PhD Defense

William de Vazelhes[1], PhD Candidate
Dr. Bin Gu[1], Supervisor
Dr. Xiaotong Yuan[2], External Examiner
Dr. Chih-Jen Lin[1], Internal Member
Dr. Karthik Nandakumar[1], Internal Member
Dr. Zhiqiang Xu[1], Internal Member

[1] Mohamed bin Zayed University of Artificial Intelligence,
[2] Nanjing University

April 17, 2024

Iterative Hard Thresholding

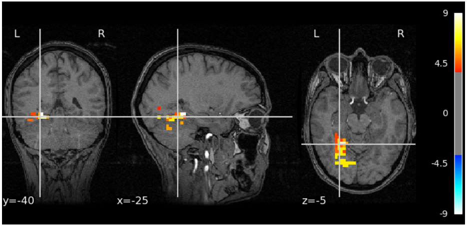
Iterative Hard Thresholding

Introduction

Sparse Optimization:

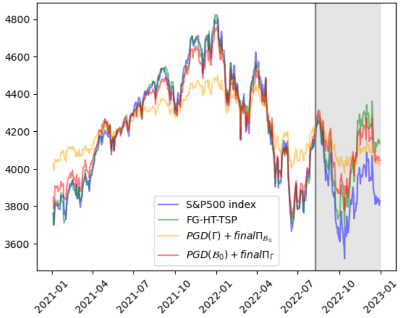
$$\min_{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_0 \leq k} f(\boldsymbol{x})$$

Application: fMRI



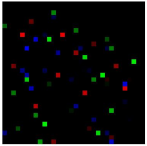
- $\boldsymbol{x}$: map of functional region of the brain (d = number of voxels)
- $f(\boldsymbol{x}) := \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$ with $y_i \in \{-1, 1\}$ standing for $\{'face','house'\}$ and $\boldsymbol{A}_{i,.}$ being the recorded activation map at time i.

Application: Index Tracking



- $\boldsymbol{x}$: amount invested in each of d stocks
- $f(\boldsymbol{x}) := \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$ with $\boldsymbol{y}_i$: S&P returns for day i, $\boldsymbol{A}_{i,j}$: return of stock j on day i

Application: Sparse Adversarial Attacks



Perturbation $\boldsymbol{x}$

'bird'

'dog'

- $\boldsymbol{x}$: perturbation of an image $\boldsymbol{z}$
- $f(\boldsymbol{x}) = \max\{F_y(\text{clip}(\boldsymbol{z}+\boldsymbol{x})) - \max_{j \neq y} F_j(\text{clip}(\boldsymbol{z}+\boldsymbol{x})), 0\}$ with y: true class of the image, F_j: prediction score for class j

The Iterative Hard Thresholding (IHT) algorithm

Algorithm 1: Iterative Hard-Thresholding (IHT)

Initialization: $\boldsymbol{x}_0$
for $t = 0, ..., T$ **do**
  $\boldsymbol{x}_{t+1} := \mathcal{H}_k(\boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t))$
end
Output: $\hat{\boldsymbol{x}_T}$:= e.g. $\boldsymbol{x}_T$ or $\arg\min_{\boldsymbol{x} \in \{\boldsymbol{x}_i\}_{t=1}^{T}} f(\boldsymbol{x}_t)$

It is a **Projected Gradient Descent** algorithm:

$$\mathcal{H}_k(\boldsymbol{x}) := \min_{\boldsymbol{y} \in \mathcal{B}_0(k)} \|\boldsymbol{y} - \boldsymbol{x}\|_2$$

$$\mathcal{B}_0(k) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_0 \leq k\}$$

Goal: Convergence Rate

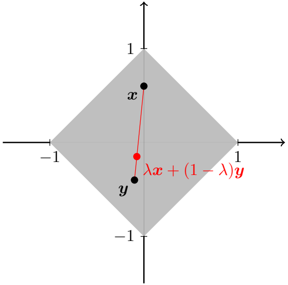
Goal: Prove Convergence Rate Why ?

- To **make sure it does not diverge**.
- To have an estimate of **how feasible it is for a large scale task**.
- To **set the hyperparameters** of the algorithm properly (e.g. η).

Warm Up: Convex Case

$$\min_{\boldsymbol{x} \in \mathcal{C}} f(\boldsymbol{x})$$

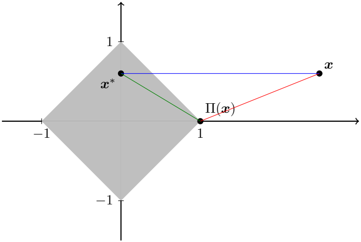
with $\mathcal{C}$ **convex** : $\forall (\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{C})^2 : \ \lambda \boldsymbol{x} + (1 - \lambda) \boldsymbol{y} \in \mathcal{C}$.



Projection onto $\mathcal{C}$

3 Point Lemma:

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \geq \|\Pi_{\mathcal{C}}(\boldsymbol{x}) - \boldsymbol{x}\|^2 + \|\Pi_{\mathcal{C}}(\boldsymbol{x}) - \boldsymbol{x}^*\|^2.$$

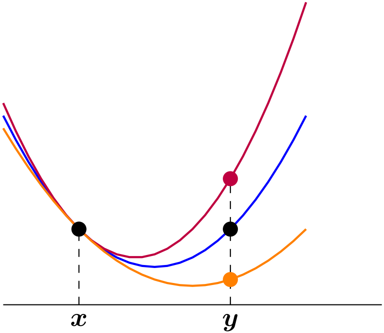


Proj. onto the ℓ_1 unit ball.

Strong Convexity and Smoothness

Assumptions: strong convexity and smoothness. $\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{C}^2$:

$$f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\nu}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \leq f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$$



Proof of Convergence (Convex Case)

Take $\eta := \frac{1}{L}$.

$$
\begin{aligned}
f(\boldsymbol{x}_t) \leq & f(\boldsymbol{x}_{t-1}) + \langle \nabla f(\boldsymbol{x}_{t-1}), \boldsymbol{x}_t - \boldsymbol{x}_{t-1} \rangle + \frac{L}{2} \|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 \\
= & f(\boldsymbol{x}_{t-1}) + \frac{L}{2} \left\| \boldsymbol{x}_t - \boldsymbol{x}_{t-1} + \frac{1}{L} \nabla f(\boldsymbol{x}_{t-1}) \right\|^2 - \frac{1}{2L} \|\nabla f(\boldsymbol{x}_{t-1})\|^2 \\
\leq & f(\boldsymbol{x}_{t-1}) + \frac{L}{2} \left\| \boldsymbol{x}^* - \boldsymbol{x}_{t-1} + \frac{1}{L} \nabla f(\boldsymbol{x}_{t-1}) \right\|^2 - \frac{L}{2} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \frac{1}{2L} \|\nabla f(\boldsymbol{x}_{t-1})\|^2 \\
= & f(\boldsymbol{x}_{t-1}) + \langle \nabla f(\boldsymbol{x}_{t-1}), \boldsymbol{x}^* - \boldsymbol{x}_{t-1} \rangle + \frac{L}{2} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}^*\|^2 - \frac{L}{2} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \\
\leq & f(\boldsymbol{x}^*) + \frac{L - \nu}{2} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}^*\|^2 - \frac{L}{2} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2
\end{aligned}
$$

Proof of Convergence (Convex Case)

$$\left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-t}[f(\mathbf{x}_t)-f(\mathbf{x}^*)] \le \left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-t}\frac{L-\nu}{2}\|\mathbf{x}_{t-1}-\mathbf{x}^*\|^2 - \left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-t}\frac{L}{2}\|\mathbf{x}_t-\mathbf{x}^*\|^2$$

$$\cdots$$

$$\left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-2}[f(\mathbf{x}_2)-f(\mathbf{x}^*)] \le \left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-2}\frac{L-\nu}{2}\|\mathbf{x}_1-\mathbf{x}^*\|^2 - \left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-2}\frac{L}{2}\|\mathbf{x}_2-\mathbf{x}^*\|^2$$

$$\left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-1}[f(\mathbf{x}_1)-f(\mathbf{x}^*)] \le \left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-1}\frac{L-\nu}{2}\|\mathbf{x}_0-\mathbf{x}^*\|^2 - \left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-1}\frac{L}{2}\|\mathbf{x}_1-\mathbf{x}^*\|^2$$

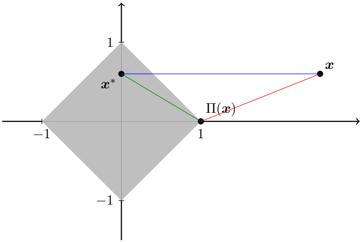
$$\sum_{t=1}^{T}\left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-t}[f(\mathbf{x}_t)-f(\mathbf{x}^*)] \le \left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-1}\frac{L-\nu}{2}\|\mathbf{x}_0-\mathbf{x}^*\|^2 - \left(\frac{\frac{L-\nu}{2}}{\frac{L}{2}}\right)^{T-t}\frac{L}{2}\|\mathbf{x}_t-\mathbf{x}^*\|^2$$

$$\boxed{f(\mathbf{x}_{\hat{T}}) - f(\mathbf{x}^*) \le C\omega^T}$$

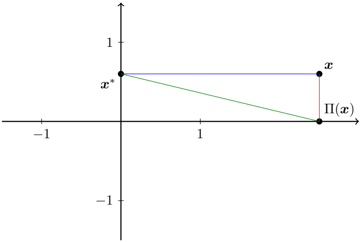
Non-Convex case: $\mathcal{C}$ is the ℓ_0 pseudo-ball

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \geq \|\mathcal{H}_k(\boldsymbol{x}) - \boldsymbol{x}\|^2 + \left(1 - \sqrt{\frac{k^*}{k}}\right) \|\mathcal{H}_k(\boldsymbol{x}) - \boldsymbol{x}^*\|^2.$$

$$\boldsymbol{x}^* \in \mathcal{B}_0(k^*), \quad k^* \leq k$$



Proj. onto the ℓ_1 unit ball.



Proj. onto the ℓ_0 unit pseudo-ball.

Non-convex case: Assumptions

Assumptions: **restricted** strong convexity and **restricted** smoothness. $\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d$ s.t. $\|\boldsymbol{x} - \boldsymbol{y}\|_0 \leq s$ $(s := 3k)$.

$$f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\nu_s}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \leq f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L_s}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

Proof of Convergence (IHT)

We take $\eta = \frac{1}{L_s}$, and $k \geq 4\kappa_s^2 k^*$, with $\kappa_s := \frac{L_s}{\nu_s} \implies \sqrt{\beta} \leq \frac{\nu_s}{2L_s}$.

$$
\begin{aligned}
f(\boldsymbol{x}_t) &\leq f(\boldsymbol{x}_{t-1}) + \langle \nabla f(\boldsymbol{x}_{t-1}), \boldsymbol{x}_t - \boldsymbol{x}_{t-1} \rangle + \frac{L_s}{2} \|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 \\
&= f(\boldsymbol{x}_{t-1}) + \frac{L_s}{2} \left\| \boldsymbol{x}_t - \boldsymbol{x}_{t-1} + \frac{1}{L_s} \nabla f(\boldsymbol{x}_{t-1}) \right\|^2 - \frac{1}{2L_s} \|\nabla f(\boldsymbol{x}_{t-1})\|^2 \\
&\leq f(\boldsymbol{x}_{t-1}) + \frac{L_s}{2} \left\| \boldsymbol{x}^* - \boldsymbol{x}_{t-1} + \frac{1}{L_s} \nabla f(\boldsymbol{x}_{t-1}) \right\|^2 - \frac{L_s}{2}(1 - \sqrt{\beta}) \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \frac{1}{2L_s} \|\nabla f(\boldsymbol{x}_{t-1})\|^2 \\
&= f(\boldsymbol{x}_{t-1}) + \langle \nabla f(\boldsymbol{x}_{t-1}), \boldsymbol{x}^* - \boldsymbol{x}_{t-1} \rangle + \frac{L_s}{2} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}^*\|^2 - \frac{L_s}{2}(1 - \sqrt{\beta}) \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \\
&\leq f(\boldsymbol{x}^*) + \frac{L_s - \nu_s}{2} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}^*\|^2 - \frac{L_s}{2}(1 - \sqrt{\beta}) \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \\
&\leq f(\boldsymbol{x}^*) + \frac{L_s - \nu_s}{2} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}^*\|^2 - \frac{2L_s - \nu_s}{4} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2
\end{aligned}
$$

Proof of Convergence (IHT)

$$\left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-t}[f(\mathbf{x}_t)-f(\mathbf{x}^*)] \leq \left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-t}\frac{L_s-\nu_s}{2}\|\mathbf{x}_{t-1}-\mathbf{x}^*\|^2 - \left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-t}\frac{2L_s-\nu_s}{4}\|\mathbf{x}_t-\mathbf{x}^*\|^2$$

$$\cdots$$

$$\left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-2}[f(\mathbf{x}_2)-f(\mathbf{x}^*)] \leq \left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-2}\frac{L_s-\nu_s}{2}\|\mathbf{x}_1-\mathbf{x}^*\|^2 - \left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-2}\frac{2L_s-\nu_s}{4}\|\mathbf{x}_2-\mathbf{x}^*\|^2$$

$$\left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-1}[f(\mathbf{x}_1)-f(\mathbf{x}^*)] \leq \left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-1}\frac{L_s-\nu_s}{2}\|\mathbf{x}_0-\mathbf{x}^*\|^2 - \left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-1}\frac{2L_s-\nu_s}{4}\|\mathbf{x}_1-\mathbf{x}^*\|^2$$

$$\sum_{t=1}^{T}\left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-t}[f(\mathbf{x}_t)-f(\mathbf{x}^*)] \leq \left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-1}\frac{L_s-\nu_s}{2}\|\mathbf{x}_0-\mathbf{x}^*\|^2 - \left(\frac{\frac{L_s-\nu_s}{2}}{\frac{2L_s-\nu_s}{4}}\right)^{T-t}\frac{2L_s-\nu_s}{4}\|\mathbf{x}_t-\mathbf{x}^*\|^2$$

$$\boxed{f(\mathbf{x}_{\hat{T}})-f(\mathbf{x}^*) \leq C\omega^T}$$

Zeroth-Order Hard-Thresholding

Zeroth-Order Hard-Thresholding

Zeroth-Order Hard-Thresholding (ZOHT)

Algorithm 2: Hard-Thresholding

Initialization: $\boldsymbol{x}_0$

for $t = 0, ..., T$ **do**

$$\boldsymbol{x}_{t+1} := \mathcal{H}_k(\boldsymbol{x}_t - \eta \overset{?}{\nabla f(\boldsymbol{x}_t)})$$

end

Output: $\hat{\boldsymbol{x}_T}$:= e.g. $\boldsymbol{x}_T$ or $\arg\min_{\boldsymbol{x} \in \{\boldsymbol{x}_i\}_{t=1}^T} f(\boldsymbol{x}_t)$

What if we don't know $\nabla f(\boldsymbol{x}_t)$? e.g. for privacy or computational reasons.

Approximating $\nabla f(x)$: two points approximation [1] [2]:

- One random direction $\boldsymbol{u}$:

$$\boldsymbol{g}_t = d\frac{f(\boldsymbol{x}_t + \mu\boldsymbol{u}) - f(\boldsymbol{x}_t)}{\mu}\boldsymbol{u} \quad \text{with} \quad \boldsymbol{u} \sim \mathrm{Uni}(\mathbb{S}_d)$$

- q random directions $\{\boldsymbol{u_i}\}_{i=1}^{q}$:

$$\boldsymbol{g}_t = \frac{d}{q}\sum_{i=1}^{q}\frac{f(\boldsymbol{x}_t + \mu\boldsymbol{u}_i) - f(\boldsymbol{x}_t)}{\mu}\boldsymbol{u}_i \;\; \text{with} \;\; \{\boldsymbol{u_i}\}_{i=1}^{q} \overset{\text{i.i.d.}}{\sim} \mathrm{Uni}(\mathbb{S}_d)$$

Curse of dimensionality: An impossibility result [5]

Under standard assumptions (strongly cvx, smooth, noisy obs.):

"$\forall$ algorithm, $\exists$ f_{adv} s.t. we need more than $O(d/\varepsilon^2)$ queries to achieve $\mathbb{E}[f_{adv}(\hat{\boldsymbol{x}}_T) - f_{adv}(\boldsymbol{x}_*)] \leq \varepsilon$"

Solutions in litterature: more assumptions on f:

- $f(\boldsymbol{x}) = g(\boldsymbol{A}\boldsymbol{x})$ with $\text{rank}(\boldsymbol{A}) << d$ [3]
- sparse/compressible gradients [4]
- What happens in our non-convex case ?

Key Insight: Error of $\boldsymbol{g}_t$ on a Support F

$$F := \mathrm{supp}(\boldsymbol{x}_t) \cup \mathrm{supp}(\boldsymbol{x}_{t-1}) \cup \mathrm{supp}(\boldsymbol{x}^*) \implies |F| = O(k).$$

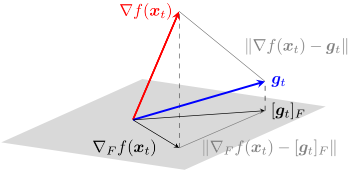
- Bias:

$$\|[\mathbb{E}\boldsymbol{g}_t]_F - [\nabla f(\boldsymbol{x}_t)]_F\|^2 \leq L^2 \epsilon_\mu \mu^2$$

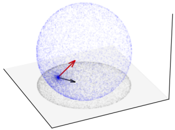
- Variance:

$$\mathbb{E}\|[\boldsymbol{g}_t]_F - \mathbb{E}[\boldsymbol{g}_t]_F\|^2 \leq \frac{\varepsilon_F}{q}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{\varepsilon_{abs}}{q}\mu^2, \text{ with } \varepsilon_F = O(k)$$

$\implies$ Dimension Independent ! (Note: we assume full smoothness here for simplicity)



$q = 1$



$q = 10^6$

ZOHT: Convergence Analysis

Proof is similar as before, except that we:

- "extract" out the error terms
- keep the constants free at the beginning, and later choose them to make things work

$$f(\mathbf{x}_t) \leq f(\mathbf{x}_{t-1}) + \frac{1}{2\eta}\|\mathbf{x}^* - \mathbf{x}_{t-1}\|^2 - \langle \nabla f(\mathbf{x}_{t-1}), \mathbf{x}_{t-1} - \mathbf{x}^* \rangle + \langle [\nabla f(\mathbf{x}_{t-1}) - \mathbf{g}_{t-1}]_F, \mathbf{x}_{t-1} - \mathbf{x}^* \rangle$$
$$- \frac{1}{2\eta}(1 - \sqrt{\beta})\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \left[\frac{L - \frac{1}{\eta} + C}{2}\right]\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \frac{1}{2C}\|[\nabla f(\mathbf{x}_{t-1}) - \mathbf{g}_{t-1}]_F\|^2$$

ZOHT: Convergence Analysis

Choose $\eta := \frac{1}{L+C} = \frac{1}{\alpha L}$, $k \geq 16\alpha^2\kappa_s^2 k^*$ $q_t := \lceil \frac{\tau}{\omega^t} \rceil$ with $\omega := 1 - \frac{1}{8\alpha\kappa_s}$ and $\tau := 16\kappa_s \frac{\varepsilon_F}{(\alpha-1)}$. Use algebraic manipulations, RSC, expression of bias and variance, and smoothness again:

$$\mathbb{E}f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{1}{2\eta}\left[\left(1 - \frac{1}{\alpha'\kappa_s}\right)\mathbb{E}\|\boldsymbol{x}^* - \boldsymbol{x}_{t-1}\|^2 - (1 - \sqrt{\beta})\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \right.$$
$$\left. + 2\eta\left(\frac{G}{2}C_3 + \frac{1}{C}\left(2C_1\|\nabla f(\boldsymbol{x}^*)\|^2 + C_2\mu^2 + C_3\right)\right)\right]$$

ZOHT: Convergence Analysis

$$\boxed{\mathbb{E}f(\hat{\boldsymbol{x}}_T) - f(\boldsymbol{x}^*) \leq F\omega^T + H\mu^2}$$

$$\text{Query Complexity} = \mathcal{O}\left(\frac{\varepsilon_F \kappa_s^3 L}{\varepsilon}\right) = \mathcal{O}\left(\frac{k\kappa_s^3 L}{\varepsilon}\right)$$

Dimension Independent !

IHT with Additional Constraints

IHT + Additional Constraints

We now consider the following problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_0 \leq k, \; \boldsymbol{x} \in \Gamma} f(\boldsymbol{x})$$

Application: e.g. Index Tracking with sector constraints.

$\Gamma = \{\boldsymbol{x} \in \mathbb{R}^d : \forall i \in [c], \|\boldsymbol{x}_{G_i}\|_1 \leq D\}$, where $\boldsymbol{x}_{G_i}$ is the restriction of $\boldsymbol{x}$ to group G_i (i.e. for $j \in [d]$, $\boldsymbol{x}_{G_i j} = \boldsymbol{x}_j$ if $j \in G_i$ and 0 otherwise).

IHT + Additional Constraints

Assumption (k-support-preserving set)

Γ is convex and for any $\boldsymbol{x} \in \mathbb{R}^d$ s.t. $\|\boldsymbol{x}\|_0 \leq k$: $supp(\Pi_\Gamma(\boldsymbol{x})) \subseteq supp(\boldsymbol{x})$.

Algorithm 3: IHT with Two-Step Proj. (TSP)

Initialization: $\boldsymbol{x}_0$

for $t = 0, ..., T$ **do**

$\quad \boldsymbol{v}_t := \mathcal{H}_k(\boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t))$

$\quad \boldsymbol{x}_{t+1} := \Pi_\Gamma(\boldsymbol{v}_t)$

end

Output: $\hat{\boldsymbol{x}_T}$:= e.g. $\boldsymbol{x}_T$ or $\arg\min_{\boldsymbol{x} \in \{\boldsymbol{x}_i\}_{t=1}^T} f(\boldsymbol{x}_t)$

Support Preserving Set and TSP

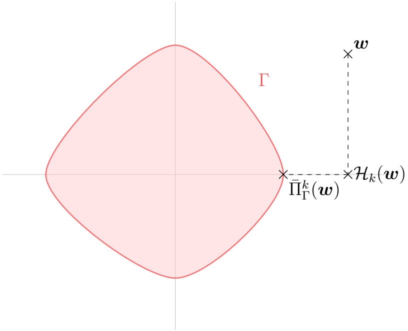


Figure: Support-preserving set and two-step projection ($d = 2, k = 1$).

$$\bar{\Pi}^k_\Gamma(\boldsymbol{w}) := \Pi_\Gamma(\mathcal{H}_k(\boldsymbol{w}))$$

3 Point Lemma with Extra Constraint

New Three (Four) - Point Lemma:

$$\|\bar{\Pi}_{\Gamma}^{k}(\boldsymbol{x}) - \boldsymbol{x}\|^2 \leq \|\boldsymbol{x} - \boldsymbol{x}^*\|^2 - \|\bar{\Pi}_{\Gamma}^{k}(\boldsymbol{x}) - \boldsymbol{x}^*\|^2 + \sqrt{\beta}\|\mathcal{H}_k(\boldsymbol{x}) - \boldsymbol{x}^*\|^2$$

Proof of Convergence

With $\rho \in (0, \frac{1}{2}]$ and $k \geq \frac{4(1-\rho)^2 L_s^2}{\rho^2 \nu_s^2} k^*$:

$$\left(f(\boldsymbol{x}_t) \leq f(\boldsymbol{x}^*) + \frac{L_s - \nu_s}{2} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}^*\|^2 - \frac{L_s}{2} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 + \frac{L_s}{2} \sqrt{\beta} \|\boldsymbol{v}_t - \boldsymbol{x}^*\|^2 \right) \times (1 - \rho)$$

$$\left(f(\boldsymbol{v}_t) \leq f(\boldsymbol{x}^*) + \frac{L_s - \nu_s}{2} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}^*\|^2 - \frac{2L_s - \nu_s}{4} \|\boldsymbol{v}_t - \boldsymbol{x}^*\|^2 \right) \times \rho$$

$$\begin{aligned} (1-\rho) f(\boldsymbol{x}_t) + \rho f(\boldsymbol{v}_t) &\leq f(\boldsymbol{x}^*) + \frac{L_s - \nu_s}{2} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}^*\|^2 - \frac{(1-\rho) L_s}{2} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \frac{\rho (L_s - \nu_s)}{2} \|\boldsymbol{v}_t - \boldsymbol{x}^*\|^2 \\ &= f(\boldsymbol{x}^*) + \frac{L_s - \nu_s}{2} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}^*\|^2 - \frac{L_s - \rho \nu_s}{2} \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 \end{aligned}$$

$$\boxed{\min_{t \in [T]} f(\boldsymbol{x}_t) \leq (1 + 2\rho) f(\boldsymbol{x}^*) + \varepsilon}$$

with $\quad T \geq \left\lceil \frac{L_s}{\nu_s} \log \left(\frac{(L_s - \nu_s) \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2\varepsilon(1-\rho)} \right) \right\rceil + 1 = \mathcal{O}(\kappa_s \log(\frac{1}{\varepsilon}))$

Proof of Convergence

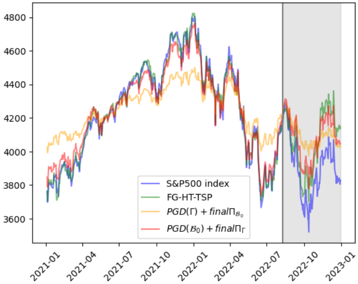
Further, if $\boldsymbol{x}^*$ is a global minimizer of f over $\mathcal{B}_0(k)$, with $\rho = 0.5$ in the expressions of k and T above:

$$\boxed{\min_{t \in [T]} f\left(\boldsymbol{x}_t\right) \leq f(\boldsymbol{x}^*) + \varepsilon.}$$

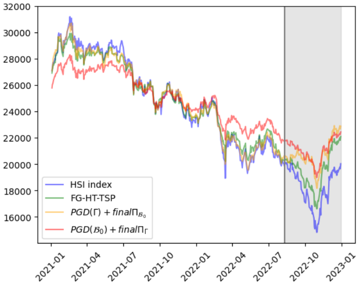
Application: Index Tracking

$$\min_{\boldsymbol{x}\in\mathcal{B}_0(k)\cap\Gamma} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|^2$$

$\Gamma = \{\boldsymbol{x}\in\mathbb{R}^d : \forall i\in[c], \|\boldsymbol{x}_{G_i}\|_1 \leq D\}$, where $\boldsymbol{x}_{G_i}$ is the restriction of $\boldsymbol{x}$ to group G_i (i.e. for $j\in[d]$, $\boldsymbol{x}_{{G_i}_j} = \boldsymbol{x}_j$ if $j\in G_i$ and 0 otherwise).



S&P500



HSI

Dual Perspective on IHT

A dual perspective on IHT: Iterative Regularization with k-Support Norm (IRKSN)

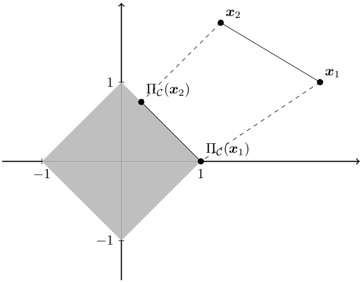
Dual Perspective on IHT

Variant of Projected Gradient Descent:
Dual Averaging (DA)[6]/(Lazy) Mirror Descent (MD)[7]/Lazy OCO[8]/Bregman Iterations [9]:

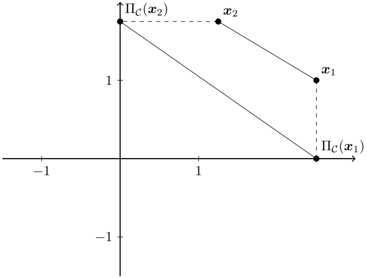
$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t - \eta_t \nabla f(\boldsymbol{x}_t)$$

$$\boldsymbol{x}_{t+1} = \mathcal{H}_k(\boldsymbol{y}_{t+1})$$

Dual Perspective on IHT



Projection onto the ℓ_1 unit ball.

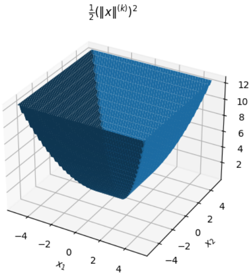


Projection onto the ℓ_0 unit pseudo-ball.

Figure: For projection onto the ℓ_1 ball, we have $\|\Pi_{\mathcal{C}}(\boldsymbol{x}_1) - \Pi_{\mathcal{C}}(\boldsymbol{x}_2)\| \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$ (contractivity), but this is not true if $\mathcal{C}$ is the ℓ_0 pseudo-ball.

Projection and Mirror Map

Contractivity of Π = Smoothness of some function ϕ
$\mathcal{H}_k(\cdot) = \partial\phi(\cdot)$ with $\phi(\cdot) = \frac{1}{2}(\|\cdot\|^{(k)})^2$ (top-k norm): but ϕ **not smooth**.



But we can take the δ-Moreau smoothing:

$$\phi_\delta(\cdot) = \left(\frac{1}{2}(\underbrace{\|\cdot\|_k^{sp}}_{k\text{-support norm (KSN)}})^2 + \frac{1}{2}(\|\cdot\|_2^2) \right)^*$$

Note on the k-support norm (KSN)

- KSN ball is tightest convex relaxation of ℓ_0 and ℓ_2 ball:

$$\{\boldsymbol{x} : \|\boldsymbol{x}\|_k^{sp} \leq D\} = \mathrm{conv}(\{\boldsymbol{x} : \|\boldsymbol{x}\|_0 \leq k\} \cap \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq D\})$$

- The proximal operator for the squared KSN is known [10].

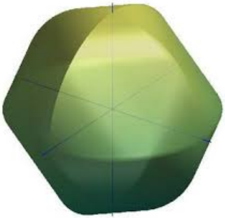


Figure: k-support norm ball (source: [11])

Dual Perspective on IHT

Algorithm becomes:

$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t - \eta_t \nabla f(\boldsymbol{x}_t)$$

$$\boldsymbol{x}_{t+1} = \operatorname{prox}_{\frac{1}{2\delta}(\|\cdot\|_k^{sp})^2}\left(\frac{\boldsymbol{y}_{t+1}}{\delta}\right)$$

Some properties:

- MD/DA Converges to $\boldsymbol{x}^*$ (not sparse in general)
- **For overparam. linear models**: **implicit bias** towards min KSN^2 $(+\delta\ell_2^2)$ solution
- BUT: may still not be k-sparse in general

IRKSN

We consider the **sparse recovery** problem:

$$\boldsymbol{y}^{\delta} = \boldsymbol{X}\boldsymbol{w}^{*} + \boldsymbol{\epsilon}$$

$$\|\boldsymbol{\epsilon}\| \leq \delta$$

Solved by ADGD [12], solving, **with early stopping**:

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) \text{ s.t. } \boldsymbol{X}\boldsymbol{w} = \boldsymbol{y}^{\delta}$$

with $f(\boldsymbol{w}) = F(\boldsymbol{w}) + \frac{\alpha}{2}\|\boldsymbol{w}\|_2{}^2$ with $F(\boldsymbol{w}) = \frac{1-\alpha}{2}(\|\boldsymbol{w}\|_k^{sp})^2$

IRKSN

Algorithm 4: IRKSN

Initialization: $\hat{\boldsymbol{v}}_0 = \hat{\boldsymbol{z}}_{-1} = \hat{\boldsymbol{z}}_0 \in \mathbb{R}^d, \gamma = \alpha \|\boldsymbol{X}\|^{-2}, \boldsymbol{x}_0 = 1$

for $t = 0, ..., T$ **do**

$\hat{\boldsymbol{w}}_t \leftarrow \text{prox}_{\alpha^{-1} F}\left(-\alpha^{-1} \boldsymbol{X}^T \hat{\boldsymbol{z}}_{\boldsymbol{t}}\right)$

$\hat{\boldsymbol{r}}_t \leftarrow \text{prox}_{\alpha^{-1} F}\left(-\alpha^{-1} \boldsymbol{X}^T \hat{\boldsymbol{v}}_t\right)$

$\hat{\boldsymbol{z}}_t \leftarrow \hat{\boldsymbol{v}}_t + \gamma\left(\boldsymbol{X} \hat{\boldsymbol{r}}_{\boldsymbol{t}} - \boldsymbol{y}^\delta\right)$

$\theta_{t+1} \leftarrow \left(1 + \sqrt{1 + 4\theta_t^2}\right)/2$

$\hat{\boldsymbol{v}}_{t+1} = \hat{\boldsymbol{z}}_t + \frac{\theta_t - 1}{\theta_{t+1}}\left(\hat{\boldsymbol{z}}_t - \hat{\boldsymbol{z}}_{t-1}\right)$

end

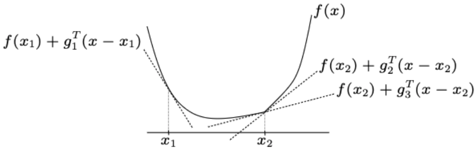
Notations

- For $S \subseteq [d]$, $\bar{S} := [d] \setminus S$
- $\boldsymbol{M}^\dagger$: Moore-Penrose pseudo-inverse [13]
- $\boldsymbol{M}_S$ column-restriction of $\boldsymbol{M}$ to support $S \subseteq [d]$, i.e. the $n \times |S|$ matrix composed of the $|S|$ columns of $\boldsymbol{M}$ of indices in S
- $\mathrm{supp}(\boldsymbol{w})$: support of $\boldsymbol{w}$ (coordinates of the non-zero components of $\boldsymbol{w}$)
- $\boldsymbol{w}_S \in \mathbb{R}^k$ restriction of $\boldsymbol{w}_S$ to a support S of size k, i.e. the sub-vector of size k formed by extracting only the components w_i with $i \in S$
- $\mathrm{sgn}(\boldsymbol{w})$ vector of signs of $\boldsymbol{w}$

Conditions for Recovery

Method	Condition on $\boldsymbol{X}$
IHT [14]	Restricted Isometry Property (RIP)
Lasso [15]	$\max_{\ell \in \bar{S}} \lvert\langle \boldsymbol{X}_S^{\dagger} \boldsymbol{x}_\ell, \operatorname{sgn}(\boldsymbol{w}_S^*)\rangle\rvert < 1$ & $\boldsymbol{X}_S$ injective
ElasticNet [16]	-
KSN pen. [11]	-
OMP [17]	RIP
SRDI [18]	$\begin{cases} \exists \gamma \in (0,1] : & \boldsymbol{X}_S^{\top} \boldsymbol{X}_S \geq n\gamma I_{d,d} \\ \exists \eta \in (0,1) : & \lVert \boldsymbol{X}_{\bar{S}} \boldsymbol{X}_S^{\dagger} \rVert_\infty \leq 1 - \eta \end{cases}$
IROSR [19]	RIP
IRCR [20]	$\max_{\ell \in \bar{S}} \lvert\langle \boldsymbol{X}_S^{\dagger} \boldsymbol{x}_\ell, \operatorname{sgn}(\boldsymbol{w}_S^*)\rangle\rvert < 1$ & $\boldsymbol{X}_S$ injective
IRKSN (ours)	$\max_{\ell \in \bar{S}} \lvert\langle \boldsymbol{X}_S^{\dagger} \boldsymbol{x}_\ell, \boldsymbol{w}_S^*\rangle\rvert < \min_{j \in S} \lvert\langle \boldsymbol{X}_S^{\dagger} \boldsymbol{x}_j, \boldsymbol{w}_S^*\rangle\rvert$

Finding Sufficient Conditions: Proof Technique



Subdifferential of the (half-squared) top-k norm:

$$\boxed{\partial \left[\frac{1}{2} \left(\| \cdot \|^{(k)} \right)^2 \right] = \mathrm{conv}(\mathcal{H}_k(\cdot))}$$

Example with $k = 1$:

$$\partial \left[\frac{1}{2} \left(\|[-1.2, 1]\|^{(k)} \right)^2 \right] = \{[-1.2, 0]\}$$

$$\partial \left[\frac{1}{2} \left(\|[-1.2, 1.2]\|^{(k)} \right)^2 \right] = \mathrm{conv}(\{[-1.2, 0], [0, 1.2]\}) = \{[-1.2\lambda, 1.2(1-\lambda)], \lambda \in [0, 1]\}$$

Sufficient conditions for recovery: comparison with ℓ_1 norm

Assumption (Conditions for recovery with ℓ_1 norm-based algorithms)

Let $\boldsymbol{w}^*$ be supported on a support $S \subset [d]$. $\boldsymbol{w}^*$ is such that:

1. $\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{y}$
2. $\boldsymbol{X}_S$ *is injective*
3. $\max_{\ell \in \bar{S}} |\langle \boldsymbol{X}_S^\dagger \boldsymbol{x}_\ell, \operatorname{sgn}(\boldsymbol{w}_S^*) \rangle| < 1$

Assumption (Conditions for recovery with IRKSN)

- $\boldsymbol{w}^*$ *k-sparse, supp*$(\boldsymbol{w}^*) = S \subset [d]$, $\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{y}$
- $\boldsymbol{w}_S^* = \arg\min_{\boldsymbol{z} \in \mathbb{R}^k : \boldsymbol{X}_S \boldsymbol{z} = \boldsymbol{y}} \|\boldsymbol{z}\|_2$
- $\max_{\ell \in \bar{S}} |\langle \boldsymbol{X}_S^\dagger \boldsymbol{x}_\ell, \boldsymbol{w}_S^* \rangle| < \min_{j \in S} |\langle \boldsymbol{X}_S^\dagger \boldsymbol{x}_j, \boldsymbol{w}_S^* \rangle|$
- *Does not need $\boldsymbol{X}_S$ to be injective !*

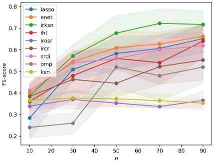
Conditions for recovery, case where $\boldsymbol{X}_S$ is injective

If $\boldsymbol{X}_S$ is injective and $\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{y}$, the conditions become:

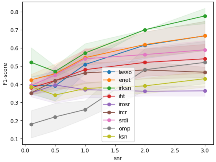
- (A) (ℓ_1-norm based): $\max_{\ell \in \bar{S}} |\langle \boldsymbol{X}_S^\dagger \boldsymbol{x}_\ell, \mathrm{sgn}(\boldsymbol{w}_S^*) \rangle| < 1$
- (B) (IRKSN): $\max_{\ell \in \bar{S}} |\langle \boldsymbol{X}_S^\dagger \boldsymbol{x}_\ell, \frac{\boldsymbol{w}_S^*}{\min_{j \in S} |\boldsymbol{w}_S^*|} \rangle| < 1$

It is possible to find examples of design matrix $\boldsymbol{X}$ and vector $\boldsymbol{w}^*$ which verify (B) but not (A): IRKSN is ensured to recover $\boldsymbol{w}^*$ there, contrary to ℓ_1 norm-based algorithms.

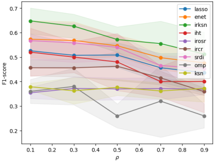
Experiments: Synthetic design matrix X



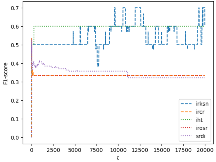
(a) F1-score vs. n



(b) F1-score vs. snr



(c) F1-score vs. ρ



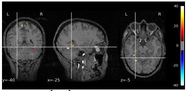
(d) F1-score vs. t

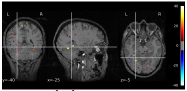
Figure: F1-score of support recovery for a correlated design matrix [20] ρ: correlation, snr: signal/noise ratio, n: num. samples.

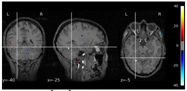
Experiments: fMRI decoding

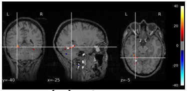
	Lasso	ElasticNet	OMP	IHT	KSN	IRKSN	IRCR	IROSR	SRDI
face'/'house'	.425	.349	.938	.2441	.247	**.2440**	.341	.381	.314
'house'/'shoe'	.528	.500	.938	.2968	.299	**.2965**	.407	.502	.357

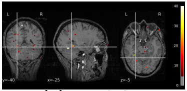
Model estimation $\|\boldsymbol{w} - \boldsymbol{w}^*\|$ ($\boldsymbol{w}^*$: obtained by EnCluDL).

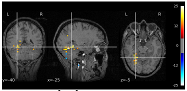

(b) Lasso

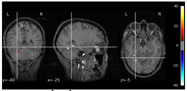

(c) Enet

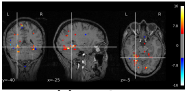

(d) OMP

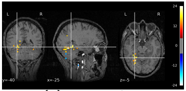

(e) SRDI

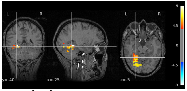

(f) IROSR


(g) IHT


(h) IRCR


(i) KSN


(j) IRKSN


(k) EnCluDL

QA

QA

References I

[1] S. Liu, P.-Y. Chen, B. Kailkhura, G. Zhang, A. O. Hero III, and P. K. Varshney, "A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications," *IEEE Signal Processing Magazine*, vol. 37, no. 5, pp. 43–54, 2020.

[2] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

References II

[3] D. Golovin, J. Karro, G. Kochanski, C. Lee, X. Song, and Q. Zhang, "Gradientless descent: High-dimensional zeroth-order optimization," in *International Conference on Learning Representations*, 2019.

[4] H. Cai, D. McKenzie, W. Yin, and Z. Zhang, "Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling," *SIAM Journal on Optimization*, vol. 32, no. 2, pp. 687–714, 2022.

[5] K. G. Jamieson, R. Nowak, and B. Recht, "Query complexity of derivative-free optimization," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.

References III

[6] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Mathematical programming*, vol. 120, no. 1, pp. 221–259, 2009.

[7] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[8] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proceedings of the 20th international conference on machine learning (icml-03)*, 2003, pp. 928–936.

References IV

[9] L. Bungert, T. Roith, D. Tenbrinck, and M. Burger, "A bregman learning framework for sparse neural networks," *Journal of Machine Learning Research*, vol. 23, no. 192, pp. 1–43, 2022.

[10] A. McDonald, M. Pontil, and D. Stamos, "Fitting spectral decay with the k-support norm," in *Artificial Intelligence and Statistics*, PMLR, 2016, pp. 1061–1069.

[11] A. Argyriou, R. Foygel, and N. Srebro, "Sparse prediction with the k-support norm," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

References V

[12] S. Matet, L. Rosasco, S. Villa, and B. L. Vu, "Don't relax: Early stopping for convex regularization," *arXiv preprint arXiv:1707.05422*, 2017.

[13] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2013.

[14] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and computational harmonic analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

References VI

[16] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[17] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on information theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[18] S. Osher, F. Ruan, J. Xiong, Y. Yao, and W. Yin, "Sparse recovery via differential inclusions," *Applied and Computational Harmonic Analysis*, vol. 41, no. 2, pp. 436–469, 2016.

References VII

[19] T. Vaskevicius, V. Kanade, and P. Rebeschini, "Implicit regularization for optimal sparse recovery," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] C. Molinari, M. Massias, L. Rosasco, and S. Villa, "Iterative regularization for convex regularizers," in *International conference on artificial intelligence and statistics*, PMLR, 2021, pp. 1684–1692.

Some images were taken from the MTH702 course at MBZUAI.