Optimization over Sparse Support-Preserving Sets: Two-Step Projection with Global Optimality Guarantees

William de Vazelhes [1], Xiao-Tong Yuan [2], Bin Gu [3]

[1]GenBio AI, work done while at MBZUAI, Abu Dhabi, UAE [2]School of Intelligence Science and Technology, Nanjing University, Suzhou, China [3]School of Artificial Intelligence, Jilin University, China

Problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} R(\boldsymbol{w}), \quad \text{s.t. } \|\boldsymbol{w}\|_0 \le k \text{ and } \boldsymbol{w} \in \Gamma. \tag{1}$$

Related Works and Overview of Contributions

Table 1. Comparison of results for Iterative Hard Thresholding with/without additional constraints. [1] $\mathcal{S}$: symmetric convex sets being sign-free or non-negative [1], $\mathcal{A}$: k-support preserving sets. [2] If a paper reports both $\|\boldsymbol{w}-\bar{\boldsymbol{w}}\|$ and $R(\boldsymbol{w})-R(\bar{\boldsymbol{w}})$, we report only the latter. $\hat{T}$: time index of the $\boldsymbol{w}$ returned by the method (e.g. $\hat{T}=\arg\min_{t\in[T]} R(\boldsymbol{w}_t)$). $\bar{\boldsymbol{w}}$: $\bar{k}$-sparse vector in Γ. Δ: System error (non-vanishing term which depends on the gradient at optimality (e.g. $\mathbb{E}_i\|\nabla R_i(\bar{\boldsymbol{w}})\|$, (see corresponding references))). [4]: $\kappa_s=\frac{L_s}{\nu_s}$ and $\kappa_{s'}=\frac{L_{s'}}{\nu_s}$ (cf. corresponding refs. for defs. of s and s'). [3] SM: Lipschitz-smooth, D: Deterministic. S: Stochastic, Z: Zeroth-Order, L: Lipschitz continuous. ♣: Notably, we could eliminate Δ from [2].

Reference	Γ^1	Convergence^2	k	Setting^3
[3]	$\mathbb{R}^d$	$R(\boldsymbol{w}_{\hat{T}}) \le R(\bar{\boldsymbol{w}})+\varepsilon$	$\Omega(\kappa_s^2\bar{k})$	D, RSS, RSC
[4]	$\mathbb{R}^d$	$\mathbb{E}\|\boldsymbol{w}_{\hat{T}}-\bar{\boldsymbol{w}}\| \le \varepsilon+\mathcal{O}(\Delta)$	$\Omega(\kappa_s^2\bar{k})$	S, RSS, RSC
[5]	$\mathbb{R}^d$	$\mathbb{E}R(\boldsymbol{w}_{\hat{T}}) \le R(\bar{\boldsymbol{w}})+\varepsilon+\mathcal{O}(\Delta)$	$\Omega(\kappa_s^2\bar{k})$	S, RSS, RSC
[6]	$\mathbb{R}^d$	$\mathbb{E}R(\boldsymbol{w}_{\hat{T}}) \le R(\bar{\boldsymbol{w}})+\varepsilon$	$\Omega(\kappa_s^2\bar{k})$	S, RSS, RSC
[2]	$\mathbb{R}^d$	$\mathbb{E}\|\boldsymbol{w}_{\hat{T}}-\bar{\boldsymbol{w}}\| \le \varepsilon+\mathcal{O}(\mu)+\mathcal{O}(\Delta)$	$\Omega(\kappa_{s'}^4\bar{k})$	S, Z, RSS', RSC
[1], [7]	$\Gamma\in\mathcal{S}$	local convergence	-	D, SM
[8]	ℓ_∞ ball around 0	local convergence	-	S, Z, L
IHT-2SP	$\Gamma\in\mathcal{A}$	$R\left(\boldsymbol{w}_{\hat{T}}\right) \le (1+2\rho)R(\bar{\boldsymbol{w}})+\varepsilon$	$\Omega\left(\frac{\kappa_s^2\bar{k}}{\rho^2}\right)$	D, RSS, RSC
HSG-HT-2SP	$\Gamma\in\mathcal{A}$	$\mathbb{E}R(\boldsymbol{w}_{\hat{T}}) \le (1+2\rho)R(\bar{\boldsymbol{w}})+\varepsilon$	$\Omega\left(\frac{\kappa_s^2\bar{k}}{\rho^2}\right)$	S, RSS, RSC
HZO-HT	$\mathbb{R}^d$	$\mathbb{E}[R(\boldsymbol{w}_{\hat{T}})-R(\bar{\boldsymbol{w}})] \le \varepsilon+\mathcal{O}(\mu)$♣	$\Omega(\kappa_{s'}^2\bar{k})$	Z, RSS', RSC
HZO-HT-2SP	$\Gamma\in\mathcal{A}$	$\mathbb{E}R(\boldsymbol{w}_{\hat{T}}) \le (1+2\rho)R(\bar{\boldsymbol{w}})+\varepsilon+\mathcal{O}(\mu)$	$\Omega\left(\frac{\kappa_{s'}^2\bar{k}}{\rho^2}\right)$	Z, RSS', RSC

The Two-Step Projection (2SP) Algorithm and Support Preserving Sets

Algorithm 1 Deterministic IHT with extra constraints (IHT-2SP)

Initialization: $\boldsymbol{w}_0$: initial value, η: learning rate, T: number of iterations
$t=1$ to T
$\boldsymbol{w}_t \leftarrow \bar{\Pi}_\Gamma^k(\boldsymbol{w}_{t-1}-\eta\nabla R(\boldsymbol{w}_{t-1}))$
Output: $\boldsymbol{w}_T$

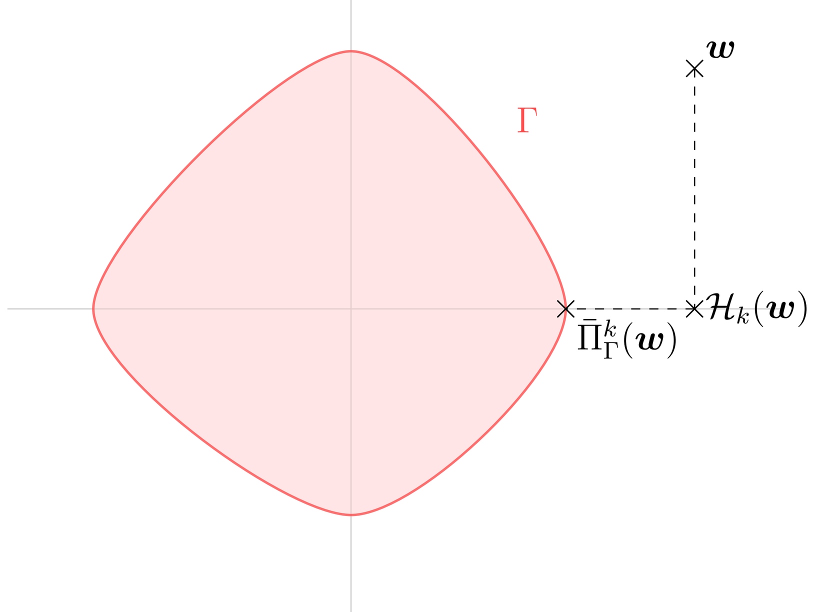
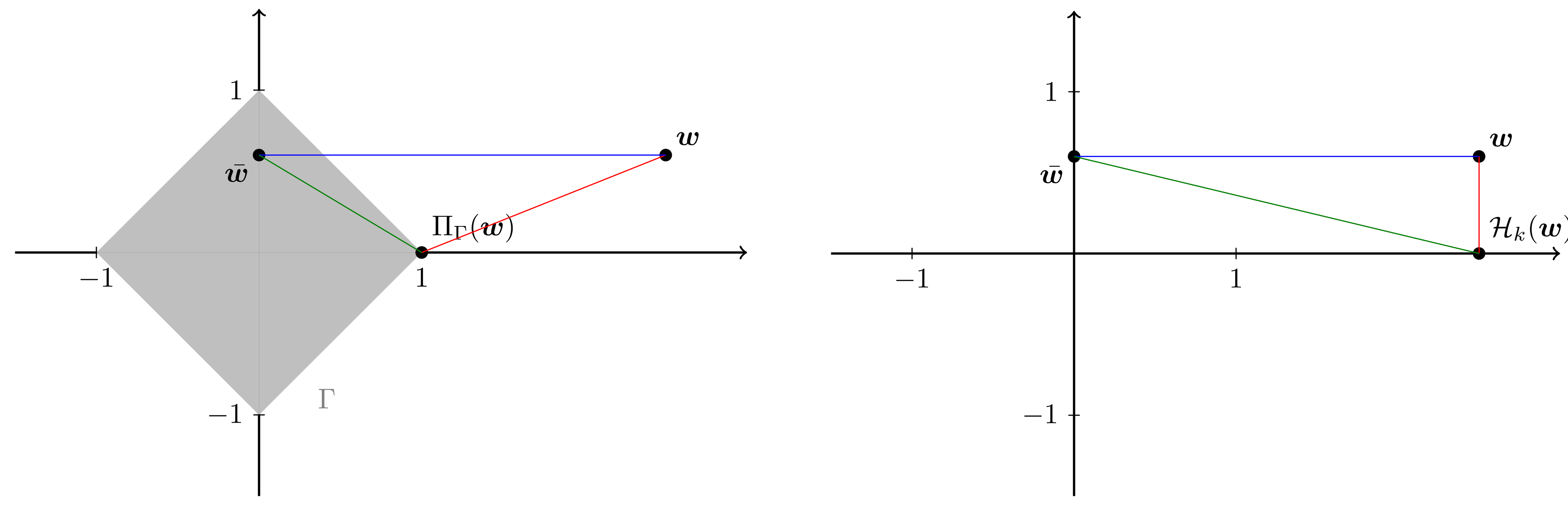


Figure 1. Support-preserving set and two-step projection ($d=2, k=1$).

Definition: $\Gamma \subseteq \mathbb{R}^d$ is k-*support-preserving*, i.e. it is convex and for any $\boldsymbol{w}\in\mathbb{R}^d$ such that $\|\boldsymbol{w}\|_0 \le k$, supp$(\Pi_\Gamma(\boldsymbol{w})) \subseteq$ supp$(\boldsymbol{w})$.

2SP and Three-Point Lemma



Convex projection:

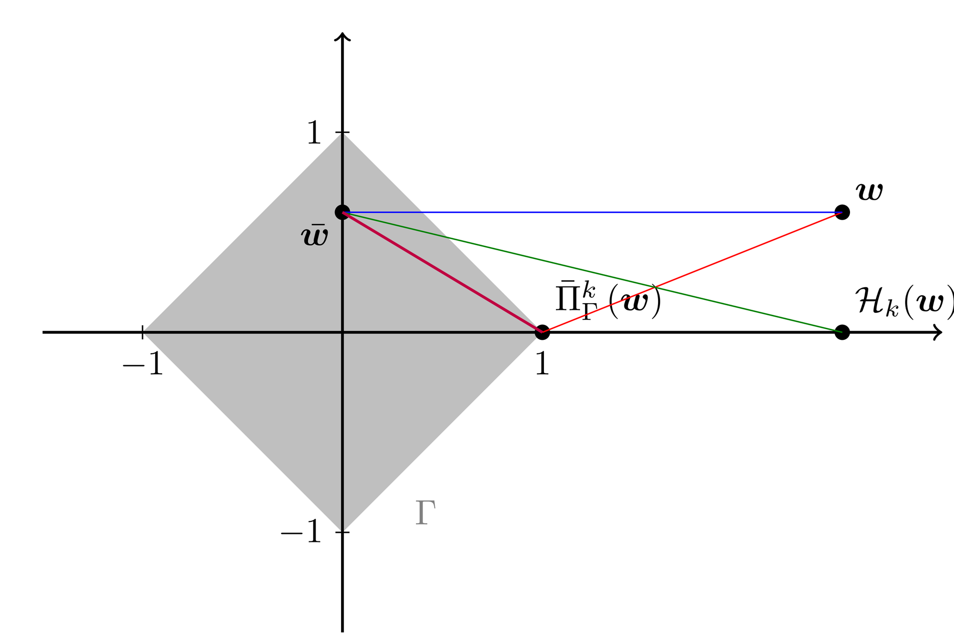
$$\|\boldsymbol{w}-\bar{\boldsymbol{w}}\|^2 \ge \|\boldsymbol{w}-\Pi_\Gamma(\boldsymbol{w})\|^2+\|\Pi_\Gamma(\boldsymbol{w})-\bar{\boldsymbol{w}}\|^2$$

Hard-thresholding:

$$\|\mathcal{H}_k(\boldsymbol{w})-\boldsymbol{w}\|^2 \le \|\boldsymbol{w}-\bar{\boldsymbol{w}}\|^2-\left(1-\sqrt{\beta}\right)\|\mathcal{H}_k(\boldsymbol{w})-\bar{\boldsymbol{w}}\|^2$$

Lemma: Constrained ℓ_0-Three-Point Suppose that Γ is k-support preserving. Consider $\boldsymbol{w},\bar{\boldsymbol{w}}\in\mathbb{R}^p$ with $\|\bar{\boldsymbol{w}}\|_0\le\bar{k}$ and $\bar{\boldsymbol{w}}\in\Gamma$. Then the following holds for any $k\ge\bar{k}$:

$$\|\bar{\Pi}_\Gamma^k(\boldsymbol{w})-\boldsymbol{w}\|^2 \le \|\boldsymbol{w}-\bar{\boldsymbol{w}}\|^2-\|\bar{\Pi}_\Gamma^k(\boldsymbol{w})-\bar{\boldsymbol{w}}\|^2+\sqrt{\beta}\|\mathcal{H}_k(\boldsymbol{w})-\bar{\boldsymbol{w}}\|^2, \text{ with } \beta:=\frac{\bar{k}}{k}.$$



Convergence Rate

Theorem: With R (L_s,s)-RSS and (ν_s,s)-RSC with $s=2k$, R non-negative (w.l.o.g.), Γ k-support preserving, $\eta=\frac{1}{L_s}$, $\bar{\boldsymbol{w}}$ any $\bar{k}$-sparse vector, $\rho\in(0,\frac{1}{2}]$, $k\ge\frac{4(1-\rho)^2L_s^2}{\rho^2\nu_s^2}\bar{k}$. Then for any $\varepsilon>0$, for $T\ge\left\lceil\frac{L_s}{\nu_s}\log\left(\frac{(L_s-\nu_s)\|\boldsymbol{w}_0-\bar{\boldsymbol{w}}\|^2}{2\varepsilon(1-\rho)}\right)\right\rceil+1=\mathcal{O}(\kappa_s\log(\frac{1}{\varepsilon}))$, the iterates of IHT-2SP satisfy:

$$\min_{t\in[T]} R(\boldsymbol{w}_t) \le (1+2\rho)R(\bar{\boldsymbol{w}})+\varepsilon.$$

References

[1] Z. Lu, "Optimization over sparse symmetric sets via a nonmonotone projected gradient method," *arXiv preprint arXiv:1509.08581*, 2015.
[2] W. de Vazelhes, H. Zhang, H. Wu, X. Yuan, and B. Gu, "Zeroth-order hard-thresholding: Gradient error vs. expansivity," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 589–22 601, 2022.
[3] P. Jain, A. Tewari, and P. Kar, "On iterative hard thresholding methods for high-dimensional m-estimation," *Advances in Neural Information Processing Systems*, vol. 27, 2014.
[4] N. Nguyen, D. Needell, and T. Woolf, "Linear convergence of stochastic iterative greedy algorithms with sparse constraints," *IEEE Transactions on Information Theory*, vol. 63, pp. 6869–6895, 2017.
[5] X. Li, R. Arora, H. Liu, J. Haupt, and T. Zhao, "Nonconvex sparse learning via stochastic optimization with progressive variance reduction," *arXiv preprint arXiv:1605.02711*, 2016.
[6] P. Zhou, X. Yuan, and J. Feng, "Efficient stochastic gradient hard thresholding," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
[7] A. Beck and N. Hallak, "On the minimization over sparse symmetric sets: Projections, optimality conditions, and algorithms," *Mathematics of Operations Research*, vol. 41, pp. 196–223, 2016.
[8] M. R. Metel, "Sparse training with lipschitz continuous loss functions and a weighted group l0-norm constraint," *Journal of Machine Learning Research*, vol. 24, pp. 1–44, 2023.