# Fast Data Driven Compressed Sensing

and application to

## compressed quantitative MRI

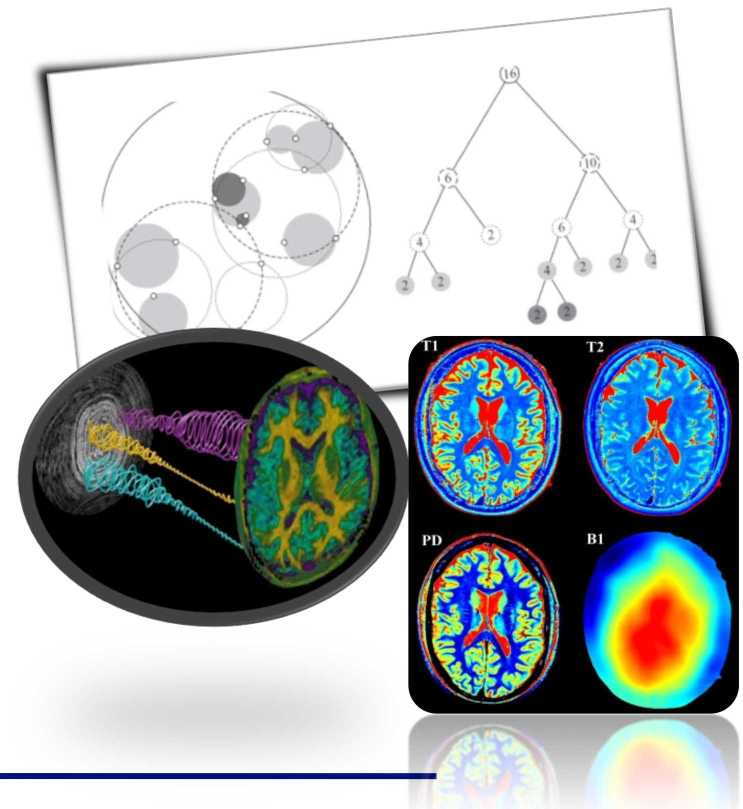
Mike Davies

& Mohammad Golbabaee

Joint work with

Zhouye Chen, Yves Wiaux

Zaid Mahbub, Ian Marshall



# Outline

- Iterative Projected Gradients (IPG)
- Approximate/inexact oracles
- Robustness of inexact IPG
- Application to data driven Compressed Sensing
  - Fast MR Fingerprinting reconstruction
  - IPG with Approximate Nearest Neighbour searches
  - Cover trees for fast ANN
- Numerical results

# Inverse problems

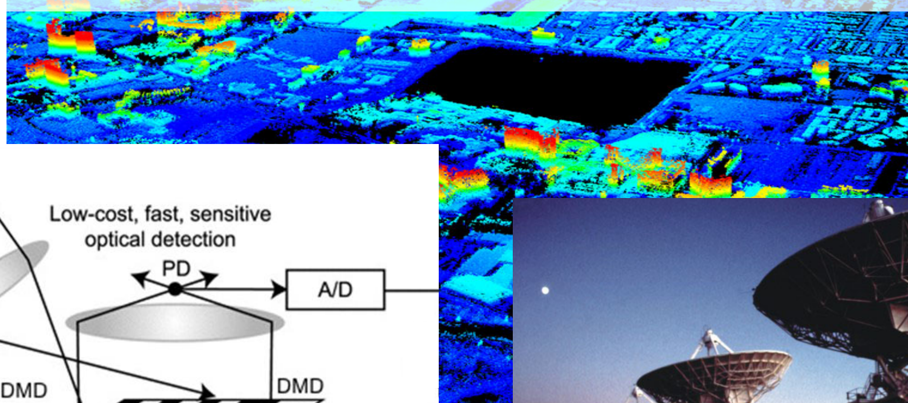
$y = Ax + w \in \mathbb{R}^m$, $\quad A : \mathbb{R}^n \to \mathbb{R}^m$ where $m \ll n$

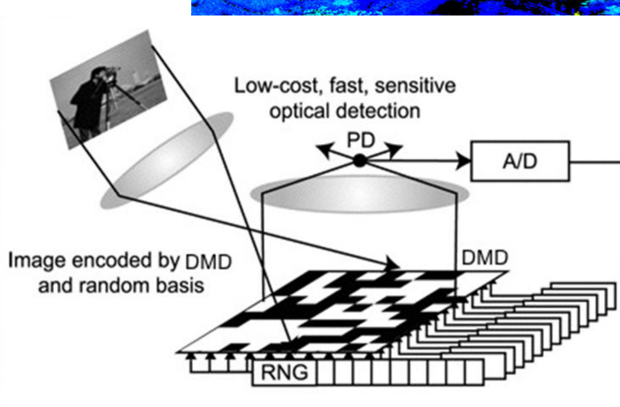
**Challenge:** Missing information

Complete measurements can be costly, time consuming and sometimes just impossible!

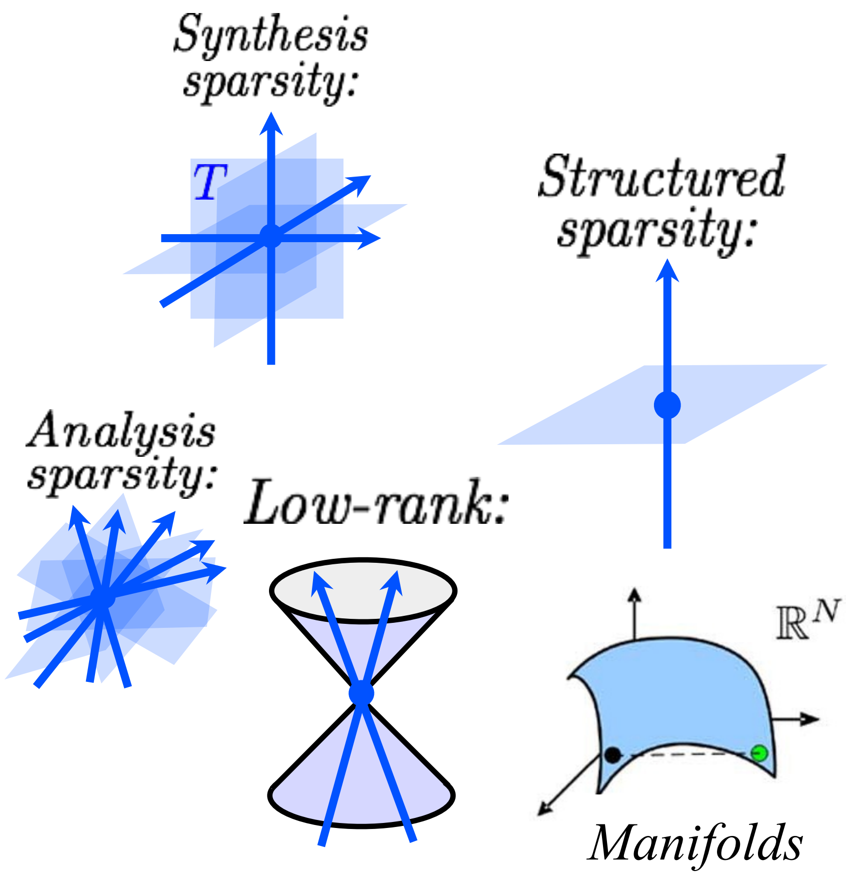
**Compressed sensing** to address the challenge

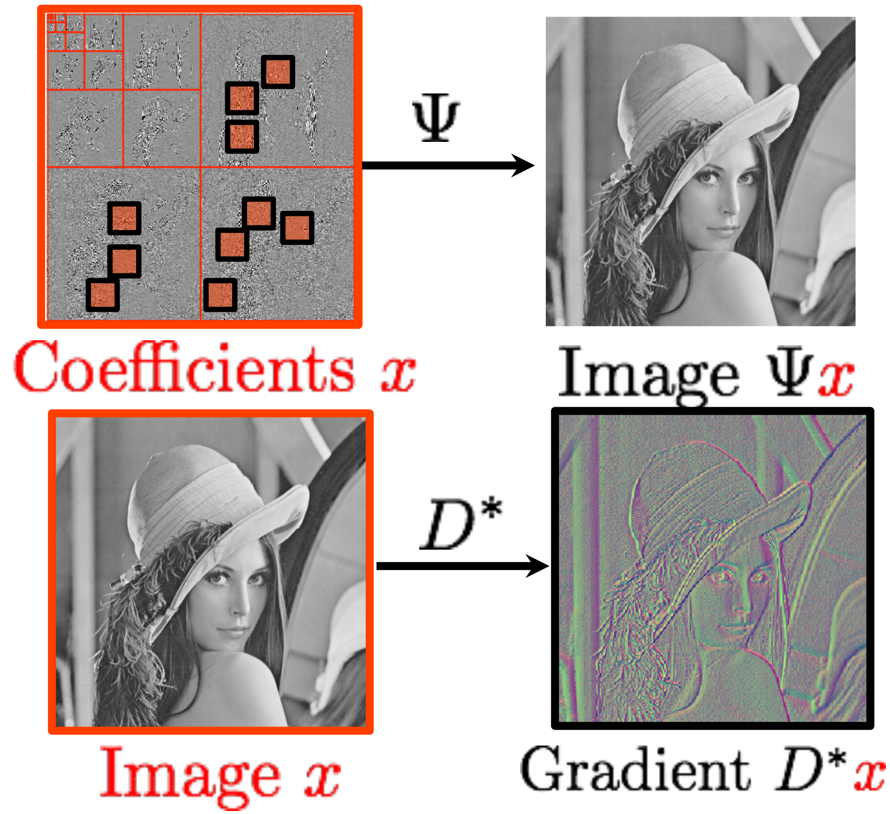
*[Donoho'06; Candès, Romberg, and Tao,'06]*





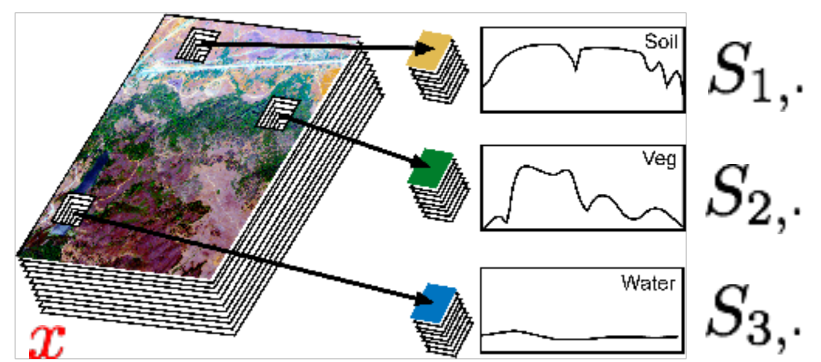

# Data models/priors




*Multi-spectral imaging:*

$$x_{i,\cdot} = \sum_{j=1}^{r} A_{i,j} S_{j,\cdot}$$



# Solving Compressed Sensing/Inv. Problems

Estimating by constrained least squares

$$x \in \arg\min \{ f(x) := \tfrac{1}{2} \|y - Ax\|_2^2 \} \quad s.t. \quad x \in C$$

!! NP-hard for most interesting models

(e.g. sparsity ***[Natarajan'95]***)

**Iterative Gradient Projection (IPG)**

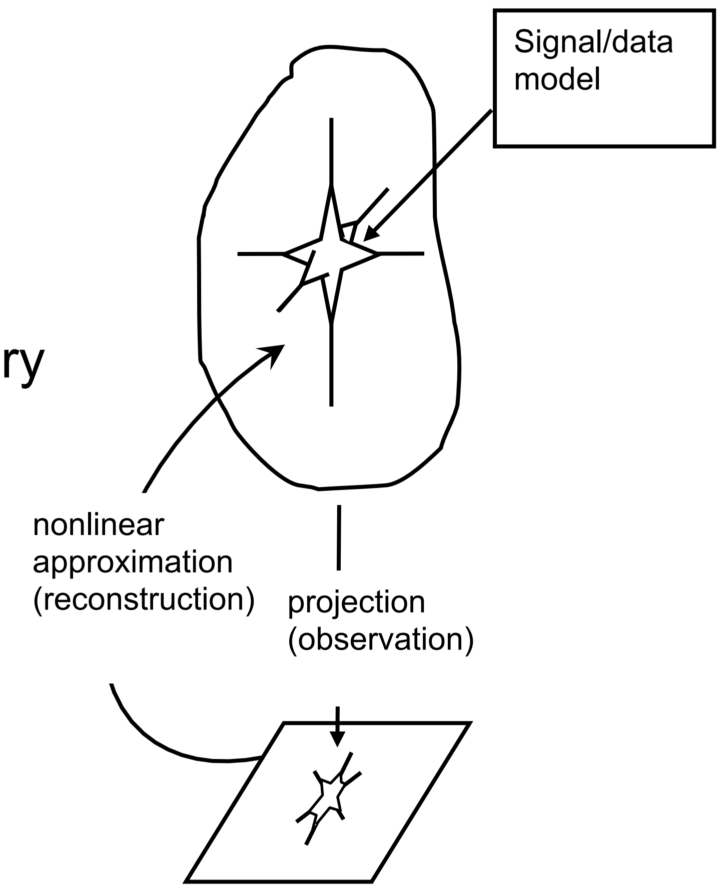
Generally proximal-gradient algorithms are very popular:

$$x^k = \boldsymbol{P}_C \left( x^{k-1} - \mu A^T (A x^{k-1} - y) \right)$$

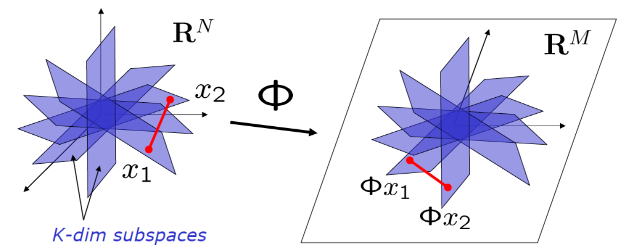
Gradient $A^T (A x^{k-1} - y)$, step size $\mu$,

Euclidean projection

$$\boldsymbol{P}_C(x) \in \arg\min \|x - x'\|_2 \quad s.t. \quad x' \in C$$



# Embedding: key to CS/IPG stability




**Bi-Lipschitz embeddable sets:** $\forall x, x' \in \mathcal{C}$

$$\alpha \|x - x'\|_2 \leq \|A(x - x')\|_2 \leq \beta \|x - x'\|_2$$

**Theorem**.*[Blumensath'11]* *For any* $(\mathcal{C}, A)$ *if holds* $\beta \leq 1.5\alpha$,

IPG → *stable & linear convergence:* $\|x^K - x_0\| \to O(w) + \tau, \quad K \sim (\log \tau^{-1})$

Global optimality even for nonconvex programs!

| **Model $\mathcal{C} \in \mathbb{R}^n$** | **O($m$)** | |
|---|---|---|
| $p$ (unstructured) points | $\theta^{-2} \log(p)$ | *[Johnson, Lindenstrauss'89]* |
| $\cup_i^L K$-flats | $\theta^{-2} (K + \log(L))$ | *[Blumensath, Davies'09]* |
| rank $r$ $(\sqrt{n} \times \sqrt{n})$ matrices | $\theta^{-2} r\sqrt{n}$ | *[Candès,Recht;Ma etal.'09]* |
| 'smooth' $d$ dim. manifold | $\theta^{-2} d$ | *[Wakin,Baraniuk'09]* |
| $K$ tree-sparse | $\theta^{-2} K$ | *[Baraniuk etal.'09]* |

Sample complexity e.g. $A$~ *i.i.d.* subgaussian

# A limitation…

Exact oracles might be too expensive or even do not exist!

**Gradient** $A^T(Ax^{k-1} - y))$

- $A$ too large to fully access or fully compute/update $\nabla f$
- Noise in communication in distributed solvers

**Projection** $\boldsymbol{P}_C(x) \in \operatorname{argmin} \|x - x'\|_2 \; s.t. \; x' \in C$

- $\boldsymbol{P}_C$ may not be analytic and requires solving an auxiliary optimization (e.g. inclusions $C = \bigcap_i C_i$, total variation ball, low-rank, tree-sparse,…)
- $\boldsymbol{P}_C$ might be NP hard! (e.g. analysis sparsity, low-rank tensor decomposition)

Is IPG robust against inexact/approximate oracles?

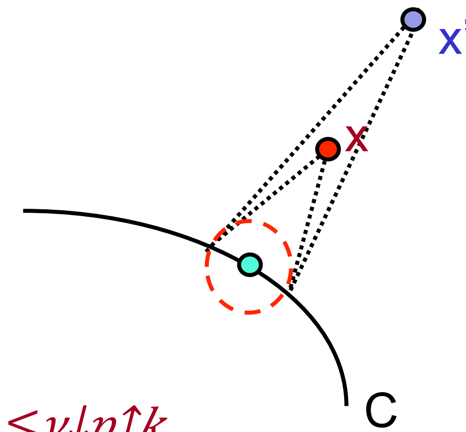
# Inexact oracles I: Fixed Precision

$$x^k = \mathbf{P}_C\left(x^{k-1} - \mu \nabla f(x^{k-1})\right)$$

Fixed Precision (FP) approximate oracles:

$$\|\nabla f(.) - \nabla f(.)\|_2 \leq \nu_g, \quad \|\mathbf{P}_C(.) - \mathbf{P}_C(.)\|_2 \leq \nu_p, \quad (\mathbf{P}_C(.) \in C)$$

**Examples:** TV ball, inclusions (e.g. Djkstra alg.), and many more… (in convex settings, Duality gap → FP proj.)




Progressive Fixed Precision (PFP) oracles:

$$\|\nabla f(.) - \nabla f(.)\|_2 \leq \nu_g^k, \quad \|\mathbf{P}(.) - \mathbf{P}(.)\|_2 \leq \nu_p^k$$

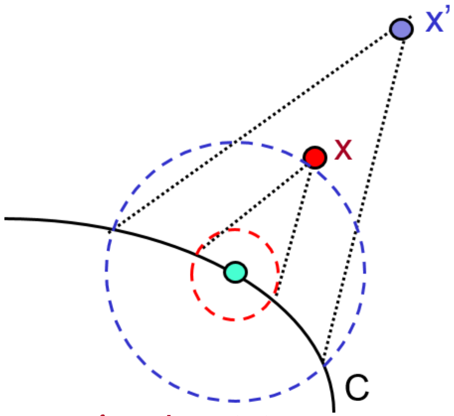
**Examples:** *Any FP oracle* with progressive refinement of the approx. levels

e.g. convex sparse CUR factorization for $\nu_p^k \sim O(1/k^3)$ ***[Schmidt etal.'11]***

# Inexact oracles II: $(1+\epsilon)$-optimal

$(1+\epsilon)$-approximate projections combined with FP/PFP gradient oracle

$$x^k = \boldsymbol{P}^{\epsilon}_{C}\left(x^{k-1} - \mu \nabla_k f(x^{k-1})\right)$$




Gradient $||\nabla_k f(.) - \nabla f(.)||_2 \leq \nu_g^k$

Projection $||\boldsymbol{P}^{\epsilon}_{C}(x) - x||_2 \leq (\boldsymbol{1}+\boldsymbol{\epsilon})||\boldsymbol{P}_{C}(x) - x||_2$

**Examples.**

Many nonconvex constraints: Cheaper low-rank proxi $\boldsymbol{P}^{\epsilon}_{C}(x) \in C$ randomized lin. algebra ***[Halko etal.'11]***, K-tree sparse signals ***[Hegde etal.'14]***, Tensor low-rank (Tucker) decomposition ***[Rauhut etal.'16]***,

*… and shortly, ANN for data driven CS!*

# **Robustness & linear convergence**
*of the*
# **inexact IPG**

# IPG with (P)FP oracles

$$x^k = \mathbf{P}_{\mathcal{C}}\left(x^{k-1} - \mu \nabla f(x^{k-1})\right)$$

**Theorem.** *For any* $(x_0 \in \mathcal{C}, \mathcal{C}, \mathcal{A})$ *if* $\beta \leq \mu^{-1} < 2\alpha_0$ *then*

$$\|x^k - x_0\| \leq \rho^k \left(\|x_0\| + \sum_{i=1}^{k} \rho^{-i} e^i\right) + \frac{2\sqrt{\beta}}{(1-\rho)\alpha_0} w$$

*where,* $\rho = \sqrt{\frac{1}{\mu\alpha_0} - 1}$ *and* $e^i = \frac{2\nu_{g^i}}{\alpha_0} + \sqrt{\frac{\nu_{p^i}}{\mu\alpha_0}}$

***Remark:***

$\rho^{-i}$ *supresses the early stages errors*

$\Rightarrow$ *use "progressive" approximations to get as good as exact!*

# IPG with (P)FP oracles

$$x^k = P_C(x^{k-1} - \mu \nabla f(x^{k-1}))$$

**Corollary I.** *After* $K=O(\log(\tau^{-1}))$ *iterations IPG-FP achieves*

$$\|x^K - x_0\| \le O(w + \nu_g + \sqrt{\nu_p}) + \tau$$

*linear convergence at rate* $\rho = \sqrt{1/\mu\alpha_0 - 1}$ *.*

**Corollary II.** *Assume* $\exists r<1 \ s.t. \ e^i = O(r^i)$*, then after* $K=O(\log(\tau^{-1}))$ *iterations IPG-PFP achieves*

$$\|x^K - x_0\| \le O(w) + \tau$$

*linear convergence at rate* $\rho = \begin{cases} \max(\rho, r) & \rho \ne r \\ \rho + \xi & \rho = r \end{cases}$ *(for any small* $\xi > 0$)

# IPG with $(1+\epsilon)$-approximate projection

$$x^k = \boldsymbol{P}_{C}^{\epsilon}\left(x^{k-1} - \mu \nabla_k f(x^{k-1})\right)$$

**Theorem.** *Assume for any* $(x_0 \in C, C, A)$ *and an* $\epsilon \geq 0$ *it holds*

$$\sqrt{2\epsilon+\epsilon^2} \leq \delta \sqrt{\alpha_0} / |||A||| \quad \text{and} \quad \beta \leq \mu^{-1} < (2-2\delta+\delta^2)\,\alpha_{x_0}$$

*Then,*

$$\|x^k - x_0\| \leq \rho^k \left(\|x_0\| + \kappa_g \sum_{i=1}^{k} \rho^{-i} \nu_g^i\right) + \kappa_z / (1-\rho)\, w$$

*where,* $\rho = \sqrt{1/(\mu\alpha_0) - 1} + \delta$, $\kappa_g = 2/\alpha_0 + \sqrt{\mu}/|||A|||\,\delta$ and $\kappa_z = 2\sqrt{\beta}/\alpha_0 + \sqrt{\mu}\,\delta$

**Remarks.**

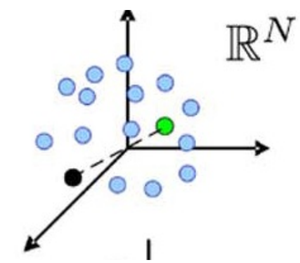
- *Requires stronger embedding cond., slower convergence!*
- *Still linear conv. &* $O(w)+\tau$ *accuracy after* $O(\log \tau^{-1})$ *iterations*
- *higher noise amplification*

# Application in data driven CS

# Data driven CS

In the absence of (semi) algebraic physical models (l0, l1, rank,…)

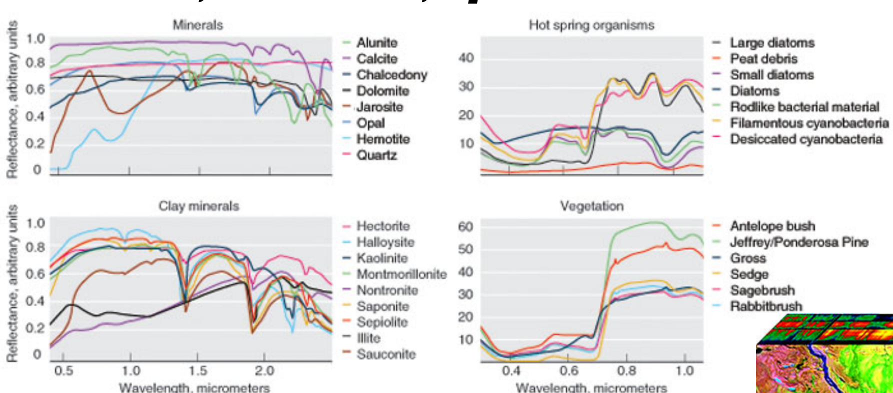
Collect a possibly large dictionary (sample the model)

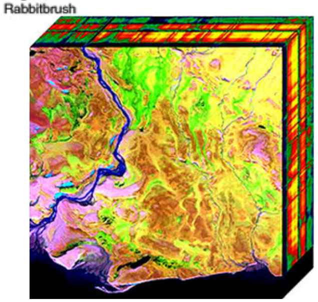


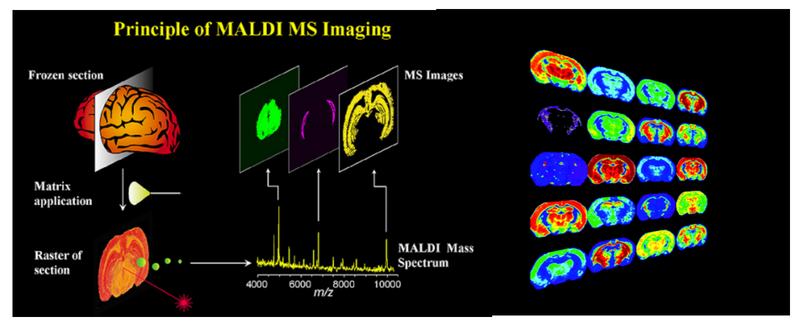

$\mathcal{C} = \cup_{i=1,\dots,d} \{\psi_i\}$ $\psi_i$ *atoms of* $\Psi \in \mathbb{R}^{n \times d}$

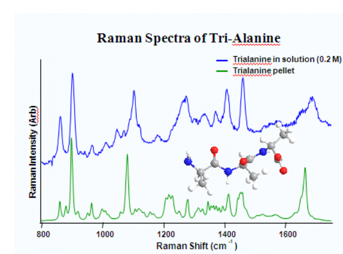
*Examples in multi-dim. imaging:*

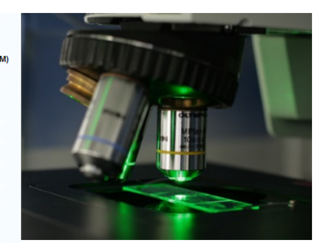
*Hyperspectral, Mass/Raman/MALDI,… spectroscopy* ***[Golbabaee et al '13; Duarte,Baraniuk'12;…]***





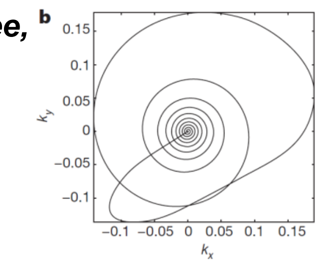



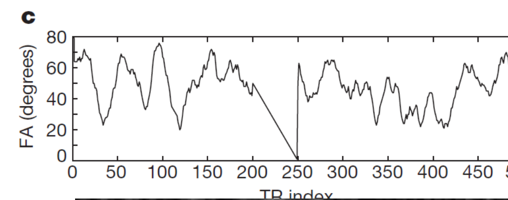



# MR Fingerprinting: fast CQ-MRI

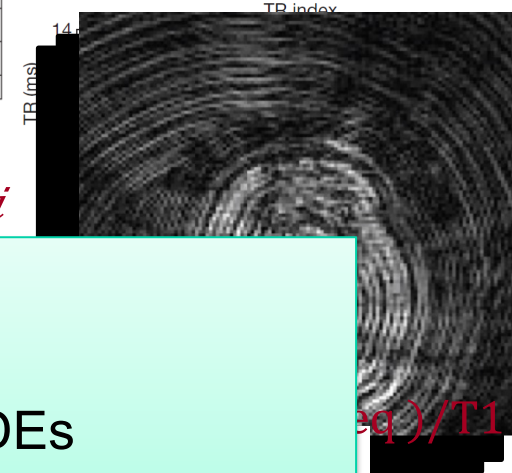
**Goal:** Measuring fast the NMR properties (relaxation times: T1, T2) ***[Ma etal.'13]***

1. Multiple random/optimized excitations (magnetic field rotation) ***[Cohen'15;Mahbub, Golbabaee, D., Marshall '16]***




2. Subsampling the k-space (per excitation)



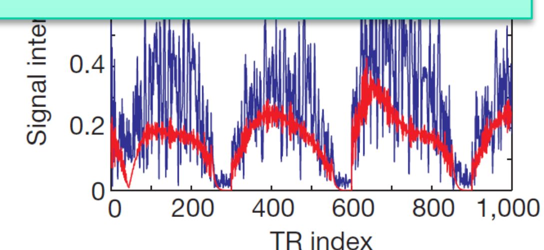
3. Construct a (huge) dictionary of "fingerprints"
   i.e.
   par

Continuous model (Bloch manifold) exists,

- but not explicit i.e. requires solving dynamic ODEs
- MRF sacrifices memory ~ enables $P \downarrow C$ computation

4. Use

   - Exact IPG ***[D. et al '15]***



# CS in product (tensor) space

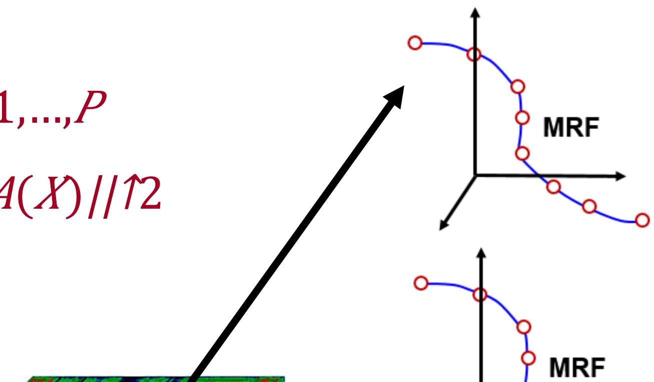
Per pixel data driven model

$C = \cup_{i=1,\dots,d}\{\psi_i\}$  $\psi_i$ atoms of $\Psi \in R^{n\times d}$

Multi dim. image: $X \in R^{n\times P}$ where $X_p \in C \ \forall p=1,\dots,P$

Inverse problem: $\min_{X_{vec}\in \prod_j C} \|y - A(X)\|^2$

Direct recovery complexity $O(P^d)$!




**IPG (LNNsearch G)**

$X_p$

Con

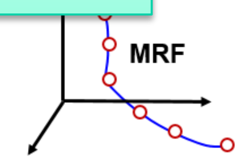
Suff

Big datasets?

MRF uses large dictionaries d~ 500K

Can we do better?

$M = O(P\, m)$, (ignoring inter channel structures)



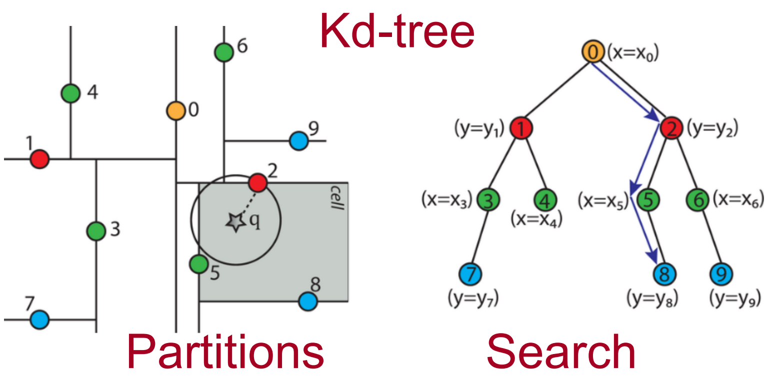

$m=$ *RIP complexity of each channel*

# Fast NN searches

**Trees**: (historical) approach to fast NN

- Hierarchical partitioning + Brand & bound

search. *e.g., kd-trees, Metric/Ball-trees,…*



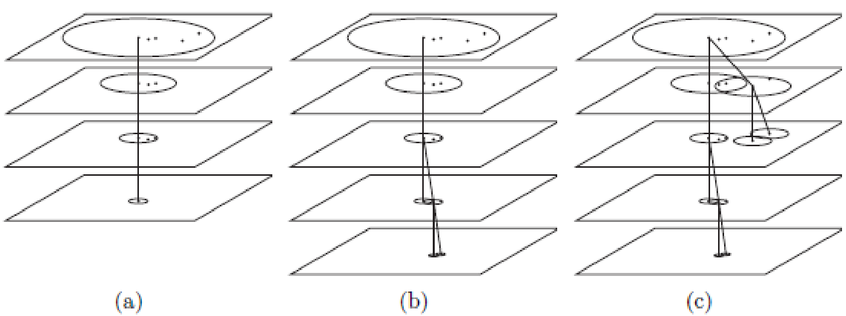

"Curse of dimensionality": exact NN in $\mathbb{R}^n$ cannot achieve $o(nd)$ in time with a reasonable memory. Approximate NN can! If datasets ~ low dim.

## Nagivating nets, Cover trees

***[Krauthgamer,Lee'04; Beygelzimmer'06]***

At scale $l = 1,\ldots,L$

- Covering (parent nodes) $\sigma 2^{-l}$
- Separation (nodes appearing at scale l) $\sigma 2^{-l}$




CT builds multi-resolution cover-nets

Brand & bound search

# CT provably good for low dim data

## Search options with cover trees

1. $(1+\epsilon)$*-ANN:* as proposed by ***[Beygelzimmer et al.'06]***

> ***Theorem.*** Cover tree $(1+\epsilon)$-ANN complexity: ***[Krauthgamer + Lee 04]***
>
> $2^{\mathcal{O}(\dim(\mathcal{C}))}\log\Delta + \epsilon^{-\mathcal{O}(\dim(\mathcal{C}))}$ *in time,* $\mathcal{O}(d)$ *space*
>
> *(typically* $\log\Delta = \mathcal{O}(\log(d))$)

*2. FP-ANN:* truncated tree $L = \lceil \log(\nu_p/\sigma) \rceil$ + exact NN

(complexity could be arbitrary large/theoretically)

*3. PFP-ANN:* progressing on truncation level $\nu_p^k = O(r^k)$, $r<1$

## Inexact IPG

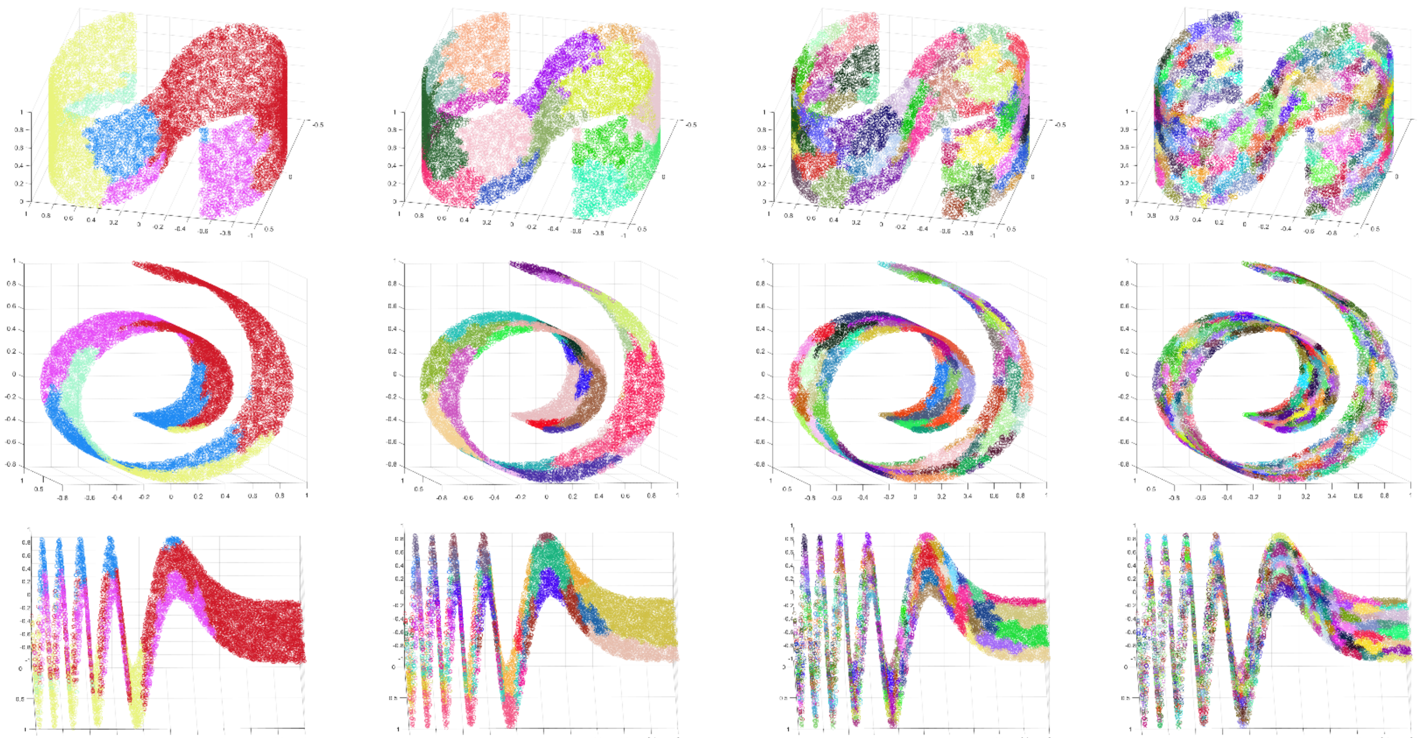
$$X_p^k = ANN_{\mathcal{C}}\left(\left[X^{k-1} - \mu A^H\left(A(X^{k-1}) - y\right)\right]_p\right), \forall p$$

# Numerical experiments 1: Toy problem

# 2D manifold data

$C=\cup_{i=1,\ldots,d}\{\psi_i\}$ $\psi_i$ atoms $\Psi\in R^{n\times d}$

CT level 2 | CT level 3 | CT level 4 | CT level 5



| Dataset | Population | Ambient dim. (N) | CT depth | CT res. |
|---|---|---|---|---|
| S-Manifold | 5’000 | 200 | 14 | 2.43E-4 |
| Swissroll | 5’000 | 200 | 14 | 1.70E-4 |
| Oscillating wave | 5’000 | 200 | 14 | 1.86E-4 |

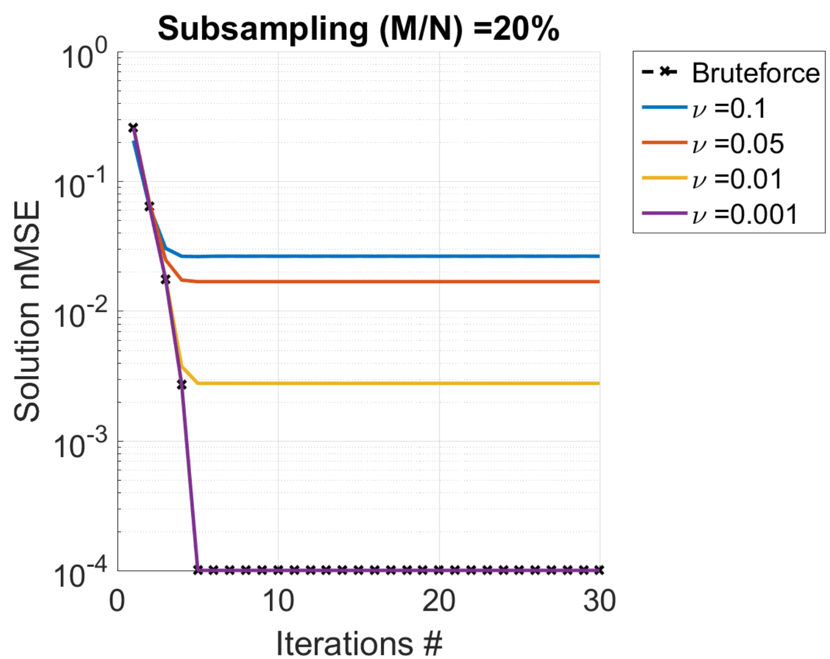
# Solution accuracy vs. Iterations (FP)

Signal: $X \in \mathbb{R}^{n \times P}$ $n=200$ , $P=50$ (randomly chosen $\in \mathcal{C}$)

$\mathrm{m}P \times nP$ i.i.d. Normal *A*, CS ratio=m/n (noiseless)

$\min_{X_{vec} \in \prod_j C} ||y - A(X)||^2 \quad \Leftrightarrow X_p^k = ANN_C\left([X^{k-1} - \mu A^H (A(X^{k-1}) - y)]_p\right), \forall p$

Swiss roll



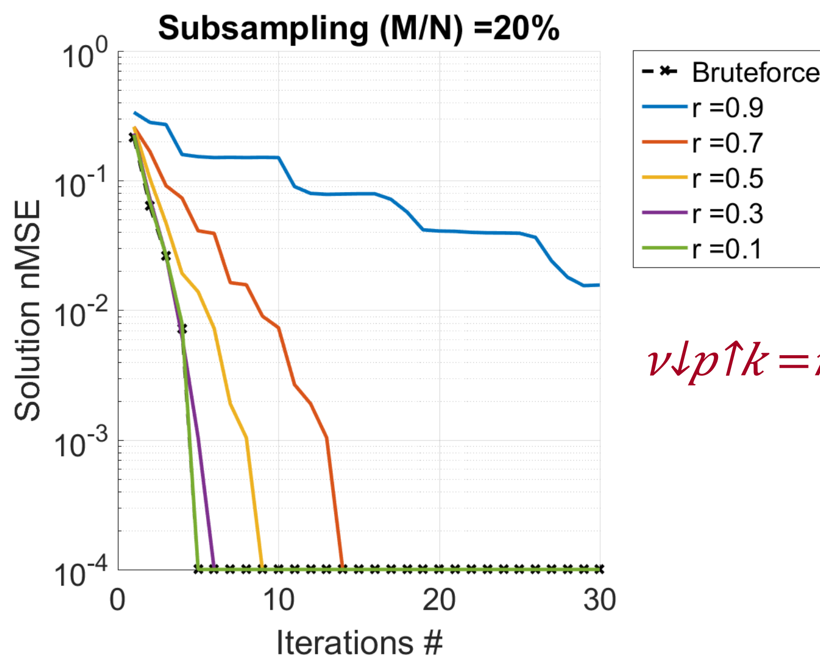
# Solution accuracy vs. Iterations (PFP)

Signal: $X\in\mathbf{R}^{n\times P}$ $n=200$, $P=50$ (randomly chosen $\in C$)

$mP\times nP$ i.i.d. Normal *A*, CS ratio=m/n (noiseless)

$$\min_{X_{vec}\in\prod_j C} \|y-A(X)\|^2 \quad \Leftrightarrow X_p^k = ANN_C([X^{k-1}-\mu A^H(A(X^{k-1})-y)]_p), \forall p$$

Swiss roll




$\nu_p^k = r^k$

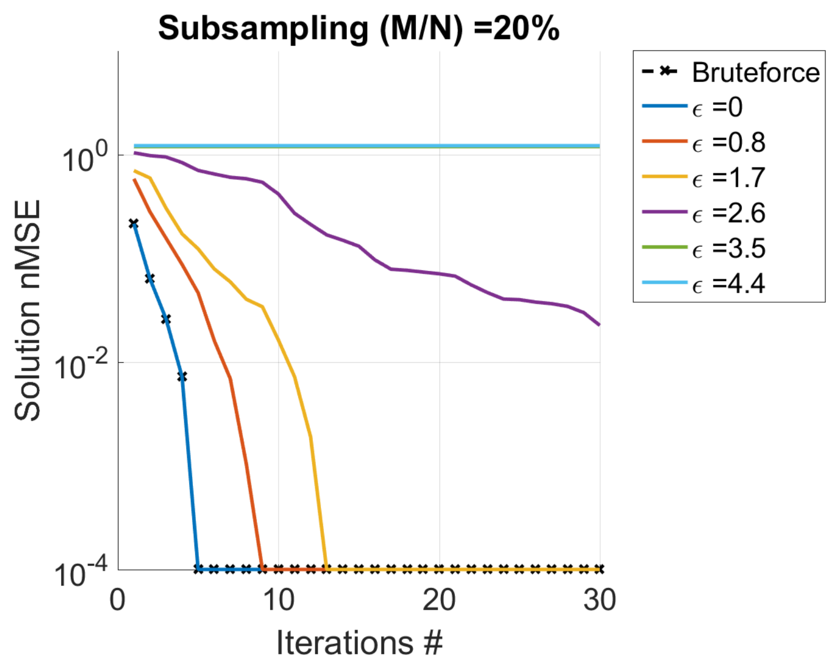
# Solution accuracy vs. Iterations $(1+\epsilon)$-ANN

Signal: $X \in \mathbf{R}^{n\times P}$ $n=200$, $P=50$ (randomly chosen $\in \mathcal{C}$)

$mP\times nP$ i.i.d. Normal *A*, CS ratio=m/n (noiseless)

$$\min_{X_{vec}\in\prod_j C} ||y-A(X)||^2 \quad \Leftrightarrow X_p^k = ANN_C\left([X^{k-1} - \mu A^H(A(X^{k-1})-y)]_p\right), \forall p$$

Swiss roll

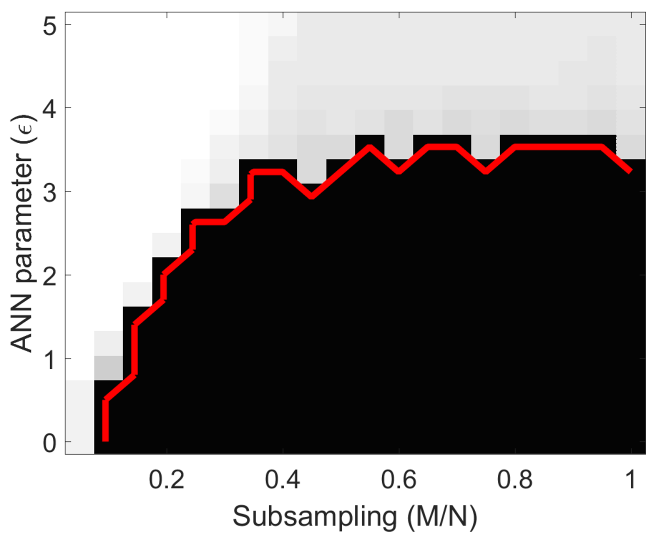


# Phase transitions

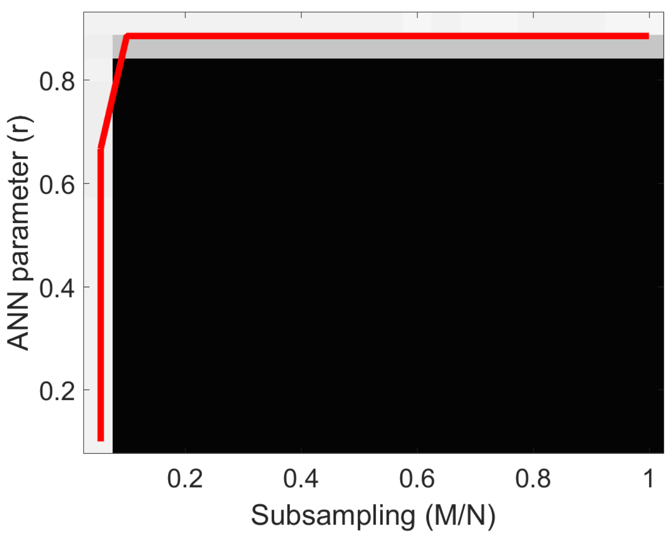
Signal: $X \in \mathbf{R}^{n\times P}$ $n=200$, $P=50$ (randomly chosen $\in \mathcal{C}$)

$\mathrm{m}P \times nP$ i.i.d. Normal *A*, CS ratio=m/n (noiseless) ~ averaged 25trials

Recovery PT: Black/white = low/high sol. nMSE, red curve = recovery region nMSE<10e-4)




$(1+\epsilon)$-ANN IPG



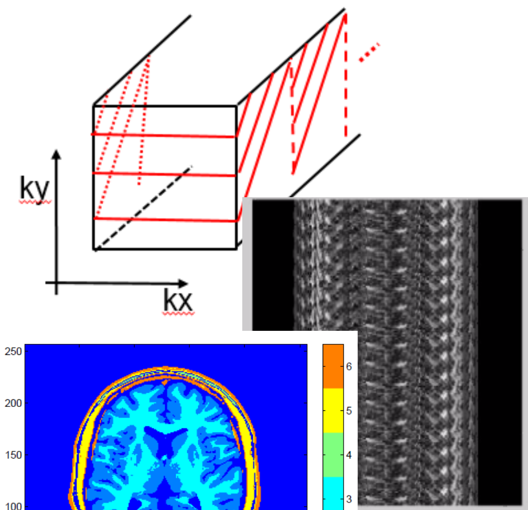

PFP-ANN IPG $\nu_{p^k} = r^k$

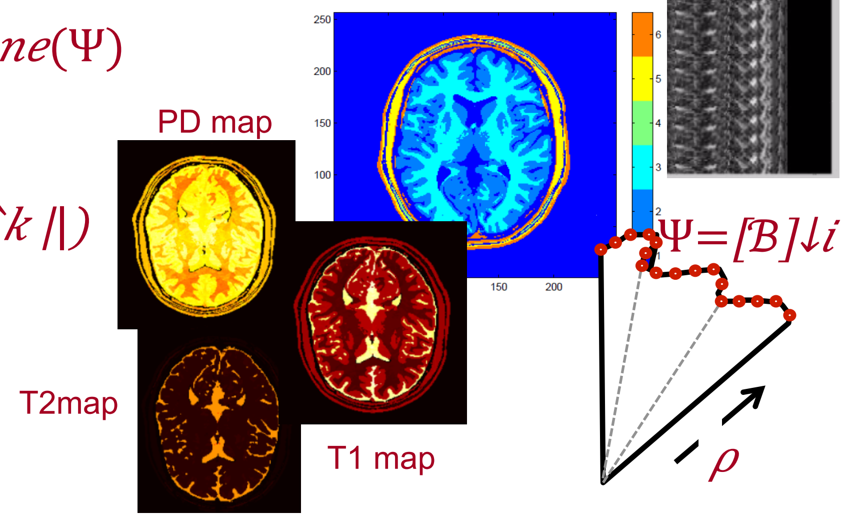
# Numerical experiments 2: MRF

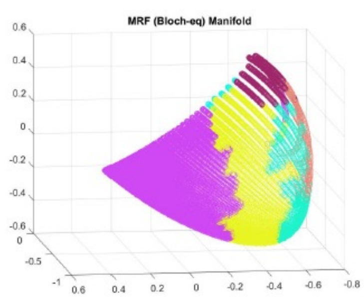
# MRF cone & EPI acquisition

- K-space subsampling: Eco Planar Imaging (EPI)
- Anatomical phantom {Grey/White matters, CSF, muscle, skin}
- Bloch eq. dictionary $\Psi \in \mathbb{C}^{512 \times \sim 50'000}$
- $\min_{X} \|y - A(X)\|^2 \quad s.t. \quad X_{vec} \in \prod_j \mathrm{cone}(\Psi)$

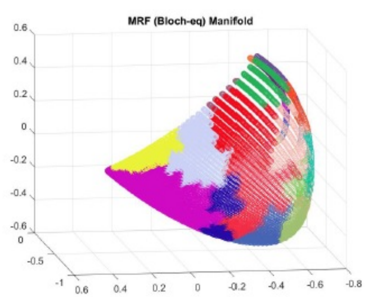
1. $a_p = X_p^{k-1} - \mu A^H (A(X_p^{k-1}) - y))$
2. $\psi_p = ANN_{normalised\{\Psi\}} (a_p^k / \|a_p^k\|)$
3. $\rho_p^k = a_p / \psi_p$
4. $X_p^k = \rho_p \psi_p, \ \forall p$



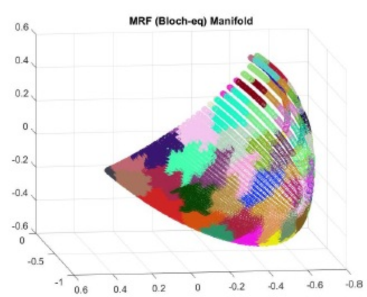



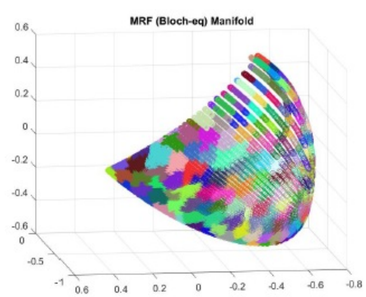
(a) Cover tree segments at scale 2



(b) Cover tree segments at scale 3



(c) Cover tree segments at scale 4

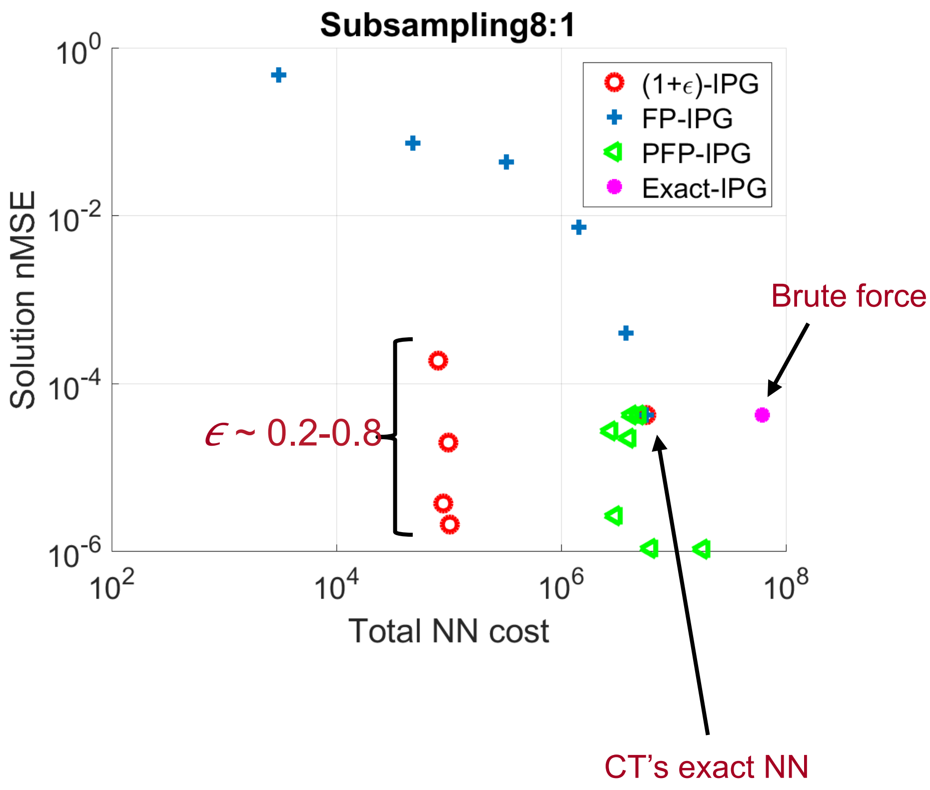


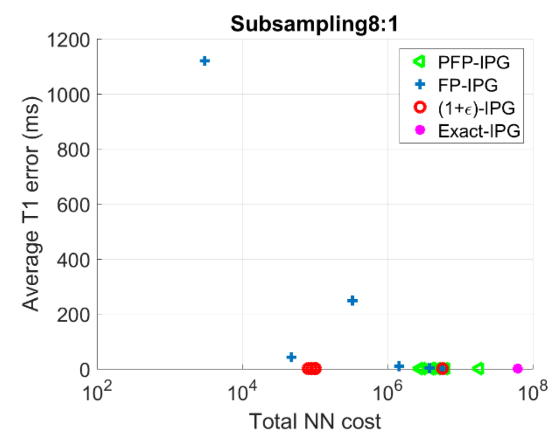
(d) Cover tree segments at scale 5

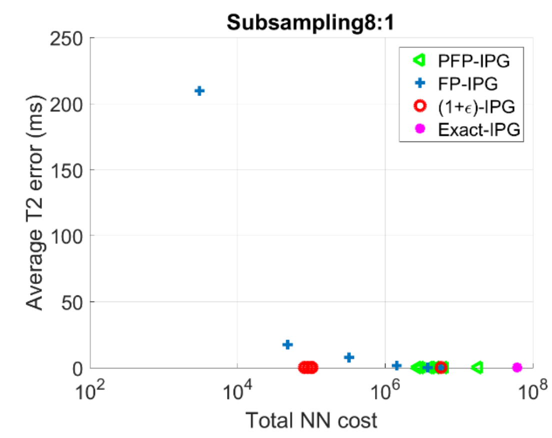
# Accuracy vs. computation

Dominant cost→ NN/ANN (since $A$ is FFT)

Projection cost = # matches calculated (i.e. visited nodes on the tree)



# Summary

- IPG robustness to inexact oracles (under embedding assumption)
- Linear convergence result:
  - PFP/$(1+\epsilon)$-oracles: same final accuracy vs. exact IPG
  - PFP: same convergence rate vs. exact IPG
  - $(1+\epsilon)$: stronger assumptions/sensitive to conditioning of *A*
- Implications in data driven CS (using ANN)
  - Cover trees for fast ANN: complexity ~ intrinsic dim(data)
- O(10e3) faster parameter estimation in MRF

*Thnx!*