Candidacy Exam

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 of Information Science & Technology

May 11, 2023

Introduction

Sparse Optimization:

$$\min_{\|\boldsymbol{x}\|_0 \leq k} f(\boldsymbol{x})$$

Applications:

- Sparse regression/classification (e.g. gene array data)
- Sparse recovery

TL;DR

Main contributions:

- ZOHT: condition on number of random directions q
- IRKSN: new conditions for linear sparse recovery

Zeroth-Order

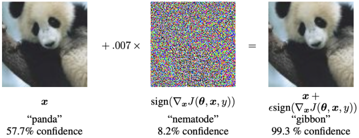
$$\min_{\boldsymbol{x} \in \mathbb{R}} f(\boldsymbol{x})$$

Gradient descent:

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

What if we don't know $\nabla f(\cdot)$, but only $f(\cdot)$?

- Black-Box Adversarial attacks [1]



- Reinforcement learning [2]

Approximate $\nabla f(x)$: two points approximation [3] [4]:

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

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

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

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

Curse of dimensionality: An impossibility result [7]

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 $\text{rank}(\boldsymbol{A}) << d$ [5]
- sparse/compressible gradients [6]

Vanilla ZO

Most ZO algorithms [4], [8], [9]:

Algorithm 1: Vanilla ZO

Initialization: η, T, $\boldsymbol{x}^{(0)}$

Output: $\boldsymbol{x}_T$.

for $t = 1, ..., T$ **do**

  Sample $\boldsymbol{u} \sim \text{Uni}(\mathbb{S}_d)$

  $\hat{\nabla} f(\boldsymbol{x}_{t-1}) \leftarrow \frac{d}{\mu}(f(\boldsymbol{x} + \mu \boldsymbol{u}) - f(\boldsymbol{x}))\,\boldsymbol{u}$;

  $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1} - \eta \hat{\nabla} f(\boldsymbol{x}_{t-1})$;

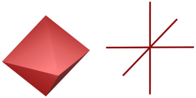
end

Note: just **1 random direction** $\boldsymbol{u}$ is sufficient (with proper η)

Zeroth-Order Hard-Thresholding: Our approach

Consider the **non-convex** ℓ_0 “ball”

$$\min_{\boldsymbol{x} \text{ s.t.} \|\boldsymbol{x}\|_0 \leq k} f(\boldsymbol{x})$$



(a) ℓ_1 (b) [10] ℓ_0

Why not ℓ_1 ? ℓ_1 is convex (impossibility result)

ZOHT: Zeroth-Order Hard-Thresholding

Algorithm 2: SZOHT (simplified)

Initialization: η, T, q, $k = \mathcal{O}(\kappa^4 k^*)$, $\boldsymbol{x}^{(0)}$
Output: $\boldsymbol{x}_T$.
for $t = 1, \ldots, T$ **do**
  for $i = 1, \ldots, q$ **do**
    Sample $\boldsymbol{u}_i \sim \mathrm{Uni}(\mathbb{S}_d)$
    $\hat{\nabla} f(\boldsymbol{x}_{t-1}; \boldsymbol{u}_i) \leftarrow \frac{d}{\mu}\left(f(\boldsymbol{x} + \mu \boldsymbol{u}_i) - f(\boldsymbol{x})\right) \boldsymbol{u}_i$;
  end
  $\hat{\nabla} f(\boldsymbol{x}_{t-1}) \leftarrow \frac{1}{q} \sum_{i=1}^{q} \hat{\nabla} f(\boldsymbol{x}_{t-1}; \boldsymbol{u}_j)$
  $\boldsymbol{x}_t \leftarrow \Phi_k(\boldsymbol{x}_{t-1} - \eta \hat{\nabla} f(\boldsymbol{x}_{t-1}))$; hard-thresholding
end

Dimension independence

- **vanilla ZO:** $O(d)$ QC comes from $\|\hat{\nabla} f(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq O(d)$
- **ZOHT:** $\|\hat{\nabla}_F f(\mathbf{x}) - \nabla_F f(\mathbf{x})\|^2 \leq O(k)$ (props. of projections [11])

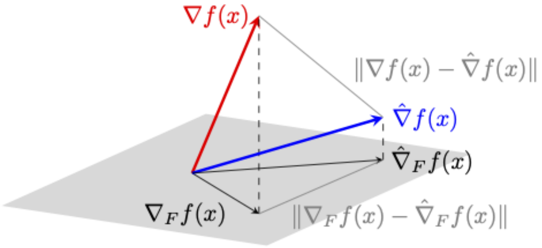


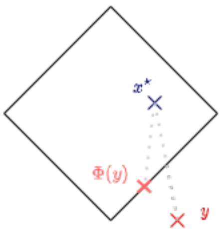
Figure: Gradient estimate and its projections

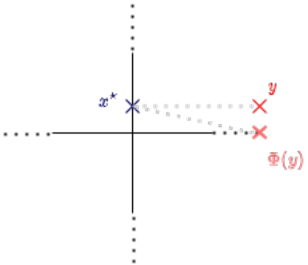
Tuning q: gradient error vs. expansivity

Main difference with vanilla ZO: projection on the ℓ_0 ball ($\mathcal{B}_{\ell_0,k}$) is not non-expansive:

$$\forall \boldsymbol{y} \in \mathbb{R}^d, \boldsymbol{x}^* \in \mathcal{B}_{\ell_0,k} : \ \|\Phi_k(\boldsymbol{y}) - \boldsymbol{x}^*\| \leq \gamma \|\boldsymbol{y} - \boldsymbol{x}^*\|$$

with $\gamma > 1$





Tuning q: gradient error vs. expansivity

Convergence rate:

$$\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^*\| \leq (\rho\gamma)^t \|\boldsymbol{x}_0 - \boldsymbol{x}^*\| + (\cdot)\sigma + (\cdot)\mu$$

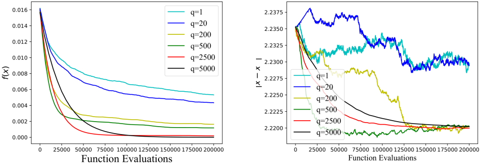
With $\eta = \frac{\nu}{(4\epsilon_{err}+1)L^2}$, $\rho = 1 - \frac{\nu^2}{(4\epsilon_{err}+1)L^2}$, $\epsilon_{err} \leq O(\frac{k}{q})$, $k \geq \frac{\rho^2 k^*}{(1-\rho^2)^2}$

Sufficient: (by inspection) valid q s.t. $\rho\gamma < 1$: $q \geq 2(3k+2)$

Necessary: minimum q so that $\exists$ valid k s.t. $\rho\gamma < 1$:

$$q \geq 4\kappa^2\sqrt{\frac{k^* d}{s_2}} > 1$$

Tuning q: gradient error vs. expansivity



(a) $f(\boldsymbol{x})$ (b) $\|\boldsymbol{x} - \boldsymbol{x}^*\|$

Figure: Sensitivity analysis

Toy XP: $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{a} \odot (\boldsymbol{x} - \boldsymbol{b})\|^2$, ($\boldsymbol{a}$ and $\boldsymbol{b}$ chosen to have $\|\nabla f(\boldsymbol{x}^*)\|$ small enough). $\eta = \frac{1}{(4\varepsilon_F + 1)}$. For $q = 1$ and 20, $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|$ does not converge.

Improvements: sampling along a random support

$\boldsymbol{u}_i \sim \text{Uni}\left(\mathbb{S}_{S}\right)$ with $S \sim \text{Uni}(\binom{[d]}{s_2})$

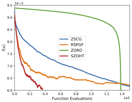
- memory efficiency (if distributed learning)
- allows to work with "restricted smoothness" only
- can improve the condition number ν/L

Experiments

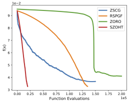
- Asset management [12], (a), (b), (c) :

$$\min_{x \in \mathbb{R}^d} \frac{\boldsymbol{x}^\top \boldsymbol{C}\boldsymbol{x}}{2\left(\sum_{i=1}^d \boldsymbol{x}_i\right)^2} + \lambda \left(\min \left\{ \frac{\sum_{i=1}^d \boldsymbol{m}_i \boldsymbol{x}_i}{\sum_{i=1}^d \boldsymbol{x}_i} - r, 0 \right\} \right)^2 \quad \text{s.t.} \quad \|\boldsymbol{x}\|_0 \leq k$$

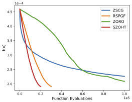
- Few pixels adv. attacks [13], (d), (e), (f) : $\min_\delta f(\boldsymbol{x} + \boldsymbol{\delta})$ such that $\|\boldsymbol{\delta}\|_0 \leq k$
- Comparison with ZSCG [14], RSPGF [15], ZORO [6] ($f(\boldsymbol{x})$ vs QC)



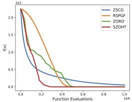
(a) port3



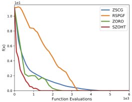
(b) port4



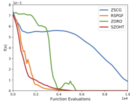
(c) port5



(d) MNIST



(e) CIFAR



(f) Imagenet

Iterative Regularization with k-support norm: A Dual Perspective on Hard-Thresholding

Original Goal: Online ℓ_0 optimization.
Attempt 1: Modify Online Convex Optimization ([16])

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

Problem: Could not get sublinear regret

Dual Perspective on IHT (contd.)

Attempt 2: Dual Averaging[17]/(Lazy) Mirror Descent[18]/Lazy OCO[16]/Bregman Iterations [19]:

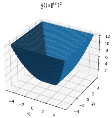
$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t - \eta_t \nabla f(\boldsymbol{x}_t)$$

$$\boldsymbol{x}_{t+1} = \Phi_k(\boldsymbol{y}_{t+1})$$

Problem: $\Phi_k(\cdot) = \partial\phi(\cdot)$ with $\phi(\cdot) = \frac{1}{2}(\|\cdot\|^{(k)})^2$ (top-k norm). ϕ not smooth...(proof cannot work)

But we can take the δ-Moreau smoothing, to get:

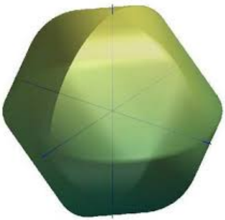
$$\phi_\delta(\cdot) = \left(\frac{1}{2}\Big(\underbrace{\|\cdot\|_k^{sp}}_{k\text{-support norm (KSN)}} \Big)^2 + \frac{1}{2}(\|\cdot\|_2^2) \right)^*$$



Note on the k-support norm

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

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



Dual Perspective on IHT (contd.)

Algorithm becomes:

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

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

Some properties: (not just online)

- Not really new now (MD/DA/BI, with just new use of KSN)
- $\boldsymbol{x}_t$ empirically "almost" sparse (prelim. XPs)
- BUT: MD, so convergence to $\boldsymbol{x}^*$ (maybe not sparse)
- **For overparam. linear models**: **implicit bias** towards min KSN^2 $(+\delta\ell_2^2)$ solution
- BUT: may still not be k-sparse

IRKSN

We simplify the problem, to make proofs easier $\rightarrow$ **sparse recovery**:

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

,

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

Solved by a tweaked version of ADGD [20], solving, **with early stopping**

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

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

Algorithm

Algorithm 3: IRKSN

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

Output: $\hat{\boldsymbol{w}}_T$

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

$$\hat{\boldsymbol{w}}_t \leftarrow \mathrm{prox}_{\alpha^{-1}F}\left(-\alpha^{-1}\boldsymbol{X}^T\hat{\boldsymbol{z}}_{\boldsymbol{t}}\right)$$
$$\hat{\boldsymbol{r}}_t \leftarrow \mathrm{prox}_{\alpha^{-1}F}\left(-\alpha^{-1}\boldsymbol{X}^T\hat{\boldsymbol{v}}_t\right)$$
$$\hat{\boldsymbol{z}}_t \leftarrow \hat{\boldsymbol{v}}_t + \gamma\left(\boldsymbol{X}\hat{\boldsymbol{r}}_{\boldsymbol{t}} - \boldsymbol{y}^\delta\right)$$
$$\boldsymbol{x}_{t+1} \leftarrow \left(1 + \sqrt{1 + 4\boldsymbol{x}_t^2}\right)/2$$
$$\hat{\boldsymbol{v}}_{t+1} = \hat{\boldsymbol{z}}_t + \frac{\boldsymbol{x}_t - 1}{\boldsymbol{x}_{t+1}}\left(\hat{\boldsymbol{z}}_t - \hat{\boldsymbol{z}}_{t-1}\right)$$

end

Recovery

Assumption

$\boldsymbol{w}^*$ *k-sparse, supp*$(\boldsymbol{w}^*) = S \subset [d]$, $\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{y}$,
$\boldsymbol{w}_S^* = \arg\min_{\boldsymbol{z}\in\mathbb{R}^k : \boldsymbol{X}_S \boldsymbol{z} = \boldsymbol{y}} \|\boldsymbol{z}\|_2$

$$\max_{\ell\in\bar{S}} |\langle \boldsymbol{X}_S^\dagger \boldsymbol{x}_\ell, \boldsymbol{w}_S^*\rangle| < \min_{j\in S} |\langle \boldsymbol{X}_S^\dagger \boldsymbol{x}_j, \boldsymbol{w}_S^*\rangle|$$

Theorem (Early Stopping Bound)

$$\|\hat{\boldsymbol{w}}_t - \boldsymbol{w}^*\|_2 \leq at\delta + bt^{-1}$$

$$\textit{with} \quad a = 4\|\boldsymbol{X}\|^{-1} \quad \textit{and} \quad b = \frac{2\|\boldsymbol{X}\|\|(\boldsymbol{X}_S^\top)^\dagger \boldsymbol{w}_S^*\|}{\alpha}$$

Comparison with ℓ_1-based recovery

Assumption (Recovery with ℓ_1 norm.)

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. $\max_{\ell \in \bar{S}} |\langle \boldsymbol{X}_S^{\dagger} \boldsymbol{x}_\ell, \mathrm{sgn}(\boldsymbol{w}_S^*) \rangle| < 1$

If $\boldsymbol{X}_S$ is injective

IRKSN

$$\max_{\ell \in \bar{S}} |\langle \boldsymbol{X}_S^{\dagger} \boldsymbol{x}_\ell, \frac{\boldsymbol{w}_S^*}{\min_{j \in S} |\boldsymbol{w}_S^*|} \rangle| < 1$$

l1 iter. reg.

$$\max_{\ell \in \bar{S}} |\langle \boldsymbol{X}_S^{\dagger} \boldsymbol{x}_\ell, \mathrm{sgn}(\boldsymbol{w}_S^*) \rangle| < 1$$

Variance Reduction

VR-SZHT introduced by Xinzhe Yuan:

Algorithm 4: Stochastic variance reduced zeroth-order Hard-Thresholding (VR-SZHT)

Initialization: η, T, $\boldsymbol{x}^0$, *SVRG update frequency* m, q, k.
Output: $\boldsymbol{x}^T$.

for $r = 1, \ldots, T$ **do**
  $\boldsymbol{x}^{(0)} = \boldsymbol{x}^{r-1}$; $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla} f_i(\boldsymbol{x}^{(0)})$; **for** $t = 0, 1, \ldots, m-1$ **do**
    Randomly sample $i_t \in \{1, 2, \ldots, n\}$; Compute ZO estimate $\hat{\nabla} f_{i_t}(\boldsymbol{x}^{(t)})$, $\hat{\nabla} f_{i_t}(\boldsymbol{x}^{(0)})$;
    $\bar{\boldsymbol{x}}^{(t+1)} = \boldsymbol{x}^{(t)} - \eta(\hat{\nabla} f_{i_t}(\boldsymbol{x}^{(t)}) - \hat{\nabla} f_{i_t}(\boldsymbol{x}^{(0)}) + \hat{\mu})$;
    $\boldsymbol{x}^{(t+1)} = \phi_k(\bar{\boldsymbol{x}}^{(t+1)})$;
  end
  $\boldsymbol{x}^r = \boldsymbol{x}^{(m)}$;
end

VR-SZHT

Removes need for minimum q: VR can compensate the variance of gradient estimate

Additional Constraints

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \mathcal{F}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}), \quad s.t. \ \|\boldsymbol{x}\|_0 \leq k \text{ and } \boldsymbol{x} \in \mathcal{S}. \tag{1}$$

Useful e.g. in **adversarial attacks**.

Structural sparsity

We may want to enforce constraints of the form:

$$\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}_{\mathcal{G}_1}\|_0 < D_1 \wedge \|\boldsymbol{x}_{\mathcal{G}_2}\|_0 < D_2\}$$

$\boldsymbol{x}_{\mathcal{G}_1}$ and $\boldsymbol{x}_{\mathcal{G}_1}$: a partition of $\boldsymbol{x}$ into coordinates from a group $\mathcal{G}_1$ and a group $\mathcal{G}_2$.

Reinforcement learning

Salimans et al. [21]: Evolution Strategies ($\approx$ ZO) very efficient for RL (esp. distributed)

BUT: dependence on d ($d >> 1$ for DNNs).

$\Longrightarrow$ Could using ZOHT reduce the dependence in d?

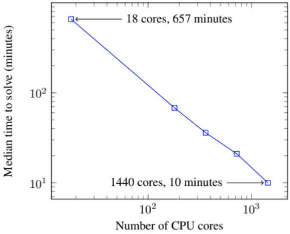


Figure 1: Time to reach a score of 6000 on 3D Humanoid with different number of CPU cores. Experiments are repeated 7 times and median time is reported.

Others

Other potential ideas:

- **Low-rank Matrices**
- **Sparse graphs**
- **Acceleration of ZOHT** [22]
- **Relaxed Assumptions** (non-RSC)
- **Lower bound** for ZOO with sparse optima
- **Non-convex regularization:** $h(\boldsymbol{x}) = \lambda(\frac{1}{2}(\|\boldsymbol{x}\|_k^{sp})^2 - \frac{1}{2}\|\boldsymbol{x}\|_2^2)$

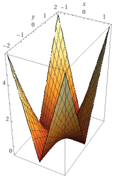


Figure: Nonconvex penalty based on the k-support norm.

QA

References I

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[3] 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.

References II

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

[5] 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.

[6] 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.

References III

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

[8] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[9] K. Balasubramanian and S. Ghadimi, "Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points," *Foundations of Computational Mathematics*, pp. 1–42, 2021.

References IV

[10] J. Francis, B. Madathil, S. N. George, and S. George, "A robust tensor-based submodule clustering for imaging data using l 1 2 regularization and simultaneous noise recovery via sparse and low rank decomposition approach," *Journal of Imaging*, vol. 7, no. 12, p. 279, 2021.

[11] S. Sykora, "Surface integrals over n-dimensional spheres," *Stan's Library*, no. Volume I, May 2005. DOI: 10.3247/sl1math05.002.

[12] T.-J. Chang, N. Meade, J. E. Beasley, and Y. M. Sharaiha, "Heuristics for cardinality constrained portfolio optimisation," *Computers & Operations Research*, vol. 27, no. 13, pp. 1271–1302, 2000.

References V

[13] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.

[14] K. Balasubramanian and S. Ghadimi, "Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

References VI

[15] S. Ghadimi, G. Lan, and H. Zhang, "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization," *Mathematical Programming*, vol. 155, no. 1, pp. 267–305, 2016.

[16] 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.

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

References VII

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

[19] 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.

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

References VIII

[21] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, "Evolution strategies as a scalable alternative to reinforcement learning," *arXiv preprint arXiv:1703.03864*, 2017.

[22] K. Axiotis and M. Sviridenko, "Iterative hard thresholding with adaptive regularization: Sparser solutions without sacrificing runtime," in *International Conference on Machine Learning*, PMLR, 2022, pp. 1175–1197.