < \rho < 1$:

$$\mathit{dist}(\boldsymbol{x}_t, \boldsymbol{x}^\star) \leq (1-\rho)^t \|\boldsymbol{x}^\star\|, \quad \forall t \in \mathbb{N}_+.$$

- **Sample complexity:** only $m = O(n)$ samples to guarantee local convergence;
- **Linear rate of convergence:** only $\log(1/\epsilon)$ iterations to reach an accuracy $\mathrm{dist}(\boldsymbol{x}_t, \boldsymbol{x}^\star)/\|\boldsymbol{x}^\star\| \leq \epsilon$.

Zhang, Zhou, Liang and **C.**, "Reshaped Wirtinger Flow and Incremental Algorithms for solving Quadratic Systems of Equations", in revision to *Journal of Machine Learning Research*.

Spectral Method for Initialization

- Key observation: consider the weighted matrix

$$\boldsymbol{Y} = \sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^*, \quad \text{where} \quad \mathbb{E}[\boldsymbol{Y}] = \lambda \boldsymbol{x}^\star (\boldsymbol{x}^\star)^*$$

  for some $\lambda > 0$.
- The top eigenvector of $\boldsymbol{Y}$ provides a good initialization (plus estimate the norm $\|\boldsymbol{x}^\star\|$) as long as $m \gtrsim n$.
- For the Gaussian model, a better initialization is obtained by truncating samples with large values.

Theorem (Chen and Candès, Zhang et.al., Wang et.al., etc...)

With high probability, the spectral method produces an initialization that satisfies

$$\mathit{dist}(\boldsymbol{x}_0, \boldsymbol{x}^\star) \leq \frac{1}{10} \|\boldsymbol{x}^\star\|$$

Zhang, Zhou, Liang and **C.**, "Reshaped Wirtinger Flow and Incremental Algorithms for solving Quadratic Systems of Equations", in revision to *Journal of Machine Learning Research*.

Performance of Spectral Methods

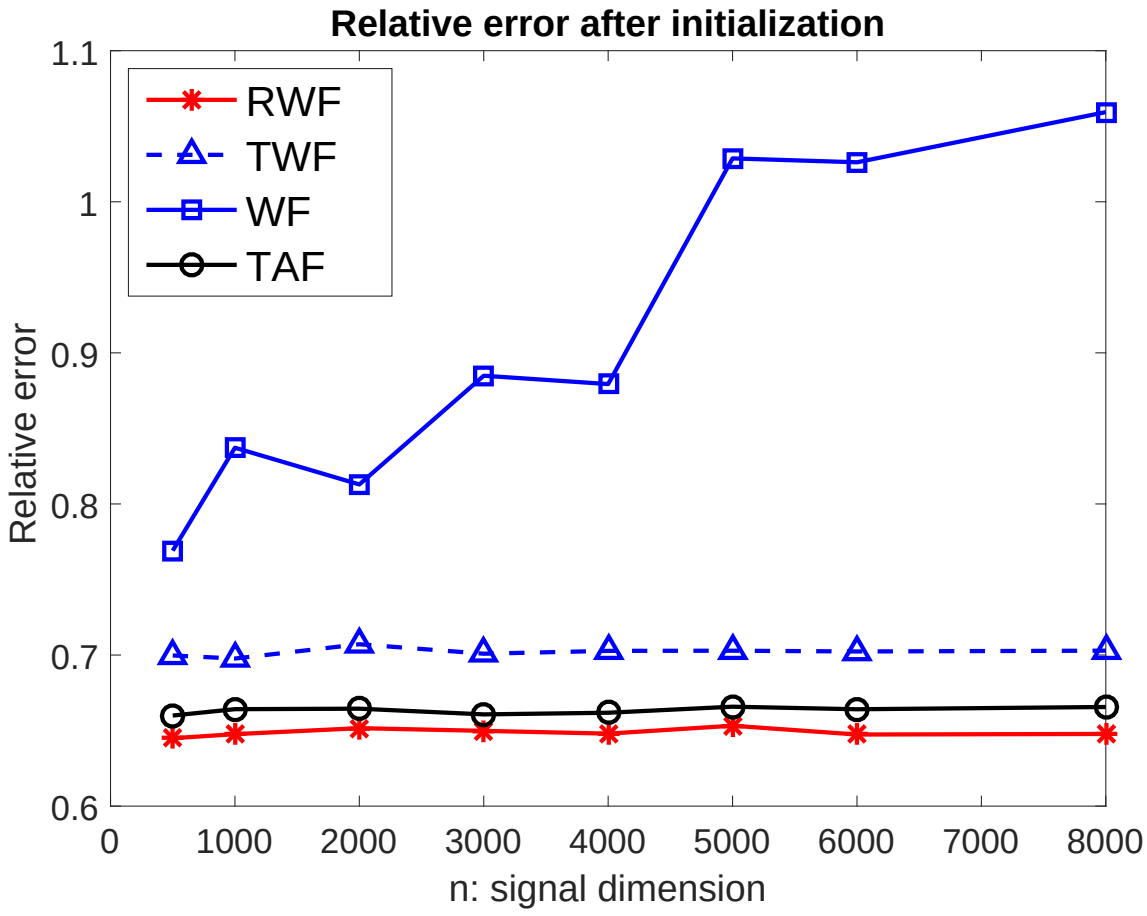
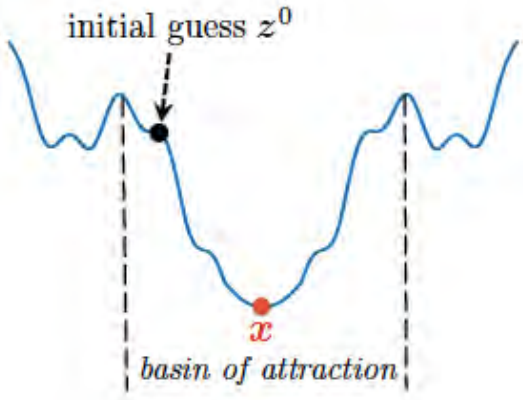


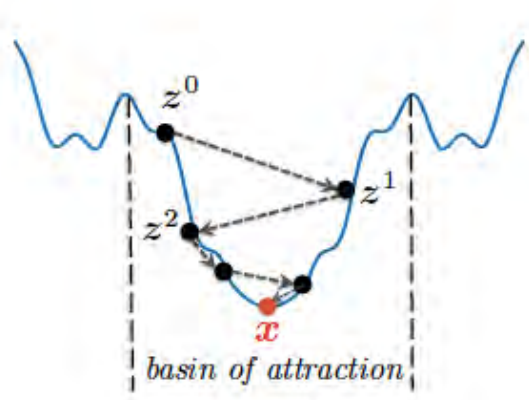
Figure: Comparison of three initialization methods with $m = 6n$ and 50 iterations using power method.

Global Convergence

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\mathbb{R}^n/\mathbb{C}^n}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} \ell(y_i;\boldsymbol{x})$$

- Initialize $\boldsymbol{z}^{(0)}$ via *spectral* methods to land in the neighborhood of the ground truth;
- Iterative update using *simple* methods such as gradient descent and alternating minimization;





Provable near-optimal performance for Gaussian measurement model:

- Statistically: $m = O(n)$ near-optimal sample complexity
- Computationally: linear convergence with near-linear run time.

Figure credit: Yuxin Chen.

Stochastic Gradient Descent

- Stochastic algorithms sometimes are in favor for memory or streaming considerations.
- Consider the stochastic gradient descent (SGD) method,

$$\begin{aligned}\boldsymbol{x}_{t+1} &= \boldsymbol{x}_t - \mu \nabla \ell(y_{i_t}; \boldsymbol{x}_t) \\ &= \boldsymbol{x}_t - \mu \left(\boldsymbol{a}_{i_t}^* \boldsymbol{x}_t - y_{i_t} \cdot \mathsf{sign}(\boldsymbol{a}_{i_t}^* \boldsymbol{x}_t) \right) \boldsymbol{a}_{i_t}\end{aligned}$$

  where i_t is drawn from $\{1, 2, \ldots, m\}$ uniformly at random.
- To fully exploit system throughput, often mini-batch version:

$$\begin{aligned}\boldsymbol{x}_{t+1} &= \boldsymbol{x}_t - \mu \nabla \ell(\boldsymbol{y}_{\Gamma_t}; \boldsymbol{x}_t) \\ &= \boldsymbol{x}_t - \mu \cdot \boldsymbol{A}_{\Gamma_t}^* \left(\boldsymbol{A}_{\Gamma_t} \boldsymbol{x}_t - \boldsymbol{y}_{\Gamma_t} \odot \mathsf{sign}(\boldsymbol{A}_{\Gamma_t} \boldsymbol{x}_t) \right),\end{aligned}$$

  where Γ_t is a subset of size K that is drawn uniformly at random from all size-K subsets of $\{1, 2, \ldots, m\}$.

Performance of SGD

Theorem (Zhang, Zhou, Liang, C., 2017)

Assume the random Gaussian measurement model. There exist some universal constants $0 < \rho, \rho_0, \nu < 1$ *and* $c_0, c_1, c_2 > 0$ *such that if* $m \geq c_0 n$ *and* $\mu = \rho_0 / n$*, then with probability at least* $1 - c_1 \exp(-c_2 m)$*, mini-batch SGD yields*

$$\mathbb{E}_{\Gamma^t}\left[\mathrm{dist}^2(\boldsymbol{x}_t, \boldsymbol{x}^\star)\right] \leq \nu \left(1 - \frac{K\rho}{n}\right)^t \|\boldsymbol{x}^\star\|^2, \quad \forall t \in \mathbb{N}_+,$$

if initialized by the spectral method, where $\mathbb{E}_{\Gamma^t}[\cdot]$ *denotes the expectation with respect to the randomness in* $\Gamma^t = \{\Gamma_1, \Gamma_2, \ldots, \Gamma_t\}$ *conditioned on the high probability event of random measurements* $\{\boldsymbol{a}_i\}_{i=1}^m$*.*

- Linear convergence of SGD is established for a non-convex and non-smooth loss function.
- The mini-batch size K trades-off the complexity per iteration and the convergence rate.

Zhang, Zhou, Liang and **C.**, "Reshaped Wirtinger Flow and Incremental Algorithms for solving Quadratic Systems of Equations", in revision to *Journal of Machine Learning Research*.

Connection to Kaczmarz Method

- The Kaczmarz method is conventionally a method for solving linear systems. We attempt to extend it to solve phase retrieval:

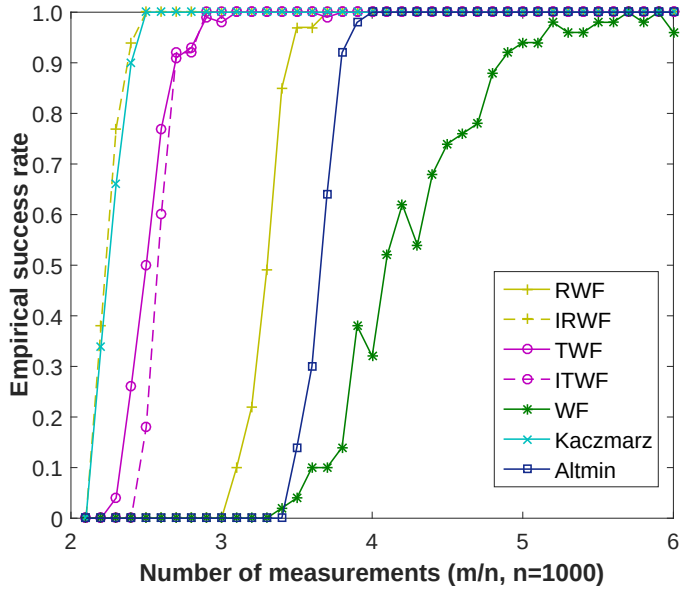
$$\begin{aligned}\boldsymbol{x}_{t+1} &= \underset{y_{i_t}=|\langle \boldsymbol{a}_{i_t}, \boldsymbol{x}\rangle|}{\operatorname{argmin}} \|\boldsymbol{x}-\boldsymbol{x}_t\|_2^2 \\ &= \boldsymbol{x}_t - \frac{1}{\|\boldsymbol{a}_{i_t}\|^2}\left(\boldsymbol{a}_{i_t}^*\boldsymbol{x}_t - y_{i_t}\cdot \operatorname{sign}(\boldsymbol{a}_{i_t}^*\boldsymbol{x}_t)\right)\boldsymbol{a}_{i_t},\end{aligned}$$

  where i_t is drawn uniformly at random from $\{1,\dots,m\}$.
- The update rule is surprisingly simple in a close form without any tuning paramters despite the nonlinear constraint.
- In fact, it becomes equivalent to SGD if we set the step size of SGD as $\mu = \frac{1}{\|\boldsymbol{a}_{i_t}\|^2} \sim \frac{1}{n}$ since $\|\boldsymbol{a}_{i_t}\|^2$ concentrates around n.
- Therefore a similar linear convergence can be established for Kaczmarz methods, and works in mini-batch as well.

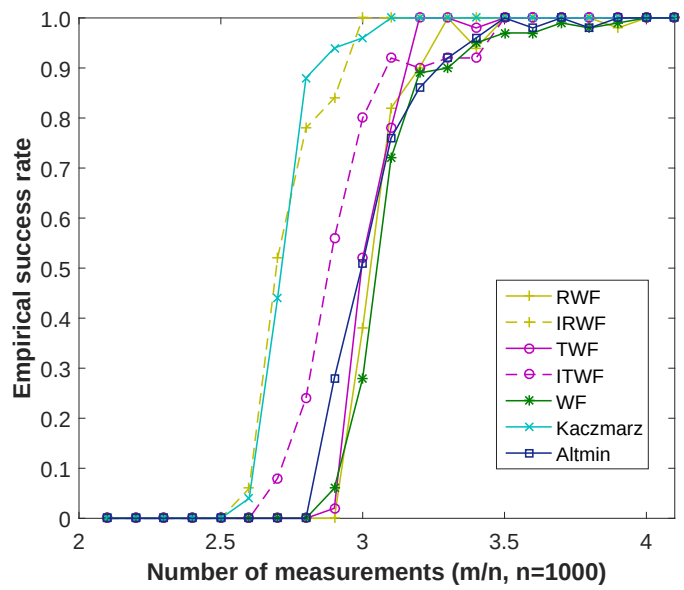
Performance on Gaussian Model

We first look at the sample complexity of a few algorithms:

- Gradient descent type algorithms: RWF (proposed loss), TWF (Poisson loss), WF (quadratic loss of intensity);
- Stochastic algorithms: IRWF (stochastic version of RWF), ITWF (stochastic version of TWF), Kaczmarz;
- Alternating Minimization (Error Reduction).



(a) Real Gaussian



(b) Complex Gaussian

Figure: The stochastic methods IRWF/Kaczmarz achieves the best sample complexity.

Computational Complexity

We next look at the computational complexity. For stochastic algorithms we cycle through the measurements several passes.

Table: Comparison of number of passes and time cost $(n = 5000, m = 8n)$.

		Real Gaussian		Complex Gaussian	
		#passes	time(s)	# passes	time(s)
Batch methods	RWF	72	**12.66**	176	**122.4**
	AltMin	**6**	79.58	**159**	9637
Stochastic methods	IRWF	9	44.77	21	233.2
	minibatch IRWF (64)	9	**8.076**	**21**	**48.58**
	Kaczmarz	9	50.68	21	248.4
	block Kaczmarz (64)	**8**	28.50	22	89.31

- A mini-batch IRWF with $K = 64$ provides best performance. It outperforms Kaczmarz by using a constant step size.

Performance on Coded Diffraction Imaging

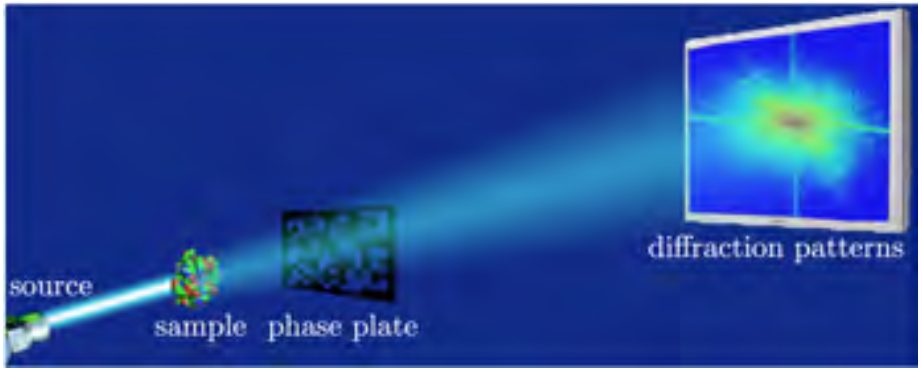


Figure: Coded diffraction imaging: a number of random masks is placed between the sample and the far field to modulate the Fourier transform.

	Algorithms	GD	SGD/Kaczmarz	AltMin
$L = 6$	#passes	140	**24**	230
	time cost(s)	110	**21.2**	167
$L = 12$	#passes	70	**8**	120
	time cost(s)	107	**13.7**	171

Table: Comparison of iterations and time cost among algorithms on Galaxy image (1920×1080), where $L = m/n$ denotes the number of CDP masks.

Robust Phase Retrieval with Outliers

What if the measurements are noisy and corrupted?

- Assume the measurements are corrupted by both *sparse outliers* and *bounded noise*:

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle| + \eta_i + w_i, \quad i = 1, \dots, m,$$

  where $\|\boldsymbol{\eta}\|_0 \le s \cdot m$ is the sparse outlier and $\boldsymbol{w}$ is bounded, $0 \le s < 1$ is the fraction of outliers.

- Outliers happen with sensor failures, malicious attacks, ...

- Goal: develop algorithms that are *oblivious* to outliers, and statistically and computationally efficient.
  - performs equally well regardless of the existence of outliers;
  - small sample size: hopefully m is linear in n;
  - large fraction of outliers: hopefully s is a small constant;
  - low computational complexity and easy to implement.

Existing Approaches are not Robust

In the presence of *arbitrary outliers*, **earlier approaches fail:**

- Spectral initialization would fail: the eigenvector of $\boldsymbol{Y}$ can be arbitrarily perturbed

$$\boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^*$$

$$\text{or} \quad \boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^* \mathbb{1}_{\{|y_i| \le \alpha_y \cdot \mathsf{mean}(\{y_i\})\}}.$$

- Gradient descent would fail: the search direction can be arbitrarily perturbed

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\mu}{m}\sum_{i=1}^{m} \nabla \ell(y_i; \boldsymbol{x}_t)$$

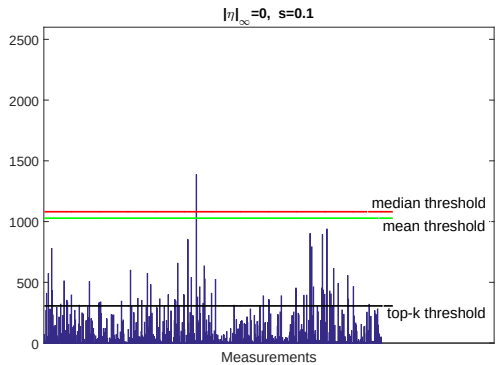
We can no longer guarantee the performance of the algorithm even with a single outlier! Need better strategies.

Median Truncation

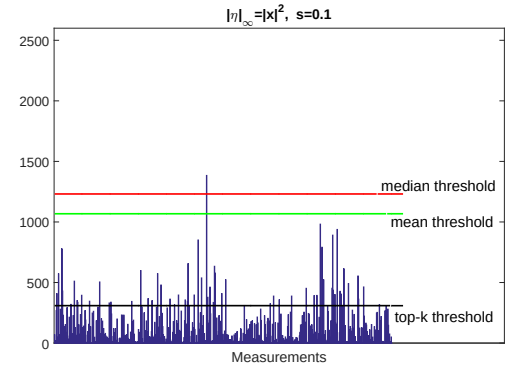
Key approach: "median-truncation": we will rule out measurements *adaptively* each iteration based on how large the sample gradient/value deviates from the median.

Median is more stable than mean and top-k truncation (which truncates a fixed amount of samples) for various levels of outliers.

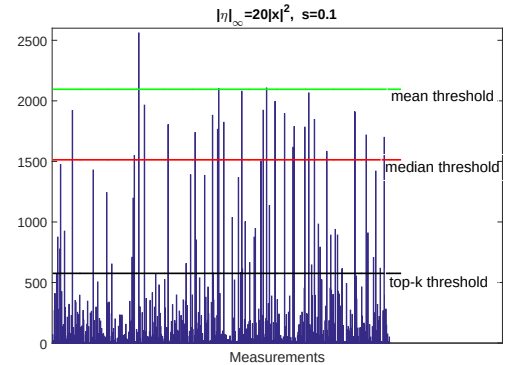
- well-known in robust statistics to be outlier-resilient;
- little appearance in high-dimensional estimation;



no outliers



small outlier magnitudes



large outlier magnitudes

Median-Truncated Gradient Descent

Median-truncated spectral initialization: Set $\boldsymbol{x}_0 := \lambda_0 \tilde{\boldsymbol{x}}_0$ where

- Direction estimation: $\tilde{\boldsymbol{x}}_0$ is the leading eigenvector of

$$\boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^* \mathbb{1}_{\{|y_i| \lesssim \text{median}(\{y_i\})\}}.$$

- Norm estimation: $\lambda_0 = \sqrt{\text{median}(\{y_i\})/0.455}$

$$y_i = |\boldsymbol{a}_i^* \boldsymbol{x}|^2 \sim \chi_1^2 \quad \text{and} \quad \mathbb{E}[\text{median}(\chi_1^2)] = 0.455$$

Median-truncated gradient descent:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\mu}{m}\sum_{i \in \mathcal{T}_t} \nabla \ell(y_i; \boldsymbol{x}_t),$$

where the set $\mathcal{T}_t$ contains samples that not deviates too much from the sample median of residual:

$$\mathcal{T}_t = \left\{ i : \ r_i^{(t)} \lesssim \text{median}(\{r_i^{(t)}\}) \right\}$$

where $r_i^{(t)} = \ell(y_i; \boldsymbol{x}_t) = |y_i - |\boldsymbol{a}_i^* \boldsymbol{x}_t||$.

Performance guarantees

Theorem (Zhang, C. and Liang, 2016)

Assume $\|\boldsymbol{w}\|_\infty \le c_1\|\boldsymbol{x}\|^2$. *Assume* $\boldsymbol{a}_i$*'s are generated with i.i.d. Gaussian entries. If* $m \gtrsim n\log n$ *and* $s \lesssim s_0$, *then with high probability, median-RWF yields*

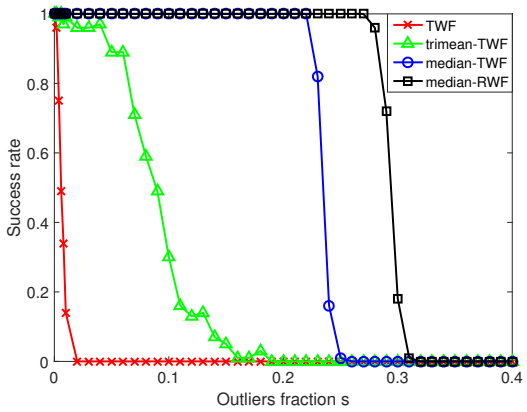
$$\operatorname{dist}(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \lesssim \frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{x}\|} + (1-\rho)^t\|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N}$$

simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n\backslash\{\boldsymbol{0}\}$ *for some* $0 < \rho < 1$.

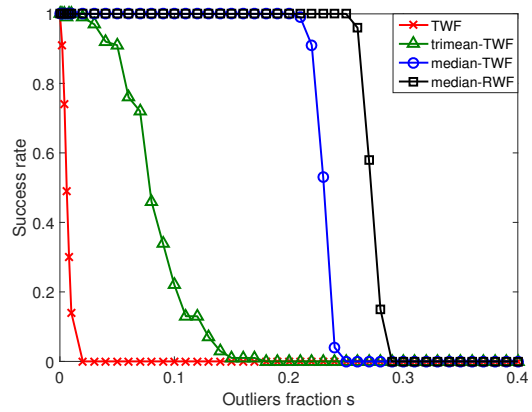
- **Exact recovery** when $\|\boldsymbol{w}\| = 0$ with slight more samples ($m = O(n\log n)$) but a constant fraction of outliers $s = O(1)$.
- **Stable recovery** with additional bounded noise;
- Resist outliers **obliviously**: no prior knowledge of outliers.
- Non-asymptotic robust recovery guarantee using median: much more involved due to the nonlinearity of median.

Numerical experiments

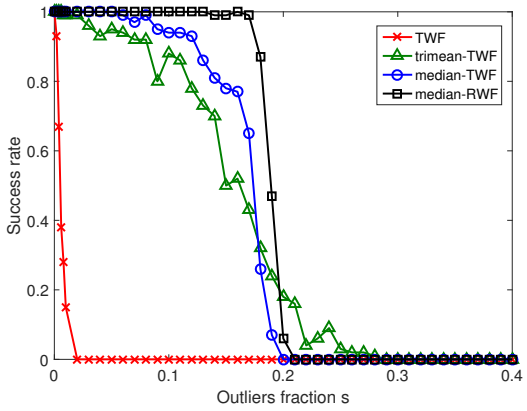
Recovery with only sparse outliers:



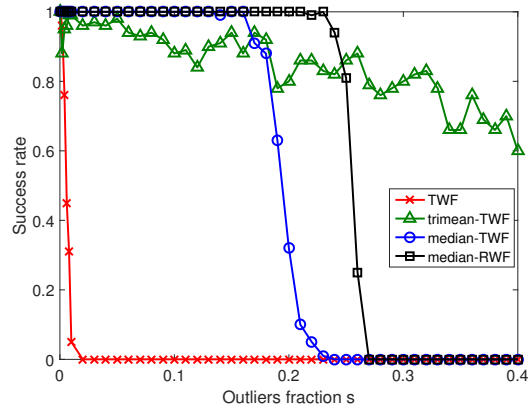
(a) $\|\boldsymbol{\eta}\|_\infty = 0.1\|\boldsymbol{x}\|^2$



(b) $\|\boldsymbol{\eta}\|_\infty = \|\boldsymbol{x}\|^2$



(c) $\|\boldsymbol{\eta}\|_\infty = 10\|\boldsymbol{x}\|^2$



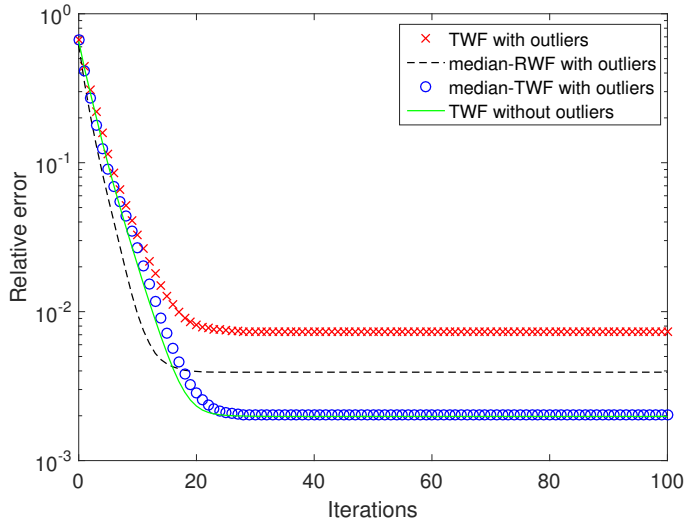
(d) $\|\boldsymbol{\eta}\|_\infty = 100\|\boldsymbol{x}\|^2$

Figure: Success rate of **exact recovery** with outliers for median-RWF, median-TWF, trimean-TWF, and TWF at different levels of outlier magnitudes.

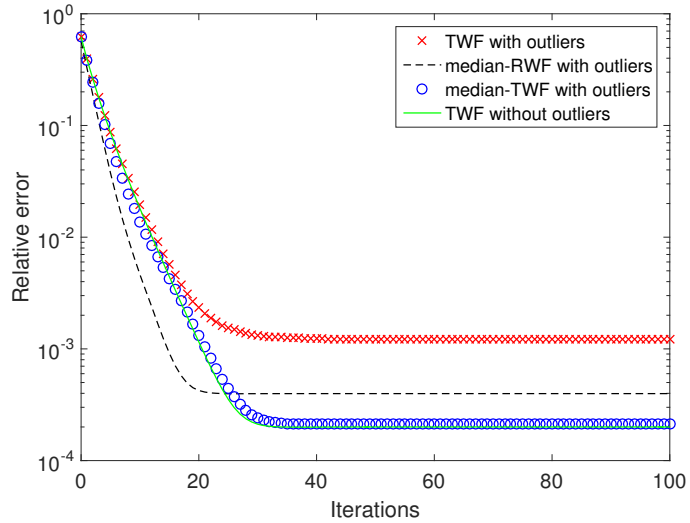
Numerical experiments

Recovery with both dense noise and sparse outliers:

- Median-TWF achieves slightly better accuracy than median-RWF.
- Moreover, median-TWF with outliers achieves almost the same accuracy of TWF without outliers.



(a) $w_{\max} = 0.01\|\boldsymbol{x}\|^2$



(b) $w_{\max} = 0.001\|\boldsymbol{x}\|^2$

Figure: The relative error with respect to the iteration count for median-TWF, median-RWF and TWF with both dense noise and sparse outliers, and TWF with only dense noise.

Conclusions

- Provable and fast-convergent algorithms for solving nonconvex signal estimation problems such as phase retrieval.
- Simple, iterative algorithms are demonstrated to perform remarkably well provided good initialization – the role of initialization is critical.
- An extension is to consider low-rank models, where

$$y_i = \|\boldsymbol{a}_i^* \boldsymbol{U}\|_2 = \boldsymbol{a}_i^*(\boldsymbol{X})\boldsymbol{a}_i, \quad \boldsymbol{U} \in \mathbb{R}/\mathbb{C}^{n\times r}$$

  for some small rank r, where $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^*$, which has a lot of applications in low-rank matrix recovery.
- Currently we're examining their performance on applications in THz imaging which appears to be very promising.

References

1. Kaczmarz Method for Solving Quadratic Equations, *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1183 - 1187, 2016.
2. Provable Non-convex Phase Retrieval with Outliers: Median Truncated Wirtinger Flow, *International Conference on Machine Learning (ICML)*, 2016.
3. Reshaped Wirtinger Flow and Incremental Algorithms for solving Quadratic Systems of Equations, Submitted to *Journal of Machine Learning Research*.

`http://www.ece.osu.edu/~chi/`