Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Sister Conferences Best Papers Track

Efficient Convex Optimization Requires Superlinear Memory (Extended Abstract)*

Annie Marsden[1], **Vatsal Sharan**[2], **Aaron Sidford**[1] and **Gregory Valiant**[1]
[1]Stanford University
[2]University of Southern California
marsden@stanford.edu, vsharan@usc.edu, {sidford, valiant}@stanford.edu

Abstract

Minimizing a convex function with access to a first order oracle—that returns the function evaluation and (sub)gradient at a query point—is a canonical optimization problem and a fundamental primitive in machine learning. Gradient-based methods are the most popular approaches used for solving the problem, owing to their simplicity and computational efficiency. These methods, however, do not achieve the information-theoretically optimal query complexity for minimizing the underlying function to small error, which are achieved by more expensive techniques based on cutting-plane methods. Is it possible to achieve the information-theoretically query complexity without using these more complex and computationally expensive methods? In this work, we use memory as a lens to understand this, and show that is is not possible to achieve optimal query complexity without using significantly more memory than that used by gradient descent.

1 Introduction

Machine learning is intricately linked with continuous optimization, and gradient-based optimization methods are the main workhorse of modern machine learning. In this work, our goal is to understand how computational considerations affect the ability to efficiently optimize a function. A natural starting point to examine this is the canonical setting of minimizing a convex function $f : \mathbb{R}^d \to \mathbb{R}$ given access to a *first-order oracle*. A first-order oracle returns the function evaluation and (sub)gradient $(f(\mathbf{x}), \nabla f(\mathbf{x}))$ when queried at any point $\mathbf{x}$. The goal is to understand how many queries to the first-order oracle are necessary to minimize the function.

Understanding the first-order query complexity for minimizing convex functions has been foundational in optimization theory [Nemirovski and Yudin, 1983]. There are methods that, given any 1-Lipschitz, convex $f : \mathbb{R}^d \to \mathbb{R}$ accessible via a first-order oracle, compute an ϵ-approximate minimizer over the unit ball with just $\mathcal{O}(\min\{\epsilon^{-2}, d\log(1/\epsilon)\})$ first order queries. This query complexity is known to be worst-case optimal [Nemirovski and Yudin, 1983].

*This paper was initially published in the 35th Annual Conference on Learning Theory (COLT 2022).

$\mathcal{O}(\epsilon^{-2})$ queries is achievable using *(sub)gradient descent*. Variants of gradient-descent are the most widely used optimization methods for modern machine learning settings, owing to their simplicity and computational efficiency. In particular, subgradient descent solves the problem using a total of $\mathcal{O}(d\epsilon^{-2})$ computation time (assuming arithmetic operations on $\mathcal{O}(\log(d/\epsilon))$-bit numbers take constant time), and only requires $\mathcal{O}(d\log(1/\epsilon))$-bits of memory.

On the other hand, building on the $O(d^2\log(1/\epsilon))$ query complexity of the well-known *ellipsoid method* [Yudin and Nemirovskii, 1976; Shor, 1977], different *cutting plane methods* achieve a query complexity of $\mathcal{O}(d\log(1/\epsilon))$, e.g. center of mass with sampling based techniques [Levin, 1965; Bertsimas and Vempala, 2004], volumetric center [Vaidya, 1989; Atkinson and Vaidya, 1995], inscribed ellipsoid [Khachiyan *et al.*, 1988; Nesterov, 1989]; these methods are perhaps less frequently used in practice and large-scale learning. This is due to the fact that they are more complex than simple gradient descent based approaches, they all use at least $\Omega(d^3\log(1/\epsilon))$-time and $\Omega(d^2\log(1/\epsilon))$ bits of memory.

Though state-of-the-art cutting plane methods have larger computational overhead compared to gradient descent and are sometimes regarded as impractical in different settings, for small enough ϵ, they give the state-of-the-art query bounds. Further, in different theoretical settings, e.g. semidefinite programming [Anstreicher, 2000], submodular optimization [McCormick, 2005] and equilibrium computation [Papadimitriou and Roughgarden, 2008], cutting-plane-methods have yielded state-of-the-art runtimes at various points of time. *This leads to the natural question of what is needed of a method to significantly outperform gradient descent and take advantage of the improved query complexity enjoyed by cutting plane methods? Can we design methods that obtain optimal query complexities while maintaining the practicality of gradient descent methods?*

In this work, we take memory as a lens to understand the computational complexity of first-order convex optimization. Not only is memory usage one of the most fundamental measures of computational complexity for an algorithm, memory considerations can also play a crucial role in contemporary learning and optimization settings. In addition, as we show in this work, using memory as a lens offers the possibility of proving a clean *unconditional* separation between simple/complex techniques—as also judged by other measures

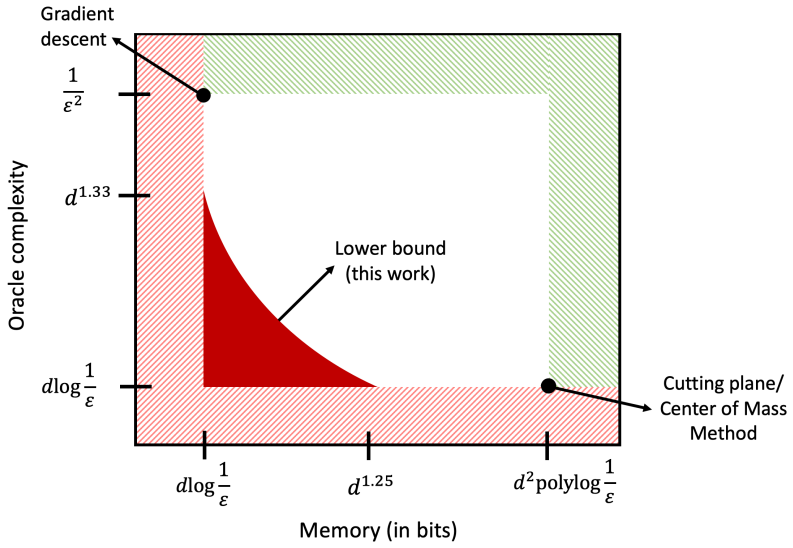

Figure 1: Tradeoffs between available memory and first-order oracle complexity for minimizing 1-Lipschitz convex functions over the unit ball (adapted from [Woodworth and Srebro, 2019]). The dashed red region corresponds to information-theoretic lower bounds on the memory and query-complexity. The dashed green region corresponds to known upper bounds. This work shows that the solid red region is not achievable for any algorithm.

such as their running time. This perspective on the role of memory in optimization is also well-articulated in the open-problem paper of [Woodworth and Srebro, 2019].

We show that memory plays a critical role in attaining optimal query complexity for convex optimization. Our main result is the following theorem which shows that any algorithm whose memory usage is sufficiently small (though still superlinear) must make polynomially more queries to a first-order oracle than cutting plane methods. Specifically, any algorithm that uses significantly less than $d^{1.25}$ bits of memory requires a polynomial factor more first order queries than the optimal $\mathcal{O}(d\log(d))$ queries achieved by quadratic memory cutting plane methods.

Theorem 1 *For some $\epsilon \geq 1/\operatorname{poly}(d)$ and any $\delta \in [0, 1/4]$ the following is true: any algorithm which outputs an ϵ-optimal point with probability at least $2/3$ given first order oracle access to any 1-Lipschitz convex function must use either at least $d^{1.25-\delta}$ bits of memory or make $\tilde{\Omega}(d^{1+\frac{4}{3}\delta})$ first order queries (where the $\tilde{\Omega}$ notation hides poly-logarithmic factors in d).*

Beyond shedding light on the complexity of a fundamental memory-constrained optimization problem, we provide several tools for establishing such lower bounds. In particular, we introduce a set of properties which are sufficient for an optimization problem to exhibit a memory-lower bound and provide an information-theoretic framework to prove these lower bounds. We hope these tools are an aid to future work on the role of memory in optimization.

This work fits within the broader context of understanding fundamental resource tradeoffs for optimization and learning. For many settings, establishing (unconditional) query/time or memory/time tradeoffs is notoriously hard—perhaps akin to P vs NP (e.g. providing time lower bounds for cutting plane methods). Questions of memory/query and memory/data tradeoffs, however, have a more information theoretic nature and hence seem more approachable. Together with the increasing importance of memory considerations in large-scale optimization and learning, there is a strong case for pinning down the landscape of such tradeoffs, which may offer a new perspective on the current suite of algorithms and inform the effort to develop new ones.

1.1 Technical Overview and Contributions

To prove Theorem 1, we provide an explicit distribution over functions that is hard for any memory-constrained randomized algorithm to optimize. Though the proof requires care and we introduce a variety of machinery to obtain it, this lower bounding family of functions is simple to state. The function is a variant of the so-called "Nemirovski" function, which has been used to show lower bounds for highly parallel non-smooth convex optimization [Nemirovski, 1994; Bubeck *et al.*, 2019; Balkanski and Singer, 2018].

Formally, our difficult class of functions for memory size M is constructed as follows: for some $\gamma > 0$ and some $N = \tilde{\mathcal{O}}\left(d^2/M\right)$ let $\mathbf{v}_1, \ldots, \mathbf{v}_N$ be unit vectors drawn i.i.d. from the d dimensional scaled hypercube $\mathbf{v}_i \overset{\text{i.i.d.}}{\sim} \text{Unif}\left(d^{-1/2}\mathcal{H}_d\right)$ and let $\mathbf{a}_1, \ldots, \mathbf{a}_{\lfloor d/2 \rfloor}$ be drawn i.i.d. from the hypercube, $\mathbf{a}_j \sim \text{Unif}\left(\mathcal{H}_d\right)$ where $\alpha\mathcal{H}_d := \{\pm\alpha\}^d$. Let $\mathbf{A} = \left(\mathbf{a}_1, \ldots, \mathbf{a}_{\lfloor d/2 \rfloor}\right)$ and define

$$F(\mathbf{x}) = (1/d^6)\max\left\{d^5\left\|\mathbf{A}\mathbf{x}\right\|_\infty - 1, \max_{i\in[N]} \mathbf{v}_i^\top \mathbf{x} - i\gamma\right\}.$$

Rather than give a direct proof of Theorem 1 using this explicit function we provide a more abstract framework which gives broader insight into which kinds of functions could lead

to non-trivial memory-constrained lower bounds, and which might lead to tighter lower bounds in the future. To that end we introduce the notion of a *memory-sensitive class* which delineates the key properties of a distribution over functions that lead to memory-constrained lower bounds. We show that for such functions, the problem of memory constrained optimization is at least as hard as the following problem of finding a set of vectors which are approximately orthogonal to another set of vectors:

Definition 2 (Orthogonal Vector Game, informal version) *Given $\mathbf{A} \in \{\pm 1\}^{d/2 \times d}$, the Player's objective is to return a set of k vectors $\{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ which satisfy*

1. *$\forall i \in [k], \mathbf{y}_i$ is approximately orthogonal to all the rows of $\mathbf{A}$: $\|\mathbf{A}\mathbf{y}_i\|_\infty / \|\mathbf{y}_i\|_2 \leq d^{-4}$.*
2. *The set of vectors $\{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ is robustly linearly independent: denoting $S_0 = \emptyset, S_i = \mathsf{span}(\mathbf{y}_1, \ldots, \mathbf{y}_i)$, $\left\|\mathrm{Proj}_{S_{i-1}}(\mathbf{y}_i)\right\|_2 / \|\mathbf{y}_i\|_2 \leq 1 - 1/d^2$,*

where the notation $\mathrm{Proj}_S(\mathbf{x})$ denotes the vector in the subspace S which is closest in $\|\cdot\|_2$ to $\mathbf{x}$. The game proceeds as follows: The Player first gets to observe $\mathbf{A}$ and store a M-bit long Message *about $\mathbf{A}$. She does not subsequently have free access to $\mathbf{A}$, but can adaptively make up to m queries as follows: for $i \in [m]$, based on* Message *and all previous queries and their results, she can request any row $i \in [d/2]$ of the matrix $\mathbf{A}$. Finally, she outputs a set of k vectors as a function of* Message *and all m queries and their results.*

Note that the Player can trivially win the game for $M \geq \Omega(dk), m = 0$ (by just storing a satisfactory set of k vectors in the Message) and for $M = 0, m = d/2$ (by querying all rows of $\mathbf{A}$). We show a lower bound that this is essentially all that is possible: *for $\mathbf{A}$ sampled uniformly at random from $\{\pm 1\}^{d/2 \times d}$, if M is a constant factor smaller than dk, then the Player must make at least $d/5$ queries to win with probability at least $2/3$.* Our analysis proceeds via an intuitive information-theoretic framework, which could have applications for showing query lower bounds for memory-constrained algorithms in other optimization and learning settings.

2 Related Work

2.1 Memory-sample Tradeoffs for Learning

There is a recent line of work to understand learning under information constraints such as limited memory or communication constraints [Balcan *et al.*, 2012; Duchi *et al.*, 2013; Zhang *et al.*, 2013; Garg *et al.*, 2014; Shamir, 2014; Arjevani and Shamir, 2015; Steinhardt and Duchi, 2015; Steinhardt *et al.*, 2016; Braverman *et al.*, 2016; Dagan and Shamir, 2018; Dagan *et al.*, 2019; Woodworth *et al.*, 2021]. Most of these results obtain lower bounds for the regime when the available memory is less than that required to store a single datapoint (with the notable exception of [Dagan and Shamir, 2018] and [Dagan *et al.*, 2019]). However the breakthrough paper [Raz, 2017] showed an exponential lower bound on the number of random examples needed for learning parities with memory as large as quadratic. Subsequent work extended and refined this result to multiple learning problems over finite fields [Moshkovitz and Moshkovitz, 2017; Beame *et al.*, 2018; Moshkovitz and Moshkovitz, 2018; Kol *et al.*, 2017; Raz, 2018; Garg *et al.*, 2018].

Most related to our line of work is [Sharan *et al.*, 2019], which considers the continuous valued learning/optimization problem of performing linear regression given access to randomly drawn examples from an isotropic Gaussian. They show that any sub-quadratic memory algorithm for the problem needs $\Omega(d \log\log(1/\epsilon)))$ samples to find an ϵ-optimal solution for $\epsilon \leq 1/d^{\Omega(\log d)}$, whereas in this regime an algorithm with memory $\tilde{\mathcal{O}}(d^2)$ can find an ϵ-optimal solution with only d examples. Since each example provides an unbiased estimate of the expected regression loss, this translates to a lower bound for convex optimization given access to a stochastic gradient oracle. However the upper bound of d examples is not a generic convex optimization algorithm/convergence rate but comes from the fact that the linear systems can be solved to the required accuracy using d examples.

There is also significant work on memory lower bounds for streaming algorithms, e.g. [Alon *et al.*, 1999; Bar-Yossef *et al.*, 2004; Clarkson and Woodruff, 2009; Dagan *et al.*, 2019], where the setup is that the algorithm only gets a single-pass over a data stream.

2.2 Lower Bounds for Convex Optimization

Starting with the early work of [Nemirovski and Yudin, 1983], there is extensive literature on lower bounds for convex optimization. Some of the key results in this area include classical lower bounds for finding approximate minimizers of Lipschitz functions with access to a subgradient oracle [Nemirovski and Yudin, 1983; Nesterov, 2003; Braun *et al.*, 2017], including recent progress on lower bounds for randomized algorithms [Woodworth and Srebro, 2016; Woodworth and Srebro, 2017; Simchowitz *et al.*, 2018; Simchowitz, 2018; Braverman *et al.*, 2020; Sun *et al.*, 2021]. For more details, we refer the reader to surveys such as [Nesterov, 2003] and [Bubeck, 2014].

2.3 Memory-limited Optimization Algorithms

While the focus of this work is lower bounds, there is a long line of work on developing memory-efficient optimization algorithms, including various techniques that leverage second-order structure via first-order methods such as Limited-memory-BFGS [Nocedal, 1980; Liu and Nocedal, 1989] and the conjugate gradient (CG) method for solving linear systems [Hestenes and Stiefel, 1952] and various nonlinear extensions of CG [Fletcher and Reeves, 1964] and methods based on subsampling and sketching the Hessian [Pilanci and Wainwright, 2017; Xu *et al.*, 2020].

3 Proof Strategy

We describe a broad family of optimization problems which may be sensitive to memory constraints (also see Fig. 2 for an overview). As suggested by the Orthogonal Vector Game (Definition 2), the primitive we leverage is that finding vectors orthogonal to the rows of a given matrix requires either

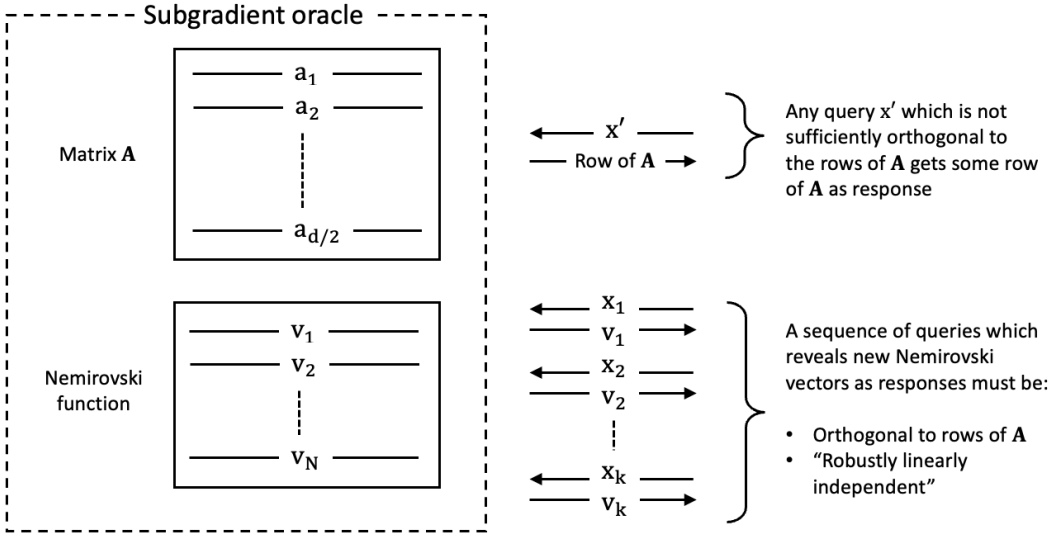

Figure 2: A high-level overview of our proof approach. The rows of $\mathbf{A}$ and the Nemirovski vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ are sampled uniformly at random from the hypercube. We show that this function class is "memory-sensitive" and has the following properties: (1) to successfully minimize the function, the algorithm must see a sufficiently large number of Nemirovski vectors, (2) to reveal new Nemivoski vectors, an algorithm must make queries which are robustly linearly independent and orthogonal to $\mathbf{A}$. Using these properties, we show that minimizing the function is at least as hard as winning the Orthogonal Vector Game (Definition 2) about N/k times. We then show memory-query tradeoffs for the Orthogonal Vector Game.

large memory or many queries to observe the rows of the matrix. With that intuition in mind, let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, let $\mathbf{A} \in \mathbb{R}^{n \times d}$, let η be a scaling parameter, and let ρ be a shift parameter; define $F_{f,\mathbf{A},\eta,\rho}(\mathbf{x})$ as the maximum of $f(\mathbf{x})$ and $\eta \left\|\mathbf{A}\mathbf{x}\right\|_\infty - \rho$:

$$F_{f,\mathbf{A},\eta,\rho}(\mathbf{x}) \coloneqq \max\left\{f(\mathbf{x}), \eta \left\|\mathbf{A}\mathbf{x}\right\|_\infty - \rho\right\}. \tag{3.1}$$

We often drop the dependence of η and ρ and write $F_{f,\mathbf{A},\eta,\rho}$ simply as $F_{f,\mathbf{A}}$. Intuitively, for large enough scaling η and appropriate shift ρ, minimizing the function $F_{f,\mathbf{A}}(\mathbf{x})$ requires minimizing $f(\mathbf{x})$ close to the null space of the matrix $\mathbf{A}$. Any algorithm which uses memory $\Omega(nd)$ can learn and store $\mathbf{A}$ in $\mathcal{O}(d)$ queries so that all future queries are sufficiently orthogonal to $\mathbf{A}$; thus this memory rich algorithm can achieve the information-theoretic lower bound for minimizing $f(\mathbf{x})$ roughly constrained to the nullspace of $\mathbf{A}$.

However, if $\mathbf{A}$ is a random matrix with sufficiently large entropy then $\mathbf{A}$ cannot be compressed to fewer than $\Omega(nd)$ bits. Thus, for $n = \Omega(d)$, an algorithm which uses only memory $o(d^{2-\delta})$ bits for some constant $0 < \delta \leq 1$ cannot remember all the information about $\mathbf{A}$. Suppose the function f is such that in order to continue to observe new information about the function, it is insufficient to submit queries that belong to some small dimensional subspace of the null space of $\mathbf{A}$. Then a memory constrained algorithm must re-learn enough information about $\mathbf{A}$ in order to find new vectors in the null space of $\mathbf{A}$ and make queries which return new information about f.

To show our lower bound, we take f in (3.1) to be the Nemirovski function:

$$f(\mathbf{x}) = \max_{i \in [N]} \left(\mathbf{v}_i^\top \mathbf{x} - i\gamma\right)$$

where the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_N$ are unit vectors drawn i.i.d. from the d dimensional scaled hypercube $\mathbf{v}_i \overset{\text{i.i.d.}}{\sim} \text{Unif}\left(d^{-1/2}\mathcal{H}_d\right)$ and we refer to them as "Nemirovski vectors". With this choice of f, the overall function $F_{f,\mathbf{A}}(\mathbf{x})$ has certain "memory-sensitive" properties. In particular, to reveal new Nemirovski vectors an algorithm cannot make queries which lie in a low-dimensional subspace. In addition, because of the $\left\|\mathbf{A}\mathbf{x}\right\|_\infty$ term in the definition of $F_{f,\mathbf{A}}(\mathbf{x})$, queries which reveal new Nemirovski vectors must also be sufficiently orthogonal to $\mathbf{A}$. Together, these properties imply that a sequence of queries which reveals a new set of Nemirovski vectors must also be winning queries for the Orthogonal Vector Game (Definition 2). This allows us to leverage our information-theoretic memory-query tradeoffs for the Orthogonal Vector Game. We show that the algorithm must reveal a sufficiently large number of Nemirovski vectors to optimize $F_{f,\mathbf{A}}(\mathbf{x})$, therefore we can repeatedly apply the lower bound for the Orthogonal Vector Game to show our final lower bound.

Acknowledgments

AM and GV were supported by NSF Awards CCF-1704417, CCF-1813049, Frontier Award 1804222 and DOE award DE-SC0019205. AM was also supported by J. Duchi's Office of Naval Research Award YIP N00014-19-2288 and NSF Award HDR 1934578. VS was supported by NSF CAREER Award CCF-2239265 and a Amazon Research Award. AS was supported by a Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-1844855, NSF Grant CCF-1955039, a PayPal research award, and a Sloan Research Fellowship.

References

[Alon *et al.*, 1999] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.

[Anstreicher, 2000] Kurt M Anstreicher. The volumetric barrier for semidefinite programming. *Mathematics of Operations Research*, 25(3):365–380, 2000.

[Arjevani and Shamir, 2015] Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. *Advances in neural information processing systems*, 28, 2015.

[Atkinson and Vaidya, 1995] David S Atkinson and Pravin M Vaidya. A cutting plane algorithm for convex programming that uses analytic centers. *Mathematical Programming*, 69(1):1–43, 1995.

[Balcan *et al.*, 2012] Maria Florina Balcan, Avrim Blum, Shai Fine, and Yishay Mansour. Distributed learning, communication complexity and privacy. In *Conference on Learning Theory*, pages 26–1. JMLR Workshop and Conference Proceedings, 2012.

[Balkanski and Singer, 2018] Eric Balkanski and Yaron Singer. Parallelization does not accelerate convex optimization: Adaptivity lower bounds for non-smooth convex minimization. *arXiv preprint arXiv:1808.03880*, 2018.

[Bar-Yossef *et al.*, 2004] Ziv Bar-Yossef, Thathachar S Jayram, Ravi Kumar, and D Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[Beame *et al.*, 2018] Paul Beame, Shayan Oveis Gharan, and Xin Yang. Time-space tradeoffs for learning finite functions from random evaluations, with applications to polynomials. In *Conference On Learning Theory*, pages 843–856. PMLR, 2018.

[Bertsimas and Vempala, 2004] Dimitris Bertsimas and Santosh Vempala. Solving convex programs by random walks. *Journal of the ACM (JACM)*, 51(4):540–556, 2004.

[Braun *et al.*, 2017] Gábor Braun, Cristóbal Guzmán, and Sebastian Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Transactions on Information Theory*, 63(7), 2017.

[Braverman *et al.*, 2016] Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020, 2016.

[Braverman *et al.*, 2020] Mark Braverman, Elad Hazan, Max Simchowitz, and Blake Woodworth. The gradient complexity of linear regression. In *Conference on Learning Theory*, pages 627–647. PMLR, 2020.

[Bubeck *et al.*, 2019] Sébastien Bubeck, Qijia Jiang, Yin-Tat Lee, Yuanzhi Li, and Aaron Sidford. Complexity of highly parallel non-smooth convex optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[Bubeck, 2014] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

[Clarkson and Woodruff, 2009] Kenneth L Clarkson and David P Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 205–214, 2009.

[Dagan and Shamir, 2018] Yuval Dagan and Ohad Shamir. Detecting correlations with little memory and communication. In *Conference On Learning Theory*, pages 1145–1198. PMLR, 2018.

[Dagan *et al.*, 2019] Yuval Dagan, Gil Kur, and Ohad Shamir. Space lower bounds for linear prediction in the streaming model. In *Conference on Learning Theory*, pages 929–954. PMLR, 2019.

[Duchi *et al.*, 2013] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[Fletcher and Reeves, 1964] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, 7(2), 1964.

[Garg *et al.*, 2014] Ankit Garg, Tengyu Ma, and Huy Nguyen. On communication cost of distributed statistical estimation and dimensionality. *Advances in Neural Information Processing Systems*, 27, 2014.

[Garg *et al.*, 2018] Sumegha Garg, Ran Raz, and Avishay Tal. Extractor-based time-space lower bounds for learning. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 990–1002, 2018.

[Hestenes and Stiefel, 1952] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6), 1952.

[Khachiyan *et al.*, 1988] Leonid G Khachiyan, Sergei Pavlovich Tarasov, and II Erlikh. The method of inscribed ellipsoids. In *Soviet Math. Dokl*, volume 37, pages 226–230, 1988.

[Kol *et al.*, 2017] Gillat Kol, Ran Raz, and Avishay Tal. Time-space hardness of learning sparse parities. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1067–1080, 2017.

[Levin, 1965] Anatoly Yur'evich Levin. An algorithm for minimizing convex functions. In *Doklady Akademii Nauk*, volume 160, pages 1244–1247. Russian Academy of Sciences, 1965.

[Liu and Nocedal, 1989] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3), 1989.

[McCormick, 2005] S Thomas McCormick. Submodular function minimization. *Handbooks in operations research and management science*, 12:321–391, 2005.

[Moshkovitz and Moshkovitz, 2017] Dana Moshkovitz and Michal Moshkovitz. Mixing implies lower bounds for space bounded learning. In *Conference on Learning Theory*, pages 1516–1566. PMLR, 2017.

[Moshkovitz and Moshkovitz, 2018] Dana Moshkovitz and Michal Moshkovitz. Entropy samplers and strong generic lower bounds for space bounded learning. In *9th Innovations in Theoretical Computer Science Conference*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[Nemirovski and Yudin, 1983] Arkadij Nemirovski and David Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

[Nemirovski, 1994] Arkadi Nemirovski. On parallel complexity of nonsmooth convex optimization. *Journal of Complexity*, 10(4):451–463, 1994.

[Nesterov, 1989] Ju E Nesterov. Self-concordant functions and polynomial-time methods in convex programming. *Report, Central Economic and Mathematic Institute, USSR Acad. Sci*, 1989.

[Nesterov, 2003] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[Nocedal, 1980] Jorge Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151), 1980.

[Papadimitriou and Roughgarden, 2008] Christos H Papadimitriou and Tim Roughgarden. Computing correlated equilibria in multi-player games. *Journal of the ACM (JACM)*, 55(3):1–29, 2008.

[Pilanci and Wainwright, 2017] Mert Pilanci and Martin J Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

[Raz, 2017] Ran Raz. A time-space lower bound for a large class of learning problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science*, pages 732–742. IEEE, 2017.

[Raz, 2018] Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. *Journal of the ACM*, 66(1):1–18, 2018.

[Shamir, 2014] Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. *Advances in Neural Information Processing Systems*, 27, 2014.

[Sharan *et al.*, 2019] Vatsal Sharan, Aaron Sidford, and Gregory Valiant. Memory-sample tradeoffs for linear regression with small error. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 890–901, 2019.

[Shor, 1977] Naum Z Shor. Cut-off method with space extension in convex programming problems. *Cybernetics*, 13(1):94–96, 1977.

[Simchowitz *et al.*, 2018] Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. Tight query complexity lower bounds for pca via finite sample deformed wigner law. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1249–1259, 2018.

[Simchowitz, 2018] Max Simchowitz. On the randomized complexity of minimizing a convex quadratic function. *arXiv preprint arXiv:1807.09386*, 2018.

[Steinhardt and Duchi, 2015] Jacob Steinhardt and John Duchi. Minimax rates for memory-bounded sparse linear regression. In *Conference on Learning Theory*, pages 1564–1587. PMLR, 2015.

[Steinhardt *et al.*, 2016] Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In *Conference on Learning Theory*, pages 1490–1516. PMLR, 2016.

[Sun *et al.*, 2021] Xiaoming Sun, David P Woodruff, Guang Yang, and Jialin Zhang. Querying a matrix through matrix-vector products. *ACM Transactions on Algorithms (TALG)*, 17(4):1–19, 2021.

[Vaidya, 1989] Pravin M Vaidya. A new algorithm for minimizing convex functions over convex sets. In *30th Annual Symposium on Foundations of Computer Science*, pages 338–343. IEEE Computer Society, 1989.

[Woodworth and Srebro, 2016] Blake E Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. *Advances in neural information processing systems*, 29, 2016.

[Woodworth and Srebro, 2017] Blake Woodworth and Nathan Srebro. Lower bound for randomized first order convex optimization. *arXiv preprint arXiv:1709.03594*, 2017.

[Woodworth and Srebro, 2019] Blake Woodworth and Nathan Srebro. Open problem: The oracle complexity of convex optimization with limited memory. In *Conference on Learning Theory*, pages 3202–3210. PMLR, 2019.

[Woodworth *et al.*, 2021] Blake E Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *Conference on Learning Theory*, pages 4386–4437. PMLR, 2021.

[Xu *et al.*, 2020] Peng Xu, Fred Roosta, and Michael W Mahoney. Newton-type methods for non-convex optimization under inexact hessian information. *Mathematical Programming*, 184(1):35–70, 2020.

[Yudin and Nemirovskii, 1976] DB Yudin and Arkadi S Nemirovskii. Informational complexity and efficient methods for the solution of convex extremal problems. *Matekon*, 13(2):22–45, 1976.

[Zhang *et al.*, 2013] Yuchen Zhang, John C Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Conference on Neural Information Processing Systems*, pages 2328–2336. Citeseer, 2013.