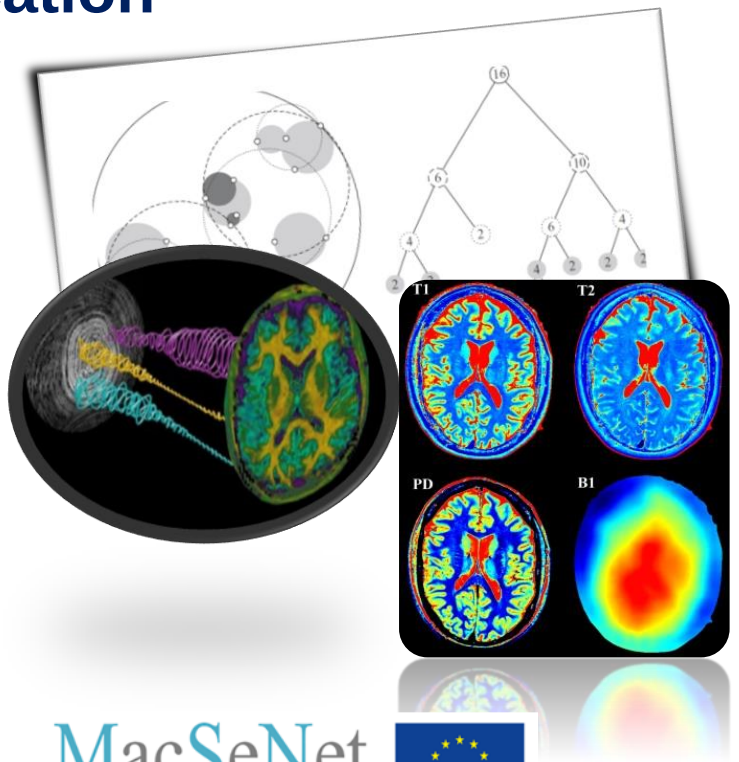
Inexact Gradient Projection and Fast Data Driven Compressed Sensing

Theory and Application

Mike Davies

& Mohammad Golbabaee



Joint work with

Arnold Benjamin, Dongdong Chen, Zaid Mahbub, Ian Marshall

Zhouye Chen, Yves Wiaux @ Heriot Watt University

Pedro Gomez, Carolin Pirkl, Marion Menzel @GE Healthcare

EPSRC Engineering and Physical Sciences Research Council

erc

MacSeNet Machine Sensing Training Network

Solving Compressed Sensing Inv. Problems

Estimation given $y = Ax + w$ by constrained LS

$$\hat{x} \in \operatorname{argmin}\left\{ f(x) := \frac{1}{2}\left|\left|y - Ax\right|\right|_2^2 \right\} \quad s.t. \;\; x \in C$$

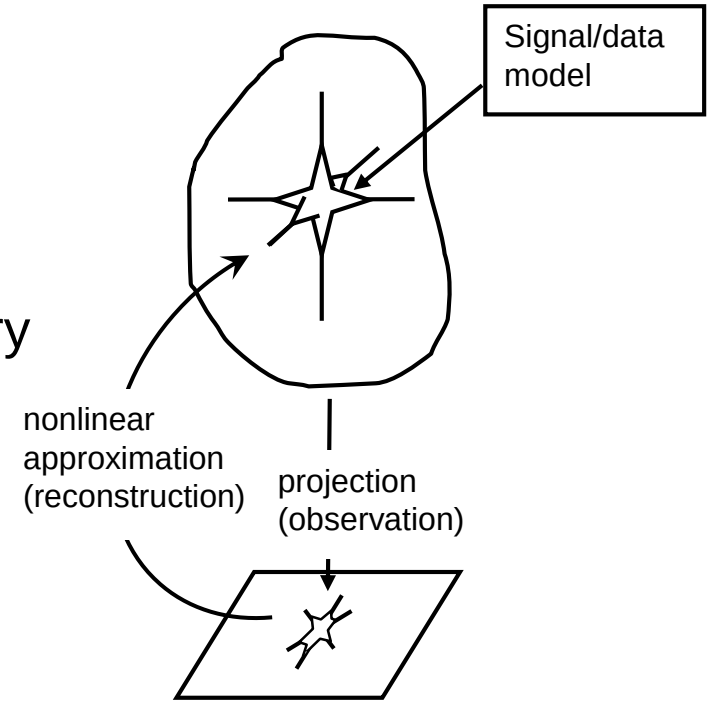
!! NP-hard for most interesting models

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

Iterative Gradient Projection (IPG)

Generally proximal-gradient algorithms are very popular:

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



Gradient $\mathrm{A}^{\mathrm{T}}(\mathrm{Ax}^{\mathrm{k-1}} - \mathrm{y}))$, step size μ,

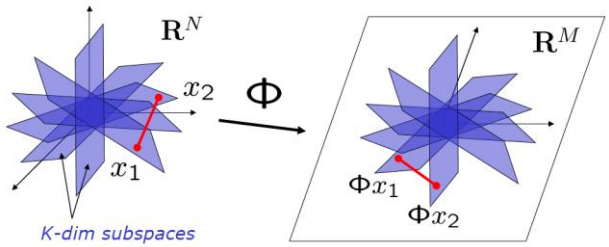
Euclidean projection

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

Embedding: key to CS/IPG stability

Bi-Lipschitz embeddable sets: $\forall x, x' \in C$

$$\alpha\left||x-x'|\right|_2 \le \left||A(x-x')|\right|_2 \le \beta\left||x-x'|\right|_2$$



Theorem ***[Blumensath'11]*** *For any* (C, A) *if holds* $\beta \le 1.5\alpha$, IPG → *stable & linear convergence:* $||x^K - x_0|| \to O(w) + \tau, \quad K \sim (\log \tau^{-1})$

Global optimality even for nonconvex programs!

Many embedding results exist, e.g. ***A~ i.i.d.*** **subGaussian**

Model $C \in R^n$	$O(m)$	
p (unstructured) points	$\theta^{-2}\log(p)$	***[Johnson, Lindenstrauss'89]***
d-dim affine subspaces	$\theta^{-2}d$	***[Sarlos '06]***
$\cup_i^L$ 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 et al.'09]***
'smooth' d dim. manifold	$\theta^{-2}d$	***[Wakin,Baraniuk'06; Clarkson 08]***

Challenges for IPG…

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

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

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

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

- $\boldsymbol{P}_C$ may not be analytic and requires solving an auxiliary optimization

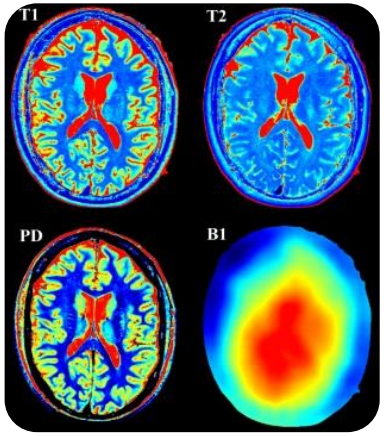
  (e.g. inclusions $C = \bigcap_i C_i$, total variation ball, low-rank, tree-sparse,…)
- $\boldsymbol{P}_C$ might be NP hard! (e.g. analysis sparsity, low-rank tensor decomposition)
- **$\boldsymbol{P}_C$ might be data driven – intrinsically approximate and requiring exhaustive testing against data model**

Is IPG robust against inexact/approximate oracles?

Motivating example:

Magnetic Resonance Fingerprinting

Quantitative MRI



Measuring the NMR properties (proton density, T1, T2) for generating anatomical (e.g. brain) maps

More information, contrast,.. than Qualitative MRI.

Relaxation times

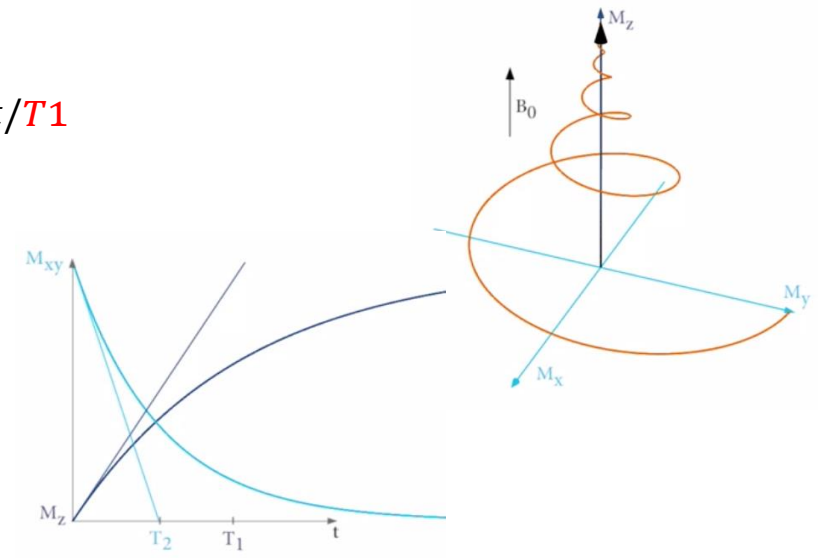
"…after excitation, nuclear spin magnetizations reach the equilibrium by two independent relaxation processes"

T1 (spin-lattice) longitudinal decay exponent

$$M_z(t) = M_{z,eq} - \left[M_{z,eq} - M_z(0)\right]e^{-t/T1}$$

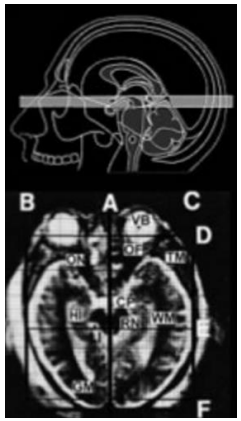
T2 (spin-spin) transverse decay exponent

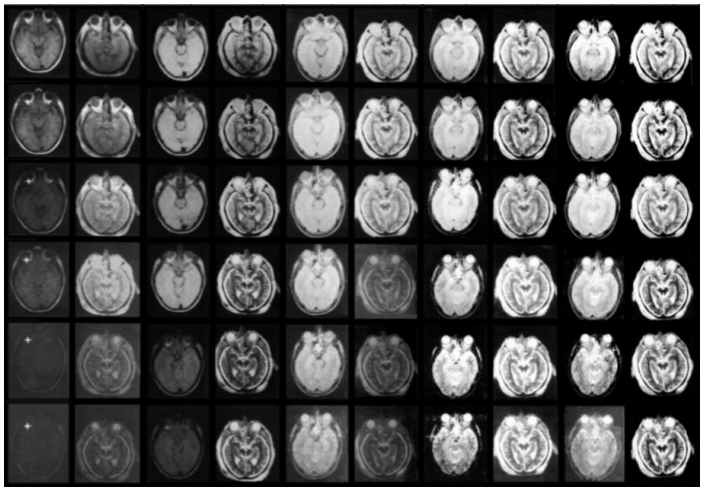
$$M_{x,y}(t) = M_{x,y}(0)e^{-t/T2}$$

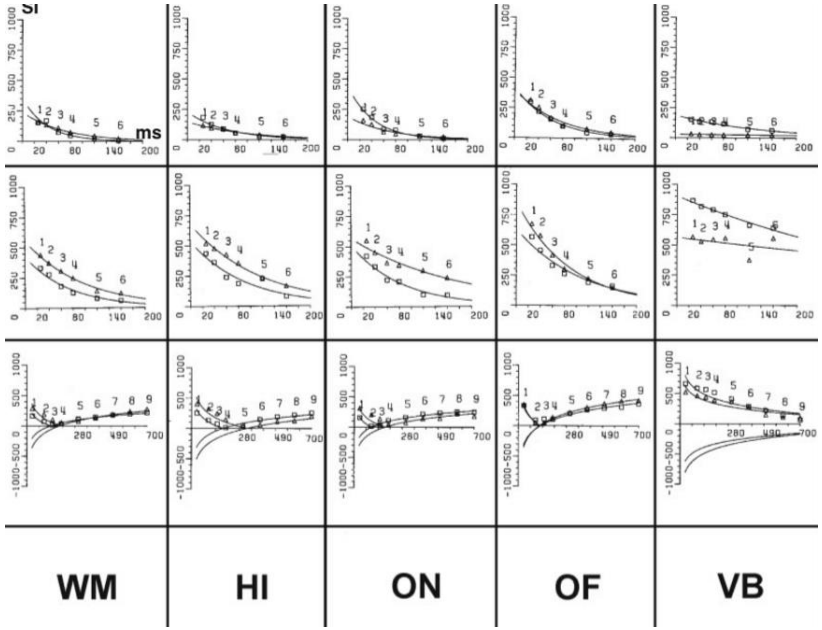


Standard approach

- Separate excitation sequences for T1 or T2
- Multiple readouts of the full k-space in different times (+ repetition)
- Equilibrium required before repetition (very long process ~ 30-45 mins)
- Curve fitting per voxel → T1,T2







WM: White matter, HI: Hippocampus, ON: Optical nerve,…

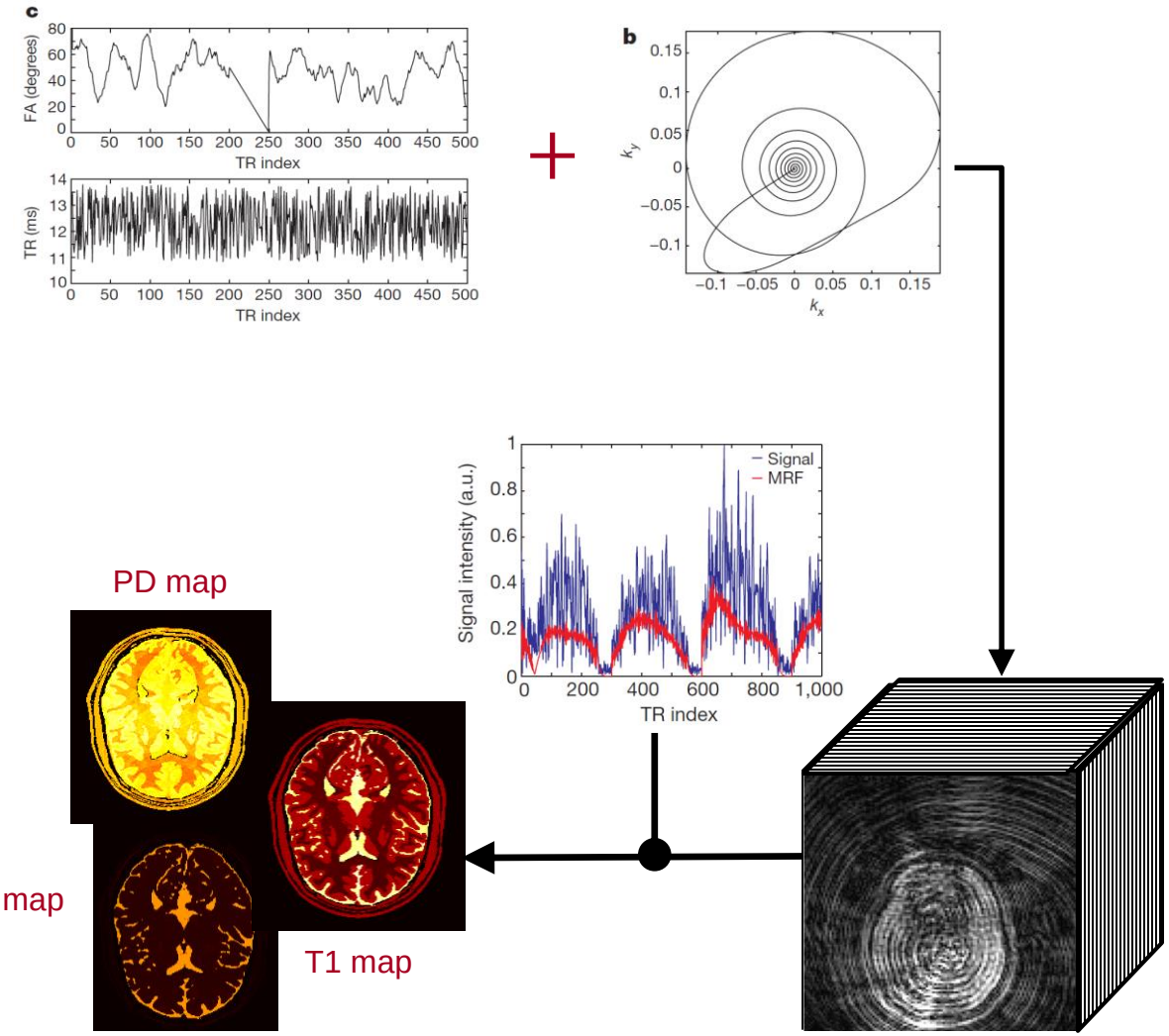
Magnetic Resonance Fingerprinting

Aim: simultaneous acquisition of all MR parameters at once

1. Continuous Excitation of magnetic spin with a rapid sequence of (random) RF pulses

2. Back project image sequence from very undersampled k-space (spiral trajectory)

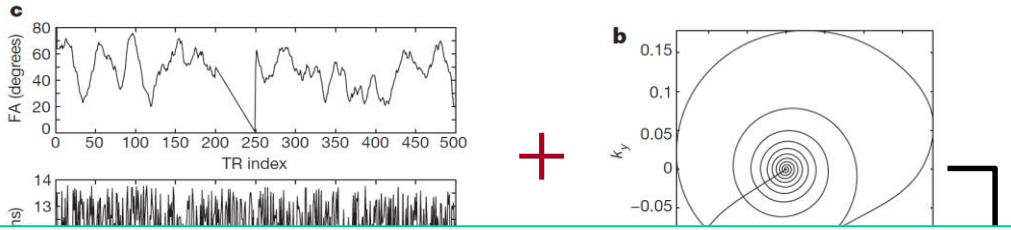
3. Use dictionary of matched filters (fingerprints) to estimate parameters per voxel sequence



Magnetic Resonance Fingerprinting

Aim: simultaneous acquisition of all MR parameters at once

1. Continuous Excitation of magnetic spin with a rapid sequence of (ra



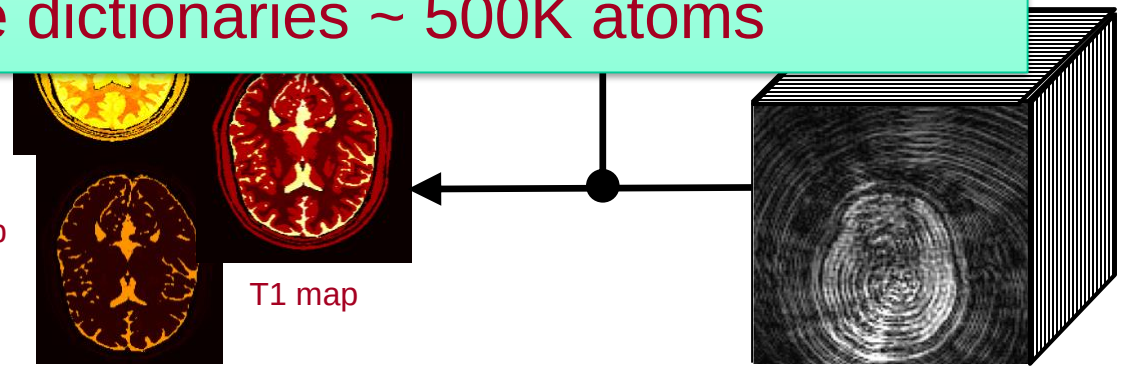
2. Ba
sec
un
(sp

Dictionary of matched filters calculated from Bloch equation response:

$$\frac{\partial\, \mathbf{m}(t)}{\partial\, t} = \mathbf{m}(t) \times \gamma\, \mathbf{B}(t) - \begin{pmatrix} \mathrm{m}^x(t)/\mathrm{T2} \\ m^y(t)/\mathrm{T2} \\ (m^z(t) - \mathrm{m_{eq}})/\mathrm{T1} \end{pmatrix}$$

Can result in large dictionaries ~ 500K atoms

3. Us
matched filters (fingerprints) to estimate parameters per voxel sequence



MRF is a product (tensor) space generalized CS

Multidimensional image data cube, $X \in \mathbb{C}^{P\times L}$, with a pixelwise data-driven model:

$$C = \cup_{i=1,\dots,d} \{\psi_i\}, \qquad \text{where } \{\psi_i\} = \Psi \in R^{d\times L}$$

Such that $X_{p,:} \in C, \ \ \forall p = 1, \dots, P$

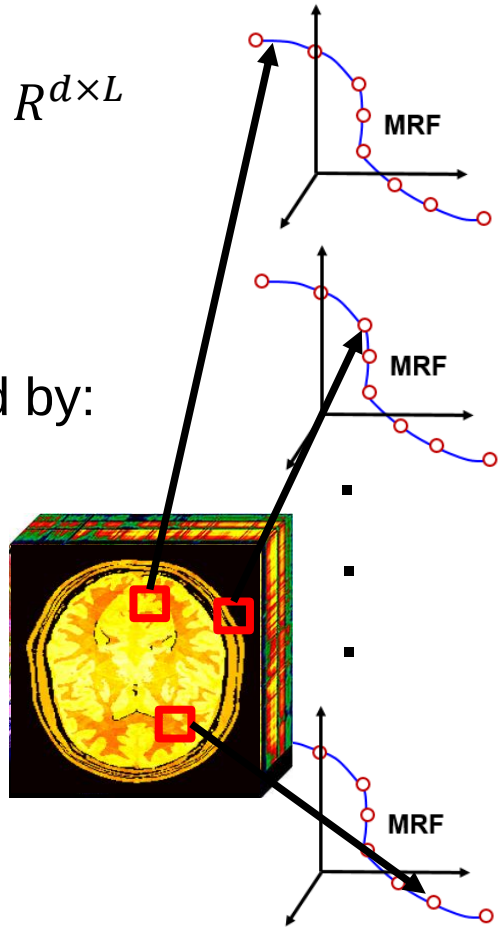
Sub (k-space) sampling each image: $Y = A(X)$ defined by:

$$Y_{:,l} = \mathrm{A}_{:,l}\mathrm{X}_{:,l} = \ S(l)F^H X_{:,l}$$

For a sequence of subsampling operators $\mathrm{S}(l)$

Inverse problem: $\underset{X\in\prod_p C(p)}{\operatorname{argmin}} \sum_l \left\| Y_{:,l} - S(l)F^H X_{:,l} \right\|^2$

Direct recovery complexity $\mathrm{O}(\mathrm{P}^{\mathrm{d}})$!



MRF via IPG

Replace direct solution by IPG step with *separable* Nearest Neighbour (NN_C) Search

IPG (I-NNsearch-G)

$$X_{p,:}^{k} = NN_C\left(\left[X^{k-1} - \mu A^H(A(X^{k-1}) - Y)\right]_{p,:}\right), \forall p$$

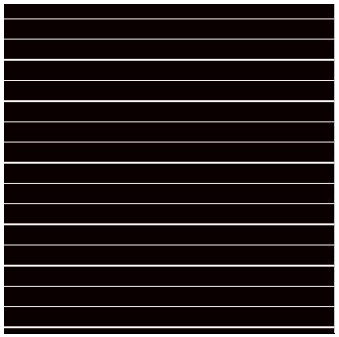
Complexity = iter $\times$ (gradient+$O(Pd)$)

(requires dictionary matching per pixel per iter.)

Sufficient RIP stability

Theorem ***[D. et al. 2014]***: random uniform subsampling of k-space each excitation (random Echo Planar Imaging) sufficient for RIP

(Spiral appears not sufficient)



Faster NN searches/Approximate searches

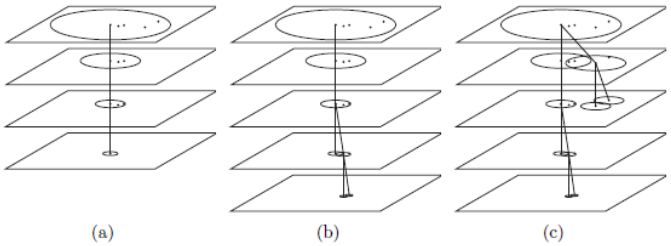
Trees: hierarchical partitioning + Brand & bound search. *e.g., kd-trees, Metric/Ball-trees,…*

Nagivating nets, Cover trees ***[Krauthgamer,Lee'04; Beygelzimmer'06]***

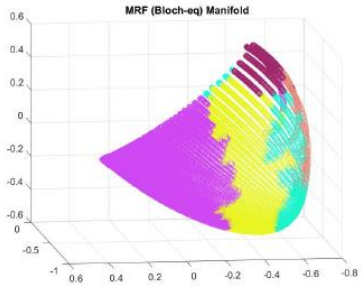
Cover tree builds a multi-resolution sequence of epsilon-nets over the data

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

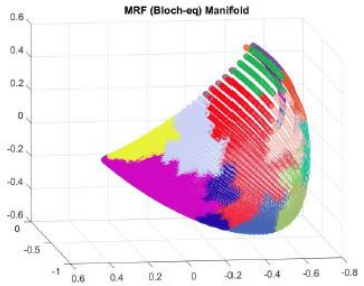
- Covering (parent nodes) $\sigma 2^{-l}$
- Separation (nodes appearing at scale l) $\sigma 2^{-l}$



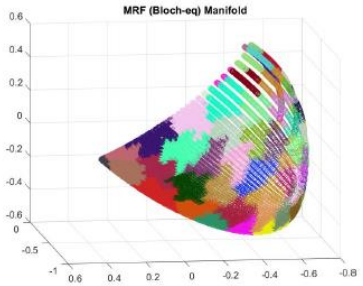
Can also be used for Approximate NN



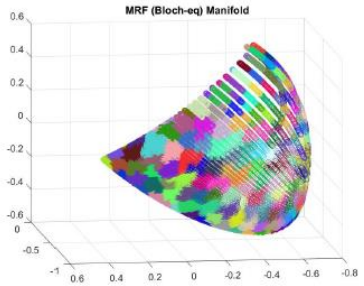
(a) Cover tree segments at scale 2



(b) Cover tree segments at scale 3



(c) Cover tree segments at scale 4



(d) Cover tree segments at scale 5

CT provably good ANN for 'low dim' data

Search options with cover trees

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

Typical $(1+\epsilon)$-ANN complexity: ***[Krauthgamer + Lee 04]***

$$2^{\mathcal{O}(\dim(C))}\log(d)+\epsilon^{-\mathcal{O}(\dim(C))} \textit{ in time, } O(d) \textit{ space}$$

2. Fixed Precision (FP)-*ANN:* truncated tree at level L + exact NN

(complexity could be arbitrary large/theoretically)

3. *Progressive (P)FP-ANN:* progressively increase truncation level

In each case leads to an **Inexact IPG**:

$$X^k_{p,:} = ANN_C\left(\left[X^{k-1}-\mu A^H(A(X^{k-1})-y)\right]_{p,:}\right), \forall p$$

Back to IPG….

Robustness & linear convergence of inexact IPG

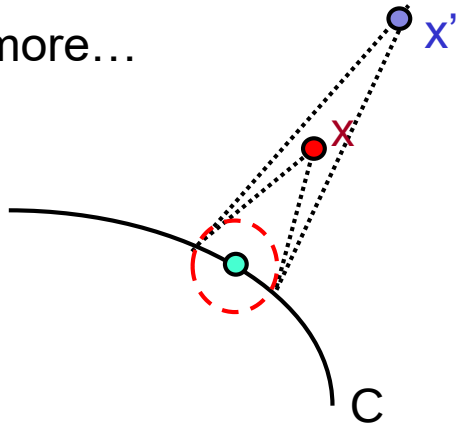
Inexact oracles I: Fixed Precision

$$x^k = \widetilde{\boldsymbol{P}}_C(x^{k-1} - \mu\, \tilde{\nabla} f(x^{k-1}))$$

Fixed Precision (FP) approximate oracles:

$$\left\|\tilde{\nabla} f(.) - \nabla f(.)\right\|_2 \leq \nu_g, \qquad \left\|\widetilde{\boldsymbol{P}_C}(.) - \boldsymbol{P}_C(.)\right\|_2 \leq \nu_p, \qquad (\ \widetilde{\boldsymbol{P}}_C(.) \in C\ )$$

Examples: TV ball, inclusions (e.g. Djkstra alg.), and many more…



Progressive Fixed Precision (PFP) oracles:

$$\left\|\tilde{\nabla} f(.) - \nabla f(.)\right\|_2 \leq \nu_g^k, \qquad \left\|\widetilde{\boldsymbol{P}}(.) - \boldsymbol{P}(.)\right\|_2 \leq \nu_p^k$$

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]***

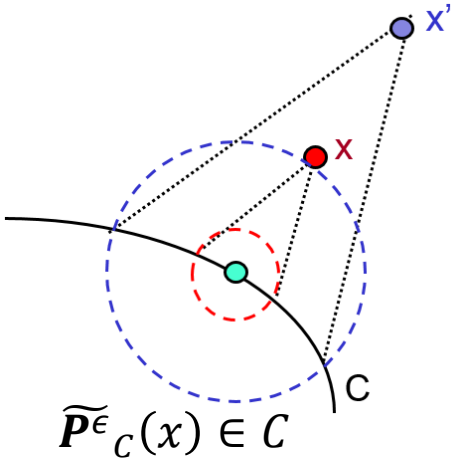
Inexact oracles II: $(1+\epsilon)$-optimal

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

$$x^k = \widetilde{\boldsymbol{P}^{\epsilon}}_C(x^{k-1} - \mu\, \widetilde{\nabla}_{\mathrm{k}}\, f(x^{k-1}))$$

Gradient $\left|\left|\tilde{\nabla}_k\, f(.) - \nabla f(.)\right|\right|_2 \leq \nu_g^k$

Projection $\left|\left|\widetilde{\boldsymbol{P}^{\epsilon}}_C(x) - x\right|\right|_2 \leq (\mathbf{1}+\boldsymbol{\epsilon})\left|\left|\boldsymbol{P}_C\,(x) - x\right|\right|_2$



Examples

Cheaper low-rank proxies based on randomized lin. algebra ***[Halko et al.'11]***, K-tree sparse signals ***[Hegde et al.'14]***, Tensor low-rank (Tucker) decomposition ***[Rauhut et al.'16]***,

IPG with (P)FP oracles

$$x^k = \widetilde{\boldsymbol{P}}_C(x^{k-1} - \mu\, \tilde{\nabla} f(x^{k-1}))$$

Theorem [Golbabaee & D.'18], *For any* $(x_0 \in C, C, A)$ *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, $\rho = \sqrt{1/\mu\alpha_0 - 1}$ *and* $e^i = \frac{2\nu_g^i}{\alpha_0} + \sqrt{\frac{\nu_p^i}{\mu\alpha_0}}$

Remark:

ρ^{-i} *supresses the early stages errors*

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

IPG with (P)FP oracles

$$x^k = \tilde{\boldsymbol{P}}_C(x^{k-1} - \mu\, \tilde{\nabla} f(x^{k-1}))$$

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

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

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*

$$\left|\left|x^K - x_0\right|\right| \leq O(w) + \tau$$

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

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

$$x^k = \widetilde{\boldsymbol{P}}_C^{\epsilon}(x^{k-1} - \mu\, \widetilde{\nabla}_{\mathrm{k}}\, f(x^{k-1}))$$

Theorem [Golbabaee & D.'18], *Assume for any* $(x_0 \in C, C, A)$ *and an* $\epsilon \geq 0$ *it holds*

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

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

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

Remarks.

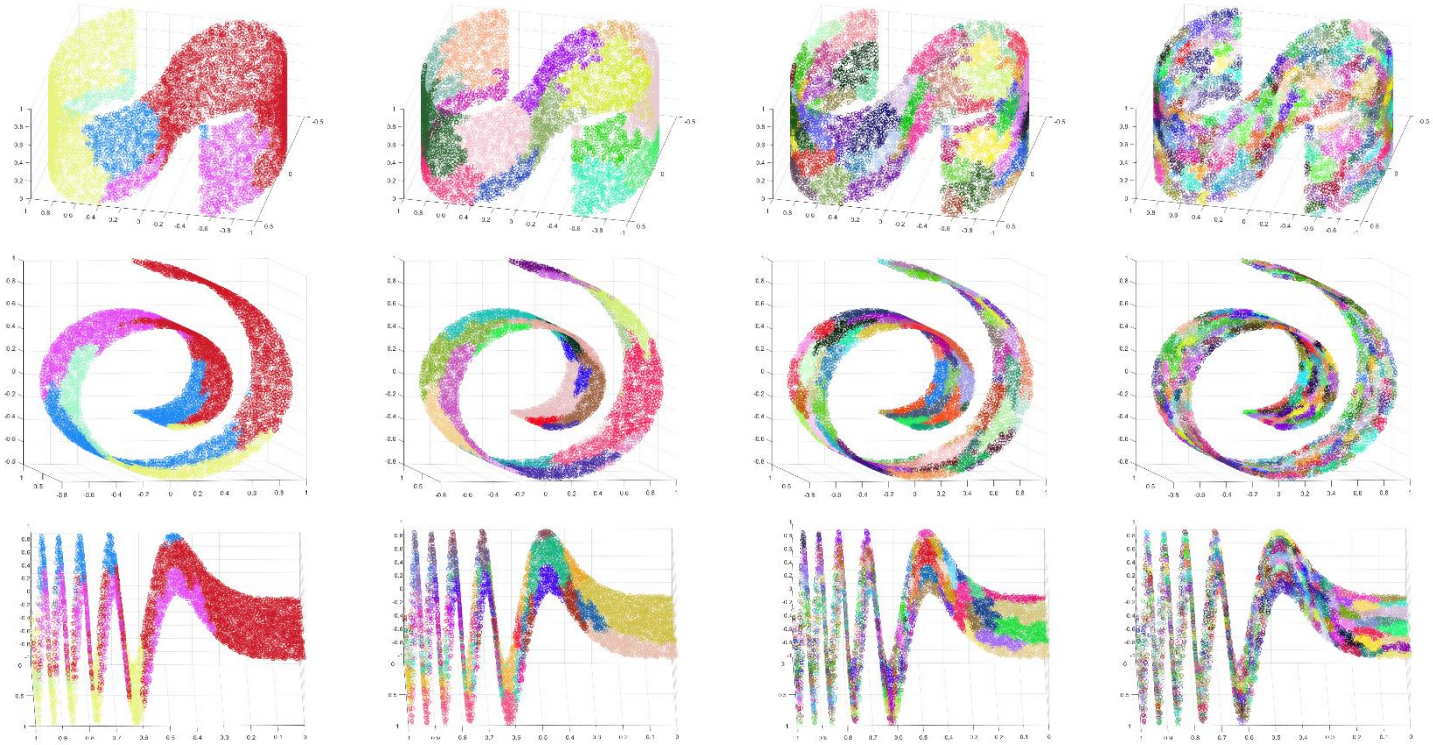
- *Requires stronger embedding cond., slower convergence!*
- *Still linear conv. &* $O(w) + \tau$ *accuracy after* $O(\log \tau^{-1})$ *iterations*
- *higher noise amplification*

Numerical experiments 1: *Toy problem*

2D manifold data

$C = \cup_{i=1,\ldots,d} \{\psi_i\}$ ψ_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

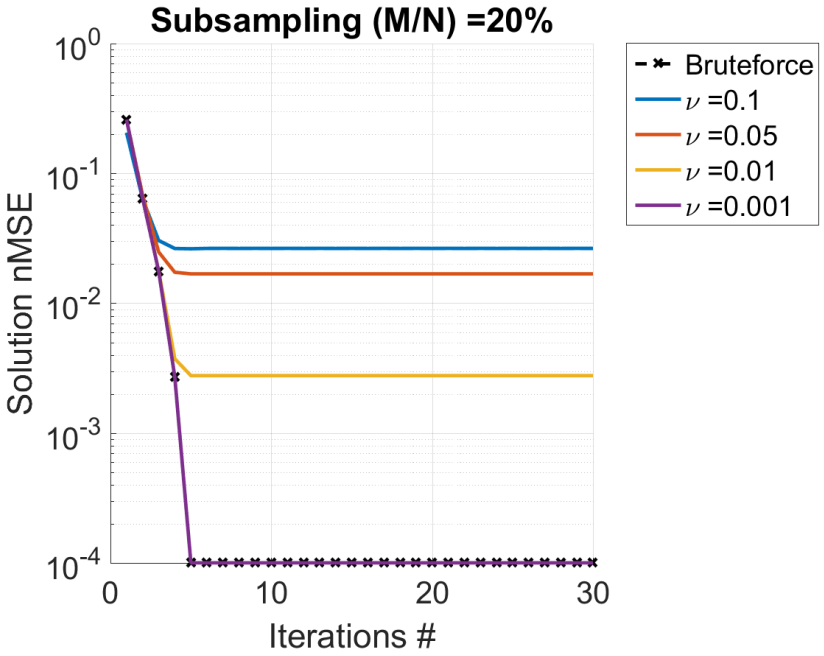
Solution accuracy vs. Iterations (FP)

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

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

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

Swiss roll



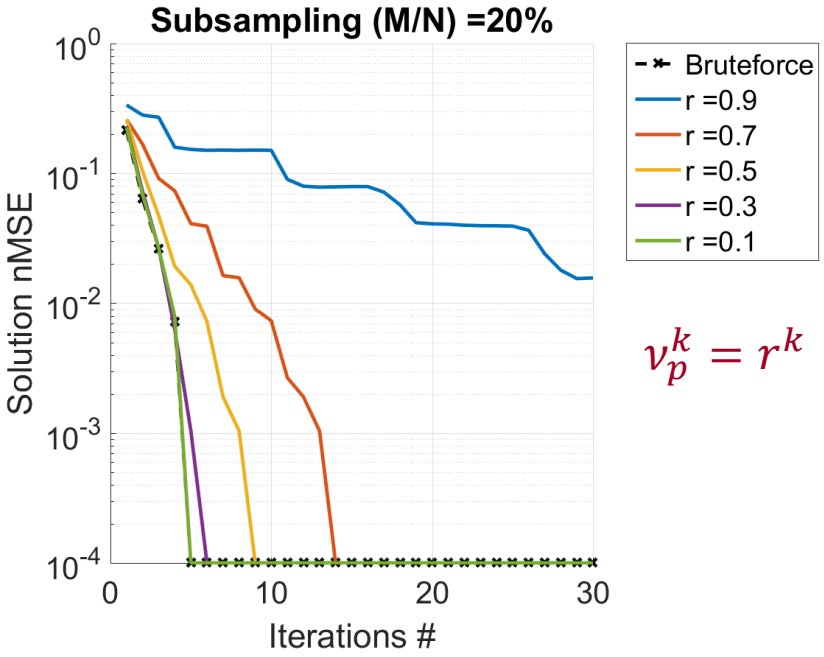
Solution accuracy vs. Iterations (PFP)

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

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

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

Swiss roll



$$v_p^k = r^k$$

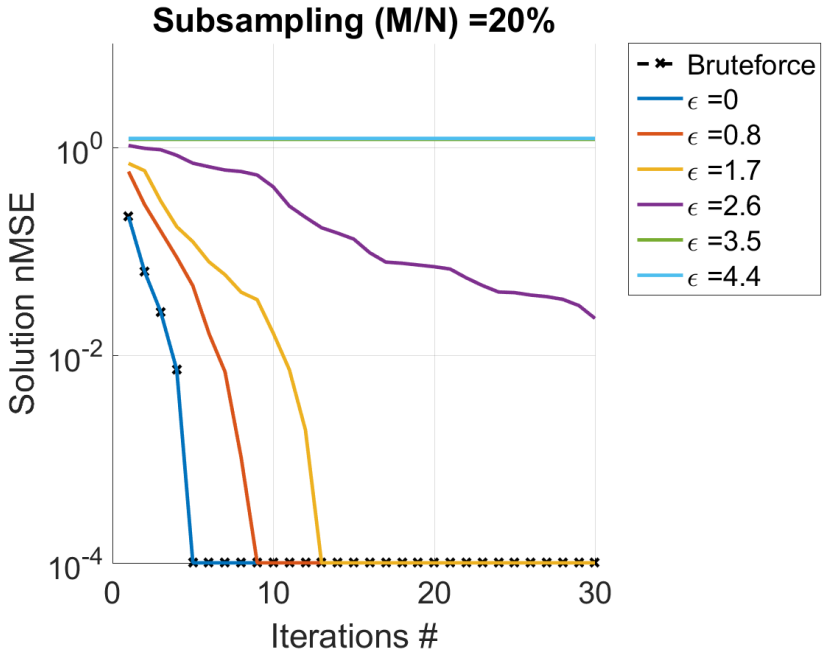
Solution accuracy vs. Iterations $(1+\epsilon)$-ANN

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

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

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

Swiss roll

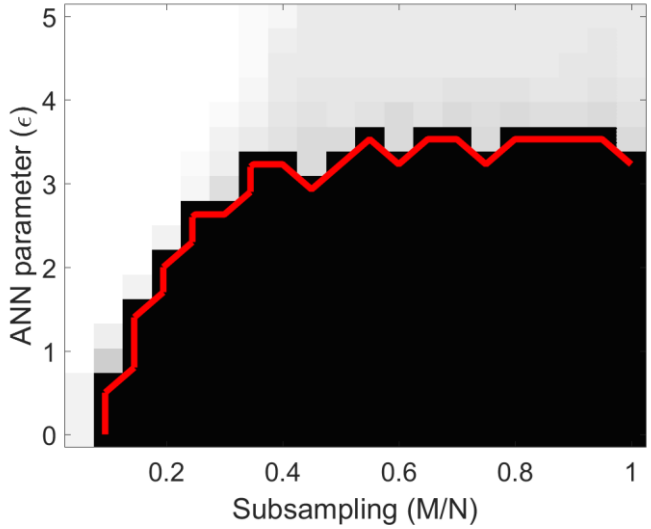


Phase transitions

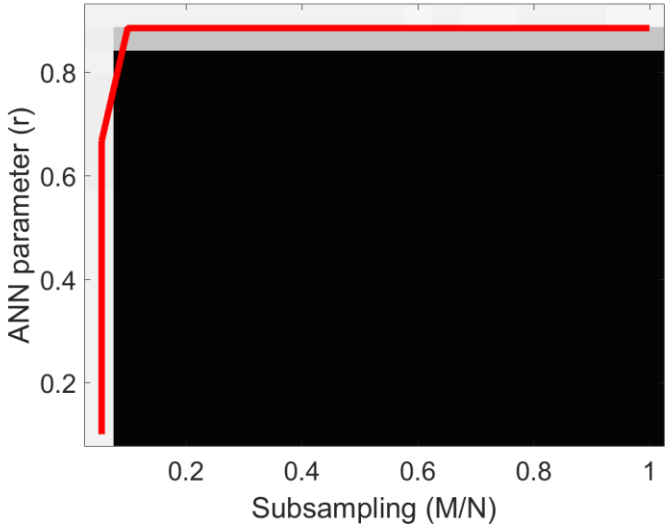
Signal: $X \in \boldsymbol{R}^{n \times P}$ $n = 200$, $P = 50$ (randomly chosen $\in \mathrm{C}$)

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

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



$(1+\epsilon)$-ANN IPG

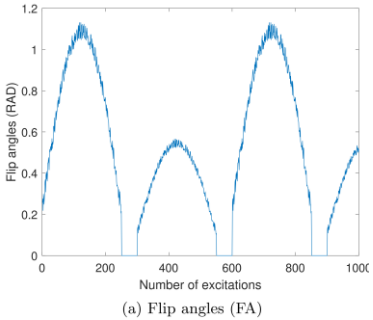


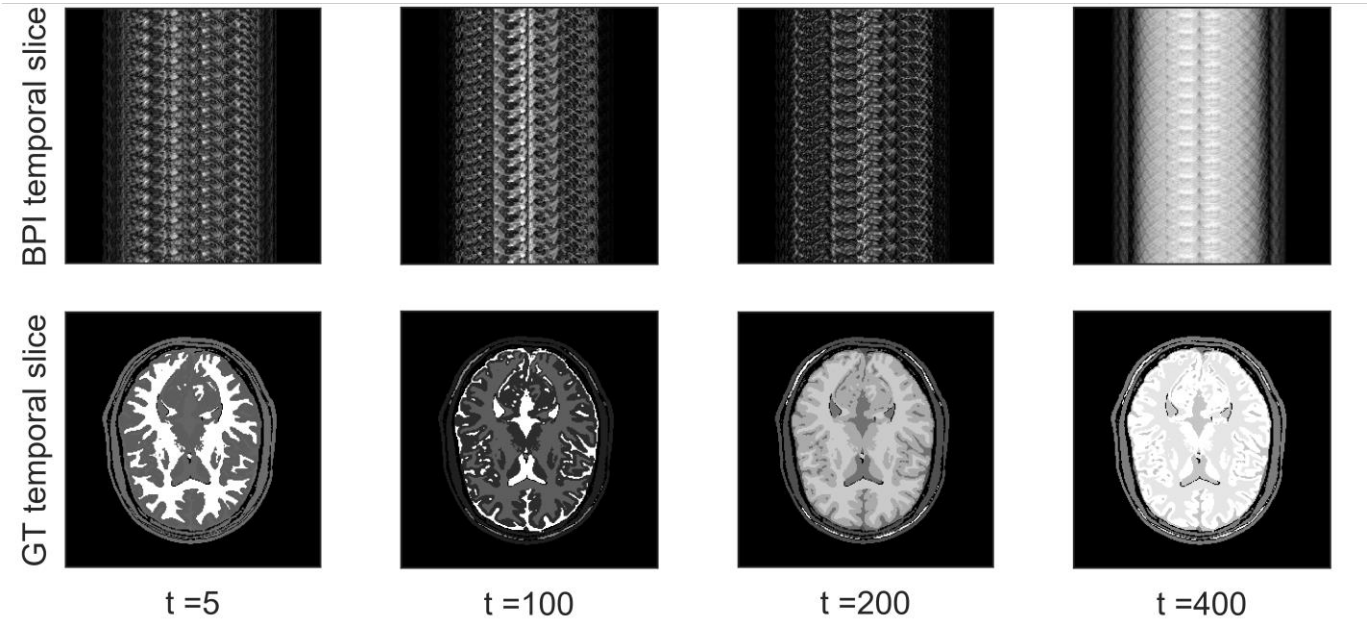
PFP-ANN IPG $v_p^k = r^k$

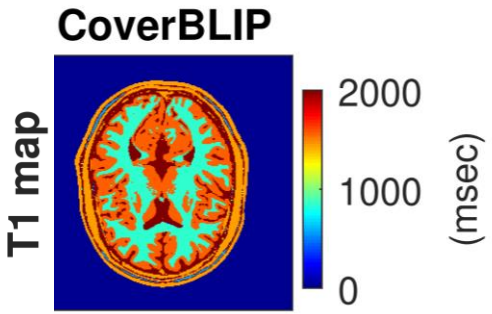
Numerical experiments 2: *MRF*

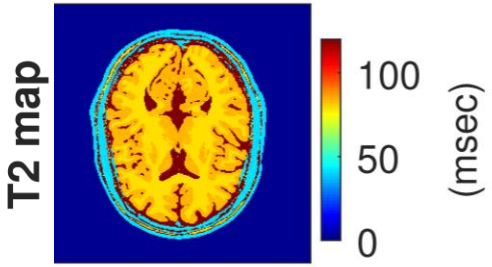
MRF set up

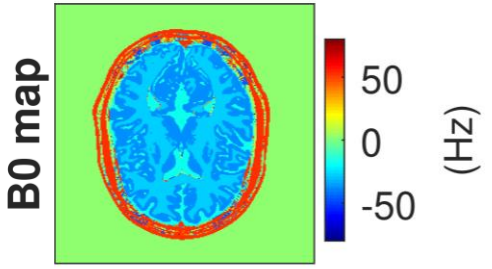
Multi-shot EPI k-space sampling: 16x undersampling per image with Pseudo-random FA excitation ***[Ma et al. '13]***







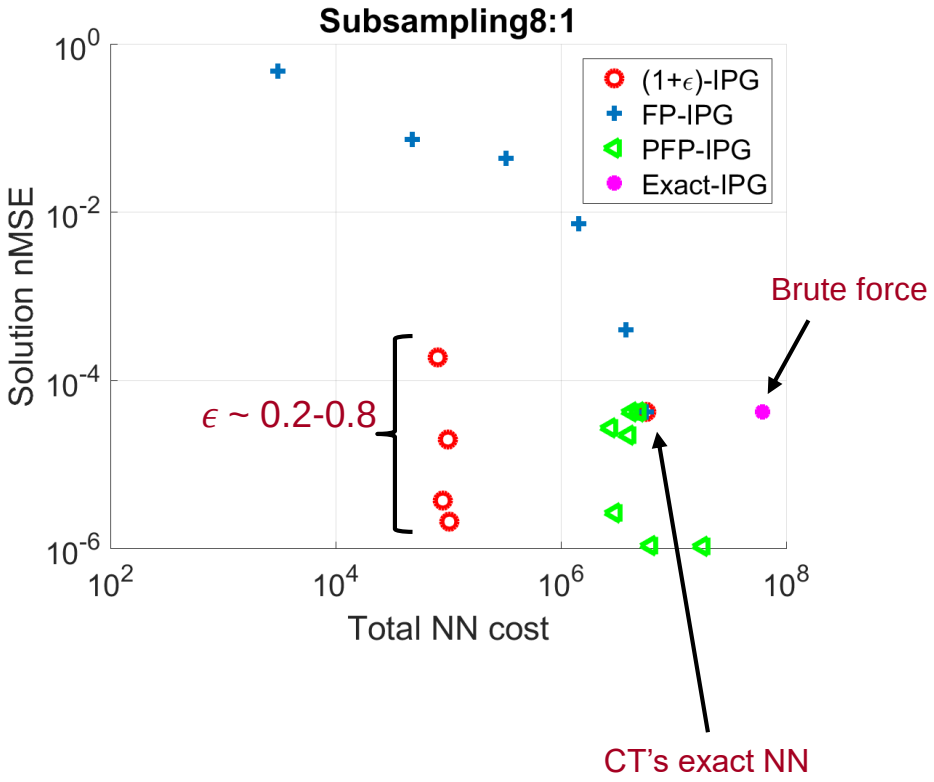


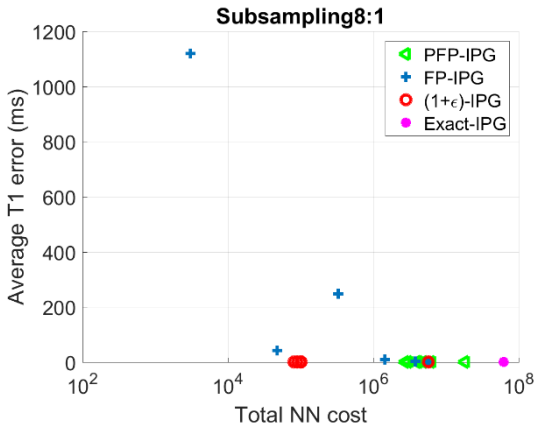


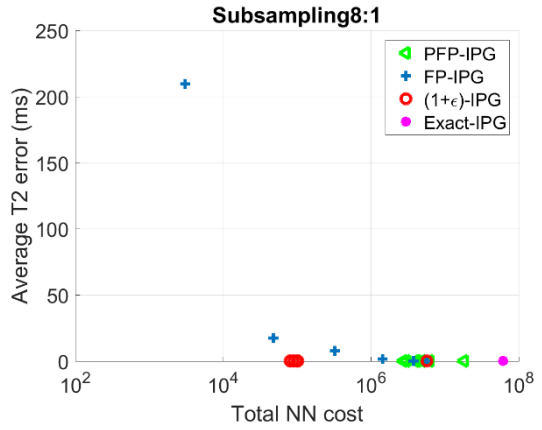
Accuracy vs. computation

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

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



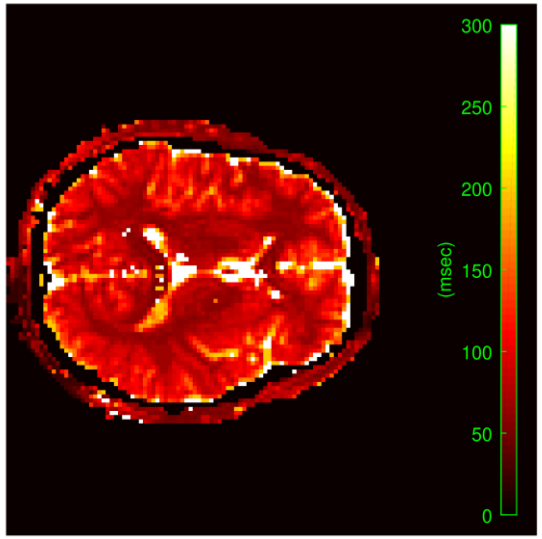




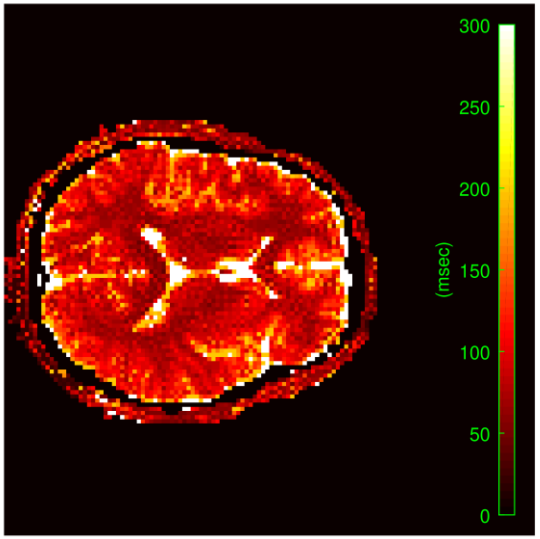
Some final thoughts…

Spiral MRF: a RIP/null space failure?

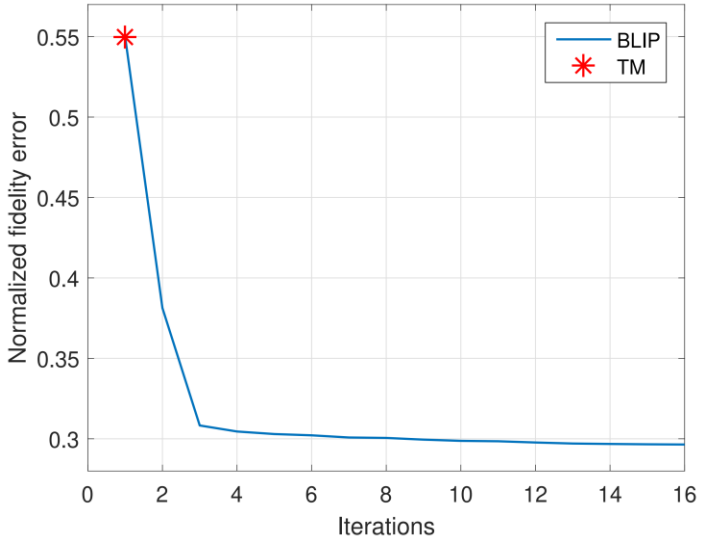
Spiral subsampling fails to fully sample the outer reaches of k-space.



(a) T2 map reconstructed by TM



(b) T2 map reconstructed by IPG



(c) Fidelity error for IPG through iterations

High frequency artefacts appear in iterated reconstruction (although data fidelity still decreases) ***[Cline et al '17; Golbabaee et al '18]***

- RIP/null space failure?

MRF-NET *[Golbabaee, Chen, Gomez, Menzel, D. '18]*

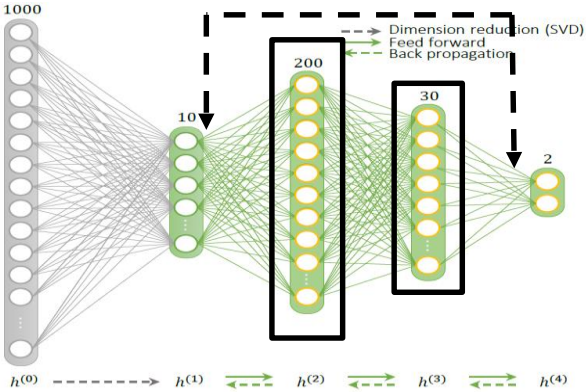
Another way to approximate the manifold projection would be to train a DNN…

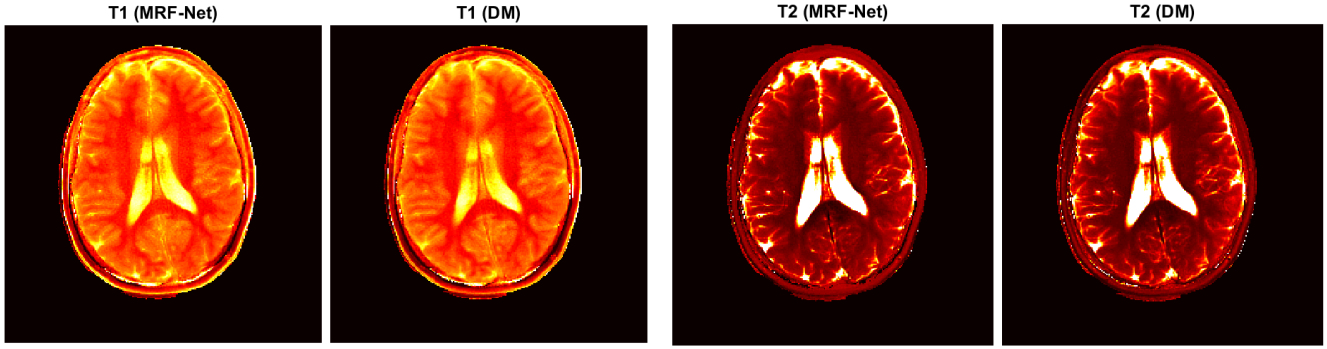
- **Reducing computation and storage**

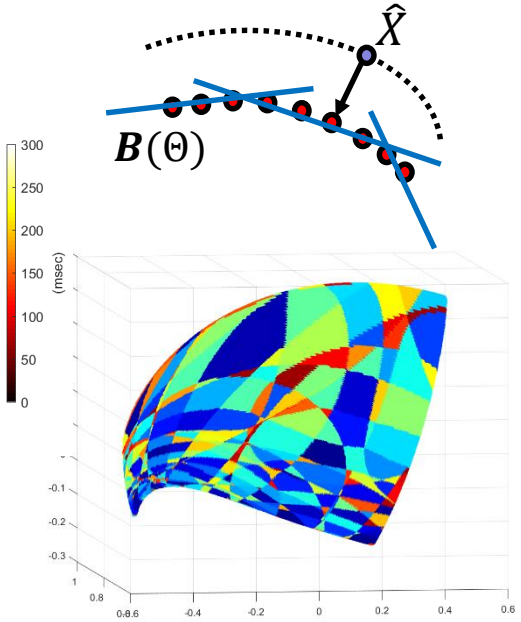
  TM is pointwise but MRF-NET is piecewise affine approximation to Bloch manifold projection (60x less storage & computation cf. subspace TM)

- **How efficient?**

  Trade-offs between # units (width/depth), approximation level & manifold conditioning

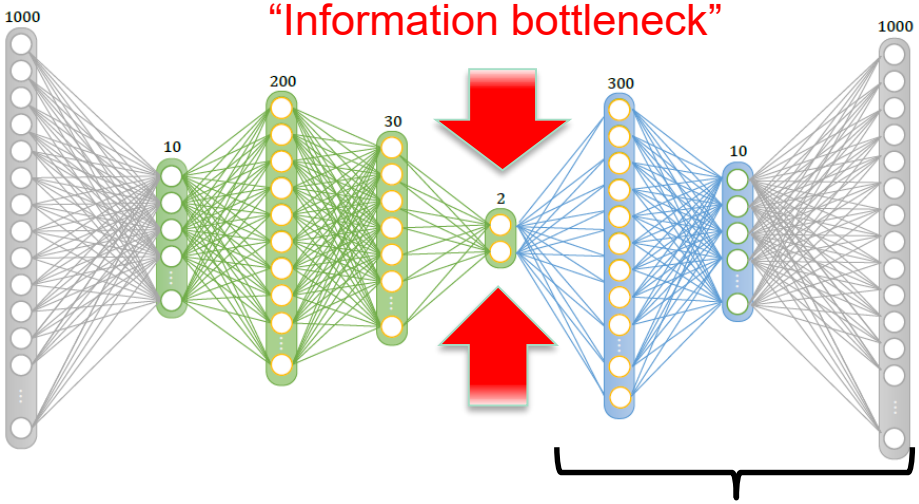






Can also be used iteratively…

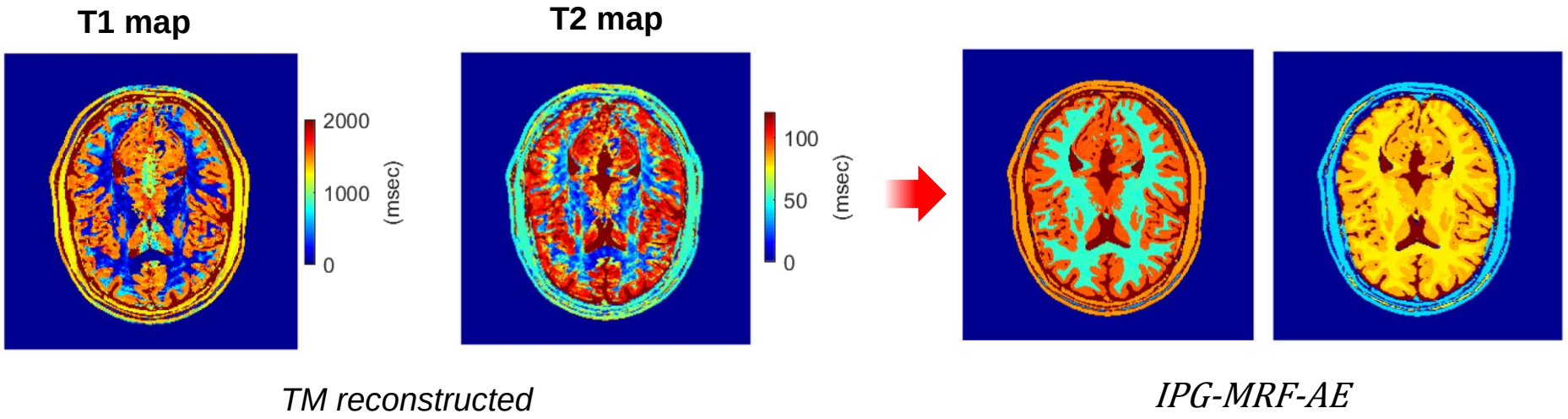
Iterative extensions: MRF-AutoEncoder



deep iterations:

$$\widehat{X^{k+1}} \leftarrow \text{MRF-AE}(\widehat{X^k} - \mu F_\omega^H(F_\omega(\hat{X}^k) - Y))$$

Promotes low-dim (manifold) temporal structures



TM reconstructed

IPG-MRF-AE

(Phantom data, 1-coil, EPI sampling low data regime)

Summary

- IPG is robust 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*
- $\mathcal{O}(10^3)$ faster parameter estimation in MRF
- Can (empirically) accommodate other approximate projection operators, e.g. DNNs

References

- M. Golbabaee, D. Chen, P. Gomez, M. Menzel, M. E. Davies, 2018 Geometry of deep learning for Magnetic Resonance Fingerprinting, http://arxiv.org/abs/1809.01749.
- M. Golbabaee, Z. Chen, Y. Wiaux and M. E. Davies, 2018, CoverBlip: an algorithm for fast Magnetic Resonance Fingerprint recovery. Preprint arXiv:1810.01967.
- M. Golbabaee, Z. Chen, Y. Wiaux, M. Davies. CoverBLIP: scalable iterative matched-filtering for MR Fingerprint recovery. ISMRM, 2018.
- M. Golbabaee and M. E. Davies, 2018, Inexact gradient projection and data-driven compressed sensing. IEEE Trans Inf Th., vol 64(10), pp. 6707-6721, ArXiv:1706.00092.
- A. Benjamin, P. A. Gómez, M. Golbabaee, T. Sprenger, M. I. Menzel, M. E. Davies, I. Marshall, 2018, Balanced multi-shot EPI for accelerated Cartesian MRF: An alternative to spiral MRF. Extended version, Submitted.
- A. J. V. Benjamin, P. Gomez, M. Golbabaee, Z. Mahbub, T. Sprenger, M. I. Menzel, M. Davies, I. Marshall, Balanced multi-shot EPI for accelerated Cartesian MR Fingerprinting: an alternative to spiral MR Fingerprinting. ISMRM, 2018.
- M. E. Davies, G. Puy, P. Vandergheynst, Y. Wiaux, 2014, A compressed sensing framework for magnetic resonance fingerprinting. SIAM J. Imaging Sci., 7(4), 2623–2656.