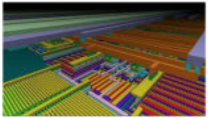
Nonconvex Optimization for High-Dimensional Learning: From Phase Retrieval to Submodular Maximization

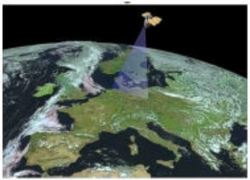
Mahdi Soltanolkotabi
Department of Electrical Engineering

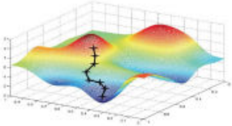
USC University of Southern California

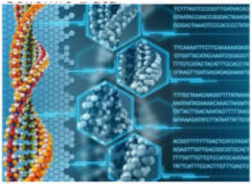
Interdisciplinary Seminar Series
North Carolina State

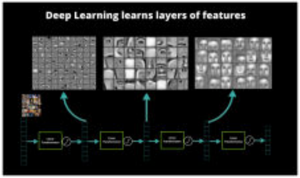
Nonconvexity is everywhere

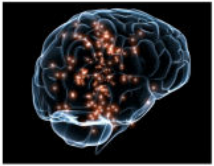


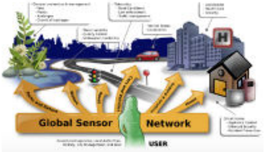












The power of convex programing

Exciting research over the last decade demonstrating the effectiveness of convex programming/greedy algorithms.

The power of convex programing

Exciting research over the last decade demonstrating the effectiveness of convex programming/greedy algorithms.

Idealogy

"when life gives you lemons, convexify"

The power of convex programing

Exciting research over the last decade demonstrating the effectiveness of convex programming/greedy algorithms.

Idealogy

"when life gives you lemons, convexify"

- Sparse use ℓ_1 norm, Low-rank use nuclear norm, etc.

convex relaxations are not perfect

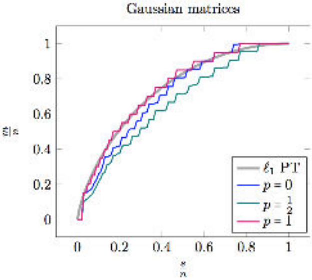
convex relaxations are not perfect

- Computation and memory: convex programs maybe inefficient

convex relaxations are not perfect

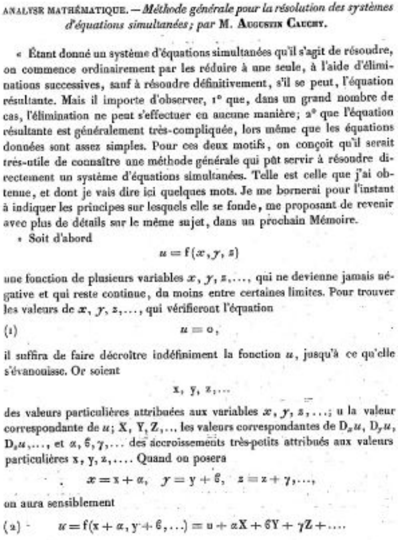
- Computation and memory: convex programs maybe inefficient

- Sometimes convex programs are inefficient in capturing the “structure” (usually require more samples)



Local search heuristics are often surprisingly effective...

Local search heuristics are often surprisingly effective...



ANALYSE MATHÉMATIQUE. — *Méthode générale pour la résolution des systèmes d'équations simultanées; par M.* AUGUSTIN CAUCHY.

« Étant donné un système d'équations simultanées qu'il s'agit de résoudre, on commence ordinairement par les réduire à une seule, à l'aide d'éliminations successives, sauf à résoudre définitivement, s'il se peut, l'équation résultante. Mais il importe d'observer, 1° que, dans un grand nombre de cas, l'élimination ne peut s'effectuer en aucune manière; 2° que l'équation résultante est généralement très-compliquée, lors même que les équations données sont assez simples. Pour ces deux motifs, on conçoit qu'il serait très-utile de connaître une méthode générale qui pût servir à résoudre directement un système d'équations simultanées. Telle est celle que j'ai obtenue, et dont je vais dire ici quelques mots. Je me bornerai pour l'instant à indiquer les principes sur lesquels elle se fonde, me proposant de revenir avec plus de détails sur le même sujet, dans un prochain Mémoire.

» Soit d'abord

$$u = f(x, y, z)$$

une fonction de plusieurs variables $x, y, z, \ldots$, qui ne devienne jamais négative et qui reste continue, du moins entre certaines limites. Pour trouver les valeurs de $x, y, z, \ldots$, qui vérifieront l'équation

$$u = 0, \tag{1}$$

il suffira de faire décroître indéfiniment la fonction u, jusqu'à ce qu'elle s'évanouisse. Or soient

$$\mathrm{x}, \mathrm{y}, \mathrm{z}, \ldots$$

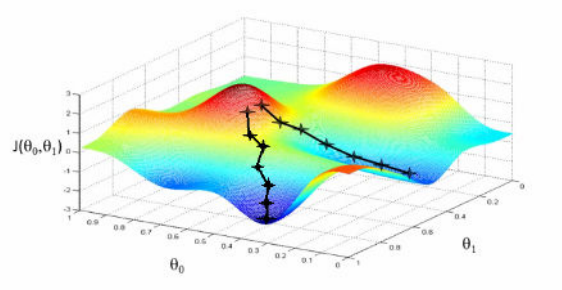
des valeurs particulières attribuées aux variables $x, y, z, \ldots$; u la valeur correspondante de u; $X, Y, Z, \ldots$ les valeurs correspondantes de $D_x u$, $D_y u$, $D_z u, \ldots$, et $\alpha, \beta, \gamma, \ldots$ des accroissements très-petits attribués aux valeurs particulières $\mathrm{x}, \mathrm{y}, \mathrm{z}, \ldots$. Quand on posera

$$x = \mathrm{x} + \alpha, \quad y = \mathrm{y} + \beta, \quad z = \mathrm{z} + \gamma, \ldots,$$

on aura sensiblement

$$u = f(\mathrm{x} + \alpha, \mathrm{y} + \beta, \ldots) = \mathrm{u} + \alpha X + \beta Y + \gamma Z + \ldots. \tag{2}$$

When should we just follow the gradient?



Two stories with a common theme

- *Story I: Structured Signal Recovery from Quadratic Measurements*

- *Story II: Submodular Maximization*

Structured Signal Recovery from Quadratic Measurements

Specific example:

Specific example:

Sparse recovery from quadratic measurements

Sparse Signal Recovery from Quadratic Measurements

Quadratic measurements from an s-sparse signal

$$y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$$

Sparse Signal Recovery from Quadratic Measurements

Quadratic measurements from an s-sparse signal

$$y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$$

Find an s-sparse signal from quadratic measurements

$$\boldsymbol{y}_r = \boldsymbol{x}^* \boldsymbol{A}_r \boldsymbol{x} \quad \text{for} \quad r = 1, 2, \ldots, m.$$

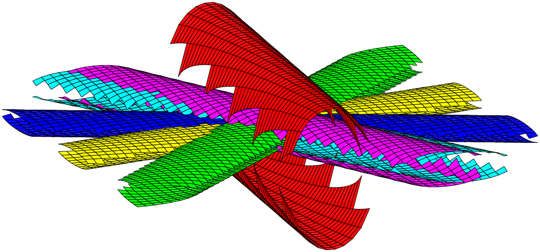
Sparse Signal Recovery from Quadratic Measurements

Quadratic measurements from an s-sparse signal

$$y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$$

Find an s-sparse signal from quadratic measurements

$$\boldsymbol{y}_r = \boldsymbol{x}^* \boldsymbol{A}_r \boldsymbol{x} \quad \text{for} \quad r = 1, 2, \ldots, m.$$

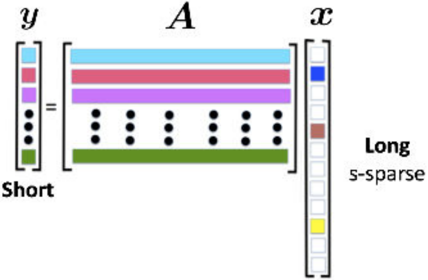


One of the universal forms of combinatorial problems, NP-hard in general.

Sparse Signal Recovery from Linear Measurements

Linear Measurements

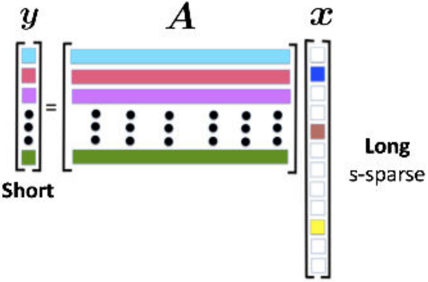
$$y_r = \langle \boldsymbol{a}_r, \boldsymbol{x} \rangle \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$



Sparse Signal Recovery from Linear Measurements

Linear Measurements

$$y_r = \langle \boldsymbol{a}_r, \boldsymbol{x} \rangle \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

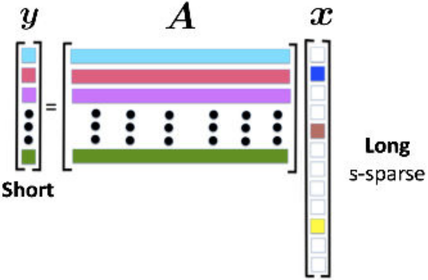


Sample complexity for uniqueness: $m \gtrsim s \log(n/s)$ generic measurements

Sparse Signal Recovery from Linear Measurements

Linear Measurements

$$y_r = \langle \boldsymbol{a}_r, \boldsymbol{x} \rangle \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

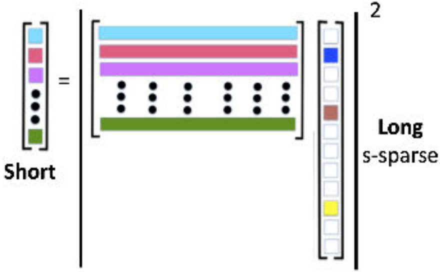


Sample complexity for uniqueness: $m \gtrsim s \log(n/s)$ generic measurements
Sample complexity of convex relaxation: $m \gtrsim s \log(n/s)$ generic measurements

Sparse Signal Recovery from Quadratic Measurements

Quadratic measurements

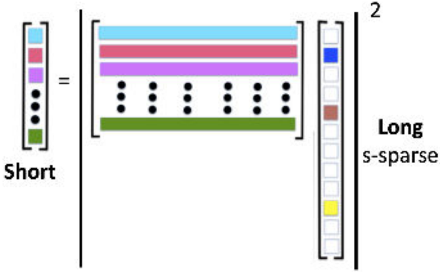
$$y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$$



Sparse Signal Recovery from Quadratic Measurements

Quadratic measurements

$$y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$$

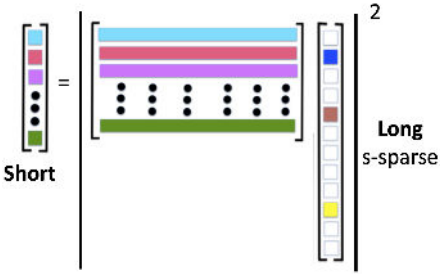


Sample complexity for uniqueness: $m \gtrsim s \log(n/s)$ generic measurements

Sparse Signal Recovery from Quadratic Measurements

Quadratic measurements

$$y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$$



Sample complexity for uniqueness: $m \gtrsim s \log(n/s)$ generic measurements
Sample complexity for exact recovery: ???????

First attempt: Convex Optimization

Semidefinite Relaxation with Sparsity

$$\min \quad \|\boldsymbol{z}\|_{\ell_0} \quad \text{subject to} \quad \boldsymbol{y}_r = |\boldsymbol{a}_r^* \boldsymbol{z}|^2 = [\mathcal{A}(\boldsymbol{z}\boldsymbol{z}^*)]_r.$$

Semidefinite Relaxation with Sparsity

$$\min \quad \|z\|_{\ell_0} \quad \text{subject to} \quad y_r = |a_r^* z|^2 = [\mathcal{A}(zz^*)]_r.$$

Lifting: $Z = zz^*$ Relax rank one constraint

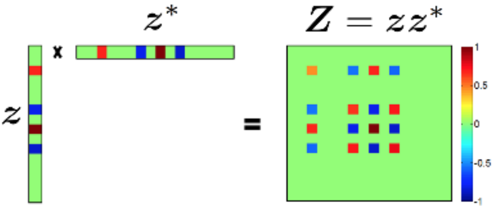
$$\min \|z\|_{\ell_0} \quad \text{subject to} \quad y = \mathcal{A}(Z) \quad \text{and} \quad Z \succeq \mathbf{0}.$$

Semidefinite Relaxation with Sparsity

$$\min \quad \|\boldsymbol{z}\|_{\ell_0} \quad \text{subject to} \quad \boldsymbol{y}_r = |\boldsymbol{a}_r^* \boldsymbol{z}|^2 = [\mathcal{A}(\boldsymbol{z}\boldsymbol{z}^*)]_r.$$

Lifting: $\boldsymbol{Z} = \boldsymbol{z}\boldsymbol{z}^*$ Relax rank one constraint

$$\min \|\boldsymbol{z}\|_{\ell_0} \quad \text{subject to} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{Z}) \quad \text{and} \quad \boldsymbol{Z} \succeq \mathbf{0}.$$

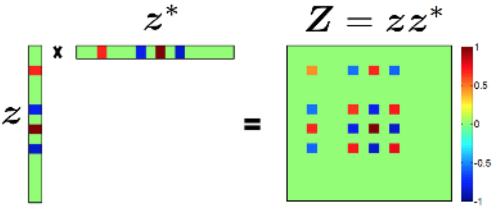


Semidefinite Relaxation with Sparsity

$$\min \quad \|\boldsymbol{z}\|_{\ell_0} \quad \text{subject to} \quad \boldsymbol{y}_r = |\boldsymbol{a}_r^* \boldsymbol{z}|^2 = [\mathcal{A}(\boldsymbol{z}\boldsymbol{z}^*)]_r.$$

Lifting: $\boldsymbol{Z} = \boldsymbol{z}\boldsymbol{z}^*$ Relax rank one constraint

$$\min \|\boldsymbol{z}\|_{\ell_0} \quad \text{subject to} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{Z}) \quad \text{and} \quad \boldsymbol{Z} \succeq \boldsymbol{0}.$$



SDP relaxation

$$\min \quad \|\boldsymbol{Z}\|_{\ell_1} \quad \textit{subject to} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{Z}) \quad \textit{and} \quad \boldsymbol{Z} \succeq \boldsymbol{0}.$$

For Phase Retrieval [Shechtman et. al. 2011, Li and Voroninski 2013].

Solving random quadratic equations

Given an s-sparse signal $\boldsymbol{x} \in \mathbb{C}^n$, measurements of the form

$$y_r = |\boldsymbol{a}_r^* \boldsymbol{x}|^2 \quad r = 1, 2, \ldots, m,$$

with $\boldsymbol{a}_r$ i.i.d. complex random vector with each entry $\sim c\mathcal{N}(0,1)$.

Solving random quadratic equations

Given an s-sparse signal $\boldsymbol{x} \in \mathbb{C}^n$, measurements of the form

$$y_r = |\boldsymbol{a}_r^* \boldsymbol{x}|^2 \quad r = 1, 2, \ldots, m,$$

with $\boldsymbol{a}_r$ i.i.d. complex random vector with each entry $\sim c\mathcal{N}(0,1)$.

Theorem (Li and Voroninski (2013))

Using $m \gtrsim s^2 \log n$ Gaussian measurements with high probability

$$\boldsymbol{x}\boldsymbol{x}^* = \arg\min \ \|\boldsymbol{Z}\|_{\ell_1} \quad \textit{subject to} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{Z}) \quad \textit{and} \quad \boldsymbol{Z} \succeq \mathbf{0}.$$

Solving random quadratic equations

Given an s-sparse signal $\boldsymbol{x} \in \mathbb{C}^n$, measurements of the form

$$y_r = |\boldsymbol{a}_r^* \boldsymbol{x}|^2 \quad r = 1, 2, \ldots, m,$$

with $\boldsymbol{a}_r$ i.i.d. complex random vector with each entry $\sim c\mathcal{N}(0,1)$.

Theorem (Li and Voroninski (2013))

Using $m \gtrsim s^2 \log n$ Gaussian measurements with high probability

$$\boldsymbol{x}\boldsymbol{x}^* = \arg\min \ \|\boldsymbol{Z}\|_{\ell_1} \quad \textit{subject to} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{Z}) \quad \textit{and} \quad \boldsymbol{Z} \succeq \boldsymbol{0}.$$

Maybe these results are not optimal...

Solving random quadratic equations

Given an s-sparse signal $\boldsymbol{x} \in \mathbb{C}^n$, measurements of the form

$$y_r = |\boldsymbol{a}_r^* \boldsymbol{x}|^2 \quad r = 1, 2, \ldots, m,$$

with $\boldsymbol{a}_r$ i.i.d. complex random vector with each entry $\sim c\mathcal{N}(0, 1)$.

Theorem (Li and Voroninski (2013))

Using $m \gtrsim s^2 \log n$ Gaussian measurements with high probability

$$\boldsymbol{x}\boldsymbol{x}^* = \arg\min \ \|\boldsymbol{Z}\|_{\ell_1} \quad \textit{subject to} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{Z}) \quad \textit{and} \quad \boldsymbol{Z} \succeq \boldsymbol{0}.$$

Maybe these results are not optimal...

Theorem (Li and Voroninski (2013), [Oymak, Jalali, Fazel, Hassibi, Eldar (2014))

With Gaussian measurements if

$$\boldsymbol{x}\boldsymbol{x}^* = \arg\min \ \|\boldsymbol{Z}\|_{\ell_1} \quad \textit{subject to} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{Z}) \quad \textit{and} \quad \boldsymbol{Z} \succeq \boldsymbol{0}.$$

holds with high probability.
Then

$$m \gtrsim \frac{s^2}{\log^2 n}.$$

Data Barriers...

$$m \gtrsim s \log(n/s) \quad \text{versus} \quad m \gtrsim \frac{s^2}{\log^2 n}$$

Uniqueness convex relaxation

Data Barriers...

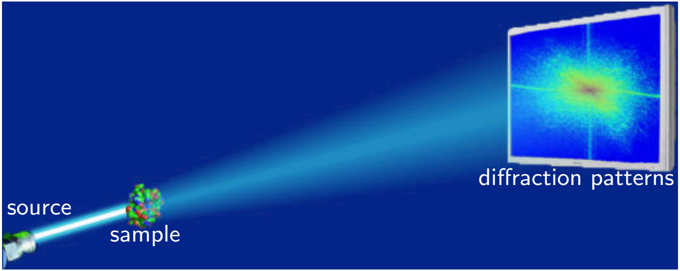
$$m \gtrsim s \log(n/s) \quad \text{versus} \quad m \gtrsim \frac{s^2}{\log^2 n}$$

Uniqueness convex relaxation

Engineering Motivation

Missing phase problem

- Detectors only record intensities of diffracted rays (magnitude measurements only!)

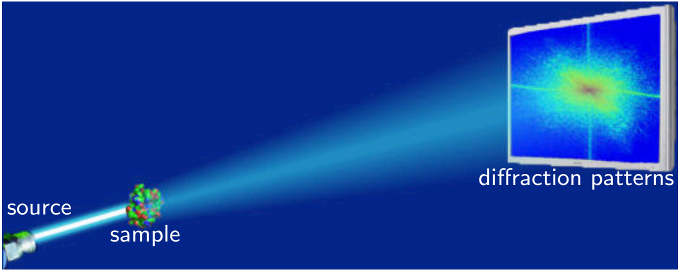


- Fraunhofer diffraction equation $\Rightarrow$ optical field at the detector $\approx$ Fourier transform

$$|\hat{x}(f_1, f_2)|^2 = \left| \int x(t_1, t_2) e^{-2\pi i (f_1 t_1 + f_2 t_2)} dt_1 dt_2 \right|^2$$

Missing phase problem

- Detectors only record intensities of diffracted rays (magnitude measurements only!)



- Fraunhofer diffraction equation $\Rightarrow$ optical field at the detector $\approx$ Fourier transform

$$|\hat{x}(f_1, f_2)|^2 = \left| \int x(t_1, t_2) e^{-2\pi i (f_1 t_1 + f_2 t_2)} dt_1 dt_2 \right|^2$$

Phase Retrieval Problem

How can we recover the phase (or equivalently signal $x(t_1, t_2)$) from $|\hat{x}(f_1, f_2)|$?

Phase retrieval (discrete 1D model)

- Phaseless measurements about $\boldsymbol{x} \in \mathbb{C}^n$

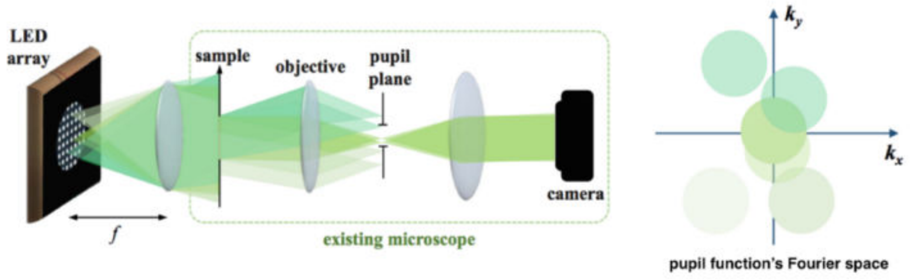
$$|\boldsymbol{f}_k^* \boldsymbol{x}|^2 = \boldsymbol{y}_k \quad k \in \{1, 2, \ldots, n\} = [n]$$

$\boldsymbol{f}_k^*$ is kth row of the DFT matrix.

- Phase retrieval is impossible, inherent ambiguity.

Resolving ambiguity?

Solution: Create diversity



(stolen from the Waller Lab)

$$y = |Ax|^2 \quad \text{where} \quad A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_L \end{bmatrix}$$

with $A_\ell \in \mathbb{C}^{n \times n}$

Data Barriers...

$$m \gtrsim s \log(n/s) \quad \text{versus} \quad m \gtrsim \frac{s^2}{\log^2 n}$$

Uniqueness　　　　　　convex relaxation

Second attempt: Nonconvex Optimization

Solving quadratic equation by non-convex optimization (no constraints)

Let $\boldsymbol{A} = \left[\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m\right]$

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \quad f(\boldsymbol{z}) := \frac{1}{2m} \sum_{r=1}^{m} \ell(\boldsymbol{y}_r, |\boldsymbol{a}_r^* \boldsymbol{z}|)$$

- Pro: operates over vectors much less intensive!
- Con: Non-convex!

Wirtinger Flow (WF)

Algorithm 1 Wirtinger Flow (WF)

Input: Measurements y_r for $r = 1, 2, \ldots, m$.

Initialization (WF-INIT):

Set $\tilde{\boldsymbol{z}}_0$ to be the eigenvector corresponding to the largest eigenvalue of

$$\boldsymbol{Y} = \frac{1}{m}\sum_{r=1}^{m} y_r \boldsymbol{a}_r \boldsymbol{a}_r^*.$$

Set $\boldsymbol{z}_0 = \left(\sqrt{\frac{1}{m}\sum_{r=1}^{m} y_r}\right)\tilde{\boldsymbol{z}}_0$.

Iterations:

for $\tau = 0$ **to** $t-1$ **do**

Set

$$\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_\tau - \frac{\mu_{\tau+1}}{\|\boldsymbol{z}_0\|_{\ell_2}^2}\left(\frac{1}{m}\sum_{r=1}^{m}\left(\left|\boldsymbol{a}_r^*\boldsymbol{z}\right|^2 - y_r\right)(\boldsymbol{a}_r\boldsymbol{a}_r^*)\boldsymbol{z}\right) := \boldsymbol{z}_\tau - \frac{\mu_{\tau+1}}{\|\boldsymbol{z}_0\|_{\ell_2}^2}\nabla f(\boldsymbol{z}_\tau).$$

end for

Output: $\hat{x} = \boldsymbol{z}_t$.

Exact Phase Retrieval by WF (Gaussian Model)

For a vector $\boldsymbol{z} \in \mathbb{C}^n$

$$\mathsf{dist}(\boldsymbol{z}, \boldsymbol{x}) = \min_{\phi \in [0, 2\pi]} \left\| \boldsymbol{z} - e^{i\phi} \boldsymbol{x} \right\|_{\ell_2}.$$

Theorem (Candes, Li, and Soltanolkotabi ('14), Soltanolkotabi ('14))

Assume $m \gtrsim n$. Using $0 \le \mu \le \mu_0/n$, with high probability

- *Initialization:*
$$\mathit{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \le \sqrt{\frac{5}{6}} \left\| \boldsymbol{x} \right\|_{\ell_2}.$$
- *After t iterations:*
$$\mathit{dist}(\boldsymbol{z}_t, \boldsymbol{x}) \le e^{-c\mu t} \cdot \mathit{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \le \sqrt{\frac{5}{6}} e^{-c\mu t} \left\| \boldsymbol{x} \right\|_{\ell_2}.$$

[Chen and Candes 2015], [Wang and Giannakis], [Zhang and Liang 2016] established $m \gtrsim n$ via variantes of Wirtinger Flow

Don't like initialization?

Don't like initialization?

Theorem (Soltanolkotabi 2017)

With $m \gtrsim n \log n$ Gaussian measurements all local optima are global optima and cubic regularization converges to a global optima in poly(n) iterations.

Earlier [Sun, Qu, Wright 2016]: All local optima are global optima with $m \gtrsim n \log^3 n$ and trust region methods converge to a global optima in poly(n) iterations.

Are local optima global optima?

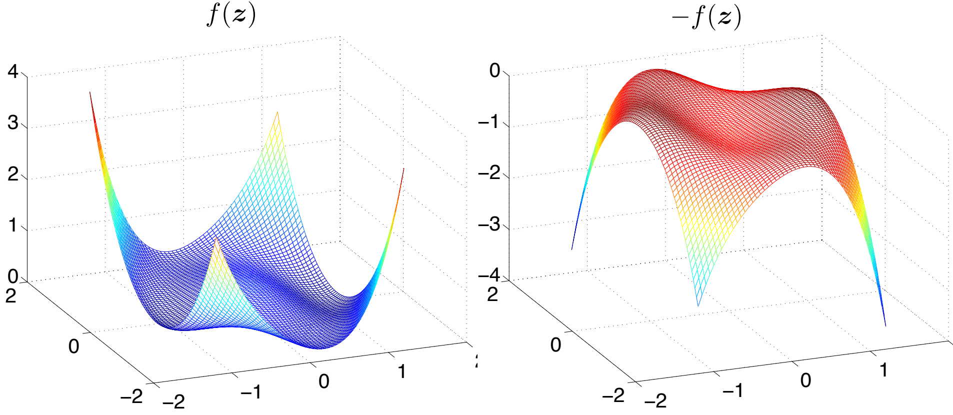
Are saddles the only problem with nonconvexity?

Are local optima global optima?

Are saddles the only problem with nonconvexity?

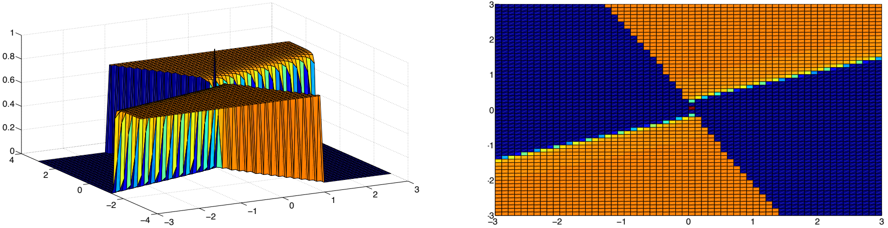
Example: $\boldsymbol{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Measurements $y_r = |\boldsymbol{a}_r^* \boldsymbol{x}|^2$, $r = 1, 2, \ldots, m$, with $m = 4$.

cost function: $f(\boldsymbol{z}) = \frac{1}{4m} \sum_{r=1}^{m} (y_r - |\boldsymbol{a}_r^* \boldsymbol{x}|^2)^2$



Which initial solutions work?

Run gradient descent $(\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_\tau - \mu \nabla f(\boldsymbol{z}_\tau))$ from different initial points.



Solving quadratic equations via Projected Wirtinger Flow (PWF)

$$\min_{\boldsymbol{z}\in\mathbb{C}^n} \quad f(\boldsymbol{z}) := \frac{1}{2m}\sum_{r=1}^{m}\left(y_r - |\boldsymbol{a}_r^*\boldsymbol{z}|^2\right)^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{z}) \le \mathcal{R}(\boldsymbol{x}).$$

Solving quadratic equations via Projected Wirtinger Flow (PWF)

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \quad f(\boldsymbol{z}) := \frac{1}{2m} \sum_{r=1}^{m} \left(y_r - \left| \boldsymbol{a}_r^* \boldsymbol{z} \right|^2 \right)^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}).$$

Follow the gradient:

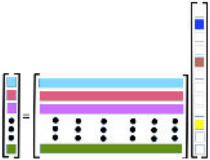
$$\boldsymbol{z}_{\tau+1} := \mathcal{P}_{\mathcal{K}} \left(\boldsymbol{z}_\tau - \mu_\tau \nabla f(\boldsymbol{z}_\tau) \right).$$

where

$$\mathcal{K} = \{ \boldsymbol{z} : \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}) \}$$

What is the sample complexity of PWF?

Simpler question: Linear inverse problems



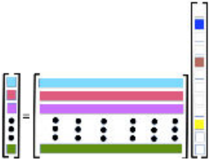
$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, and $\boldsymbol{x} \in \mathbb{R}^n$ with $m << n$.

$$\hat{\boldsymbol{x}} = \operatorname*{argmin}_{\boldsymbol{z}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}\|_{\ell_2}^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}).$$

When is $\hat{\boldsymbol{x}} = \boldsymbol{x}$? m?

What is the sample complexity of PWF?

Simpler question: Linear inverse problems



$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, and $\boldsymbol{x} \in \mathbb{R}^n$ with $m << n$.

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{z}}{\text{argmin}} \ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}\|_{\ell_2}^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}).$$

When is $\hat{\boldsymbol{x}} = \boldsymbol{x}$? m?

Theorem (Chandrasekaran, Recht, Parrilo, and Willskey 2012-Amelunxen, Lotz, McCoy, Tropp 2014)

For i.i.d. normal matrices as long as

$$m \approx m_0(\mathcal{R}, \boldsymbol{x}),$$

then with high probability $\hat{\boldsymbol{x}} = \boldsymbol{x}$

e.g. for an s-sparse signal $m \geq 2s \log(n/s)$

What is the sample complexity of PWF? (local)

Let $\boldsymbol{a}_r \in \mathbb{R}^n$ be i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2$ for $r = 1, 2, \ldots, m$.

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \quad f(\boldsymbol{z}) := \frac{1}{2m} \sum_{r=1}^{m} \left(y_r - |\boldsymbol{a}_r^* \boldsymbol{z}|^2 \right)^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}).$$

Follow the gradient: $\boldsymbol{z}_{\tau+1} := \mathcal{P}_{\mathcal{K}} \left(\boldsymbol{z}_\tau - \mu_\tau \nabla f(\boldsymbol{z}_\tau) \right)$ with $\mathcal{K} = \{\boldsymbol{z} : \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x})\}$.

Theorem (Soltanolkotabi 2017)

Assume $m \gtrsim m_0 \log n$. Using $0 \leq \mu \leq \mu_0 / n$, with high probability Starting from any initial point $\boldsymbol{z}_0$ obeying

$$\textit{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \leq \sqrt{\frac{5}{6}} \|\boldsymbol{x}\|_{\ell_2},$$

we have

$$\textit{dist}(\boldsymbol{z}_t, \boldsymbol{x}) \leq e^{-c\mu t} \cdot \textit{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \leq \sqrt{\frac{5}{6}} e^{-c\mu t} \|\boldsymbol{x}\|_{\ell_2}.$$

What is the sample complexity of PWF? (local)

Let $\boldsymbol{a}_r \in \mathbb{R}^n$ be i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2$ for $r = 1, 2, \ldots, m$.

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \quad f(\boldsymbol{z}) := \frac{1}{2m} \sum_{r=1}^{m} \left(y_r - |\boldsymbol{a}_r^* \boldsymbol{z}|^2 \right)^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}).$$

Follow the gradient: $\boldsymbol{z}_{\tau+1} := \mathcal{P}_{\mathcal{K}} \left(\boldsymbol{z}_\tau - \mu_\tau \nabla f(\boldsymbol{z}_\tau) \right)$ with $\mathcal{K} = \{ \boldsymbol{z} : \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}) \}$.

Theorem (Soltanolkotabi 2017)

Assume $m \gtrsim m_0 \log n$. Using $0 \leq \mu \leq \mu_0/n$, with high probability Starting from any initial point $\boldsymbol{z}_0$ obeying

$$\text{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \leq \sqrt{\frac{5}{6}} \, \|\boldsymbol{x}\|_{\ell_2},$$

we have

$$\text{dist}(\boldsymbol{z}_t, \boldsymbol{x}) \leq e^{-c\mu t} \cdot \text{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \leq \sqrt{\frac{5}{6}} e^{-c\mu t} \, \|\boldsymbol{x}\|_{\ell_2}.$$

- e.g. for sparsity $m \gtrsim 2s \log(n/s) \log n$
- previous known result for local neighborhood via Thresholded WF $m \gtrsim s^2 \log n$ [Cai, Li, Ma 2015]

What is the sample complexity of PWF? (global)

Let $\boldsymbol{a}_r \in \mathbb{R}^n$ be i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2$ for $r = 1, 2, \ldots, m$.

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \quad f(\boldsymbol{z}) := \frac{1}{2m} \sum_{r=1}^{m} \left(y_r - |\boldsymbol{a}_r^* \boldsymbol{z}|^2 \right)^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}).$$

Follow the gradient: $\boldsymbol{z}_{\tau+1} := \mathcal{P}_{\mathcal{K}} \left(\boldsymbol{z}_\tau - \mu_\tau \nabla f(\boldsymbol{z}_\tau) \right)$ with $\mathcal{K} = \{\boldsymbol{z} : \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x})\}$.

Theorem (Soltanolkotabi 2017)

With $m \gtrsim m_0 \log n$ Gaussian measurements all local optima are global optima and cubic regularization converges in poly(n) iterations.

What is the sample complexity of PWF? (global)

Let $\boldsymbol{a}_r \in \mathbb{R}^n$ be i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2$ for $r = 1, 2, \ldots, m$.

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \quad f(\boldsymbol{z}) := \frac{1}{2m} \sum_{r=1}^{m} \left(y_r - |\boldsymbol{a}_r^* \boldsymbol{z}|^2 \right)^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}).$$

Follow the gradient: $\boldsymbol{z}_{\tau+1} := \mathcal{P}_{\mathcal{K}} \left(\boldsymbol{z}_\tau - \mu_\tau \nabla f(\boldsymbol{z}_\tau) \right)$ with $\mathcal{K} = \{ \boldsymbol{z} : \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}) \}$.

Theorem (Soltanolkotabi 2017)

With $m \gtrsim m_0 \log n$ Gaussian measurements all local optima are global optima and cubic regularization converges in poly(n) iterations.

- e.g. for sparsity $m \gtrsim 2s \log(n/s) \log n$

Removing logs and other things...

Removing logs and other things...

Let $\boldsymbol{a}_r \in \mathbb{R}^n$ be i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2$ for $r = 1, 2, \ldots, m$.

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \quad f(\boldsymbol{z}) := \frac{1}{2m} \sum_{r=1}^{m} \left(\sqrt{y_r} - |\boldsymbol{a}_r^* \boldsymbol{z}| \right)^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}).$$

Follow the "gradient": $\boldsymbol{z}_{\tau+1} := \mathcal{P}_{\mathcal{K}} \left(\boldsymbol{z}_\tau - \mu_\tau \nabla f(\boldsymbol{z}_\tau) \right)$ with $\mathcal{K} = \{ \boldsymbol{z} : \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}) \}$.

Theorem (Soltanolkotabi 2017)

Assume $m \gtrsim m_0$. *Using* $0 \leq \mu \leq \mu_0$, *with high probability Starting from any initial point* $\boldsymbol{z}_0$ *obeying*

$$\textit{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \leq \sqrt{\frac{5}{6}} \, \|\boldsymbol{x}\|_{\ell_2},$$

we have

$$\textit{dist}(\boldsymbol{z}_t, \boldsymbol{x}) \leq e^{-c\mu t} \cdot \textit{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \leq \sqrt{\frac{5}{6}} e^{-c\mu t} \, \|\boldsymbol{x}\|_{\ell_2}.$$

Removing logs and other things...

Let $\boldsymbol{a}_r \in \mathbb{R}^n$ be i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2$ for $r = 1, 2, \ldots, m$.

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \quad f(\boldsymbol{z}) := \frac{1}{2m} \sum_{r=1}^{m} \left(\sqrt{y_r} - |\boldsymbol{a}_r^* \boldsymbol{z}| \right)^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}).$$

Follow the "gradient": $\boldsymbol{z}_{\tau+1} := \mathcal{P}_{\mathcal{K}} \left(\boldsymbol{z}_\tau - \mu_\tau \nabla f(\boldsymbol{z}_\tau) \right)$ with $\mathcal{K} = \{ \boldsymbol{z} : \mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x}) \}$.

Theorem (Soltanolkotabi 2017)

Assume $m \gtrsim m_0$. *Using* $0 \leq \mu \leq \mu_0$, *with high probability*
Starting from any initial point $\boldsymbol{z}_0$ *obeying*

$$\mathit{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \leq \sqrt{\frac{5}{6}} \, \|\boldsymbol{x}\|_{\ell_2},$$

we have

$$\mathit{dist}(\boldsymbol{z}_t, \boldsymbol{x}) \leq e^{-c\mu t} \cdot \mathit{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \leq \sqrt{\frac{5}{6}} e^{-c\mu t} \, \|\boldsymbol{x}\|_{\ell_2}.$$

- This result also holds for nonconvex regularizers!

Connecting sample complexity to mini-max denoising

Theorem (Soltanolkotabi 2016)

For any set $\mathcal{K}$, as long as

$$m \geq c \max_{\sigma} \frac{\mathbb{E}\left\|\mathcal{P}_{\mathcal{K}}(\boldsymbol{x}+\sigma\boldsymbol{z})-\boldsymbol{x}\right\|_{\ell_2}^2}{\sigma^2}$$

PWF works.

[Oymak and Hassibi 2014-Oymak, Recht, Soltanolkotabi 2016] + [Amelunxen, Lotz, McCoy, Tropp 2014] shows equivalence between min-max denoising and data complexity of linear inverse problems

Theoretical implications

- signal with entries ± 1

Theoretical implications

- signal with entries ± 1
  no problem best Gaussian denoiser is actually $tanh$

Theoretical implications

- signal with entries ± 1
  no problem best Gaussian denoiser is actually $tanh$
- optimization over integers?

Theoretical implications

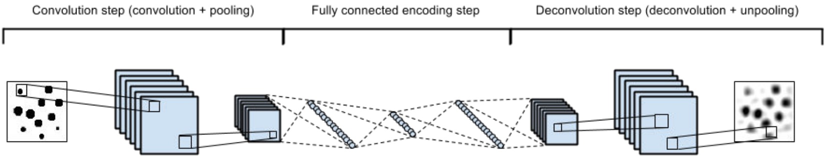
- signal with entries ± 1
  no problem best Gaussian denoiser is actually $tanh$
- optimization over integers?
  no problem just threshold to the closest integer...
- many others

Implications for imaging systems

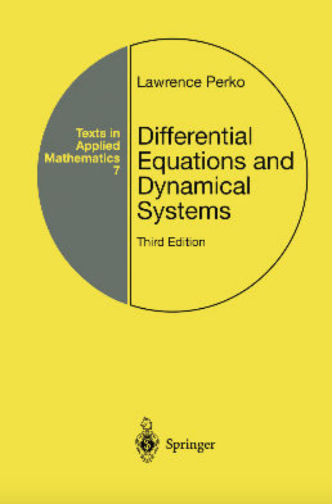
What projection or non-linear shrinkage should you use?

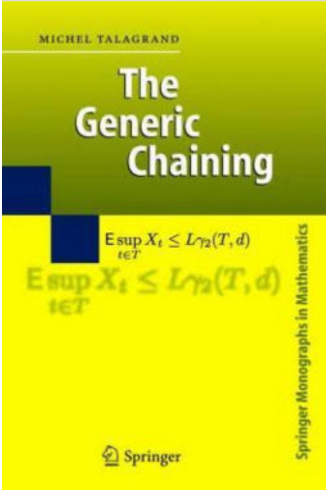
Implications for imaging systems

What projection or non-linear shrinkage should you use?
We use GDS file from IBM add Gaussian noise and just learn the best denoiser...



Tools


Lawrence Perko
Texts in Applied Mathematics 7
Differential Equations and Dynamical Systems
Third Edition
Springer


MICHEL TALAGRAND
The Generic Chaining
$\mathsf{E}\sup_{t\in T} X_t \le L\gamma_2(T,d)$
$\mathsf{E}\sup_{t\in T} X_t \le L\gamma_2(T,d)$
Springer Monographs in Mathematics
Springer

Regularity condition?

$$\langle \nabla f(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x} \rangle \geq \frac{1}{\alpha} \|\boldsymbol{z} - \boldsymbol{x}\|_{\ell_2}^2 + \frac{1}{\beta} \|\nabla f(\boldsymbol{z})\|_{\ell_2}^2$$

Regularity condition?

$$\langle \nabla f(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x} \rangle \geq \frac{1}{\alpha} \|\boldsymbol{z} - \boldsymbol{x}\|_{\ell_2}^2 + \frac{1}{\beta} \|\nabla f(\boldsymbol{z})\|_{\ell_2}^2$$

Not really ...

Proof Sketch

$$\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_\tau - \mu_\tau \nabla f(\boldsymbol{z}_\tau).$$

Want to prove

$$\|\boldsymbol{z}_{\tau+1} - \boldsymbol{x}\|_{\ell_2} \leq \frac{1}{2} \|\boldsymbol{z}_\tau - \boldsymbol{x}\|_{\ell_2}$$

Proof Sketch

$$\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_{\tau} - \mu_{\tau} \nabla f(\boldsymbol{z}_{\tau}).$$

Want to prove

$$\|\boldsymbol{z}_{\tau+1} - \boldsymbol{x}\|_{\ell_2} \leq \frac{1}{2} \|\boldsymbol{z}_{\tau} - \boldsymbol{x}\|_{\ell_2}$$

Define the stochastic process

$$X_{\boldsymbol{u},\boldsymbol{z}} = \frac{\boldsymbol{u}^T \left(\boldsymbol{z} - \mu \nabla f(\boldsymbol{z})\right)}{\|\boldsymbol{z} - \boldsymbol{x}\|_{\ell_2}}$$

Proof Sketch

$$\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_{\tau} - \mu_{\tau} \nabla f(\boldsymbol{z}_{\tau}).$$

Want to prove

$$\|\boldsymbol{z}_{\tau+1} - \boldsymbol{x}\|_{\ell_2} \leq \frac{1}{2} \|\boldsymbol{z}_{\tau} - \boldsymbol{x}\|_{\ell_2}$$

Define the stochastic process

$$X_{\boldsymbol{u},\boldsymbol{z}} = \frac{\boldsymbol{u}^T \left(\boldsymbol{z} - \mu \nabla f(\boldsymbol{z})\right)}{\|\boldsymbol{z} - \boldsymbol{x}\|_{\ell_2}}$$

We prove that for all $\boldsymbol{u} \in \mathbb{R}^n$ and $\boldsymbol{z}$ obeying $\mathcal{R}(\boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{x})$

$$\sup_{\boldsymbol{u} \in \mathbb{S}^{n-1}, \boldsymbol{z} \in \mathcal{K}} X_{\boldsymbol{u},\boldsymbol{z}} \leq \frac{1}{2}$$

Submodular Maximization

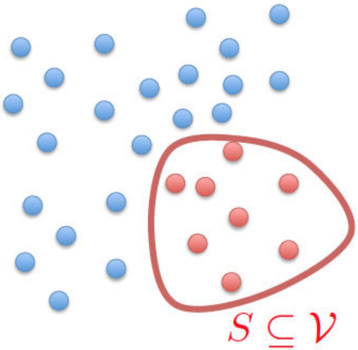
Collaborators: Hamed Hassani and Amin Karbasi

Submodular Maximization

Collaborators: Hamed Hassani and Amin Karbasi

(Introductory figures/slides stolen from Stefanie Jegelka and Andreas Krause)

Set Function Maximization



- ground set $\mathcal{V}$
- (scoring) function
  $F : 2^{\mathcal{V}} \to \mathbb{R}_+$

$$\boxed{\max \quad F(S)}$$

Maximizing monotone functions

$$\max_{\mathcal{S} \subset \mathcal{V}} F(\mathcal{S}) \quad \text{subject to} \quad |\mathcal{S}| \leq k$$

Maximizing monotone functions

$$\max_{\mathcal{S}\subset\mathcal{V}} F(\mathcal{S}) \quad \text{subject to} \quad |\mathcal{S}| \leq k$$

Greedy algorithm

- $\mathcal{S}_0 =$
- for $i = 0, 1, \ldots, k-1$

$$\begin{aligned} e^* &= \arg\max_{e\in\mathcal{V}/\mathcal{S}_i} F(\mathcal{S}_i \cup \{e\}) \\ \mathcal{S}_{i+1} &= \mathcal{S}_i \cup \{e^*\} \end{aligned}$$

Theory for greedy

$$\max_{\mathcal{S} \subset \mathcal{V}} F(\mathcal{S}) \quad \text{subject to} \quad |\mathcal{S}| \leq k$$

Theory for greedy

$$\max_{\mathcal{S}\subset\mathcal{V}} F(\mathcal{S}) \quad \text{subject to} \quad |\mathcal{S}| \leq k$$

Theorem (Nemhauser, Fisher, Wolsey '78)

F monotone submodular. Then solution of greedy obeys

$$F(\hat{\mathcal{S}}) \geq \left(1 - \frac{1}{e}\right) F(\mathcal{S}^*)$$

Theory for greedy

$$\max_{\mathcal{S} \subset \mathcal{V}} F(\mathcal{S}) \quad \text{subject to} \quad |\mathcal{S}| \leq k$$

Theorem (Nemhauser, Fisher, Wolsey '78)

F monotone submodular. Then solution of greedy obeys

$$F(\hat{\mathcal{S}}) \geq \left(1 - \frac{1}{e}\right) F(\mathcal{S}^*)$$

No poly-time algorithm can do better than that!

Why not just use greedy

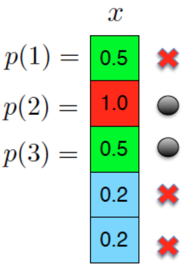
- Many cases don't have exact function evaluations
- Greedy takes $O(nk)$ time. What if n is large?
- What if the function is not submodular

Making things continuous

sample item e with probability x_e

$$f_M(x) = \mathbb{E}_{S \sim x}\left[F(S)\right]$$

$$= \sum_{S \subseteq \mathcal{V}} F(S) \prod_{e \in S} x_e \prod_{e \notin S} (1 - x_e)$$



Basis for continuous greedy [Vondrak et. al.]

Just follow the gradient

$$\boldsymbol{x}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{x}_{\tau} + \mu_{\tau} \nabla f_M(\boldsymbol{x}_{\tau})\right)$$

where

$$\mathcal{K} = \{\boldsymbol{z} \in \mathbb{R}_{+}^{n} : \sum_{i=1}^{n} z_i = k \quad 0 \leq z_i \leq 1\}$$

How well does it work?

$$\max_{\mathcal{S}\subset\{1,2,\ldots,n\}} F(\mathcal{S}) = \mathsf{logdet}(\boldsymbol{I} + \boldsymbol{A}_{\mathcal{S},\mathcal{S}}) \quad \text{subject to} \quad |\mathcal{S}| \leq k$$

How well does it work?

$$\max_{\mathcal{S} \subset \{1,2,\ldots,n\}} F(\mathcal{S}) = \mathsf{logdet}(\boldsymbol{I} + \boldsymbol{A}_{\mathcal{S},\mathcal{S}}) \quad \text{subject to} \quad |\mathcal{S}| \leq k$$

Greedy: 67.1 Gradient Descent: 74.81

Stochastic Methods

Assume access to a stochastic oracle

$$\mathbb{E}[\boldsymbol{g}_t] = \nabla f_M(\boldsymbol{x}_t).$$

Run

$$\boldsymbol{x}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{x}_\tau + \mu_\tau \boldsymbol{g}_\tau\right)$$

where

$$\mathcal{K} = \{\boldsymbol{z} \in \mathbb{R}^n_+ : \sum_{i=1}^{n} z_i = k \quad 0 \leq z_i \leq 1\}$$

Some theory

$$\boldsymbol{x}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{x}_\tau + \mu_\tau g_\tau\right)$$

Theorem (Stochastic Gradient Method)

Assumptions

- $R^2 = \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{K}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell_2}^2$
- *f_M is L-smooth, monotone and multinear extension of submodular*
- *stochastic oracle $\boldsymbol{g}_t$ obeying*

$$\mathbb{E}[\boldsymbol{g}_t] = \nabla f_M(\boldsymbol{x}_t) \quad \textit{and} \quad \mathbb{E}\left[\,\|\boldsymbol{g}_t - \nabla f_M(\boldsymbol{x}_t)\|_{\ell_2}^2\,\right] \le \sigma^2.$$

Run stochastic gradient updates with $\mu_t = \frac{1}{L+\frac{\sigma}{R}\sqrt{t}}$. Then,

$$\mathbb{E}[f_M(\boldsymbol{x}_T)] \ge \mathrm{OPT}\left(\frac{1}{2} - (\frac{R^2 L}{T} + 2\frac{R\sigma}{\sqrt{T}})\right).$$

Some theory

$$\boldsymbol{x}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{x}_\tau + \mu_\tau g_\tau\right)$$

Theorem (Stochastic Gradient Method)

Assumptions

- $R^2 = \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{K}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell_2}^2$
- f_M *is* L*-smooth, monotone and multinear extension of submodular*
- *stochastic oracle* $\boldsymbol{g}_t$ *obeying*

$$\mathbb{E}[\boldsymbol{g}_t] = \nabla f_M(\boldsymbol{x}_t) \quad \textit{and} \quad \mathbb{E}\left[\,\|\boldsymbol{g}_t - \nabla f_M(\boldsymbol{x}_t)\|_{\ell_2}^2\,\right] \le \sigma^2.$$

Run stochastic gradient updates with $\mu_t = \frac{1}{L+\frac{\sigma}{R}\sqrt{t}}$*. Then,*

$$\mathbb{E}[f_M(\boldsymbol{x}_T)] \ge \mathrm{OPT}\left(\frac{1}{2} - (\frac{R^2L}{T} + 2\frac{R\sigma}{\sqrt{T}})\right).$$

- With Mirror descent can ensure L is constant
- Can get better approximation ratio starting from 0

Conclusion

- Convex relaxations may be inefficient in terms of sample complexity
- discussed results towards breaking this barrier
- a lot of exciting barriers to think about e.g. planted clique
- interesting directions for bridging the gap between discrete and continuous optimization

References

- Phase retrieval
  - Phase retrieval via Wirtinger flow: Theory and algorithms E. J. Candes, X. Li, and M. Soltanolkotabi
  - Algorithms and theory for clustering and non-convex quadratic programming. M. Soltanolkotabi 2014.
  - Structured signal recovery from quadratic measurements: breaking data barriers via nonconvex optimization, M. Soltanolkotabi, 2017
- Low-rank matrix recovery
  - Low-rank Solutions of Linear Matrix Equations via Procrustes Flow. S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht 2015.
- Sharp time-data tradeoffs for (non)convex projected gradients
  - Sharp Time–Data Tradeoffs for Linear Inverse Problems. S. Oymak, B. Recht and M. Soltanolkotabi
  - Fast and reliable parameter estimation from nonlinear observations. S. Oymak and M. Soltanolkotabi, 2016.
- Submodular maximization
  - Stochastic gradient methods for submodular maximization H. Hassani, M. Soltanolkotabi, and A. Karbasi.

Thanks!

When should we just follow the gradient?