Recipes for Hard Thresholding Methods

A. Kyrillidis[1] and V. Cevher[1,2]

[1]Laboratory for Information and Inference Systems (EPFL) [2]Idiap Research Institute
{anastasios.kyrillidis,volkan.cevher}@epfl.ch

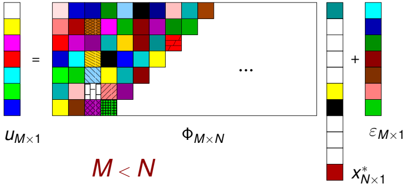
http://lions.epfl.ch/

Grant Acknowledgements:
KEEP project through DARPA KeCoM, Marie Curie reintegration.
VC also acknowledges his Faculty Fellowship position at Rice University.

Problem Statement

- **Underdetermined linear regression**:

$$\boldsymbol{y} = \Phi \mathbf{X}^* + \varepsilon,$$



Some applications

1. Compressed Sensing
2. Medical Imaging
3. Computational Biology
4. Communications

- **Goal**: find signal $\mathbf{X}^*$ that generated the set of measurements.
- $M < N \rightarrow$ Non-trivial nullspace of $\Phi \rightarrow$ ill-posed problem $\rightarrow$ more information needed:

$$x^* \text{ is } K\text{-sparse signal},$$

where $K < M$.

Restricted Isometry Property

- Sparsity: not enough by itself.
- Conditions on $\boldsymbol{\mathcal{A}}$: stable embedding, null space property, spark, unique representation property, exact recovery condition, etc.
- In this work $\rightarrow$ stable embedding.
- Let Σ_K^N: union-of-subspaces with at most K-nonzero entries in N-dimensions.

Restricted Isometry Property (RIP)

Let $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{M \times N}$ and $K < N$ be an integer number. Then, $\boldsymbol{\mathcal{A}}$ satisfies the restricted isometry property with constant δ_K iff

$$(1 - \delta_K) \leq \frac{\|\boldsymbol{\mathcal{A}}\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2} \leq (1 + \delta_K),$$

is satisfied for any $x \in \Sigma_K^N$.

- Real data example - $\Phi_{300 \times 1000}$ i.i.d. $\sim \mathcal{N}(0, 1/M)$: $\frac{\|\boldsymbol{\mathcal{A}}\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2} \in [\sim 0.5, \sim 2.5] \Rightarrow$

  Non-Symmetric

Restricted Isometry Property

- Sparsity: not enough by itself.
- Conditions on $\boldsymbol{\mathcal{A}}$: stable embedding, null space property, spark, unique representation property, exact recovery condition, etc.
- In this work $\rightarrow$ stable embedding.
- Let Σ_K^N: union-of-subspaces with at most K-nonzero entries in N-dimensions.

Non-Symmetric RIP

Let $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{M \times N}$ and α_K, β_K be two positive numbers. Then, $\boldsymbol{\mathcal{A}}$ satisfies the non-symmetric restricted isometry property with constants α_K, β_K iff

$$\alpha_K \leq \frac{\|\boldsymbol{\mathcal{A}}\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2} \leq \beta_K,$$

is satisfied for any $x \in \Sigma_K^N$.

Hard Thresholding Methods

l_0-"norm" minimization	l_1-norm minimization
$\hat{x} = \arg\min_{x : x \in \Sigma_K^N} f(x)$	$\hat{x} = \arg\min_{x : \|x\|_1 \leq \lambda} f(x)$

$\longleftarrow f(x) \triangleq \|\boldsymbol{y} - \boldsymbol{\mathcal{A}}\mathbf{X}\|_2^2$

- Both approaches are computationally attractive $\rightarrow$ l_0-"norm" minimization allows the use of model-CS framework (e.g., incorporate structured sparsity models) [BaraniukCevherDuarteHedge10].
- We focus on: *Iterative Hard Thresholding (IHT)* algorithm [NowakFigueiredo98, KingsburyReeves03, DaubechiesDefriseDeMol04; BlumensathDavies08, ...]
- Let $H_K(y) = \arg\min_{x : x \in \Sigma_K^N} \|x - y\|_2^2$, $\mathcal{X}_{(\cdot)} = \text{supp}(x_{(\cdot)})$ and $\nabla f(x) = -2\boldsymbol{\mathcal{A}}^*(\boldsymbol{y} - \boldsymbol{\mathcal{A}}\mathbf{X})$.

Algorithm 1: IHT Algorithm

Input: $\boldsymbol{y}$, $\boldsymbol{\mathcal{A}}$, K, Tolerance, MaxIterations
Initialize: $\mathbf{X}_0 \leftarrow \mathbf{X}_{\text{init}}$, $\mathcal{X}_0 \leftarrow \mathcal{X}_{\text{init}}$, $i \leftarrow 0$
repeat
  $b \leftarrow x_i - \frac{\mu_i}{2} \nabla f(x_i)$ *(Update current estimate)*
  $x_{i+1} \leftarrow H_K(b)$ *(Best K-term approximation)*
  $i \leftarrow i + 1$
until $\|\mathbf{X}_i - \mathbf{X}_{i-1}\|_2 \leq \text{Tolerance}\|\mathbf{X}_i\|_2$ or MaxIterations.

Convergence Guarantees - prior work

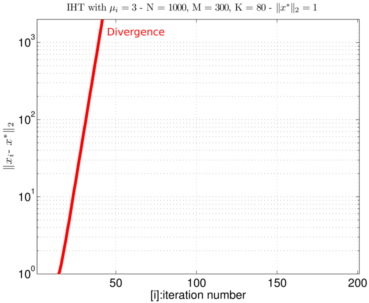
- General convergence formula:

$$\|x_{i+1} - x^*\|_2 \leq \rho\|x_i - x^*\|_2 + \gamma\|\varepsilon\|_2,$$

where

1. $\rho \rightarrow$ convergence rate,
2. $\gamma \rightarrow$ approximation guarantee.

Reference	Convergence Guarantee ($\rho < 1$)	Assumptions
[BlumensathDavies09]	$\delta_{3K} < 1/\sqrt{8}$	$\|\mathcal{A}\|_{2\rightarrow 2}^2 < 1$, $\mu_i = 1, \forall i$
[Foucart10]	$\delta_{3K} < 1/2$ or $\delta_{2K} < 1/4$	$\|\mathcal{A}\|_{2\rightarrow 2}^2 < 1$, $\mu_i = 1, \forall i$



- Trade-off: the faster you converge, the worse approximation guarantee you get.
- Majority of prior work $\rightarrow$ optimizing convergence speed, not approximation guarantee.

Convergence Guarantees - prior work

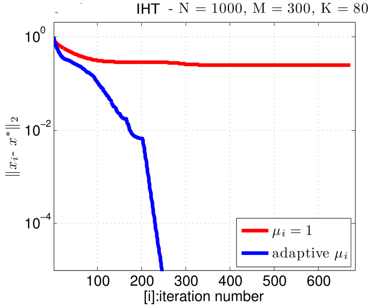
- General convergence formula:

$$\|x_{i+1} - x^*\|_2 \leq \rho\|x_i - x^*\|_2 + \gamma\|\varepsilon\|_2,$$

where

1. $\rho \rightarrow$ convergence rate,
2. $\gamma \rightarrow$ approximation guarantee.

Reference	Convergence Guarantee ($\rho < 1$)	Assumptions
[BlumensathDavies09]	$\delta_{3K} < 1/\sqrt{8}$	$\|\boldsymbol{\mathcal{A}}\|_{2\rightarrow 2}^2 < 1$, $\mu_i = 1, \forall i$
[Foucart10]	$\delta_{3K} < 1/2$ or $\delta_{2K} < 1/4$	$\|\boldsymbol{\mathcal{A}}\|_{2\rightarrow 2}^2 < 1$, $\mu_i = 1, \forall i$



- Trade-off: the faster you converge, the worse approximation guarantee you get.
- Majority of prior work $\rightarrow$ optimizing convergence speed, not approximation guarantee.

In this work...

We present:

- Basic “ingredients” of Hard Thresholding methods.
- Optimal/efficient strategies under various problem assumptions.
- General IHT template (ALgebraic PursuitS, dubbed as ALPS) that “mixes” the ingredients into one framework.

Step Size Selection

Step Size Selection

$$x_{i+1} = H_K\Big(\underbrace{x_i - \frac{\mu_i}{2}\nabla f(x_i)}_{=b}\Big)$$

- Key observation:

  "x_{i+1} is the *best* K-sparse approximation to b"

- Convergence proof of IHT:

$$\|x_{i+1} - b\|_2^2 \leq \|x^* - b\|_2^2 \Rightarrow$$

$$\|x_{i+1} - x^*\|_2 \leq 2\|\mathbb{I} - \mu_i \boldsymbol{\mathcal{A}}_T^* \boldsymbol{\mathcal{A}}_T\|_{2\to 2}\|x_i - x^*\|_2 + 2\mu_i\sqrt{1 + \delta_{2K}}\|\boldsymbol{\varepsilon}\|_2 \qquad (*)$$

  where $T = \text{supp}(x^*) \cup \text{supp}(x_{i+1}) \cup \text{supp}(x_i)$.

- Fundamental operator in IHT convergence rate proofs:

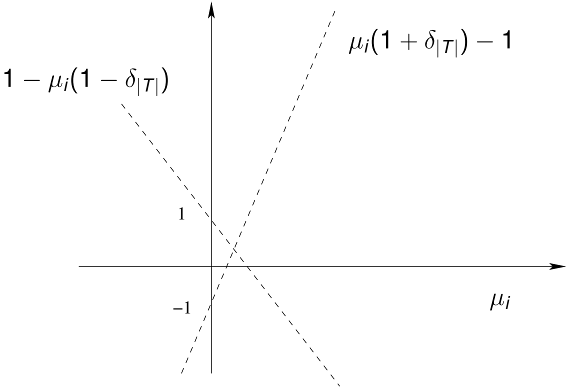
$$\|\mathbb{I} - \mu_i \boldsymbol{\mathcal{A}}_T^* \boldsymbol{\mathcal{A}}_T\|_{2\to 2} \leq \max\left\{\mu_i \lambda_{\max}(\boldsymbol{\mathcal{A}}_T^* \boldsymbol{\mathcal{A}}_T) - 1,\ \ 1 - \mu_i \lambda_{\min}(\boldsymbol{\mathcal{A}}_T^* \boldsymbol{\mathcal{A}}_T)\right\}$$

Constant Step Size Selection

- Symmetric RIP:

$$\lambda(\boldsymbol{\mathcal{A}}_T^*\boldsymbol{\mathcal{A}}_T) \in [1-\delta_{|T|}, 1+\delta_{|T|}]$$

- $\min_{\mu_i} \|\mathbb{I} - \mu_i \boldsymbol{\mathcal{A}}_T^*\boldsymbol{\mathcal{A}}_T\|_{2\rightarrow 2} \leq \min_{\mu_i} \max\left\{\mu_i\left(1+\delta_{|T|}\right) - 1,\ \ 1 - \mu_i\left(1-\delta_{|T|}\right)\right\}$



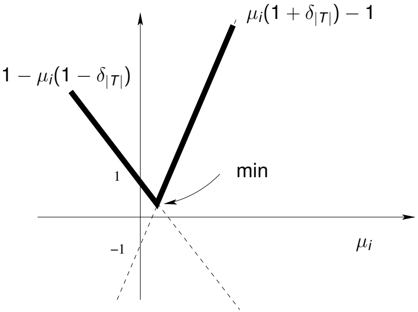
Constant Step Size Selection

Lemma 1

Given the symmetric RIP assumption: $(1-\delta_K) \leq \frac{\|\boldsymbol{\mathcal{A}}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \leq (1+\delta_K), \forall x \in \Sigma_K^N$, the step size μ_i that implies the fastest convergence rate in $(*)$ amounts to

$$\mu_i = 1, \forall i = \{1, 2, \ldots, \},$$

where $\rho = 2\delta_{3K} < 1 \Rightarrow \delta_{3K} < 1/2$ and $\gamma = 2\sqrt{1+\delta_{2K}}$.



Constant Step Size Selection

- For non-symmetric RIP with known upper/lower bounds, $(*)$ becomes:

$$\|x_{i+1} - x^*\|_2 \leq 2\|\mathbb{I} - \mu_i \boldsymbol{\mathcal{A}}_T^* \boldsymbol{\mathcal{A}}_T\|_{2\to 2}\|x_i - x^*\|_2 + 2\mu_i \sqrt{\beta_{2K}}\|\varepsilon\|_2. \qquad (**)$$

Corollary 1

Given non-symmetric RIP with known upper/lower bounds:

$$\alpha_K \leq \frac{\|\boldsymbol{\mathcal{A}}\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2} \leq \beta_K, \quad \forall x \in \Sigma_K^N,$$

the step size μ_i that implies the fastest convergence rate in $(**)$ amounts to

$$\mu_i = \frac{2}{\alpha_{3K} + \beta_{3K}}, \forall i = \{1, 2, \ldots, \},$$

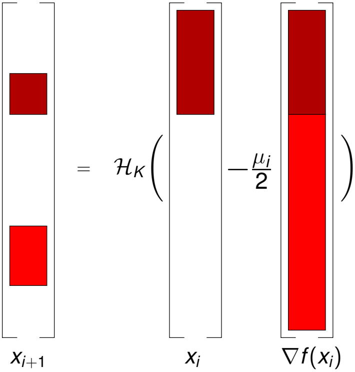
where $\rho = \frac{2(\beta_{3K} - \alpha_{3K})}{\alpha_{3K} + \beta_{3K}} < 1 \Rightarrow \beta_{3K} < 3\alpha_{3K}$ and $\gamma = \frac{2\sqrt{\beta_{2K}}}{\alpha_{3K} + \beta_{3K}}$.

Potential pitfalls in step size selection

- No knowledge about RIP condition.
- Can we leverage the ideas from convex optimization?
  1. At each iteration, pick a conservative (small) value for $\mu_i \rightarrow$ premature termination of the algorithm.
- Perform binary search over step size μ_i:
  1. No knowledge about RIP bounds $\rightarrow$ we may miss the "sweet" μ_i range that leads to convergence near the true vector.

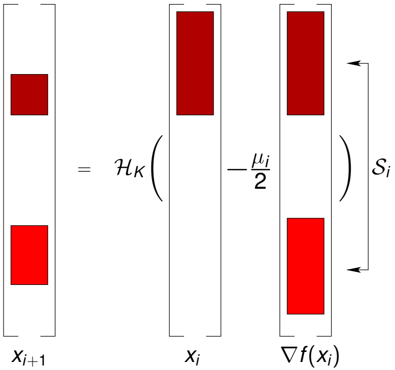
Adaptive Step Size Selection

- Let $\bar{\mathcal{X}}_i = \text{supp}(H_K(\nabla_{\mathcal{X}_i^c} f(x_i)))$.
- Key observation: x_{i+1} contains non-zero elements at positions from the set $\mathcal{S}_i = \mathcal{X}_i \cup \bar{\mathcal{X}}_i,\ |\mathcal{S}_i| \leq 2K$.



Adaptive Step Size Selection

- Let $\bar{\mathcal{X}}_i = \mathrm{supp}(H_K(\nabla_{\mathcal{X}_i^c} f(x_i)))$.
- Key observation: x_{i+1} contains non-zero elements at positions from the set $\mathcal{S}_i = \mathcal{X}_i \cup \bar{\mathcal{X}}_i,\ |\mathcal{S}_i| \leq 2K$.



- $H_K(\cdot) \Longrightarrow \mathcal{O}(K \log K)$ vs. $\mathcal{O}(N \log N)$ complexity.
- More sophisticated method: median method with $\mathcal{O}(K)$ amortized complexity.

Adaptive Step Size Selection

- Let $\bar{\mathcal{X}}_i = \text{supp}(H_K(\nabla_{\mathcal{X}_i^c} f(x_i)))$.
- Key observation: x_{i+1} contains non-zero elements at positions from the set $\mathcal{S}_i = \mathcal{X}_i \cup \bar{\mathcal{X}}_i, \ |\mathcal{S}_i| \leq 2K$.
- Thus:

$$x_{i+1} = H_K\Big(\underbrace{x_i - \frac{\mu_i}{2} \nabla_{\mathcal{S}_i} f(x_i)}_{=b} \Big).$$

- Observe $b \in \Sigma_{2K}^N$ since $|\mathcal{S}_i| \leq 2K$.
- Calculate step size that minimizes the objective value, i.e.

$$\mu_i = \arg\min_{\mu} \|\boldsymbol{y} - \boldsymbol{\mathcal{A}} b\|_2^2 = \frac{\|\nabla_{\mathcal{S}_i} f(x_i)\|_2^2}{\|\boldsymbol{\mathcal{A}} \underbrace{\nabla_{\mathcal{S}_i} f(x_i)}_{2K-\text{sparse}}\|_2^2}.$$

- RIP:

$$\frac{1}{1+\delta_{2K}} \leq \mu_i \leq \frac{1}{1-\delta_{2K}}$$

- Non-symmetric RIP:

$$\frac{1}{\beta_{2K}} \leq \mu_i \leq \frac{1}{\alpha_{2K}}$$

Convergence guarantees

Theorem 1 [Iteration Invariant]

Assume $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{M \times N}$ satisfies:

$$\alpha_K \leq \frac{\|\boldsymbol{\mathcal{A}}\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2} \leq \beta_K, \quad \forall x \in \Sigma_K^N,$$

where α_K, β_K are unknown. Then, in the worst case scenario, IHT with adaptive step size selection approximates the true K-sparse signal $\mathbf{X}^*$ with convergence rate ρ according to:

$$\|\mathbf{X}_{i+1} - \mathbf{X}^*\|_2 \leq \rho \|\mathbf{X}_i - \mathbf{X}^*\|_2 + \gamma \|\varepsilon\|_2,$$

where $\rho = 2\max\{\frac{\beta_{3K}}{\alpha_{2K}} - 1, 1 - \frac{\alpha_{3K}}{\beta_{2K}}\} < 1$ and $\gamma = \frac{2\sqrt{\beta_{2K}}}{\alpha_{2K}}$.

- If symmetrc RIP holds: $(1 - \delta_K) \leq \frac{\|\boldsymbol{\mathcal{A}}\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2} \leq (1 + \delta_K), \ \forall x \in \Sigma_K^N$, then

$$\|\mathbf{X}_{i+1} - \mathbf{X}^*\|_2 \leq \rho \|\mathbf{X}_i - \mathbf{X}^*\|_2 + \frac{2\sqrt{1 + \delta_{2K}}}{1 - \delta_{2K}} \|\varepsilon\|_2,$$

where $\rho \triangleq \frac{\delta_{3K} + \delta_{2K}}{1 - \delta_{2K}} < 1 \Rightarrow \delta_{3K} < 1/5$ (worst-case scenario - $\mu_i = \frac{1}{1 - \delta_{2K}}$).

Ingredients

Memory

Memory

$$x_{i+1} = H_K(x_i - \frac{\mu_i}{2}\nabla_{\mathcal{S}_i} f(x_i))$$

- Idea: why not use information from previous estimates (x_{i-1}, x_{i-2}, etc.)?
- Nesterov's one-memory utilization scheme over convex sets:

$$\tau_i = \frac{\alpha_i(1-\alpha_i)}{\alpha_i^2 + \alpha_{i+1}}, \tag{1}$$

where $\alpha_0 \in (0,1)$ and $\alpha_{i+1} \in (0,1)$ is computed as the root of

$$\alpha_{i+1}^2 = (1-\alpha_{i+1})\alpha_i^2 + q\alpha_{i+1}, \text{ for } q \triangleq \frac{\lambda_{\min}(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}})}{\lambda_{\max}(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}})}. \tag{2}$$

- In our case:

$$x_i = H_K(y_i - \frac{\mu_i}{2}\nabla_{\mathcal{S}_i} f(y_i)), \quad y_{i+1} = x_i + \tau_i(x_i - x_{i-1})$$

where $\mathcal{Y}_i = \text{supp}(y_i)$ and $\mathcal{S}_i = \mathcal{Y}_i \cup \text{supp}(H_K(\nabla_{\mathcal{Y}_i^c} f(y_i)))$ with $|\mathcal{S}_i| \leq 3K$.

Nesterov's scheme

- Nesterov's scheme is not optimal when the algorithm is linear convergent.

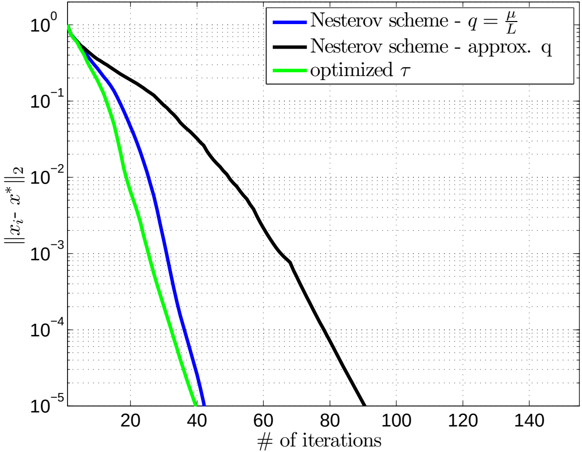


Figure: ALPS convergence rate example using memory. $N = 2000, M = 600, K = 120$. Blue and black lines represent Nesterov's τ_i selection scheme with $q = \frac{\lambda_{\text{min}}(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}})}{\lambda_{\text{max}}(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}})}$ and $q \sim \frac{\mu_i^{\text{min}}}{\mu_i^{\text{max}}}$, respectively; green line represents the proposed momentum step size selection.

Momentum step size selection

- Momentum step size selection:

  (i) Constant τ_i, e.g. $\tau_i = 1/2, \forall i \rightarrow$ No additional computational cost,

  (ii) Nesterov's scheme $\tau_i = (a_i - 1)/(a_{i+1})$ where $a_{i+1} = \frac{1+\sqrt{4a_i^2+1}}{2} \rightarrow$ No additional computational cost,

  (iii) Objective minimizer τ_i: $\tau_i = \arg\min_\tau \|\boldsymbol{y} - \boldsymbol{\mathcal{A}} y_{i+1}\|_2^2 \rightarrow$ No additional computational cost (!!!):

$$
\begin{aligned}
\tau_i &= \arg\min_{\tau} \|\boldsymbol{y} - \boldsymbol{\mathcal{A}} y_{i+1}\|_2^2 \\
&= \arg\min_{\tau} \|(\boldsymbol{y} - \boldsymbol{\mathcal{A}} x_i) - \tau \boldsymbol{\mathcal{A}}(x_i - x_{i-1})\|_2^2 \\
&= \frac{\langle \boldsymbol{y} - \boldsymbol{\mathcal{A}}\mathbf{X}_i, \boldsymbol{\mathcal{A}}\mathbf{X}_i - \boldsymbol{\mathcal{A}}\mathbf{X}_{i-1} \rangle}{\|\boldsymbol{\mathcal{A}}\mathbf{X}_i - \boldsymbol{\mathcal{A}}\mathbf{X}_{i-1}\|_2^2}
\end{aligned}
$$

where $\boldsymbol{\mathcal{A}}\mathbf{X}_i, \boldsymbol{\mathcal{A}}\mathbf{X}_{i-1}$ are previously computed - similar memory-based “tricks” in [Blumensath10].

Ingredients

Gradient updates over restricted support sets

Gradient updates over restricted support sets

- Proxy vector $b = x_i - \frac{\mu_i}{2}\nabla_{\mathcal{S}_i} f(x_i)$: gradient descent over support set $\mathcal{S}_i$.
- Alternatively, b can be computed as the minimizer:

$$b = \underset{x:\text{supp}(x)\subseteq\mathcal{S}_i}{\arg\min} \|\boldsymbol{y} - \boldsymbol{\mathcal{A}}\mathbf{X}\|_2^2.$$

- Based on Hard Thresholding Pursuit [Foucart10] and Subspace Pursuit [WeiMilenkovic09], we can futher refine $x_{i+1} = H_K(b)$ by

$$x_{i+1} = x_{i+1} - \frac{\bar{\mu}_i}{2}\nabla_{\mathcal{X}_{i+1}} f(x_{i+1}), \qquad \text{where } \bar{\mu}_i = \frac{\|\nabla_{\mathcal{X}_{i+1}} f(x_{i+1})\|_2^2}{\|\boldsymbol{\mathcal{A}}\nabla_{\mathcal{X}_{i+1}} f(x_{i+1})\|_2^2},$$

or

$$x_{i+1} = \underset{x:\text{supp}(x)\subseteq\mathcal{X}_{i+1}}{\arg\min} \|\boldsymbol{y} - \boldsymbol{\mathcal{A}}\mathbf{X}\|_2^2.$$

Experiments

(Cookbook: ALgebraic PursuitS (ALPS)
Please check http://lions.epfl.ch/ALPS/)

Execution time

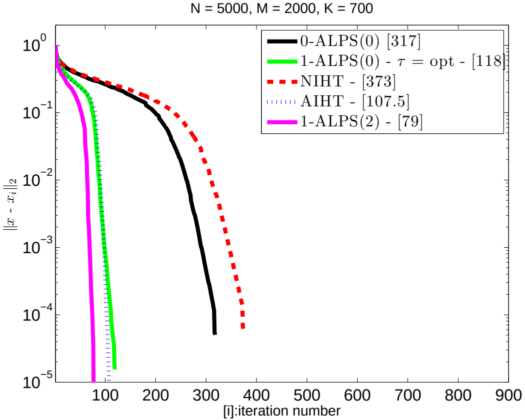


Figure: Median error per iter. - [median # of iter.]
#-ALPS(0): adaptive μ_i with # memory, NIHT: Normalized IHT, AIHT: NIHT with Double Relaxation, 1-ALPS(2): adaptive μ_i and additional gradient update.

Complexity per iter.

0-ALPS(0)	$\mathcal{O}(MN) + 3\mathcal{O}(MK)$
*NIHT	$\mathcal{O}(MN) + 2\mathcal{O}(MK)$
*AIHT	$\mathcal{O}(MN) + 3\mathcal{O}(MK)$
1-ALPS(0)	$\mathcal{O}(MN) + 3\mathcal{O}(MK)$
1-ALPS(2)	$2\mathcal{O}(MN)+5\mathcal{O}(MK)$

of $H_K(\cdot)$ per iter.

0-ALPS(0)	2
*NIHT	2
*AIHT	3
1-ALPS(0)	2
1-ALPS(2)	2

Table: (*) Best case scenario

- Less # of iterations $\nRightarrow$ faster algorithm.

- Remark: model-based projection $\mathcal{M}_K(\cdot) \Rightarrow \#$ of thresholding operations does matter!

Phase transitions

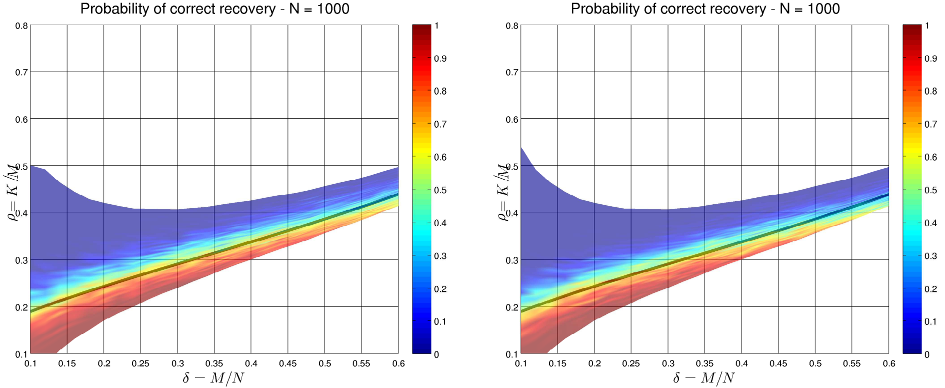


Figure: Empirical phase transition performance of 1-ALPS(0) (left column) and AIHT (right column) algorithms. $\rho = K/M$ and $\delta = M/N$ where $0 \leq \rho, \delta \leq 1$. A signal recovery with solution $\hat{\mathbf{X}}$ is considered successful provided that $\|\hat{\mathbf{X}} - \mathbf{X}^*\|_2 < 10^{-6}$. Solid black line denotes the theoretical l_1 minimization phase transition curve.

Phase transitions

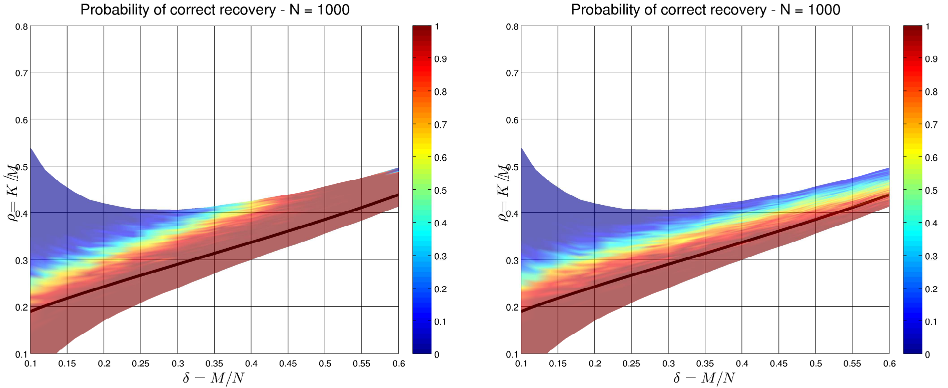


Figure: Empirical phase transition performance of HTP with the proposed step size selection (left column) and HTP with NIHT step size selection (right column). $\rho = K/M$ and $\delta = M/N$ where $0 \leq \rho, \delta \leq 1$. A signal recovery with solution $\hat{\mathbf{X}}$ is considered successful provided that $\|\hat{\mathbf{X}} - \mathbf{X}^*\|_2 < 10^{-6}$. Solid black line denotes the theoretical l_1 minimization phase transition curve.

Conclusions

- Basic “ingredients” of IHT method:
  1. Step size μ_i,
  2. Memory,
  3. Additional gradient updates on restricted support sets.
- Step size selection μ_i: different strategies for different problem assumptions.
- Memory: usage of memory leads to faster convergence rate with (almost) no additional computational cost.
- Additional gradient updates over restricted support sets $\rightarrow$ better phase transition performance.

Thank you

Bon Appetit!

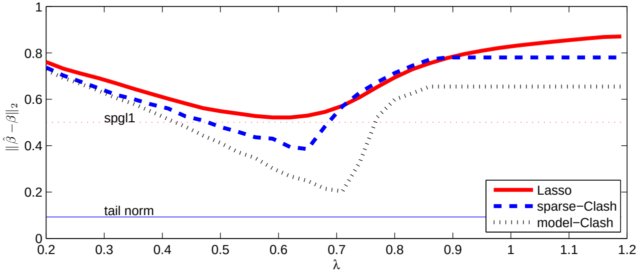
References

- Please check out the ALPS homepage for further information and codes:

  http://lions.epfl.ch/ALPS

- Also, check out the CLASH (Combinatorial selection and Least Absolute SHrinkage) algorithm:

  http://lions.epfl.ch/CLASH



Adaptive Step Size Selection - prior work

- Normalized Iterative Hard Thresholding [BlumensathDavies10].

Algorithm 2: NIHT Algorithm

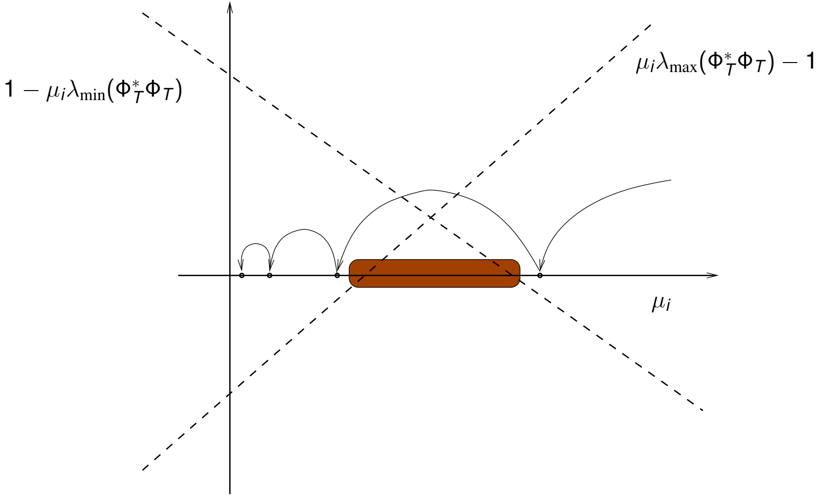
Input: u, Φ, K, Stopping Criteria
Initialize: $x_0 \leftarrow x_{\text{init}}$, $\mathcal{X}_0 \leftarrow \mathcal{X}_{\text{init}}$, $i \leftarrow 0$
repeat
  $\mu_i = \frac{\|\nabla_{\mathcal{X}_i} f(x_i)\|_2^2}{\|\Phi \nabla_{\mathcal{X}_i} f(x_i)\|_2^2}$ (Step size selection)
  $b \leftarrow x_i - \frac{\mu_i}{2} \nabla f(x_i)$ (Update current estimate)
  $\hat{x}_{i+1} \leftarrow H_K(b)$ (Best K-term approximation)
  if $\text{supp}(\hat{x}_{i+1}) = \mathcal{X}_i$ then
    $\mathcal{X}_{i+1} = \mathcal{X}_i$
  else
    Iterate by decreasing μ_i until specific step size criteria are met.
  end if
  $i \leftarrow i + 1$
until Stopping criteria are met.

Adaptive Step Size Selection - prior work

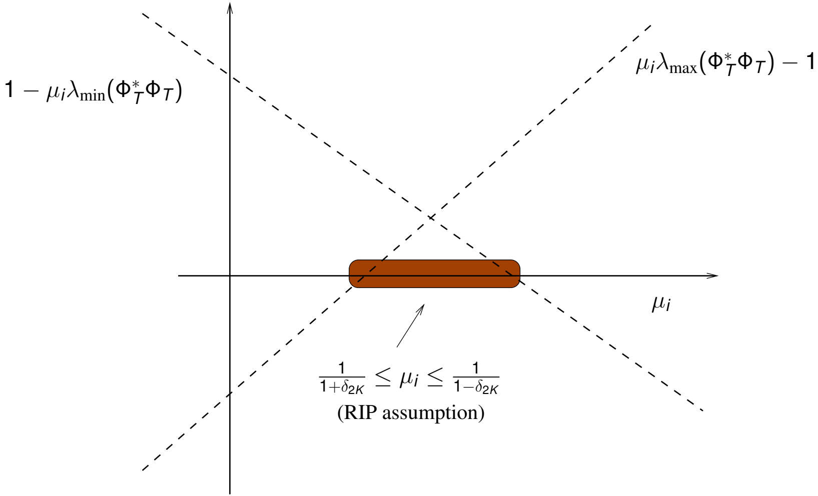
- Normalized Iterative Hard Thresholding [BlumensathDavies10].

Adaptive Step Size Selection - prior work

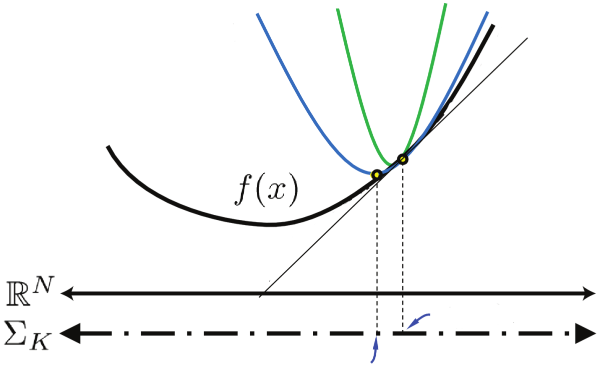
- Why NIHT with binary search over μ_i may not work?



Adaptive step size selection



Adaptive step size selection - convex optimization approach



	Convex-based
Solution via	Convex relaxation $\|\cdot\|_* + \|\cdot\|_1$,...
Criteria example	$\min_{\|\mathbf{y}-\mathcal{A}(\mathbf{L}+\mathbf{M})\|_2 \leq \sigma} \|\mathbf{L}\|_* + \lambda\|\mathbf{M}\|_1$
Algorithms	(S)PCP[1,2,3], CPCP[1,2,3,4], SVT[1,3], ...

	Greedy-based
Solution via	Non-convex projections, ...
Criteria example	$\min_{\text{rank}(\mathbf{L})\leq k, \|\mathbf{M}\|_0 \leq s} \|\mathbf{y} - \mathcal{A}(\mathbf{L}+\mathbf{M})\|_2^2$
Algorithms	SpaRCS[1,2,3,4], GoDec[1,2], SVP[1,3], ...

	Manifold-based
Solution via	Manifold Trust regions, subspace identification, ...
Criteria example	$\min_{\text{rank}(\mathbf{US})\leq k, \|\mathbf{M}\|_0 \leq s} \|\mathbf{y} - \mathcal{A}(\mathbf{US}+\mathbf{M})\|_2^2$
Algorithms	RTRMC[1], GROUSE[1], GRASTA[1], ...

[1]MC, [2]RPCA, [3]ARM, [4]handles CS data