Statistical Issues in Searches: Photon Science Response

Rebecca Willett, Duke University

Photon science seen this week

Applications

- tomography
- ptychography
- nanochrystalography
- coherent diffraction imaging

Challenges

- high dimensional signals
- underdetermined problems
- limited number of photons
- complex forward models

Generic problem statement

- f^* is the signal of interest
- Before hitting our detector, the signal is transformed

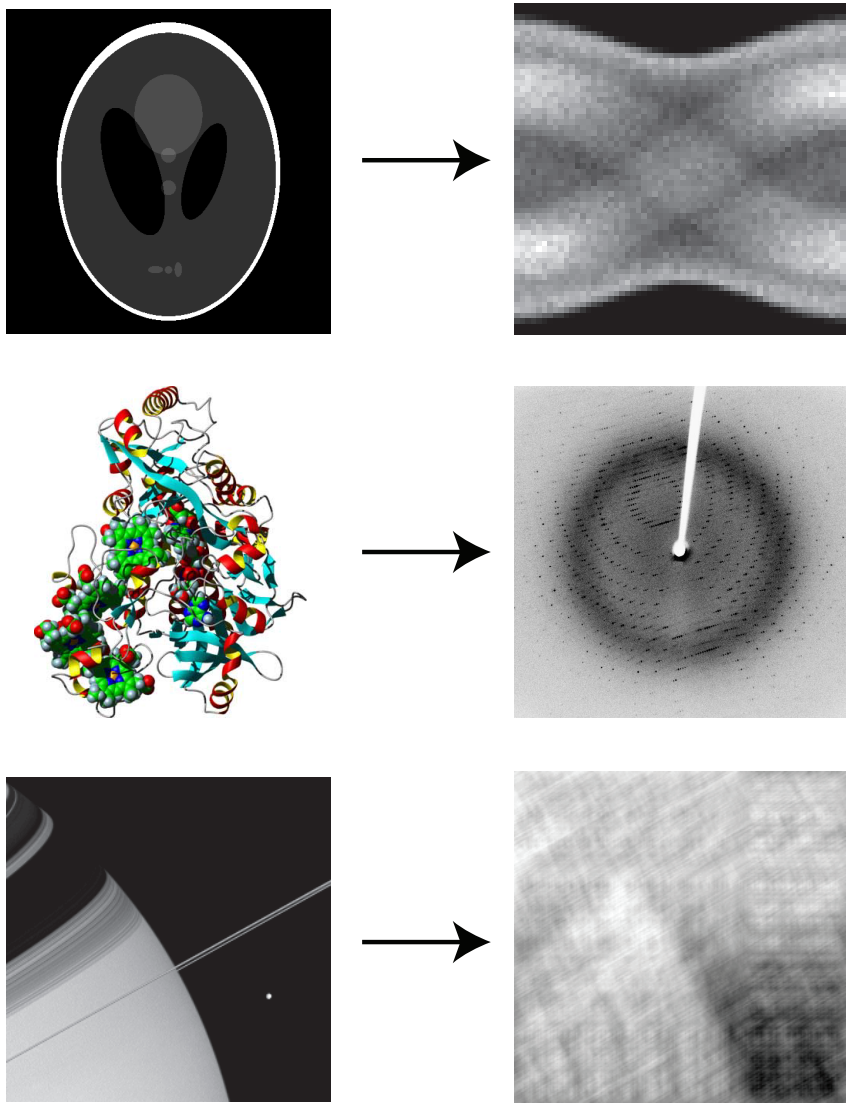
$$f^* \mapsto \mathcal{A}(f^*)$$

(tomographic projections, diffraction patterns, coded aperture)

- We observe

$$y \sim \text{Poisson}(\mathcal{A}(f^*))$$

- Goal is to infer f^* from y



Key questions

We infer f^* via

$$\widehat{f} = \underset{f \in \mathcal{F}}{\arg\min} - \log p(y|\mathcal{A}(f)) + \text{pen}(f)$$

where

- $-\log p(y|\mathcal{A}(f))$ is the negative Poisson log likelihood
- $\text{pen}(f)$ is a penalty / regularizer / negative log prior
- $\mathcal{F}$ is a class of candidate estimates

We'd like to understand the following:

- How does performance depend on $\mathcal{A}$?
- What should we use for $\text{pen}(f)$?
- How do we solve the optimization problem?

First a case study

Linear operator Convex objective	Nonlinear operator Convex objective
Linear operator Nonconvex objective	Nonlinear operator Nonconvex objective

For now, consider the case where

- $\mathcal{A}(f) = Af$ is a (possibly underdetermined) linear operator
- $\mathrm{pen}(f)$ is convex and measures sparsity

An oracle inequality[1]

Theorem

If $y \sim \text{Poisson}(f^*)$, where $\|f^*\|_1 = I$, and

$$\widehat{f} \triangleq \arg\min_{f \in \mathcal{F}} \{-\log p(y|f) + \ell(f)\}$$

where $\ell(f) \triangleq$ prefix codelength(f), then (omitting constants)

$$\underbrace{\mathbb{E}H^2\left(\frac{\widehat{f}}{I}, \frac{f^*}{I}\right)}_{\text{risk}} \preceq \min_{f \in \mathcal{F}} \left\{ \underbrace{\left\|\frac{f}{I} - \frac{f^*}{I}\right\|_2^2}_{\text{approximation error}} + \underbrace{\frac{\ell(f)}{I}}_{\text{estimation error}} \right\}$$

(Similar oracle inequalities exist without the codelength perspective.)

Codelengths correspond to generalized notions of sparsity.

[1] Li & Barron 2000, Massart 2003, Willett & Nowak 2005, Baraud & Birge 2006-2011, Bunea Tsybakov & Wegkamp 2007

A myriad of tradeoffs

We want to choose $\mathcal{F}$ and $\ell(\cdot)$ to make this bound as small as possible for large families of possible f^*

- Choosing a small model class $\mathcal{F}$ gives short codelengths but large approximation errors (bias)
- Choosing a rich model class makes the choice of coding scheme challenging
  - we want $\ell(\cdot)$ convex so we can perform search quickly
  - we want shorter codes for estimators which conform better to prior information

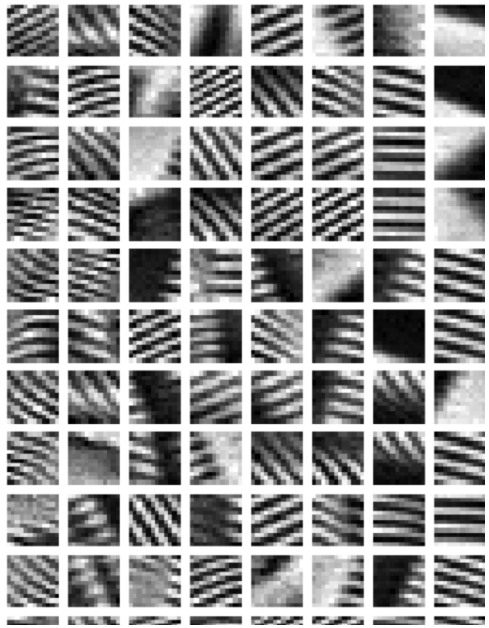
Codes and sparsity

Encode which pixels contain stars and the brightness of those stars

Codes and sparsity

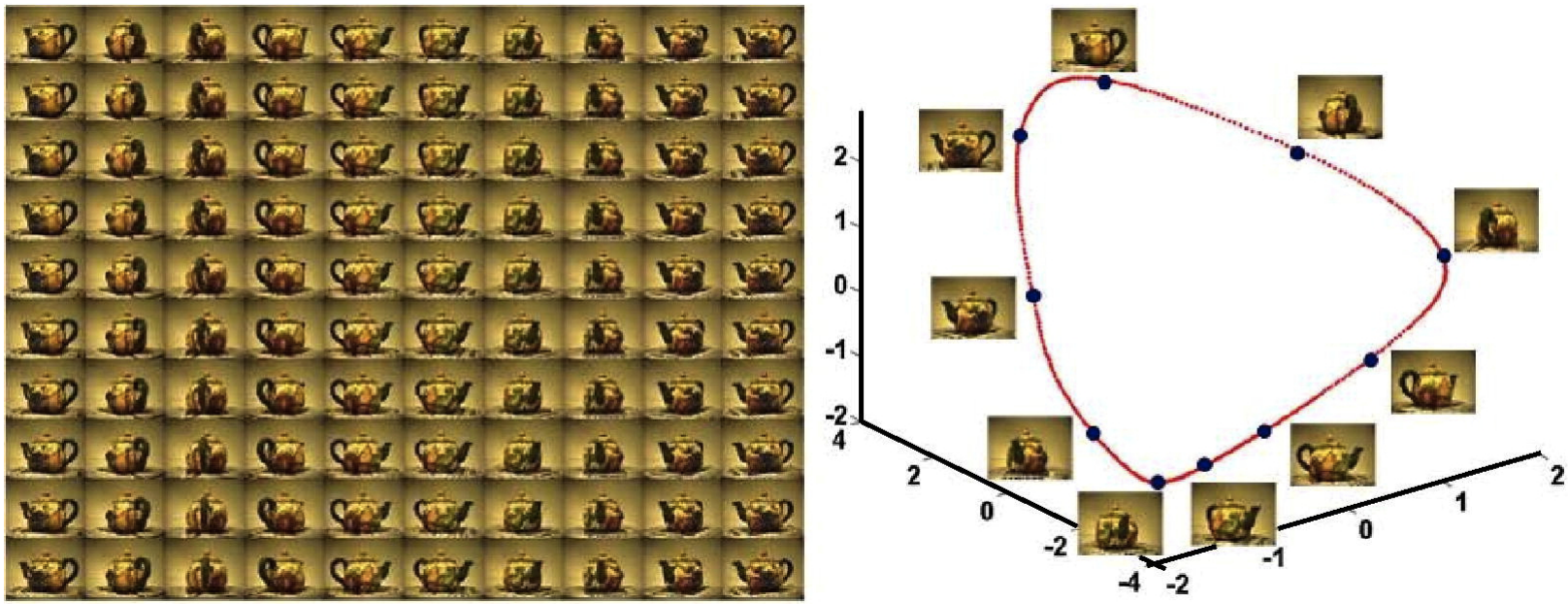
Encode which wavelet coefficients are non-zero and their magnitudes

Codes and sparsity



Encode image patches as linear combination of a small number of representative patches – encode which representatives used and their weights

Codes and sparsity



Encode location of object along a low-dimensional manifold[2]

[2] Fu *et al.* 2007

An oracle inequality

Theorem

If $y \sim \text{Poisson}(f^*)$, where $\|f^*\|_1 = I$, and

$$\widehat{f} \triangleq \arg\min_{f \in \mathcal{F}} \{-\log p(y|f) + \ell(f)\}$$

where $\ell(f) \triangleq$ prefix codelength(f), then (omitting constants)

$$\underbrace{\mathbb{E}H^2\left(\frac{\widehat{f}}{I}, \frac{f^*}{I}\right)}_{\text{risk}} \preceq \min_{f \in \mathcal{F}} \left\{ \underbrace{\left\| \frac{f}{I} - \frac{f^*}{I} \right\|_2^2}_{\text{approximation error}} + \underbrace{\frac{\ell(f)}{I}}_{\text{estimation error}} \right\}$$

(Similar oracle inequalities exist without the codelength perspective.)

This framework leads to guarantees on estimation performance for all f^* with sparse approximations.

Scheffé's identity

We focus on the squared Hellinger distance:

$$H^2(f_1, f_2) \triangleq \int \left(\sqrt{f_1(x)} - \sqrt{f_2(x)}\right)^2 dx$$

The squared Hellinger distance helps bound the L_1 error:

$$H^2(f_1, f_2) \leq \|f_1 - f_2\|_1 \leq 2H(f_1, f_2)$$

which is important due to

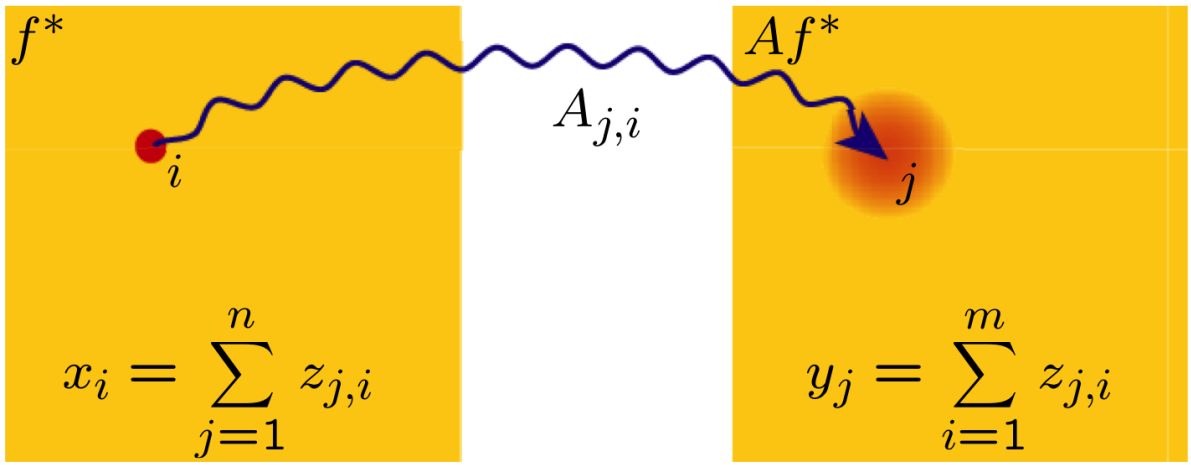
Sheffé's identity

$$\sup_{B \in \mathcal{B}} \left| \int_B f_1 - \int_B f_2 \right| = \frac{1}{2} \|f_1 - f_2\|_1$$

The L_1 distance gives a bound on the absolute different of probability measures of any Borel-measurable subset of the density's domain.

Poisson inverse problems

$A_{j,i}$ corresponds to the probability of a photon originating at location i hitting a detector at location j.



Hardware systems can only aggregate events / photons, not measure differences:

$$A_{j,i} \geq 0$$

The total number of observed events / photons cannot exceed the total number of events:

$$\sum_{j=1}^{n} (Af)_j \leq \sum_{i=1}^{m} f_i$$

Oracle inequality revisited

Theorem

Assume

$$c\|f_1 - f_2\|_2^2 \leq \|Af_1 - Af_2\|_2^2 \leq C\|f_1 - f_2\|_2^2$$

for all $f_1, f_2 \in \mathcal{F} \bigcup f^*$. If $y \sim \text{Poisson}(Af^*)$ and

$$\widehat{f} \triangleq \arg\min_{f \in \mathcal{F}} \{-\log p(y|Af) + \ell(f)\}$$

where $\ell(f) \triangleq$ prefix codelength(f), then (omitting constants)

$$\underbrace{\mathbb{E}\left\|\frac{f^*}{I} - \frac{\widehat{f}}{I}\right\|_2^2}_{\text{risk}} \preceq \min_{f \in \mathcal{F}} \left\{ \underbrace{\frac{C}{c}\left\|\frac{f}{I} - \frac{f^*}{I}\right\|_2^2}_{\text{approximation error}} + \underbrace{\frac{\ell(f)}{cI}}_{\text{estimation error}} \right\}$$

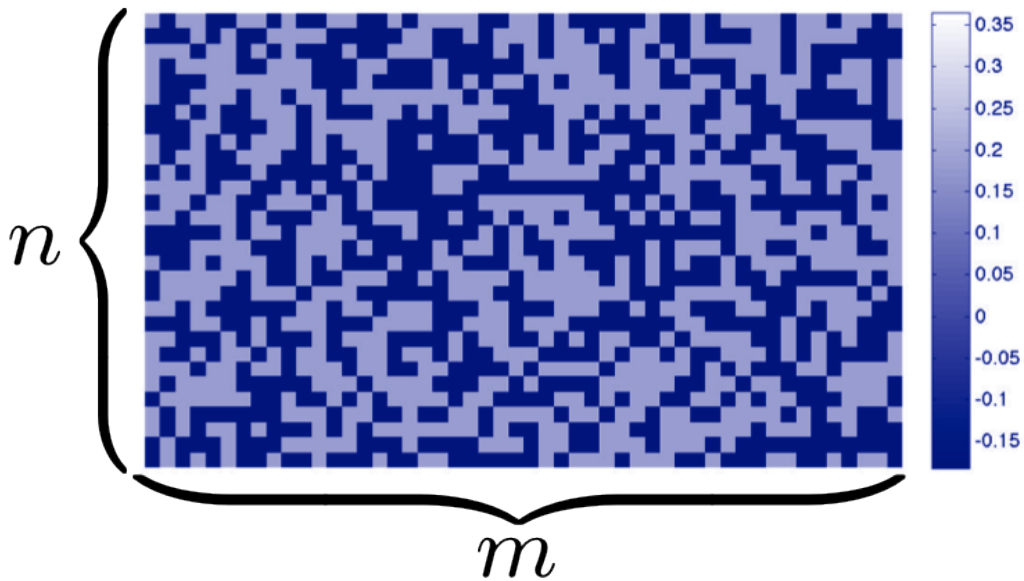
Similar results hold for ℓ_1 norms.

Big questions

1. If we have complete control over A, can we design our measurement system to get accurate high-resolution estimates from relatively few photodetectors? (compressed sensing)
2. When we have limited control over A, what does the theory tell us about various design tradeoffs?

Random compressed sensing matrices

Let $\widetilde{A}$ be an $n \times m$ Bernoulli ensemble matrix, so each $\widetilde{A}_{i,j}$ is $\frac{-1}{\sqrt{n}}$ or $\frac{1}{\sqrt{n}}$ with equal probability. $\widetilde{A}$ then satisfies the "restricted isometry property" (RIP, so $C \approx c \approx 1$) with high probability for n sufficiently large.



We would *like* to observe $\widetilde{A}f^*$.

Incorporating physical constraints

Elements of A must be nonnegative $\Longrightarrow$ we shift $\widetilde{A}$ so all elements are nonnegative

Total measured intensity $\|Af^*\|_1$ must not exceed the total incident intensity $\|f^*\|_1 \Longrightarrow$ we must scale $\widetilde{A}$ to ensure flux preservation.

Shifted and scaled measurements

- "Original" sensing matrix

$$\widetilde{A} \in \left\{ \frac{-1}{\sqrt{n}}, \frac{1}{\sqrt{n}} \right\}^{n \times m}$$

- Our shifted and scaled version is

$$A = \frac{1}{2\sqrt{n}} \left(\widetilde{A} + \frac{1}{\sqrt{n}} \right)$$

- So we measure

$$Af^* = \frac{\widetilde{A} f^*}{2\sqrt{n}} + \frac{I}{2n},$$

where $\widetilde{A} f^*$ is the ideal signal and $I \triangleq \|f^*\|_1$ is the total signal intensity

Compressible signals

Assume f^* is compressible in some basis Φ; that is,

$$f^* = \Phi\theta^* \qquad \text{and} \qquad \frac{\|\theta^* - \theta^{(k)}\|_2}{I} = O(k^{-\alpha})$$

where $\theta^{(k)}$ is the best k-term approximation to θ^*.

Theorem

There exist a finite set of candidate estimators $\mathcal{F}_I$ and $c \in (0, I/n)$ with the property

$$Af \succeq c \quad \forall f \in \mathcal{F}_I$$

and a prefix code

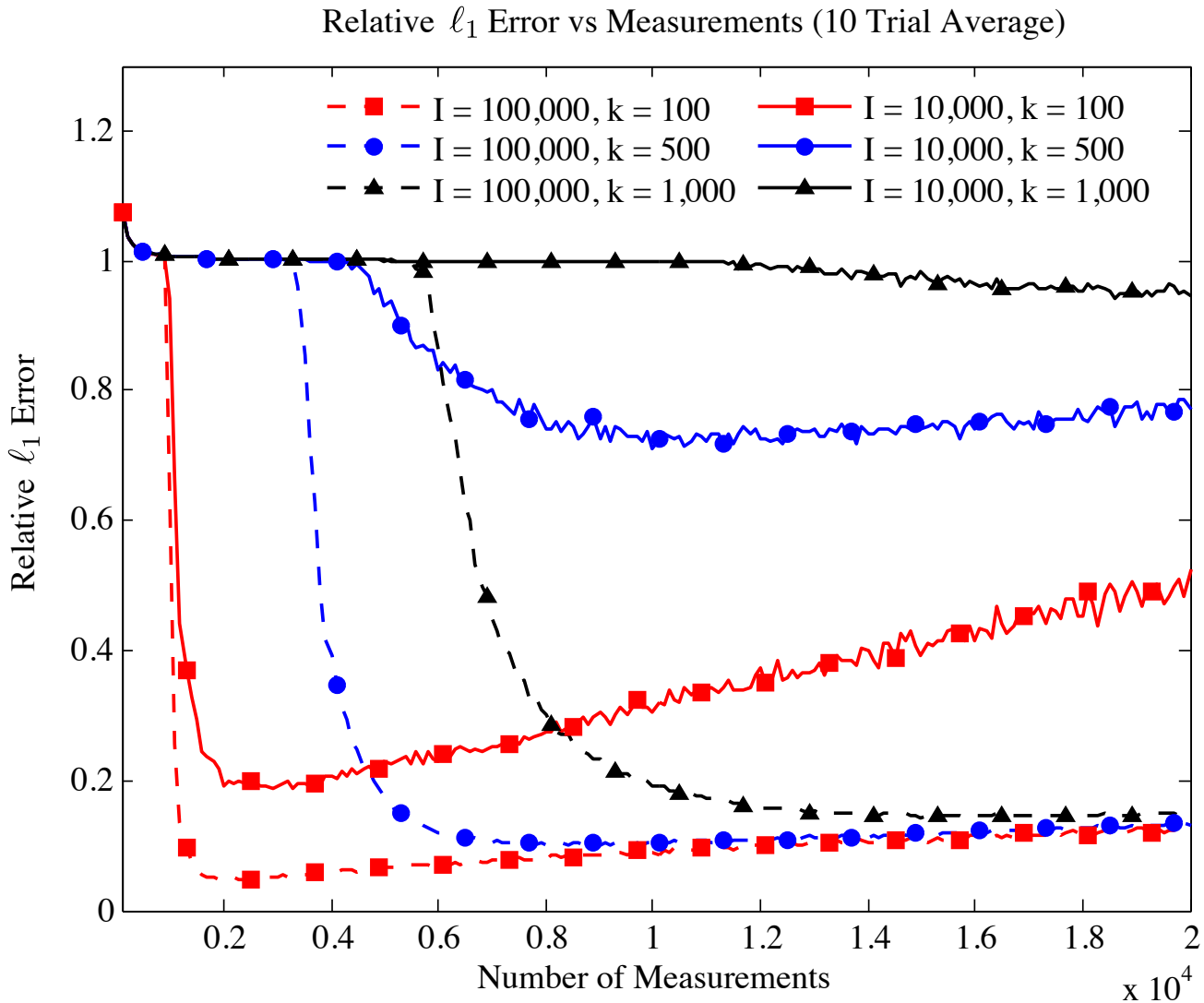
$$\ell(f) \propto \|\Phi^T f\|_0 \log(m)$$

such that, for n sufficiently large

$$\mathbb{E}\left\|\frac{\widehat{f}}{I} - \frac{f^*}{I}\right\|_2^2 = O\left[n\left(\frac{\log(m)}{I}\right)^{\frac{2\alpha}{2\alpha+1}} + \frac{\log(m/n)}{n}\right].$$

(Recall m is length of f, n is number of detectors, and I is the total intensity.)

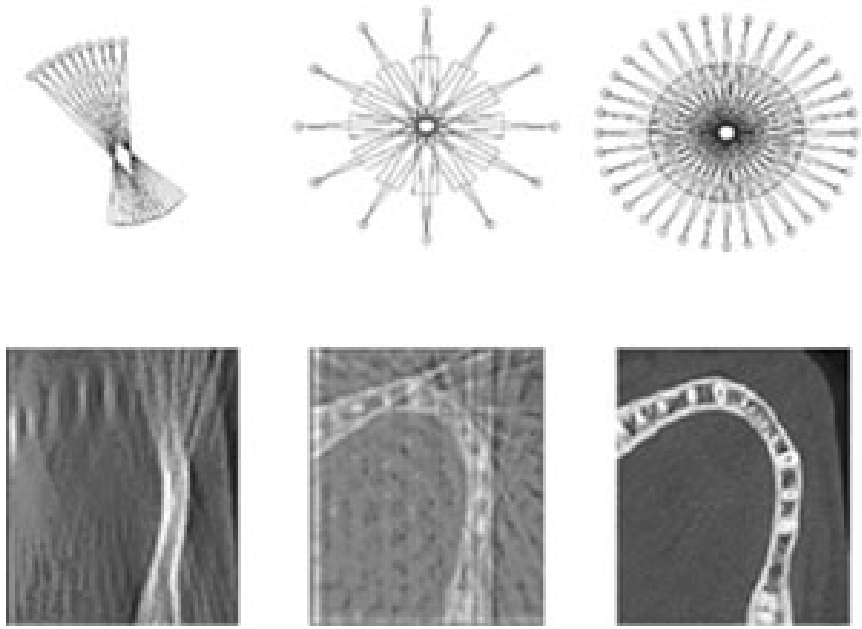
Simulation results, $m = 10^5$



Exercising limited control over A

These bounds can also provide insight when we have a fixed signal intensity or photon budget, and want to choose the measurement system A to optimize the tradeoff between **measurement diversity** and **photon scarcity**.

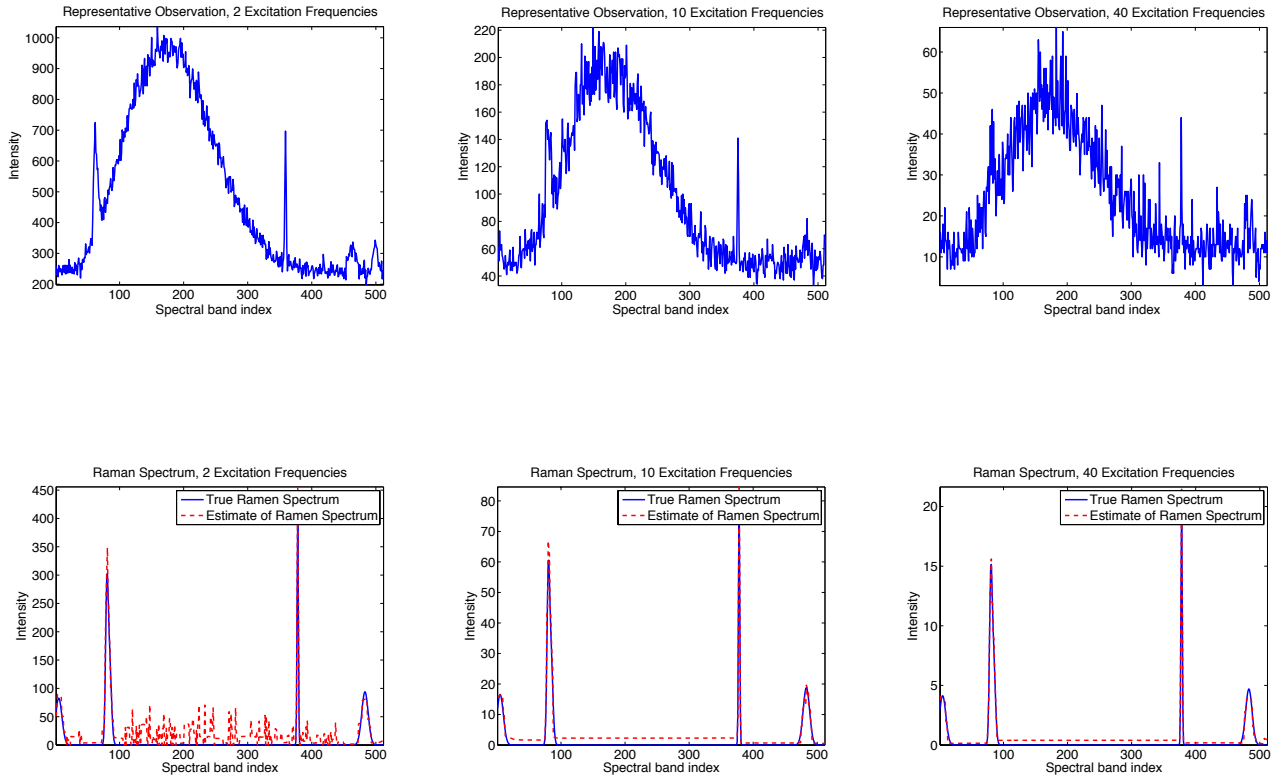
Measurement diversity and photon scarcity



In limited- or sparse-angle tomography, how many and which different angles should be used?[3]

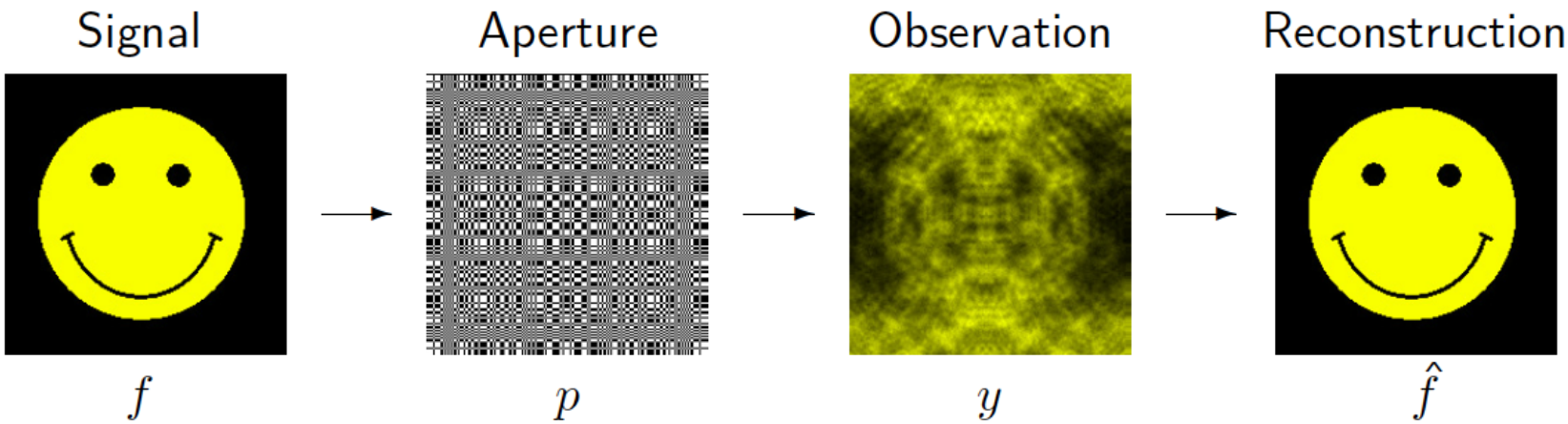
[3] Cederlund *et al.* 2009

Measurement diversity and photon scarcity



In Shifted Excitation Raman Spectroscopy, how many and which different excitation frequencies should be used to separate Raman spectrum from fluorescent background?

Measurement diversity and photon scarcity



In coded aperture or structured illumination imaging, which mask patterns best facilitate high-resolution image reconstruction?

Linear operator Convex objective	Nonlinear operator Convex objective
Linear operator Nonconvex objective	Nonlinear operator Nonconvex objective

We know a lot about linear systems and convex objectives. A good strategy for the other regimes is to find a problem relaxation such that

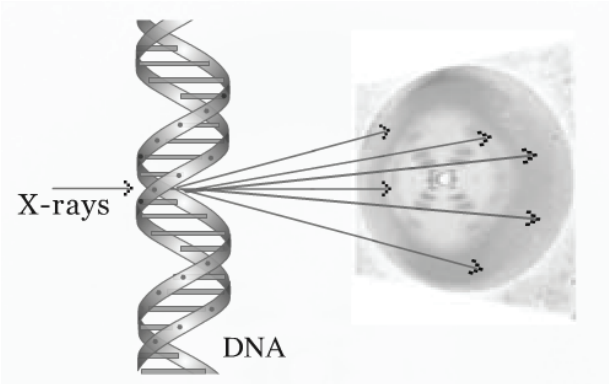
(a) the relaxed problem is linear and convex and

(b) the solution to the relaxed problem closely corresponds to the solution to the original problem

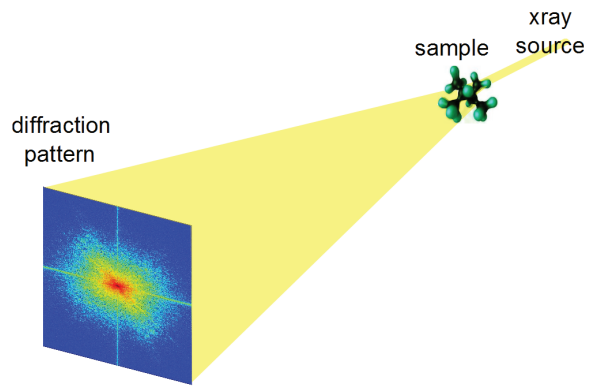
The following slides are derived from slides made by Thomas Strohmer, UC Davis

Phase retrieval

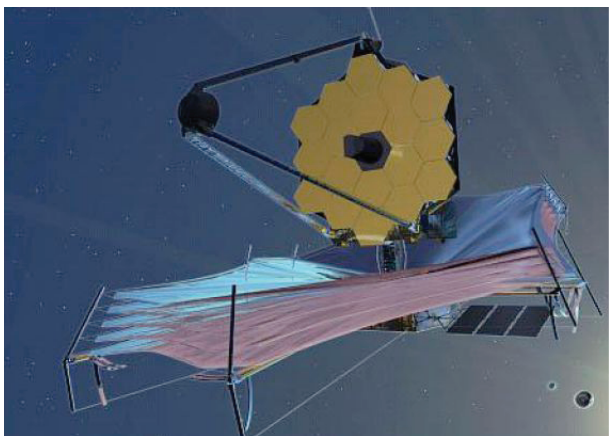
- X-ray diffraction of DNA (Photo 51 by Rosalind Franklin)



- X-ray crystallography



- Optical alignment (*e.g.*, James Web Space Telescope)



At the core of phase retrieval lies the problem:

We want to recover a function $f(t)$ from intensity measurements of its Fourier transform, $|\hat{f}(\omega)|^2$.

- Without further information about f, the phase retrieval problem is ill-posed. We can either impose additional properties of f or take more measurements (or both)

At the core of phase retrieval lies the problem:

We want to recover a function $f(t)$ from intensity measurements of its Fourier transform, $|\hat{f}(\omega)|^2$.

- Without further information about f, the phase retrieval problem is ill-posed. We can either impose additional properties of f or take more measurements (or both)
- We want an efficient phase retrieval algorithm based on a rigorous mathematical framework, for which we can guarantee exact recovery, and which is also provably stable in presence of noise.
- Want flexible framework that does not require any prior information about the function (signal, image,...), yet can incorporate additional information if available.

General phase retrieval problem

Suppose we have $f^* \in \mathbb{C}^m$ or $\mathbb{C}^{m_1 \times m_2}$ about which we have quadratic measurements of the form

$$y = \mathcal{A}(f^*) = \{|\langle a_k, f^* \rangle|^2 : k = 1, 2, \ldots, m\}.$$

Phase retrieval:

$$\begin{array}{ll} \text{find} & f \\ \text{obeying} & \mathcal{A}(f) = \mathcal{A}(f^*) := y. \end{array}$$

Goals:

- Find measurement vectors $\{a_k\}_{k \in \mathcal{I}}$ such that f^* is uniquely determined by $\{|\langle a_k, f^* \rangle|\}_{k \in \mathcal{I}}$.
- Find an algorithm that reconstructs f^* from $\{|\langle a_k, f^* \rangle|\}_{k \in \mathcal{I}}$.

Standard approach: Oversampling

- Sufficient oversampling of the diffraction pattern in the Fourier domain gives uniqueness for generic two- or higher-dim. signals in combination with support constraints
- Alternating projections [Gerchberg-Saxton, Fienup]
  Alternatingly enforce object constraint in spatial domain and measurement constraint in Fourier domain
- Seems to work in certain cases but has many limitations and problems.
- No proof of convergence to actual solution!

Lifting

Following [Balan, Bodman, Casazza, Edidin, 2007], we will interpret quadratic measurements of f as linear measurements of the rank-one matrix $F := ff^H$:

$$|\langle a_k, f\rangle|^2 = \mathrm{Tr}(f^H a_k a_k^H f) = \mathrm{Tr}(A_k F)$$

where A_k is the rank-one matrix $a_k a_k^H$. Define linear operator A that maps pos.sem.def. matrix F into $\{\mathrm{Tr}(A_k F)\}_{k=1}^n$.

Lifting

Following [Balan, Bodman, Casazza, Edidin, 2007], we will interpret quadratic measurements of f as linear measurements of the rank-one matrix $F := ff^H$:

$$|\langle a_k, f\rangle|^2 = \mathrm{Tr}(f^H a_k a_k^H f) = \mathrm{Tr}(A_k F)$$

where A_k is the rank-one matrix $a_k a_k^H$. Define linear operator A that maps pos.sem.def. matrix F into $\{\mathrm{Tr}(A_k F)\}_{k=1}^n$.
Now, the phase retrieval problem is equivalent to

$$\begin{array}{ll} \text{find} & F \\ \text{subject to} & A(F) = y \\ & F \succeq 0 \\ & \mathrm{rank}(F) = 1 \end{array} \qquad \text{(RANKMIN)}$$

Having found F, we factorize F as ff^H to obtain the phase retrieval solution (up to global phase factor).

Phase retrieval as convex problem?

We need to solve:

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(F) \\ \text{subject to} & A(F) = y \\ & F \succeq 0. \end{array} \qquad \text{(RANKMIN)}$$

Note that $A(F)$ is highly underdetermined, thus cannot just invert A to get F.
Rank minimization problems are typically NP-hard.

Phase retrieval as convex problem?

We need to solve:

$$\begin{array}{ll} \text{minimize} & \text{rank}(F) \\ \text{subject to} & A(F) = y \\ & F \succeq 0. \end{array} \qquad \text{(RANKMIN)}$$

Note that $A(F)$ is highly underdetermined, thus cannot just invert A to get F.
Rank minimization problems are typically NP-hard.
Use trace norm as convex surrogate for the rank functional [Beck '98, Mesbahi '97], giving the semidefinite program:

$$\begin{array}{ll} \text{minimize} & \text{trace}(F) \\ \text{subject to} & A(F) = y \\ & F \succeq 0. \end{array} \qquad \text{(TRACEMIN)}$$

A new methodology for phase retrieval

Lift up the problem of recovering a vector from quadratic constraints into that of recovering a rank-one matrix from affine constraints, and relax the combinatorial problem into a convenient convex program.

A new methodology for phase retrieval

Lift up the problem of recovering a vector from quadratic constraints into that of recovering a rank-one matrix from affine constraints, and relax the combinatorial problem into a convenient convex program.

PhaseLift

A new methodology for phase retrieval

Lift up the problem of recovering a vector from quadratic constraints into that of recovering a rank-one matrix from affine constraints, and relax the combinatorial problem into a convenient convex program.

PhaseLift

But when (if ever) is the trace minimization problem equivalent to the rank minimization problem?

When is phase retrieval a convex problem?

Theorem: [Candès-Strohmer-Voroninski '11]

Let f^* in $\mathbb{R}^m$ or $\mathbb{C}^m$ and suppose we choose the measurement vectors $\{a_k\}_{k=1}^n$ independently and uniformly at random on the unit sphere of $\mathbb{C}^m$ or $\mathbb{R}^m$. If $n \geq c\, m \log m$, where c is a constant, then PhaseLift recovers f^* exactly from $\{|\langle a_k, f^* \rangle|^2\}_{k=1}^n$ with probability at least $1 - 3e^{-\gamma \frac{n}{m}}$, where γ is an absolute constant.

Note that the "oversampling factor" $\log m$ is rather minimal!

This is the first result of its kind: phase retrieval can provably be accomplished via convex optimization.

Stability in presence of noise

Assume we observe

$$y_i = |\langle f, a_i \rangle|^2 + \nu_i,$$

where ν_i is a noise term with $\|\boldsymbol{\nu}\|_2 \leq \epsilon$.
Consider the solution to

$$\begin{array}{ll} \text{minimize} & \operatorname{trace}(F) \\ \text{subject to} & \|A(F) - y\|_2 \leq \epsilon \\ & F \succeq 0. \end{array}$$

Theorem: [Candès-Strohmer-Voroninski '11]

Under the same assumptions as in the other theorem, the solution to the noisy, the solution $\hat{f}$ computed via PhaseLift obeys

$$\|\hat{f} - f^*\|_2 \leq C_0 \, min \, (\|f^*\|_2, \epsilon / \|f^*\|_2)$$

Noise-aware framework

Suppose we observe

$$y \sim \mathrm{Poisson}(\mathcal{A}(f^*))$$

Our convex formulation suggests to solve

$$\begin{array}{ll} \text{minimize} & -\log p(y|A(F)) + \lambda \operatorname{trace}(F) \\ \text{subject to} & F \succeq 0. \end{array}$$

We can easily include additional constraints frequently used in phase retrieval, such as support constraints or positivity.

Limitations of our theory

Theorems are not yet completely practical, since most phase retrieval problems involve diffraction, *i.e.*, Fourier transforms, and not unstructured random measurements.

Conclusions

- Oracle inequalities provide valuable insight into tradeoffs related to
  - sensing system design and measurement diversity,
  - photon limitations,
  - generalized sparsity, and
  - optimization methods
- Nonlinear, nonconvex problems notoriously difficult to solve (hard to find global optimum)
- Convex relaxations may help alleviate this challenge