Adaptive and Oblivious Randomized Subspace Methods for High-Dimensional Optimization: Sharp Analysis and Lower Bounds

Jonathan Lacotte, Mert Pilanci

arXiv:2012.07054v1 [cs.IT] 13 Dec 2020

Abstract

We propose novel randomized optimization methods for high-dimensional convex problems based on restrictions of variables to random subspaces. We consider oblivious and data-adaptive subspaces and study their approximation properties via convex duality and Fenchel conjugates. A suitable adaptive subspace can be generated by sampling a correlated random matrix whose second order statistics mirror the input data. We illustrate that the adaptive strategy can significantly outperform the standard oblivious sampling method, which is widely used in the recent literature. We show that the relative error of the randomized approximations can be tightly characterized in terms of the spectrum of the data matrix and Gaussian width of the dual tangent cone at optimum. We develop lower bounds for both optimization and statistical error measures based on concentration of measure and Fano's inequality. We then present the consequences of our theory with data matrices of varying spectral decay profiles. Experimental results show that the proposed approach enables significant speed ups in a wide variety of machine learning and optimization problems including logistic regression, kernel classification with random convolution layers and shallow neural networks with rectified linear units.

Index Terms

Convex optimization, Dimension reduction, Random subspaces, Randomized singular value decomposition, Kernel methods.

I. Introduction

HIGH-dimensional optimization problems are becoming ever more common in applications such as computer vision, natural language processing, robotics, medicine, genomics, seismology or weather forecasting, where the volume of the data keeps increasing at a rapid rate. It is also standard practice to use high-dimensional representations of data measurements such as random Fourier features [2] or pre-trained neural networks' features [3], [4]. In this work, we are interested in solving a convex optimization problem of the form

$$x^* := \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ F(x) := f(Ax) + \frac{\lambda}{2}\|x\|_2^2 \right\}, \tag{1}$$

where $A \in \mathbb{R}^{n\times d}$ is a data matrix and f is a convex function. Such convex optimization problems are typically formulated to fit a linear prediction model, or, they may occur as the subroutine of an optimization method, e.g., proximal optimization [5]; prox-linear method for convex-smooth composite objectives [6]. Moreover, several standard non-convex neural network training problems can be equivalently stated as convex optimization problem in higher dimensions [7]–[10]. In the large-scale setting

A preliminary version of this work was presented in part [1] at the Advances in Neural Information Processing Systems 32 (NeurIPS 2019).

J. Lacotte and M. Pilanci are with the Electrical Engineering Department at Stanford University.

$d \gg 1$, random projections are an effective way of performing dimensionality reduction [11]–[16], and the common practice is to employ oblivious sampling or sketching matrices, which are typically randomized and fixed ahead of time. Furthermore, sketches can be iteratively applied [17]–[20] or averaged in independent trials [21]–[23] to reduce the approximation error. However, it is not clear whether one can do better by adapting the sketching matrices to data. In fact, *we will show that adaptive sketching matrices can significantly improve the approximation quality of the optimal solution of a convex smooth optimization problem, and we will characterize the recovery error in terms of the smoothness of the objective function and the spectral decay of the data matrix. Furthermore, we establish lower-bounds on the performance of the oblivious sketch that exhibit its fundamental limitations.*.

Although the oracle complexity of first-order optimization methods has the property of being dimension-free [24], the cost of forming gradients and manipulating data matrices may be computationally prohibitive in large-scale settings, let alone second-order methods involving hessian computations. As a result, many sketching-based algorithms [20], [25]–[28] have been specifically designed to address the computational issues of the Newton method by reducing the cost of solving the linear Newton system. The Newton sketch algorithm [20] addresses the case where $n \gg d$ and f is separable, and solves an approximate Newton system based on a sketch of the data matrix. Stochastic Dual Newton Ascent [25] requires knowledge of a global upper bound on the Hessian and then solves an approximate Newton system using random principal sub-matrices of that upper bound. The Randomized Subspace Newton (RSN) method [28] uses, at each iteration, an approximate descent direction $S(S^\top HS)^\dagger S^\top g$ where $S^\top$ is a $m \times d$ random embedding with $m \ll d$ and H and g are respectively the Hessian and gradient at the current iterate. The Randomized Block Cubic Newton method [26] addresses block-separable convex optimization problems and combines the ideas of randomized coordinate descent [29] and cubic regularization [30].

In this work, we take a perspective which is agnostic to the optimization algorithm. Our goal is to formulate a low-dimensional optimization problem of the form

$$\alpha^* \in \operatorname*{argmin}_{\alpha \in \mathbb{R}^m} \left\{ F(S\alpha) = f(AS\alpha) + \frac{\lambda}{2}\|S\alpha\|_2^2 \right\}, \tag{2}$$

where S is a $d \times m$ random embedding, and then to construct a (potentially nonlinear) recovery map $\varphi : \mathbb{R}^m \to \mathbb{R}^d$ such that $\widehat{x} := \varphi(\alpha^*)$ is a close approximation of x^*, as measured by the relative recovery error

$$\frac{\|\widehat{x} - x^*\|_2}{\|x^*\|_2}. \tag{3}$$

The low-dimensional formulation (2) draws connections with the aforementioned randomized Newton methods: using the linear reconstruction map $\varphi(\alpha) = S\alpha$, the Newton method applied to (2) yields the update $\alpha_{t+1} = \alpha_t - \eta_t (S^\top \nabla^2 F(S\alpha_t) S)^\dagger S^\top \nabla F(S\alpha_t)$, i.e., $S\alpha_{t+1} = S\alpha_t - \eta_t S(S^\top \nabla^2 F(S\alpha_t) S)^\dagger S^\top \nabla F(S\alpha_t)$ which is the update of the RSN method. However, our perspective is different, as one is free to use any optimization method to solve for α^* and any reconstruction map φ.

Our approach falls within the scope of *random subspace optimization methods*, where we restrict the range of the optimization variable to a random lower-dimensional subspace. In the context of convex smooth optimization, the authors of [31], [32] propose to project the d-dimensional features of the data matrix A using an *oblivious* random embedding $S \in \mathbb{R}^{d \times m}$, chosen independently of the data. Based on the solution $\alpha_\dagger^*$ of a low-dimensional optimization problem similar to (2), they compute the dual solution $y^* = \nabla f(AS\alpha_\dagger^*)$ and then set $\widehat{x} := -\lambda^{-1} A^\top y^*$. Interestingly, the resulting recovery error is smaller than the error of the linear estimator $S\alpha_\dagger^*$, although these guarantees hold under some arguably restrictive assumptions (e.g., low-rank data matrix A, or, x^* approximately lies in the span of the top singular vectors of A). On the other hand, such a linear estimator

has been considered in [33] in the context of generalization bounds for empirical risk minimization of convex Lipschitz loss functions. They use an *adaptive* (data-dependent) random embedding S of the form $S = A^\top \widetilde{S}$, where $\widetilde{S}$ is itself oblivious. The authors study the generalization error of the approximate solution, and relate it to the norm of the tail singular values of A. Our approach draws connections with these two works: similarly to [33], we will consider adaptive random embeddings of the form $S = A^\top \widetilde{S}$; similarly to [31], [32], we will consider a non-linear dual mapping φ and study the recovery error of $\varphi(\alpha^*)$. As for the intriguing connection between the RSN method and the linear recovery map $\varphi(\alpha) = S\alpha$, the dual recovery map has been shown [34] to be equivalent to the Hessian sketch [35] applied to the Fenchel dual program.

The dual recovery map has also been analyzed in the specific context of sparse recovery. The authors of [31], [32] establish that, with an oblivious random embedding S, accurate recovery is guaranteed for a sketch size scaling at least as the sparsity level of x^* (i.e., its number of non-zero entries), under the assumption that the support set of x^* includes the most important coordinates of the data matrix (see Theorem 4 in [32]). In the context of support vector machines (SVM) classification where the dual solution may be sparse (as measured by the number of support vectors), the authors of [36], [37] propose to add a sparsity-promoting regularization term to the dual of the low-dimensional problem (2), and provide recovery guarantees in terms of the sparsity of the dual solution.

Besides accurate recovery in convex smooth optimization through a low-dimensional formulation, random subspace optimization with oblivious embeddings has been used and analyzed for SVM-classification and the preservation of margins [38]–[40], for large-scale trust region problems with oblivious embeddings [41], for scaling up linear systems solvers [42] and for statistically optimal prediction through kernel ridge regression [43]. In the latter work, the authors approximate an empirical kernel matrix K by sketching its columns with an oblivious embedding $\widetilde{S}$. For a kernel based on a finite-dimensional feature space, i.e., $K = \Phi\Phi^\top$, the sketch $K\widetilde{S}$ satisfies $K\widetilde{S} = \Phi S$ where $S = \Phi^\top \widetilde{S}$. Hence, this corresponds to a sketch of the data matrix Φ with the adaptive embedding S.

Adaptive embeddings of the form $S = A^\top \widetilde{S}$ (where $\widetilde{S}$ is itself oblivious) are reminiscent of randomized low-rank approximations methods [44]–[47]. More precisely, let us denote $P_S := S(S^\top S)^\dagger S^\top$ the linear projector onto the range of S and $P_S^\perp := I - P_S$ the projector onto its orthogonal complement. Then, for commonly used embeddings $\widetilde{S} \in \mathbb{R}^{n\times m}$ such as Gaussian or the subsampled randomized Hadamard transform (SRHT), it holds with high probability that

$$\|P_S^\perp A^\top\|_2 \lesssim \|A - A_m\|_2 + \frac{1}{\sqrt{m}}\|A - A_m\|_F\,, \tag{4}$$

where A_m is the best rank-m approximation of A. Unsurprisingly, the aforementioned generalization error guarantees based on adaptive sketching proposed in [33] depend on S through the critical quantity $\|P_S^\perp A^\top\|_2$, and so do the guarantees we develop in this manuscript. In other words, our randomized adaptive subspace optimization method is an approximate way of restricting the optimization variable to the span of the top m singular vectors of A. Naturally, using the deterministic embedding $S = V_m$, where V_m is the matrix of top m right singular vectors of A, could yield stronger guarantees than our adaptive sketch method. However, there are strong computational benefits in using a randomized embedding [44] instead of an exact singular value decomposition (SVD) algorithm, and the right combination of this randomized method along with the dual recovery map is not yet understood in the case of convex smooth optimization.

Our work relates to the considerable amount of literature on randomized approximations of high-dimensional kernel matrices K. A popular approach consists of building a low-rank factorization of the matrix K, using a random subset of its columns [48]–[51]. The so-called *Nystrom method* has proven to be effective empirically [52], and many research efforts have been devoted

to improving and analyzing the performance of its many variants (e.g., uniform column sub-sampling, leverage-score based sampling), especially in the context of regularized regression [53]–[55]. For conciseness of this manuscript, we will not consider explicitly column subsampling matrices, although most of our results may extend to them.

A. Notations and assumptions

We work with a data matrix $A \in \mathbb{R}^{n\times d}$ and we refer to n as the sample size and d the features' dimension. We denote by $\rho \leqslant \min\{n, d\}$ the rank of A, and by $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_\rho > 0$ its non-zero singular values. Given an embedding $S \in \mathbb{R}^{d\times m}$, we define the linear projector P_S onto the range of S and the linear projector $P_S^\perp$ onto the orthogonal complement of the range of S as

$$P_S := S(S^\top S)^\dagger S^\top\,, \qquad P_S^\perp := I_d - P_S\,, \tag{5}$$

where the superscript † denotes the pseudo-inverse. For a matrix $M \in \mathbb{R}^{p\times q}$ with arbitrary dimensions $p,\, q \geqslant 1$, we denote by $\|M\|_2$ its operator norm (i.e., its largest singular value), and by $\|M\|_F$ its Frobenius norm. For a real number $\eta \geqslant 1$, we denote the L_η-norm of a vector $w \in \mathbb{R}^p$ as $\|w\|_\eta = (\sum_{i=1}^p |w_i|^\eta)^{\frac{1}{\eta}}$, and the corresponding unit L_η-ball as $\mathcal{B}_\eta^p := \{w \in \mathbb{R}^p \mid \|w\|_\eta \leqslant 1\}$. We introduce several measures of the size of a set $T \subset \mathbb{R}^p$: its radius is $\mathrm{rad}(T) := \sup_{t\in T} \|t\|_2$, its diameter is $\mathrm{diam}(T) := \sup_{t,t'\in T} \|t - t'\|_2$, and its Gaussian width is

$$\omega(T) := \mathbb{E}_g \left\{ \sup_{t\in T} \langle t, g\rangle \right\}\,, \tag{6}$$

where $g \sim \mathcal{N}(0, I_p)$.

We work with a real-valued objective function f which is defined over $\mathbb{R}^n$. Unless stated otherwise, we will assume the function f to be convex, differentiable and μ-strongly smooth for some $\mu > 0$, i.e.,

$$\|\nabla f(y) - \nabla f(x)\|_2 \leqslant \mu \cdot \|y - x\|_2 \tag{7}$$

for any $x, y \in \mathbb{R}^n$. We introduce the convex (or Fenchel) conjugate of f, defined as

$$f^*(z) := \sup_{w\in\mathbb{R}^n} \{\langle w, z\rangle - f(w)\}\,. \tag{8}$$

We recall a few results from convex analysis (we refer the reader to the books [56], [57] for more background and details). The domain of f^* is defined as $\mathrm{dom}\, f^* := \{z \in \mathbb{R}^n \mid f^*(z) < +\infty\}$ and is a closed convex set. The function f^* is convex. Smoothness of f implies that f^* is $(1/\mu)$-strongly convex, i.e., $\langle y - z, g_y - g_z\rangle \geqslant \frac{1}{\mu}\cdot\|y - z\|_2^2$ for any $y,\, z \in \mathrm{dom}\, f^*$ and for any $g_y \in \partial f^*(y)$ and $g_z \in \partial f^*(z)$, where $\partial f^*(\cdot)$ denotes the subgradient operator of f^*. We denote by $\mathrm{int\,dom}\, f^*$ the interior of the domain of f^*.

We work with an arbitrary regularization parameter $\lambda > 0$. For comparing our guarantees with other methods, we will typically assume that $\frac{\lambda}{\mu}$ is within the range of the eigenvalues of the matrix $AA^\top$, i.e., $\frac{\lambda}{\mu} \approx \sigma_K^2$ for some threshold rank $1 \leqslant K \leqslant \rho$. This corresponds to a common (or desirable) choice of λ in practice, when the goal is either to improve the condition number of the primal program (1) for numerical stability purposes (a.k.a. Tikhonov regularization), or, to discard the effect of the small singular values of A (i.e., noise) for statistical estimation (a.k.a. ℓ_2-shrinkage).

For arbitrary dimensions $p \leqslant q$, we say that $U \in \mathbb{R}^{p\times q}$ is a (partial) Haar matrix in $\mathbb{R}^q$ if $UU^\top = I_p$ and the range of U is uniformly distributed among the p-dimensional subspaces of $\mathbb{R}^q$.

B. Randomized sketches

Given a dimension $p \in \{n, d\}$ and a *sketch size* $m \leqslant p$, we work with two oblivious classes of m-dimensional random embeddings $S \in \mathbb{R}^{p\times m}$, namely, Gaussian embeddings with independent and identically distributed (i.i.d.) Gaussian entries $S_{ij} \sim \mathcal{N}(0, 1/m)$, and, the SRHT [58]–[61], defined as $S = \sqrt{\frac{p}{m}} \cdot D \cdot H \cdot R$ where $R \in \mathbb{R}^{p\times m}$ is a column subsampling matrix, $D \in \mathbb{R}^{p\times p}$ is a diagonal matrix with independent entries uniformly sampled from $\{\pm 1\}$ and $H \in \mathbb{R}^{p\times p}$ is the Walsh-Hadamard transform. The k-th Walsh-Hadamard transform $H \equiv H_k$ is obtained by the recursion $H_0 = [1]$ and $H_{k+1} = \begin{bmatrix} H_k & H_k \\ H_k & -H_k \end{bmatrix}$, so that the dimension of H_k is $2^k \times 2^k$. This requires the dimension p to be a power of 2. If not, a standard practice for sketching a matrix M with p columns is to form the matrix $\widetilde{M} = [M, 0]$ which has an additional number of $\widetilde{p} - p$ columns filled with zeros, where $\widetilde{p}$ is the smallest power of 2 greater than p (note that $\widetilde{p} \leqslant 2p$), and to use the sketch $\widetilde{M}S$ with a SRHT $S \in \mathbb{R}^{\widetilde{p}\times m}$. For conciseness, in our formal statements involving a SRHT, we will implicitly assume that the relevant dimension p is a power of 2, although this does not restrict the applicability of our results thanks to the aforementioned zero-padding trick. In contrast to a Gaussian embedding, the SRHT verifies the orthogonality property $S^\top S = \frac{p}{m} \cdot I_m$. Furthermore, the recursive structure of the Walsh-Hadamard transform enables fast computation of a sketch MS, in time $\mathcal{O}(\text{nnz}(M) \log m)$ where $\text{nnz}(M)$ is the number of non-zero entries of M (see, for instance, [45] for details), as opposed to a Gaussian embedding which requires time $\mathcal{O}(\text{nnz}(M)m)$ (using classical matrix multiplication).

C. Oblivious vs Adaptive Sketches

Data adaptive sketches of the form $S = A^\top \widetilde{S}$, where $\widetilde{S}$ is an oblivious sketching matrix, aim to mirror the correlation structure of the original data matrix A. For instance, for a Gaussian embedding $\widetilde{S}$, we observe that the columns of $S = A^\top \widetilde{S}$ are distributed as $S_i \sim \mathcal{N}(0, A^T A)$. Furthermore, in many statistical applications, assuming that the rows of A are independent (and normalized) data sample vectors in $\mathbb{R}^d$, the matrix $A^T A$ corresponds to the $d \times d$ empirical covariance matrix. Therefore, one can expect the adaptive sketch to provide a more faithful summary of the data. However, this increased accuracy comes with the price of an additional pass over the dataset to compute the product $A^\top \widetilde{S}$, which is typically negligible.

II. An overview of our contributions

We introduce the Fenchel dual of the primal program (1),

$$z^* \in \operatorname*{argmin}_{z\in\mathbb{R}^n} \left\{ f^*(z) + \frac{1}{2\lambda}\|A^\top z\|_2^2 \right\}, \tag{9}$$

and the Fenchel dual of the sketched primal program (2),

$$y^* \in \operatorname*{argmin}_{y\in\mathbb{R}^n} \left\{ f^*(y) + \frac{1}{2\lambda}\|P_S A^\top y\|_2^2 \right\}. \tag{10}$$

Proposition 1 (Strong Fenchel duality)**.** *There exist a unique* primal *solution $x^* \in \mathbb{R}^d$ to* (1) *and a unique* dual *solution $z^* \in$ dom f^* to* (9)*, and these solutions are related through the Karush-Kuhn-Tucker (KKT) conditions $x^* = -\frac{A^\top z^*}{\lambda}$ and $z^* = \nabla f(Ax^*)$. If the function f is strictly convex, then $z^* \in$ int dom f^*.*

Proposition 2 (Strong Fenchel duality on sketched program)**.** *There exist a sketched primal solution $\alpha^* \in \mathbb{R}^m$ to* (2) *and a sketched dual solution $y^* \in$ dom f^* to* (10)*, and these solutions are related through the KKT conditions $S^\top S\alpha^* = -\frac{S^\top A^\top y^*}{\lambda}$ and $y^* = \nabla f(AS\alpha^*)$. If the function f is strictly convex, then $y^* \in$ int dom f^*.*

Strong Fenchel duality is critical to understand the influence of the right-sketch AS on the high-dimensional problem (1). By duality, the right-sketch is identical to employing a left-sketch $P_S A^T$ for the dual problem (10) where $P_S = S(S^T S)^\dagger S^T$ is the range space projector of S. As it will be shown, the sketch is significantly more accurate when the adaptive embedding $S = A^\top \widetilde{S}$ is employed.

A. Low-dimensional estimators

Based on a low-dimensional solution α^* to (2), we focus on two candidate estimators of x^*. The most natural candidate is given by the linear mapping

$$\widehat{x}^{(0)} := S\alpha^* \,, \tag{11}$$

which we refer to as the *zero-order* estimator. On the other hand, the first-order optimality conditions $x^* = G_\lambda(x^*)$ where $G_\lambda(x) := -\frac{1}{\lambda} A^\top \nabla f(Ax)$ suggest the estimator

$$\widehat{x}^{(1)} := G_\lambda(S\alpha^*) \tag{12}$$

and we refer to it as the *first-order* estimator of x^*. We note that $\widehat{x}^{(1)} = G_\lambda(\widehat{x}^{(0)}) = \widehat{x}^{(0)} - \frac{1}{\lambda}\nabla F(\widehat{x}^{(0)})$, i.e., the first-order estimator $\widehat{x}^{(1)}$ is the result of applying a gradient step correction to $\widehat{x}^{(0)}$ with step size $1/\lambda$. Moreover, it holds that

$$x^* = -\lambda^{-1} A^\top z^* \,, \qquad \widehat{x}^{(1)} = -\lambda^{-1} A^\top y^* \,. \tag{13}$$

In contrast to $\widehat{x}^{(0)}$, the first-order estimator $\widehat{x}^{(1)}$ is the result of a linear mapping of the sketched dual solution y^*. Except for a quadratic function f, this corresponds to a non-linear transformation of α^*. The estimator $\widehat{x}^{(0)}$ is computed solely based on the sketched data matrix AS, whereas $\widehat{x}^{(1)}$ uses an additional call to A through the mapping G_λ. These observations would suggest a better performance for $\widehat{x}^{(1)}$.

B. Main contribution

Our analysis involves a canonical geometric object in convex analysis, namely, the *tangent cone* of the domain of f^* at the dual solution z^*, defined as

$$\mathcal{T}_{z^*} := \{t \cdot (y - z^*) \mid t \geqslant 0,\ y \in \mathrm{dom} f^*\} \,. \tag{14}$$

The tangent cone is the intersection of the supporting hyperplanes of the domain of f^* at z^*. A critical quantity in our error guarantees is the maximal singular value $\mathcal{Z}_f$ of the matrix $P_S^\perp A^\top$ restricted to the spherical cap

$$\mathcal{C}_{z^*} := \mathcal{T}_{z^*} \cap \mathcal{B}_2^n \tag{15}$$

and formally defined as

$$\mathcal{Z}_f := \sup_{\Delta \in \mathcal{C}_{z^*}} \|P_S^\perp A^\top \Delta\|_2 \,. \tag{16}$$

Our main result is the following deterministic upper bound on the relative recovery error of the first-order estimator $\widehat{x}^{(1)}$, which shows in particular that $\widehat{x}^{(1)}$ has better performance than $\widehat{x}^{(0)}$, at least by a factor one half.

Theorem 1 (Recovery error of the first-order estimator)**.** *Let $S \in \mathbb{R}^{d\times m}$ be an embedding matrix, and let α^* be a minimizer of the sketched program* (2)*. Under the condition $\lambda \geqslant 2\mu\mathcal{Z}_f^2$, it holds that*

$$\frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2} \leqslant \sqrt{\frac{\mu}{2\lambda}} \cdot \mathcal{Z}_f \cdot \min\left\{1, \frac{\|\widehat{x}^{(0)} - x^*\|_2}{\|x^*\|_2}\right\}. \tag{17}$$

Naturally, we have $\sup_{\Delta\in\mathcal{T}_{z^*}\cap\mathcal{B}_2^n} \|P_S^\perp A^\top\Delta\|_2 \leqslant \sup_{\Delta\in\mathcal{B}_2^n} \|P_S^\perp A^\top\Delta\|_2$, i.e., $\mathcal{Z}_f \leqslant \|P_S^\perp A^\top\|_2$. When z^* is an extreme point of the tangent cone $\mathcal{T}_{z^*}$, then one may expect $\mathcal{T}_{z^*}$ to have a small size and $\mathcal{Z}_f \ll \|P_S^\perp A^\top\|_2$. We discuss such instances in Section III. On the other hand, when z^* belongs to the interior of the domain of f^*, then $\mathcal{C}_{z^*} = \mathcal{B}_2^n$ so that $\mathcal{Z}_f = \|P_S^\perp A^\top\|_2$. For an adaptive embedding $S = A^\top\widetilde{S}$, we can then leverage well-known upper bounds [44], [45] on the residual error $\|P_S^\perp A^\top\|_2$ of the form

$$\|P_S^\perp A^\top\|_2 \lesssim R_{m/2}(A)\,, \tag{18}$$

where the *spectral residual* $R_\delta(A)$ of the matrix A at level $\delta > 0$ is defined as

$$R_\delta(A) := \sigma_{\lfloor\delta\rfloor+1} + \frac{1}{\sqrt{\lfloor\delta\rfloor}} \cdot \sqrt{\sum_{j=\lfloor\delta\rfloor+1}^{\rho} \sigma_j^2}\,. \tag{19}$$

As an immediate consequence, we establish in Theorem 2 in Section III high-probability upper bounds on the recovery error,

$$\frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2} \lesssim \sqrt{\frac{\mu}{\lambda}} \cdot R_{m/2}(A)\,, \tag{20}$$

provided that $\lambda \gtrsim \mu R_{m/2}^2(A)$.

C. Comparison to existing work and oblivious sketching

The recovery method most related to ours is based on the first-order estimator $\widehat{x}_\dagger^{(1)} = G_\lambda(Q\alpha_\dagger^*)$ introduced in [31], [32], where $\alpha_\dagger^* := \operatorname{argmin}_{\alpha\in\mathbb{R}^m}\{f(AQ\alpha) + \frac{\lambda}{2}\|\alpha\|_2^2\}$ and $Q \in \mathbb{R}^{d\times m}$ is an oblivious embedding with i.i.d. Gaussian entries $\mathcal{N}(0, 1/m)$. In contrast to (2), the regularization term does not involve the sketching matrix Q, and this incurs performance guarantees different from ours: under several stringent assumptions (see Theorem 6 in Section IV for details), a *best-case* upper bound on the recovery error of $\widehat{x}_\dagger^{(1)}$ is given as

$$\frac{\|\widehat{x}_\dagger^{(1)} - x^*\|_2}{\|x^*\|_2} \lesssim \sqrt{\frac{d_{\lambda/\mu}}{m}}\,, \tag{21}$$

where the *effective dimension* $d_{\lambda/\mu}$ – a critical quantity in many contexts for sketching-based algorithms [62]–[64] – is defined as

$$d_{\lambda/\mu} := \frac{\operatorname{trace}(D_{\lambda/\mu})}{\|D_{\lambda/\mu}\|_2}\,, \tag{22}$$

and $D_{\lambda/\mu} := A(\frac{\lambda}{\mu}I_d + A^\top A)^{-1}A^\top$. The characteristic measure of error for our adaptive estimator $\widehat{x}^{(1)}$ is $\sqrt{\frac{\mu}{\lambda}} \cdot R_{m/2}(A)$ as opposed to $\sqrt{\frac{d_{\lambda/\mu}}{m}}$ for the oblivious estimator $\widehat{x}_\dagger^{(1)}$. In Table I, we compare these two measures for spectral decays which are common [43], [65] in machine learning, e.g., polynomial $\sigma_j \asymp j^{-\frac{1+\nu}{2}}$ or exponential $\sigma_j \asymp e^{-\frac{\nu j}{2}}$, for some $\nu > 0$ (the proofs of these results involve simple summations and we leave details to the reader). The ratio $\sqrt{\frac{d_{\lambda/\mu}}{m}}$ scales in terms of m as $\mathcal{O}(m^{-\frac{1}{2}})$ regardless of the spectral decay, whereas the term $\sqrt{\frac{\mu}{\lambda}} \cdot R_{m/2}(A)$ has the better scaling $\mathcal{O}(m^{-\frac{1+\nu}{2}})$ for a polynomial decay and $\mathcal{O}(e^{-\frac{\nu m}{2}})$ for an exponential decay.

TABLE I
COMPARISON OF THE CHARACTERISTIC MEASURES OF ERRORS FOR POLYNOMIAL AND EXPONENTIAL SPECTRAL DECAYS. WE ASSUME THAT $\frac{\lambda}{\mu} \asymp \sigma_K^2$ FOR SOME THRESHOLD RANK $1 \leqslant K \leqslant \rho$, AND THAT $m \geqslant 2K$.

Estimator	Embedding	Characteristic error	Polynomial $\sigma_j \asymp j^{-\frac{1+\nu}{2}}$	Exponential $\sigma_j \asymp e^{-\frac{-\nu j}{2}}$
$\widehat{x}^{(1)}$ (Ours)	Adaptive, Gaussian	$\sqrt{\frac{\mu}{\lambda}} R_{\frac{m}{2}}(A)$	$\left(\frac{2K}{m}\right)^{\frac{1+\nu}{2}}$	$e^{-\nu(\frac{m}{2}-K)}$
$\widehat{x}_{\dagger}^{(1)}$ ([31], [32])	Oblivious, Gaussian	$\sqrt{\frac{d_{\lambda/\mu}}{m}}$	$\sqrt{\frac{K}{m}}$	$\sqrt{\frac{K}{m}}$

D. Statement of additional contributions

In addition to our main result (Theorems 1 and 2) on the recovery error of $\widehat{x}^{(1)}$, we have the following contributions. In Section III, by leveraging the equivalence between adaptive sketching of the data matrix A and oblivious sketching of the Gram matrix $AA^\top$, we extend our results to kernel methods (Theorem 3). In Section IV, we establish lower bounds (Theorems 4 and 5) on the recovery error of the estimators $\widehat{x}^{(0)}$ and $\widehat{x}^{(1)}$ with oblivious embeddings. We show that these recovery errors are bounded away from 0 unless $m \approx d$, which would defeat the purpose of sketching. In Section V, we provide a prototype algorithm for adaptive sketching (Algorithm 1), and we show that our proposed low-dimensional formulation is at least as numerically stable as the original program (1). Furthermore, we extend Algorithm 1 to an iterative version (Algorithm 2) with the following guarantee (see Theorem 7). Based on a single sketch AS, it returns after T iterations a solution $\widehat{x}_T^{(1)}$ which satisfies with high probability

$$\frac{\|\widehat{x}_T^{(1)} - x^*\|_2}{\|x^*\|_2} \leqslant \left(\frac{\mu \mathcal{Z}_f^2}{2\lambda}\right)^{\frac{T}{2}}, \tag{23}$$

provided that $\lambda \geqslant 2\mu\mathcal{Z}_f^2$. Consequently, one can construct an approximation $\widehat{x}_T^{(1)}$ with linear convergence rate $\sqrt{\frac{\mu\mathcal{Z}_f^2}{2\lambda}}$ based on a single sketch of the data matrix A, a number T of matrix-vector multiplications of the form $A^\top \nabla f(\cdot)$ (which is equivalent to a gradient call to the function F), and through solving a number T of low-dimensional optimization programs. In Section VI, we extend our analysis to the case of a non-smooth objective function f. We show how to construct a first-order estimator $\widehat{x}^{(1)}$ based on a low-dimensional optimization program and which satisfies the following high-probability guarantee (see Theorem 8),

$$\|\widehat{x}^{(1)} - x^*\|_2 \lesssim \frac{1}{\lambda} \cdot \mathcal{Z}_f \,. \tag{24}$$

Lastly, we show in Theorem 9 and Corollary 2 that, as $\lambda \to 0$, the zero-order estimator $\widehat{x}^{(0)}$ with an oblivious embedding achieves the minimax rate of optimality for a fundamental statistical problem, namely, estimating the mean of a Gaussian distribution $\mathcal{N}(Ax_{\text{pl}}, \frac{\sigma^2}{n} I_n)$ under the smoothness assumption $\|x_{\text{pl}}\|_2 \leqslant 1$. In contrast to our main results, this suggests different benefits of the linear reconstruction map and oblivious embeddings for small values of the regularization parameter λ.

III. SMOOTH, CONVEX OPTIMIZATION IN ADAPTIVE RANDOM SUBSPACES

A. Restricted singular value with adaptive random embeddings

According to Theorem 1, the relative recovery error depends on the restricted singular value $\mathcal{Z}_f$, and we provide next an upper bound on $\mathcal{Z}_f$ in the case of an adaptive Gaussian embedding. Let $A = U\Sigma V^\top$ be a thin a singular value decomposition of A, where $U \in \mathbb{R}^{n\times\rho}$ and $V \in \mathbb{R}^{d\times\rho}$ have orthonormal columns, and $\Sigma \in \mathbb{R}^{\rho\times\rho}$ is the diagonal matrix of the non-zero singular values of A in non-increasing order. For a given target rank $1 \leqslant k \leqslant \frac{\rho}{2}$, we define $\Sigma_k := \text{diag}\{\sigma_1, \ldots, \sigma_k\}$,

$\Sigma_{\rho-k} := \mathrm{diag}\{\sigma_{k+1}, \dots, \sigma_\rho\}$, the matrix $U_k \in \mathbb{R}^{n\times k}$ with the first k columns of U and $U_{\rho-k} \in \mathbb{R}^{n\times(\rho-k)}$ with its last $(\rho-k)$ columns.

Lemma 1 (Restricted singular value with adaptive embeddings)**.** *Let $S = A^\top \widetilde{S}$ where $\widetilde{S} \in \mathbb{R}^{n\times m}$ has i.i.d. Gaussian entries and $m = 2k$ for some target rank $1 \leqslant k \leqslant \frac{\rho}{2}$. Then, it holds that*

$$\mathcal{Z}_f \leqslant c'_g \cdot rad(U_k^\top \mathcal{C}_{z^*}) \cdot R_k(A) + \sup_{\Delta\in\mathcal{C}_{z^*}} \left\|\Sigma_{\rho-k} U_{\rho-k}^\top \Delta\right\|_2 , \tag{25}$$

with probability at least $1 - 6e^{-k}$, where c'_g is a universal constant which satisfies $c'_g \leqslant 25$. This implies in particular that, with probability at least $1 - 6e^{-k}$,

$$\mathcal{Z}_f \leqslant \|P_S^\perp A^\top\|_2 \leqslant c_g \cdot R_k(A) , \tag{26}$$

where c_g is a universal constant which satisfies $c_g \leqslant 26$.

The upper bound (25) on $\mathcal{Z}_f$ involves the quantities $\mathrm{rad}(U_k^\top \mathcal{C}_{z^*})$ and $\sup_{\Delta\in\mathcal{C}_{z^*}} \left\|\Sigma_{\rho-k} U_{\rho-k}^\top \Delta\right\|_2$ which are deterministic (they do not depend on the randomness of S). These quantities depend on the coupling between the matrix U and the spherical cap $\mathcal{C}_{z^*}$ which may vary based on the problem at hand. Under some (idealized) assumptions, we are able to characterize their typical values, and thus a bound on $\mathcal{Z}_f$ which decouples A and the spherical cap $\mathcal{C}_{z^*}$.

Lemma 2. *We assume the matrix $U \in \mathbb{R}^{n\times\rho}$ of left singular vectors of A to be a random Haar matrix in $\mathbb{R}^n$, and the dual solution z^* to be independent of U. Under the hypotheses of Lemma 1, there exists universal constants $c_0, c_1, c_2 > 0$ such that*

$$\mathcal{Z}_f \leqslant c_0 \cdot \frac{\omega(\mathcal{C}_{z^*}) + \sqrt{m}}{\sqrt{n}} \cdot R_{m/2}(A) , \tag{27}$$

with probability at least $1 - c_1 e^{-c_2 m}$.

The upper bound $\frac{\omega(\mathcal{C}_{z^*})+\sqrt{m}}{\sqrt{n}}$ on the ratio $\frac{\mathcal{Z}_f}{R_{m/2}(A)}$ decreases down to $\frac{\omega(\mathcal{C}_{z^*})}{\sqrt{n}}$ as m decreases down to $\omega^2(\mathcal{C}_{z^*})$, and then plateaus at $\frac{\omega(\mathcal{C}_{z^*})}{\sqrt{n}}$. The smaller the Gaussian width $\omega(\mathcal{C}_{z^*})$, the more favorable the statistical-computational trade-off in terms of the sketch size. For instance, suppose that the domain of f^* is the L_1-ball (e.g., ℓ_∞-regression; see Example 2), and denote by s^* the number of non-zero entries of the dual solution z^*. It is known (see, for instance, Section 2.2.2 in [16]) that $\mathcal{C}_{z^*} \subseteq \left\{\Delta \in \mathbb{R}^n \mid \|\Delta\|_1 \leqslant 2\sqrt{s^*}\right\}$, and thus, $\omega(\mathcal{C}_{z^*}) \lesssim \sqrt{s^* \cdot \log n}$. Consequently,

$$\mathcal{Z}_f \lesssim \frac{\sqrt{s^* \cdot \log n} + \sqrt{m}}{\sqrt{n}} \cdot R_{m/2}(A) . \tag{28}$$

The inequality $\|P_S^\perp A^\top\| \leqslant c_g R_k(A)$ in (26) is well-known: it is in fact a simplified statement of the result of Corollary 10.9 in [44]. We do not aim to derive a high-probability bound on $\mathcal{Z}_f$ with the SRHT as this would involve a different machinery of technical arguments, but a high-probability bound on $\|P_S A^\top\|_2$ already exists in the literature.

Lemma 3 (SRHT spectral residual [45])**.** *Let $k \geqslant 2$ be a target rank and pick the sketch size $m := 19(\sqrt{k}+4\sqrt{\log n})^2 \log(kn)$. Let $S = A^\top \widetilde{S}$ where $\widetilde{S} \in \mathbb{R}^{n\times m}$ is a SRHT. Then, it holds with probability at least $1 - \frac{5}{n}$ that $\|P_S^\perp A^\top\|_2 \leqslant c_s \cdot R_k(A)$ for some universal constant $c_s \leqslant 5$.*

Gaussian width based results established in this section complement the analysis of left-sketching in terms of the Gaussian width of the tangent cone [35]. However, the results are quite different due to the extra regularization terms, the projection matrix $P_S^\perp$ appearing in the dual problem and the effect of the spectral residual term $R_k(A)$.

B. Recovery error in randomized adaptive random subspaces

Combining the upper bound (17) on the recovery error of $\widehat{x}^{(1)}$ along with the results of Lemmas 1 and 3, we obtain the following high-probability guarantees in terms of the spectral residual of A.

Theorem 2 (High-probability upper bound on $\widehat{x}^{(1)}$ with adaptive sketching)**.** *For a Gaussian embedding, under the hypotheses of Lemma 1 and provided that $\lambda \geqslant 2\mu c_g^2 R_k^2(A)$, it holds with probability at least $1 - 6e^{-k}$ that*

$$\frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2} \leqslant \sqrt{\frac{c_g^2 \mu}{2\lambda}} \cdot R_k(A) \cdot \min\left\{1, \frac{\|\widehat{x}^{(0)} - x^*\|_2}{\|x^*\|_2}\right\}. \tag{29}$$

For a SRHT embedding, under the hypotheses of Lemma 3 and provided that $\lambda \geqslant 2\mu c_s^2 R_k^2(A)$, it holds with probability at least $1 - \frac{5}{n}$ that

$$\frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2} \leqslant \sqrt{\frac{c_s^2 \mu}{2\lambda}} \cdot R_k(A) \cdot \min\left\{1, \frac{\|\widehat{x}^{(0)} - x^*\|_2}{\|x^*\|_2}\right\}. \tag{30}$$

The regime $\lambda/\mu \ll R_k^2(A)$ corresponds to a sketch size which is too small relatively to the regularization parameter λ and the spectral decay of A. In this regime, it can be shown that $\frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2} \lesssim \frac{\mu}{\lambda} \cdot R_k(A)$, and this is a weaker bound for small values of $\frac{\lambda}{\mu}$.

C. Extension to kernel methods

The first-order optimality conditions $x^* = -\lambda^{-1} A^\top \nabla f(Ax^*)$ yield that $x^* \in \mathrm{range}(A^\top)$. Therefore, the primal program (1) can be solved through the kernel formulation

$$w^* \in \operatorname*{argmin}_{w \in \mathbb{R}^n} \left\{ f(Kw) + \frac{\lambda}{2} w^\top K w \right\}, \tag{31}$$

where $K = AA^\top$, and then setting $x^* = A^\top w^*$. Given an embedding $\widetilde{S} \in \mathbb{R}^{n \times m}$, we consider the sketched version of (31),

$$\alpha_K^* \in \operatorname*{argmin}_{\alpha \in \mathbb{R}^m} \left\{ f(K\widetilde{S}\alpha) + \frac{\lambda}{2} \alpha^\top \widetilde{S}^\top K \widetilde{S} \alpha \right\}. \tag{32}$$

The sketched program (32) is equivalent to (2) with the adaptive embedding $S = A^\top \widetilde{S}$, i.e., $\alpha^* = \alpha_K^*$ and $\widehat{x}^{(1)} = A^\top \widehat{w}^{(1)}$ where $\widehat{w}^{(1)} := -\lambda^{-1} \nabla f(K\widetilde{S}\alpha_K^*)$. In other words, we have equivalence between adaptive sketching of the data matrix A and oblivious sketching of the Gram matrix K. This equivalence naturally extends to any kernel method with a smooth loss function. We recall the concepts necessary to the exposition of our results, and we refer the reader to the books [65]–[67] for more details and background. Given a measurable space Ω endowed with a probability distribution $\mathbb{P}$, we consider the space $L^2(\Omega, \mathbb{P})$ of real-valued functions over Ω which are square-integrable with respect to $\mathbb{P}$, and we let $\mathcal{H} \subset L^2(\Omega, \mathcal{P})$ be a reproducing kernel Hilbert space (RKHS) with reproducing kernel $\mathcal{K} : \Omega \times \Omega \to \mathbb{R}$ and associated norm $\|\cdot\|_{\mathcal{H}}$. Given $\omega_1, \ldots, \omega_n \in \Omega$, we aim to solve the (infinite-dimensional) kernel program $h^* := \mathrm{argmin}_{h \in \mathcal{H}} f(\{h(\omega_i)\}_{i=1}^n) + \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2$. This kernel program occurs in many widely used machine learning contexts (e.g., kernel ridge regression, kernel support vector machines with smooth hinge loss or kernel logistic regression). The representer theorem [68] states that h^* belongs to the span of the functions $\mathcal{K}(\cdot, \omega_1), \ldots, \mathcal{K}(\cdot, \omega_n)$, i.e., there exists $w^* \in \mathbb{R}^n$ such that $h^* = \sum_{i=1}^n \mathcal{K}(\cdot, \omega_i) w_i^*$, and one can solve instead the finite-dimensional program (31) with the empirical kernel matrix $K := \{\mathcal{K}(\omega_i, \omega_j)\}_{i,j}$. Based on a low-dimensional solution α_K^* to (32), we define the zero- and first-order estimators of h^* as $\widehat{h}^{(0)} := \sum_{i=1}^n \mathcal{K}(\cdot, \omega_i) \widehat{w}_i^{(0)}$ where $\widehat{w}^{(0)} := \widetilde{S}\alpha_K^*$,

and, $\widehat{h}^{(1)} := \sum_{i=1}^n \mathcal{K}(\cdot, \omega_i)\widehat{w}_i^{(1)}$ where $\widehat{w}^{(1)} := -\lambda^{-1}\nabla f(K\widetilde{S}\alpha_K^*)$. Let K_h be a square-root of K, i.e., $K = K_h K_h^\top$, and we introduce its corresponding restricted singular value

$$\mathcal{Z}_{f,K} := \sup_{\Delta \in \mathcal{C}_{z_K^*}} \|P^\perp_{K_h^\top \widetilde{S}} K_h^\top \Delta\|_2 \,. \tag{33}$$

where $z_K^* := \operatorname{argmin}_z \left\{ f^*(z) + \frac{1}{2\lambda} z^\top K z \right\}$. We recall that $\| \sum_{i=1}^n \mathcal{K}(\cdot, \omega_i) w_i \|_{\mathcal{H}} = \sqrt{w^\top K w}$ for any $w \in \mathbb{R}^n$. We obtain the following recovery guarantee as a function of the spectral decay of $\mathcal{Z}_{f,K}$.

Theorem 3. *Let $\widetilde{S} \in \mathbb{R}^{n \times m}$ be an embedding matrix, and let α_K^* be a minimizer of the sketched kernel program* (32)*. Under the condition $\lambda \geqslant 2\mu \mathcal{Z}_{f,K}^2$, it holds that*

$$\frac{\|\widehat{h}^{(1)} - h^*\|_{\mathcal{H}}}{\|h^*\|_{\mathcal{H}}} \leqslant \sqrt{\frac{\mu}{2\lambda}} \cdot \mathcal{Z}_{f,K} \cdot \min\left\{1, \frac{\|\widehat{h}^{(0)} - h^*\|_{\mathcal{H}}}{\|h^*\|_{\mathcal{H}}}\right\} . \tag{34}$$

Proof. Define $x^* := \operatorname{argmin}_x f(K_h x) + \frac{\lambda}{2}\|x\|_2^2$, and $S := K_h^\top \widetilde{S}$. Note that $\alpha_K^* \in \operatorname{argmin}_{\alpha \in \mathbb{R}^m} f(K_h S\alpha) + \frac{\lambda}{2}\|S\alpha\|_2^2$. Set $\widehat{x}^{(1)} = -\lambda^{-1} K_h^\top \nabla f(K_h S\alpha_K^*)$ and $\widehat{x}^{(0)} = S\alpha_K^*$. Then, using the results of Theorem 1 with $A = K_h$, we obtain $\frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2} \leqslant \sqrt{\frac{\mu}{2\lambda}} \mathcal{Z}_{f,K} \min\left\{1, \frac{\|\widehat{x}^{(0)} - x^*\|_2}{\|x^*\|_2}\right\}$, provided that $\lambda \geqslant 2\mu\mathcal{Z}_{f,K}^2$. We conclude by using the identities $\|\widehat{h}^{(1)} - h^*\|_{\mathcal{H}} = \|\widehat{x}^{(1)} - x^*\|_2$, $\|\widehat{h}^{(0)} - h^*\|_{\mathcal{H}} = \|\widehat{x}^{(0)} - x^*\|_2$ and $\|h^*\|_{\mathcal{H}} = \|x^*\|_2$. □

Similarly to Theorem 2, the results of Theorem 3 along with the concentration bounds in Lemmas 1 and 3 yield high-probability bounds on the recovery error $\frac{\|\widehat{h}^{(1)} - h^*\|_{\mathcal{H}}}{\|h^*\|_{\mathcal{H}}}$ in terms of the spectral decay of the kernel class $\mathcal{K}$. Typical decays encountered in practice are polynomial (e.g., Sobolev kernel) and exponential (e.g., Gaussian kernel).

Oblivious sketching of kernel matrices with Gaussian embeddings or the SRHT has already been considered in [43], in the context of kernel ridge regression: the authors analyze the statistical performance of the zero-order estimator $\widehat{h}^{(0)}$ as measured by the in-sample predictive norm $\sqrt{\sum_{i=1}^n (\widehat{h}(\omega_i) - h^*(\omega_i))^2}$. We contribute to this set of results by showing that the first-order estimator $\widehat{h}^{(1)}$ has better recovery error than $\widehat{h}^{(0)}$ in the RKHS norm. We leave as an open problem a more extensive comparison of $\widehat{h}^{(1)}$ and $\widehat{h}^{(0)}$ in the predictive norm.

IV. Smooth, convex optimization in oblivious random subspaces

A. Limited performance of $\widehat{x}^{(0)}$ and $\widehat{x}^{(1)}$ with oblivious sketching

We first provide an upper bound on the restricted singular value $\mathcal{Z}_f$ for an oblivious Gaussian embedding, which is significantly weaker than in the adaptive case.

Lemma 4 (Restricted singular value with oblivious embeddings)**.** *Let $S \in \mathbb{R}^{d \times m}$ be a matrix with i.i.d. Gaussian entries. Then, it holds with probability at least $1 - 2e^{-(d-m)}$ that*

$$\mathcal{Z}_f \leqslant c \cdot \left(\frac{\omega(A^\top \mathcal{C}_{z^*})}{\sqrt{d}} + \sqrt{\frac{d-m}{d}} \cdot \|A\|_2 \right) , \tag{35}$$

for some universal constant $c > 0$.

Even for a small width $\omega(A^\top \mathcal{C}_{z^*})$, the upper bound (35) scales at least as $\sqrt{\frac{d-m}{d}} \cdot \|A\|_2$, and this is large unless $d \approx m$, which defeats the purpose of sketching. This suggests a limited performance of the estimators $\widehat{x}^{(0)}$ and $\widehat{x}^{(1)}$ with oblivious embeddings. We formalize this statement by deriving lower bounds on their respective recovery error. For conciseness, we

focus on the expected relative error, where the expectation is taken with respect to the randomness of the embedding matrix S.

Theorem 4 (Lower bound on the recovery error of $\widehat{x}^{(0)}$ with oblivious sketching)**.** *It holds for both Gaussian embeddings and the SRHT that*

$$\mathbb{E}_S\left\{\frac{\|\widehat{x}^{(0)}-x^*\|_2^2}{\|x^*\|_2^2}\right\} \geqslant 1-\frac{m}{d}\,. \tag{36}$$

Proof. Note that $\widehat{x}^{(0)} = P_S\widehat{x}^{(0)}$ and thus, $\|\widehat{x}^{(0)}-x^*\|_2^2 = \|P_S(\widehat{x}^{(0)}-x^*)\|_2^2 + \|P_S^\perp x^*\|_2^2$, i.e., $\|\widehat{x}^{(0)}-x^*\|_2^2 \geqslant \|P_S^\perp x^*\|_2^2$. Using that $\mathbb{E}_S\|P_S^\perp x^*\|_2^2 = (1-\frac{m}{d})\|x^*\|_2^2$ for both embeddings yields the claim. □

The estimator $\widehat{x}^{(0)}$ lies in a low-dimensional oblivious random subspace and does not recover the residual projection of x^* onto $P_S^\perp$: the error $\frac{\|P_S^\perp x^*\|_2^2}{\|x^*\|_2^2} \approx 1-\frac{m}{d}$ is large unless $m \approx d$. We provide next a lower bound on the *worst-case* recovery error of $\widehat{x}^{(1)}$, for which we assume the function f to be γ-strongly convex, i.e., $f(y) \geqslant f(x) + \langle \nabla f(x), y-x\rangle + \frac{\gamma}{2}\|y-x\|_2^2$ for any $x, y \in \mathbb{R}^d$. Although a similar result could hold for the SRHT, the proof would involve different technical arguments and we specialize our result to Gaussian embeddings for conciseness.

Theorem 5 (Lower bound on the recovery error of $\widehat{x}^{(1)}$ with oblivious sketching)**.** *Let $S \in \mathbb{R}^{d\times m}$ be a matrix with i.i.d. Gaussian entries. Then, it holds that*

$$\sup_{f\in\mathcal{F}_{\gamma,\mu}} \mathbb{E}_S\left\{\frac{\|\widehat{x}^{(1)}-x^*\|_2^2}{\|x^*\|_2^2}\right\} \geqslant \left(1-\frac{m}{d}\right)^3 \cdot \frac{\sigma_1^4}{(\sigma_1^2+\frac{2\lambda}{\gamma})^2}\,, \tag{37}$$

where $\mathcal{F}_{\gamma,\mu}$ is the set of real-valued functions defined over $\mathbb{R}^n$, which are γ-strongly convex and μ-smooth.

As for the zero-order estimator, the first-order estimator $\widehat{x}^{(1)}$ with oblivious embeddings has a (worst-case) recovery error bounded away from 0, unless $m \approx d$. In both lower bounds (36) and (37), the limiting factor $1-\frac{m}{d}$ is primarily due to the bias $\mathbb{E}\{P_S^\perp\} = \left(1-\frac{m}{d}\right) I_d \neq 0$, which we aim to address next.

B. Improved performance through unbiased oblivious embeddings

We consider a variant of the sketched primal program (2), given by

$$\alpha_\dagger^* := \operatorname*{argmin}_{\alpha_\dagger\in\mathbb{R}^m}\left\{f(AQ\alpha_\dagger) + \frac{\lambda}{2}\|\alpha_\dagger\|_2^2\right\}, \tag{38}$$

where $Q \in \mathbb{R}^{d\times m}$ is a sketching matrix. The low-dimensional formulations (2) and (38) only differ in the choice of the regularization term, that is, $\frac{\lambda}{2}\|S\alpha\|_2^2$ versus $\frac{\lambda}{2}\|\alpha_\dagger\|_2^2$. We define the corresponding zero- and first-order estimators based on the (unique) low-dimensional solution $\alpha_\dagger^*$ as

$$\widehat{x}_\dagger^{(0)} := Q\alpha_\dagger^*\,, \qquad \widehat{x}_\dagger^{(1)} := -\frac{1}{\lambda}A^\top\nabla f(AQ\alpha_\dagger^*)\,. \tag{39}$$

The next result relates the two low-dimensional programs (2) and (38) more precisely. For an embedding matrix S, we denote by $U_S\Sigma_S V_S^\top$ a thin SVD of S, and we define its *whitened* version as

$$Q_S := U_S V_S^\top\,, \tag{40}$$

Lemma 5. *Let $S \in \mathbb{R}^{d\times m}$ be an embedding matrix, and let $\alpha_\dagger^*$ be the low-dimensional solution of* (38) *with the whitened matrix Q_S. Then, it holds that the vector $\alpha^* := V_S\Sigma_S^{-1}V_S^\top\alpha_\dagger^*$ is a solution of* (2) *with the embedding matrix S. Furthermore, the corresponding zero- and first-order estimators of α^* and $\alpha_\dagger^*$ are respectively equal, i.e.,*

$$\widehat{x}^{(0)} = \widehat{x}_\dagger^{(0)}\,, \qquad \widehat{x}^{(1)} = \widehat{x}_\dagger^{(1)}\,. \tag{41}$$

Proof. We have by first-order optimality conditions that $Q_S^\top A^\top \nabla f(AQ_S\alpha_\dagger^*) + \lambda\alpha_\dagger^* = 0$. Multiplying by $V_S\Sigma_S V_S^\top$ and plugging-in the definition of α^*, we obtain $S^\top A^\top\nabla f(AS\alpha^*) + \lambda S^\top S\alpha^* = 0$, i.e., α^* is a solution of (2). On the other hand, we have $Q_S\alpha_\dagger^* = U_S V_S^\top V_S\Sigma_S V_S^\top\alpha^* = S\alpha^*$, which further implies that $\widehat{x}^{(0)} = \widehat{x}_\dagger^{(0)}$ and $\widehat{x}^{(1)} = \widehat{x}_\dagger^{(1)}$. □

The low-dimensional formulation (2) is equivalent to (38) with the whitened matrix Q_S, in the sense that they yield the same zero- and first-order estimators. In addition to the whitened matrices Q_S which are in general biased (e.g., $\mathbb{E}\{Q_SQ_S^\top\} = \frac{m}{d}\cdot I_d$ for an oblivious Gaussian embedding S), the formulation (38) can incorporate any sketching matrix Q, and in particular, random matrices such that $\mathbb{E}\{QQ^\top\} = I_d$. This yields a larger class of zero- and first-order estimators, which may overcome the aforementioned limited performance of $\widehat{x}^{(0)}$ and $\widehat{x}^{(1)}$ with oblivious embeddings. The estimator $\widehat{x}_\dagger^{(1)}$ with an unbiased oblivious Gaussian embedding has already been considered in [31], and we recall their main result (Theorem 6 in [31]).

Theorem 6 (Recovery error of the unbiased oblivious estimator $\widehat{x}_\dagger^{(1)}$)**.** *Suppose that the function f is separable and that the solution x^* lies in the span of the top k right singular vectors of A. Let $Q \in \mathbb{R}^{d\times m}$ be a matrix with i.i.d. Gaussian entries $\mathcal{N}(0, 1/m)$. Then, provided that $m \geqslant 32d_{\lambda/\mu}\log(2d/\delta)$, it holds that*

$$\frac{\|\widehat{x}_\dagger^{(1)} - x^*\|_2}{\|x^*\|_2} \leqslant \sqrt{128\log(2d/\delta)}\cdot\sqrt{\frac{d_{\lambda/\mu}}{m}}\cdot\left(1 + \sqrt{\frac{\lambda}{\mu\sigma_k^2}}\right)\,, \tag{42}$$

with probability at least $1-\delta$.

The best-case upper bound (42) scales as $\sqrt{\frac{d_{\lambda/\mu}}{m}}$. That is, the sketch size m needs to scale at least as the effective dimension which can be much smaller than the ambient dimension d. This significantly improves on the guarantees for $\widehat{x}^{(0)}$ and $\widehat{x}^{(1)}$ with oblivious embeddings. Furthermore, the oblivious zero-order estimator $\widehat{x}_\dagger^{(0)}$ has worse performance than $\widehat{x}_\dagger^{(1)}$ (see Theorem 3 in [31]). The assumptions of Theorem 6 in [31] have been slightly relaxed by the same authors (see Theorem 2 and Corollary 3 in [32]): provided that the projection of x^* onto the subspace orthogonal to the top k right singular vectors of A has small enough norm, then the recovery error scales as $\sqrt{\frac{k}{m}}$.

Let us now compare the guarantees for the oblivious estimator $\widehat{x}_\dagger^{(1)}$ and for the adaptive estimator $\widehat{x}^{(1)}$. Besides being independent of any assumption on x^*, the upper bound (29) scales as $\sqrt{\frac{\mu}{\lambda}}\cdot R_{m/2}(A)$. In light of the comparison between the characteristic quantities $\sqrt{\frac{d_{\lambda/\mu}}{m}}$ and $\sqrt{\frac{\mu}{\lambda}}\cdot R_{m/2}(A)$ provided in Table I, the guarantees for the adaptive estimator $\widehat{x}^{(1)}$ are in general stronger than the best-case guarantee (42) for the unbiased oblivious estimator $\widehat{x}_\dagger^{(1)}$.

In light of the results of Theorems 1, 2, 4, 5 and 6, we obtain that among the different possible estimators $\widehat{x}^{(0)}$ and $\widehat{x}^{(1)}$ with oblivious and adaptive sketching, and, $\widehat{x}_\dagger^{(0)}$ and $\widehat{x}_\dagger^{(1)}$ with unbiased oblivious sketching, the strongest guarantees are obtained for the first-order estimator $\widehat{x}^{(1)}$ with an adaptive embedding. We compare numerically the performance of the adaptive estimator $\widehat{x}^{(1)}$ and the unbiased oblivious estimator $\widehat{x}_\dagger^{(1)}$ with Gaussian embeddings and for polynomial and exponential decays. We use $n = 1000$, $d = 2000$ and we generate data matrices with respective spectral decay $\sigma_j = \sqrt{n}e^{-0.05j}$ and $\sigma_j = \sqrt{n}j^{-1}$. We consider two convex smooth loss functions: the logistic function $f(w) = n^{-1}\sum_{i=1}^n \log(1 + e^{-y_iw_i})$ where $y \in \{\pm 1\}^n$, and, a second loss function $f(w) = (2n)^{-1}\sum_{i=1}^n (w_i)_+^2 - 2w_iy_i$ (where $a_+ := \max\{a, 0\}$ for $a \in \mathbb{R}$) which can be seen as the

convex relaxation of the penalty $\frac{1}{2}\|w_+ - y\|_2^2$ for fitting a shallow neural network with a ReLU non-linearity [7], [9], [69]. We report results in Figure 1. The adaptive estimator $\widehat{x}^{(1)}$ has better empirical performance than the oblivious one $\widehat{x}_\dagger^{(1)}$.

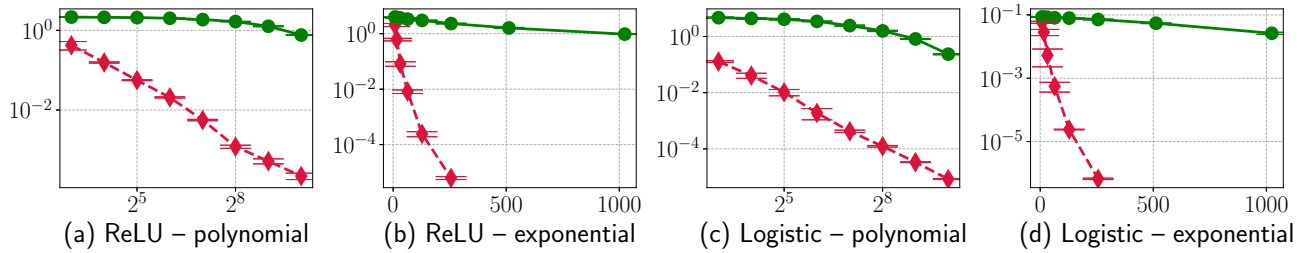


Fig. 1. Relative recovery error versus sketching dimension $m \in \{2^k \mid 3 \leqslant k \leqslant 10\}$ of $\widehat{x}^{(1)}$ (red diamonds) and $\widehat{x}_\dagger^{(1)}$ (green circles), for the ReLU-type and logistic loss functions, and the exponential and polynomial decays. We use $\lambda = 10^{-4}$ for all simulations. Results are averaged over 10 trials. Bar plots show (twice) the empirical standard deviations.

V. Algorithms for smooth convex objectives in adaptive random subspaces

A. Prototype algorithm for adaptive sketching

A standard quantity to characterize whether a convex program can be solved efficiently is its condition number [56], which, for the primal (1) and sketched program (2), is respectively given by

$$\kappa := \sup_{x\in\mathbb{R}^d} \frac{\lambda + \sigma_1\big(A^\top \nabla^2 f(Ax) A\big)}{\lambda + \sigma_d(A^\top \nabla^2 f(Ax) A)}\,, \qquad \kappa_S := \sup_{\alpha\in\mathbb{R}^m} \frac{\sigma_1\big(S^\top(\lambda I_d + A^\top \nabla^2 f(AS\alpha)A)S\big)}{\sigma_m(S^\top(\lambda I_d + A^\top \nabla^2 f(AS\alpha)A)S)}\,. \tag{43}$$

The latter can be significantly larger than κ, up to $\kappa_S \approx \kappa \cdot \frac{\sigma_1(S^\top S)}{\sigma_m(S^\top S)} \gg \kappa$. According to Lemma 5, we can solve instead the optimization problem (38) with the whitened matrix Q_S, and we have $\widehat{x}^{(1)} = -\lambda^{-1} A^\top \nabla f(AQ_S\alpha_\dagger^*)$. Fortunately, the re-scaled sketched program (38) with $Q = Q_S$ is numerically well-conditioned: in fact, it is even better conditioned than the original primal program (1).

Proposition 3. *The condition number $\kappa_\dagger$ of the re-scaled sketched program* (38)

$$\kappa_\dagger := \sup_{\alpha\in\mathbb{R}^m} \frac{\lambda + \sigma_1\big(Q_S^\top A^\top \nabla^2 f(AQ_S\alpha) AQ_S\big)}{\lambda + \sigma_m\big(Q_S^\top A^\top \nabla^2 f(AQ_S\alpha) AQ_S\big)} \tag{44}$$

satisfies $\kappa_\dagger \leqslant \kappa$ almost surely.

Proof. Fix $\alpha \in \mathbb{R}^m$. Using $\|Q_S\|_2 \leqslant 1$, we obtain

$$\sigma_1(Q_S^\top A^\top \nabla^2 f(AQ_S\alpha) AQ_S) \leqslant \sigma_1(A^\top \nabla^2 f(AQ_S\alpha) A)\,,$$
$$\sigma_m(Q_S^\top A^\top \nabla^2 f(AQ_S\alpha) AQ_S) \geqslant \sigma_d(A^\top \nabla^2 f(AQ_S\alpha) A)\,.$$

Consequently,

$$\frac{\lambda + \sigma_1(Q_S^\top A^\top \nabla^2 f(AQ_S\alpha) AQ_S)}{\lambda + \sigma_m(Q_S^\top A^\top \nabla^2 f(AQ_S\alpha) AQ_S)} \leqslant \frac{\lambda + \sigma_1(A^\top \nabla^2 f(AQ_S\alpha) A)}{\lambda + \sigma_d(A^\top \nabla^2 f(AQ_S\alpha) A)}\,.$$

Taking the supremum over $\alpha \in \mathbb{R}^m$ in both sides of the latter inequality, we obtain $\kappa_\dagger \leqslant \sup_{\alpha\in\mathbb{R}^m} \frac{\lambda+\sigma_1(A^\top \nabla^2 f(AQ_S\alpha)A)}{\lambda+\sigma_d(A^\top \nabla^2 f(AQ_S\alpha)A)}$. We conclude using the fact that the latter right-hand side is smaller than κ. □

Algorithm 1 is decomposed into three steps: forming the sketch AQ_S, solving the low-dimensional program (38), and, mapping $\alpha_\dagger^*$ to $\widehat{x}^{(1)}$. The last step is, in general, relatively cheap, as it only requires a matrix-vector multiplication with $A^\top$

Algorithm 1: Prototype algorithm for adaptive sketching in the smooth case.

Input: Data matrix $A \in \mathbb{R}^{n\times d}$, random matrix $\widetilde{S} \in \mathbb{R}^{n\times m}$ and regularization parameter $\lambda > 0$.

Compute the sketching matrix $S = A^\top \widetilde{S}$.

Compute a thin SVD $S = U_S \Sigma_S V_S^\top$ and set $Q_S = U_S V_S^\top$.

Solve the convex optimization problem (38) with $Q = Q_S$, and return $\widehat{x}^{(1)} = -\frac{1}{\lambda} A^\top \nabla f\left(AQ_S \alpha_\dagger^*\right)$.

and a gradient call to f. In total, the algorithm requires three passes over the entire data matrix A. Depending on the choice and structure of the random embedding, the sketching part has different computational costs, as discussed next. We denote by $\mathrm{nnz}(A)$ the number of non-zero entries of A.

1) Computational complexity for Gaussian embeddings: Forming $S = A^\top \widetilde{S}$ takes time $\mathcal{O}(m{\cdot}\mathrm{nnz}(A))$. The cost of computing the SVD of S is $\mathcal{O}(d \cdot m^2)$. The matrix multiplication $A \cdot Q_S$ takes time $\mathcal{O}(m \cdot \mathrm{nnz}(A))$. Therefore, the total complexity is given by

$$\mathcal{O}(2 \cdot m \cdot \mathrm{nnz}(A) + d \cdot m^2)\,. \tag{45}$$

For a dense matrix A, this results in $\mathcal{O}(2mnd + dm^2) = \mathcal{O}(2mnd)$ floating point operations, and the cost is dominated by the sketching part. For a sparse enough matrix A with $\mathrm{nnz}(A) \lesssim dm$, the total cost is $\mathcal{O}(dm^2)$.

2) Computational complexity for the SRHT: Differently from dense and unstructured embeddings, forming $S = A^\top \widetilde{S}$ takes time $\mathcal{O}(\log m \cdot \mathrm{nnz}(A))$. The total time complexity is then given by

$$\mathcal{O}(\log m \cdot \mathrm{nnz}(A) + m \cdot \mathrm{nnz}(A) + d \cdot m^2) = \mathcal{O}(m \cdot \mathrm{nnz}(A) + d \cdot m^2)\,, \tag{46}$$

which is always smaller than the cost of Gaussian embeddings. For a dense matrix A, this results in $\mathcal{O}(mnd)$ floating point operations, and this is half the cost with Gaussian embeddings. For a sparse enough matrix A with $\mathrm{nnz}(A) \lesssim dm$, the cost similarly scales as $\mathcal{O}(dm^2)$.

B. Improved algorithms

1) Iterative method and almost exact recovery of the optimal solution: The estimator $\widehat{x}^{(1)}$ satisfies a guarantee of the form $\|\widehat{x}^{(1)} - x^*\|_2 \leqslant \varepsilon \|x^*\|_2$ with high probability, and with $\varepsilon < 1$ provided that m is large enough relatively to $\frac{\lambda}{\mu}$ and the spectral decay of A. Here, we extend Algorithm 1 to an iterative version which takes advantage of this error contraction, and which does not incur additional memory requirements, at the expense of additional time complexity.

Algorithm 2: Prototype iterative method for adaptive sketching in the smooth case

Input: Data matrix $A \in \mathbb{R}^{n\times d}$, random matrix $\widetilde{S} \in \mathbb{R}^{n\times m}$, iterations number T, regularization parameter $\lambda > 0$.

Compute the matrices $Q_S \in \mathbb{R}^{d\times m}$ and AQ_S as in Algorithm 1. Set $\widehat{x}_0^{(1)} = 0$.

for $t = 1, 2, \ldots, T$ **do**

Solve the low-dimensional convex optimization problem

$$\alpha_{\dagger,t}^* := \operatorname*{argmin}_{\alpha_\dagger \in \mathbb{R}^m} \left\{ f(AQ_S\alpha_\dagger + A\widehat{x}_{t-1}^{(1)}) + \frac{\lambda}{2}\|\alpha_\dagger + Q_S^\top \widehat{x}_{t-1}^{(1)}\|_2^2 \right\}. \tag{47}$$

Update the solution $\widehat{x}_t^{(1)} = -\frac{1}{\lambda} A^\top \nabla f(AQ_S\alpha_{\dagger,t}^* + A\widehat{x}_{t-1}^{(1)})$.

end

Return the last iterate $\widehat{x}_T^{(1)}$.

A key advantage of Algorithm 2 is that, at each iteration, the same sketching matrix S is used, i.e., the matrices Q_S and AQ_S need to be computed only once, at the beginning of the procedure. The output $\widehat{x}_T^{(1)}$ satisfies the following recovery property, whose empirical benefits are illustrated in Figure 2.

Theorem 7. *After T iterations of Algorithm 2, provided that $\lambda \geqslant 2\mu\mathcal{Z}_f^2$, it holds that*

$$\frac{\|\widehat{x}_T^{(1)} - x^*\|_2}{\|x^*\|_2} \leqslant \left(\frac{\mu\mathcal{Z}_f^2}{2\lambda}\right)^{\frac{T}{2}}. \tag{48}$$

For an adaptive Gaussian embedding, under the hypotheses of Lemma 1 and provided that $\lambda \geqslant 2\mu c_g^2 R_k^2(A)$, it holds with probability at least $1 - 6e^{-k}$ that

$$\frac{\|\widehat{x}_T^{(1)} - x^*\|_2}{\|x^*\|_2} \leqslant \left(\frac{c_g^2\mu R_k^2(A)}{2\lambda}\right)^{\frac{T}{2}}. \tag{49}$$

For an adaptive SRHT, under the hypotheses of Lemma 3 and provided that $\lambda \geqslant 2\mu c_s^2 R_k^2(A)$, it holds with probability at least $1 - \frac{5}{n}$ that

$$\frac{\|\widehat{x}_T^{(1)} - x^*\|_2}{\|x^*\|_2} \leqslant \left(\frac{c_s^2\mu R_k^2(A)}{2\lambda}\right)^{\frac{T}{2}}. \tag{50}$$

Let us compare the *oracle* complexities of Algorithm 2 and first-order methods applied to the high-dimensional program (1) under the assumption that $\frac{\lambda}{\mu\sigma_1^2(A)} \ll 1$ is small, so that the high-dimensional objective function F in (1) is ill-conditioned. After a number T of matrix-vector multiplications of the form $A^\top\nabla f(\cdot)$ (which is equivalent to a gradient call to F), Algorithm 2 returns an approximate solution $\widehat{x}_T^{(1)}$ whose relative error is upper bounded by $\left(\frac{\mu\mathcal{Z}_f^2}{2\lambda}\right)^{\frac{T}{2}}$. In comparison, given a similar budget of T gradient calls to the objective function F, first-order methods applied to the high-dimensional program (1) return an approximate solution $\widetilde{x}$ whose relative recovery error is upper bounded by $(1 - \frac{\lambda}{\mu\sigma_1^2(A)})^{\frac{T}{2}}$ (see, for instance, Theorem 3.10 in [24]). The latter convergence rate is close to 1, whereas the convergence rate of the sequence $\widehat{x}_T^{(1)}$ is bounded away from 1 provided that the sketch size is large enough.

2) Shrinking the spectral residual with the power method: An immediate extension of Algorithms 1 and 2 consists in using the so-called *power method* [44]. Given $q \in \mathbb{N}$, the adaptive sketching matrix at power q is defined as $S := (A^\top A)^q A^\top \widetilde{S}$. The larger q, the smaller the approximation error $\|P_S^\perp A^\top\|_2$. More precisely, for a Gaussian embedding $\widetilde{S}$ with $m = 2k$, we have according to Corollary 10.10 in [44] that

$$\|P_S^\perp A^\top\|_2 \lesssim \sigma_k \left(1 + \sqrt{\frac{1}{k}\sum_{j>k}\left(\frac{\sigma_j}{\sigma_k}\right)^{2(2q+1)}}\right)^{\frac{1}{2q+1}} \tag{51}$$

with high probability. The above right-hand side decreases as q increases. Of practical interest are data matrices A of the form $A = \overline{A} + W$, where $\overline{A}$ is a signal with a fast spectral decay, and W is a noise matrix with a relatively small and flat spectral profile. The power method reduces the noise contribution to the spectrum, i.e., it shrinks the error factor $\left(1 + \sqrt{\frac{1}{k}\sum_{j>k}\left(\frac{\sigma_j}{\sigma_k}\right)^{2(2q+1)}}\right)^{\frac{1}{2q+1}}$. Our algorithms readily incorporate the power method, and we illustrate its empirical benefits in Figure 2.

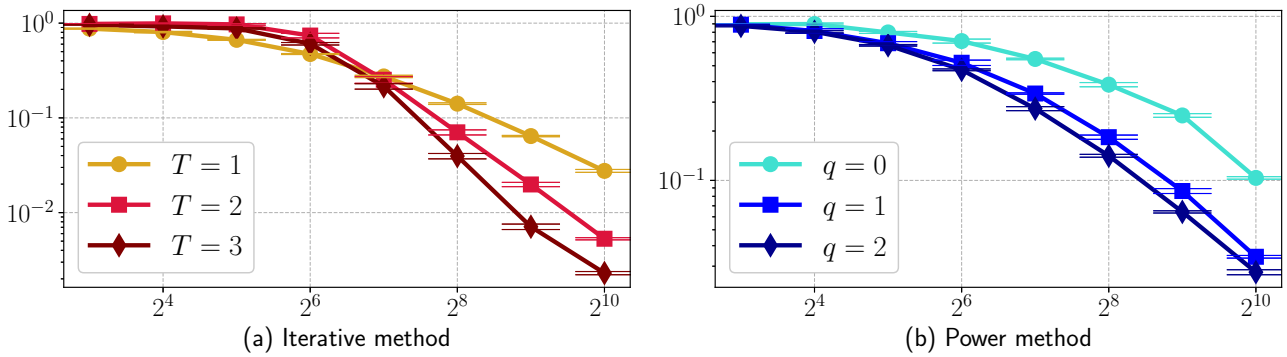


Fig. 2. Relative recovery error versus sketching dimension $m \in \{2^k \mid 3 \leqslant k \leqslant 10\}$ of adaptive Gaussian sketching for (a) the iterative method (Algorithm 2) and (b) the power method. We use the MNIST dataset with 50000 training images and 10000 testing images, and we map images through 10000-dimensional through random cosines ψ which approximate the Gaussian kernel [2], i.e., $\langle \psi(a), \psi(a') \rangle \approx \exp(-\gamma \|a - a^2\|_2^2)$, with $\gamma = 0.02$. We perform binary logistic regression for even-vs-odd classification of digits. For the iterative method, we use the sketching matrix $S = (A^\top A)^2 A^\top \widetilde{S}$, where $\widetilde{S}$ is Gaussian i.i.d. That is, we run the iterative method on top of the power method, with $q = 2$. We use $\lambda = 10^{-5}$. Results are averaged over 20 trials. Bar plots show (twice) the empirical standard deviations.

VI. Non-smooth, convex optimization in adaptive random subspaces

We extend our analysis to the case where the function f is Lipschitz continuous but not necessarily smooth nor even differentiable. In contrast to the smooth case, the dual map $\partial f(AQ_S\alpha_\dagger^*)$ is set-valued and this makes the recovery through dual mapping more challenging. On the other hand, according to Proposition 1, when the function f is strictly convex, the dual solution z^* lies within the interior of the domain of f^* and consequently, $\mathcal{Z}_f = \|P_S^\perp A^\top\|_2$. As strict convexity holds for most smooth convex objective functions f of practical interest, one may instead expect the tangent cone $\mathcal{T}_{z^*}$ to have a small size in the case of a non-smooth objective function.

A. Undeterminate estimator through dual mapping

Theorem 8 (Deterministic upper bound for non-smooth objectives)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-Lipschitz but not necessarily smooth nor differentiable. Let $S \in \mathbb{R}^{d\times m}$ be an embedding matrix, and y^* be any sketched dual solution, and set $\widehat{x}^{(1)} := -\lambda^{-1} A^\top y^*$. Then, for any $\lambda > 0$, it holds that*

$$\|\widehat{x}^{(1)} - x^*\|_2 \leqslant 2 \cdot \frac{L}{\lambda} \cdot \sqrt{\mathcal{Z}_f^2 + \frac{1}{2}\,\mathcal{Z}_f \|P_S^\perp A^\top\|_2} \leqslant \sqrt{6} \cdot \frac{L}{\lambda} \cdot \|P_S^\perp A^\top\|_2\,. \tag{52}$$

Further, under the additional assumption that $\inf_w f(w) > -\infty$, we have the improved upper bound

$$\|\widehat{x}^{(1)} - x^*\|_2 \leqslant \sqrt{6} \cdot \frac{L}{\lambda} \cdot \mathcal{Z}_f\,. \tag{53}$$

The upper bound (53) is, for most cases of interest, weaker than (17) as it scales as $\mathcal{O}(\frac{\mathcal{Z}_f}{\lambda})$ in contrast to the bound $\mathcal{O}(\frac{\mathcal{Z}_f}{\sqrt{\lambda}})$ in the smooth case: for small λ, the former scaling is worse. Furthermore, the upper bound (53) controls the recovery error $\|\widehat{x}^{(1)} - x^*\|_2$ whereas the upper bound in the smooth case controls the relative recovery error $\frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2}$. The latter is contractive and enables an iterative version for (almost) exact recovery of x^*, whereas this approach does not readily extend to the non-smooth case. More importantly, when f is smooth and thus differentiable, the mapping between a low-dimensional solution $\alpha_\dagger^*$ and the estimator $\widehat{x}^{(1)} = -\lambda^{-1} A^\top \nabla f(AQ_S\alpha_\dagger^*)$ is well-defined. However, in the non-differentiable case, we only know that the sketched dual solution y^* belongs to the set $\partial f(AQ_S\alpha_\dagger^*)$, i.e., $\widehat{x}^{(1)} \in -\lambda^{-1} A^\top \partial f(AQ_S\alpha_\dagger^*)$, and one cannot directly compute $\widehat{x}^{(1)}$ based on the low-dimensional solution $\alpha_\dagger^*$ and the dual mapping. Furthermore, picking an arbitrary subgradient $g \in \partial f(AQ_S\alpha_\dagger^*)$ yields an estimator $\widehat{x} = -\lambda^{-1} A^\top g$ with weaker recovery guarantees.

Corollary 1. *Suppose that* $\inf_w f(w) > -\infty$. *Let* $\alpha_\dagger^*$ *be a minimizer of the sketch primal program* (38) *with* $Q = Q_S$. *Pick any estimator* $\widehat{x}$ *in the set* $-\lambda^{-1}A^\top \partial f(AQ_S\alpha_\dagger^*)$. *It holds that*

$$\|\widehat{x} - x^*\|_2 \leqslant \sqrt{6} \cdot \frac{L}{\lambda} \cdot \mathcal{Z}_f + \mathit{diam}\big(\lambda^{-1}A^\top \partial f(AQ_S\alpha_\dagger^*)\big) \ . \tag{54}$$

Proof. Fix $\widehat{x} \in -\lambda^{-1}A^\top \partial f(AQ_S\alpha_\dagger^*)$. The first-order estimator $\widehat{x}^{(1)} = -\lambda^{-1}A^\top y^*$ also belongs to $-\lambda^{-1}A^\top \partial f(AQ_S\alpha_\dagger^*)$, so that $\|\widehat{x} - \widehat{x}^{(1)}\|_2 \leqslant \operatorname{diam}\Big(\lambda^{-1}A^\top \partial f(AQ_S\alpha_\dagger^*)\Big)$. According to Theorem 8, we also have $\|\widehat{x}^{(1)} - x^*\|_2 \leqslant \sqrt{6} \cdot \frac{L}{\lambda} \cdot \mathcal{Z}_f$. The claim then follows from the triangular inequality $\|\widehat{x} - x^*\|_2 \leqslant \|\widehat{x} - \widehat{x}^{(1)}\|_2 + \|\widehat{x}^{(1)} - x^*\|_2$. □

According to Corollary 1, an arbitrary estimator $\widehat{x} \in -\lambda^{-1}A^\top \partial f(AQ_S\alpha_\dagger^*)$ may perform poorly, even when $\mathcal{Z}_f$ is small, as soon as the diameter of the set $\lambda^{-1}A^\top \partial f(AQ_S\alpha_\dagger^*)$ is relatively large. One may wonder whether this upper bound is tight: although we do not provide a lower bound, we carried out extensive numerical simulations with some commonly used non-smooth loss functions, and we obtained poor performance when we picked the easiest-to-compute subgradient in $\partial f(AQ_S\alpha_\dagger^*)$ (see Figure 3).

B. Resolving the indeterminacy of the set-valued dual mapping through the sketched dual program

Based on an adaptive sketch AQ_S and a low-dimensional solution $\alpha_\dagger^*$, the dual map yields a set of candidate approximate solutions $-\lambda^{-1}A^\top \partial f(AQ_S\alpha_\dagger^*)$. We aim to resolve this indeterminacy by computing the optimal subgradient $y^* \in \partial f(AQ_S\alpha_\dagger^*)$ and thus the first-order estimator $\widehat{x}^{(1)} = -\lambda^{-1}A^\top y^*$.

1) Easy-to-compute subgradient set $\partial f(AQ_S\alpha_\dagger^*)$ *and restricted sketched dual program:* We propose to (i) compute a low-dimensional solution $\alpha_\dagger^*$, (ii) to compute the subgradient set $\partial f(AQ_S\alpha_\dagger^*)$, and finally, (iii) to solve the *restricted* sketched dual program

$$y^* \in \operatorname*{argmin}_{y \in \partial f(AQ_S\alpha_\dagger^*)} \left\{ f^*(y) + \frac{1}{2\lambda}\|Q_S^\top A^\top y\|_2^2 \right\} . \tag{55}$$

This procedure is especially relevant when the function f is separable and when there are only a few number k of indices $i \in \{1, \dots, n\}$ such that the function f is not partially differentiable at $(AS\alpha_\dagger^*)_i$. In this case, the restricted sketched dual program (55) only involves k dual variable's coordinates. This is reminiscent of the Stochastic Dual Newton Ascent (SDNA) method [25] which selects at each iteration a random subset of coordinates of the dual variable y. Differently, we use the low-dimensional solution $\alpha_\dagger^*$ and the subgradient set $\partial f(AQ_S\alpha_\dagger^*)$ to determine a subset of coordinates of y to optimize and which yields the exact optimal solution y^*. Moreover, our perspective is, again, agnostic to the choice of the optimization algorithm, whereas SDNA is itself an optimization method.

More generally, the restricted sketched dual program is relevant for practical cases where (i), given a low-dimensional solution $\alpha_\dagger^*$, computing the subgradient set $\partial f(AQ_S\alpha_\dagger^*)$ can be done efficiently, and (ii) the subgradient set $\partial f(AQ_S\alpha_\dagger^*)$ is, in some sense, small. We discuss some examples below.

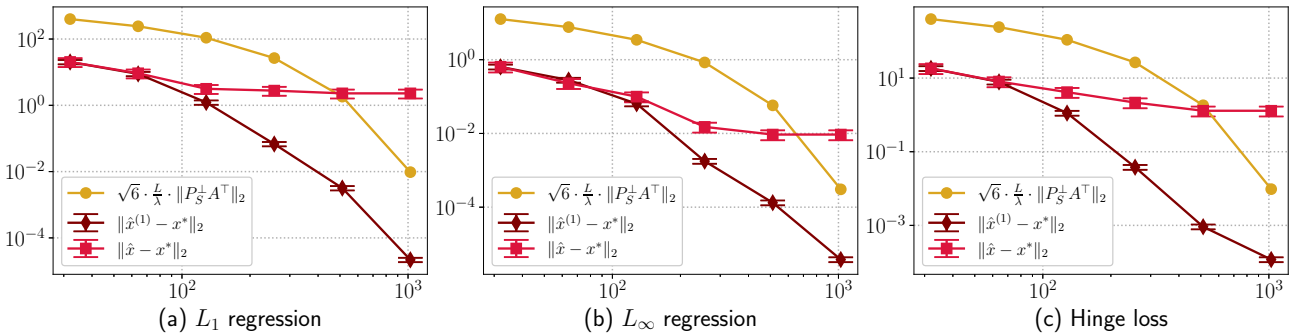
Example 1 (L_1-regression)**.** *Given* $b \in \mathbb{R}^n$, *we consider the objective function* $f(w) = \|w - b\|_1$. *The subgradient set of* f *at some* $w \in \mathbb{R}^n$ *is the Cartesian product* $\prod_{i=1}^n \mathcal{I}_i$, *where* $\mathcal{I}_i = [-1, 1]$ *if* $w_i = b_i$ *and* $\mathcal{I}_i = \{\mathit{sign}(w_i - b_i)\}$ *otherwise. The restricted sketched dual program* (55) *only involves the variables* y_i *for the indices* i *such that* $(AS\alpha_\dagger^*)_i = b_i$.

Example 2 (L_∞-regression)**.** *Given some target vector* $b \in \mathbb{R}^n$, *consider the objective function* $f(w) = \|w - b\|_\infty$. *The subgradient set of* f *at some* $w \in \mathbb{R}^n$ *is the convex hull of the vectors* $\mathit{sign}(w_I - b_I) \cdot e_I$, *where* $e_1, \dots, e_n$ *is the canonical*

basis of $\mathbb{R}^n$ *and for all* $I \in \operatorname{argmax}_{i=1,\dots,n} |w_i - b_i|$. *The restricted sketched dual program* (55) *only involves the variables* y_i *for the indices* i *such that* $|(AS\alpha_\dagger^*)_i - b_i| = \|AS\alpha_\dagger^* - b\|_\infty$.

Example 3 (Support vector machines)**.** *Given* $b \in \{\pm 1\}^n$, *we consider the hinge loss* $f(w) = \sum_{i=1}^n \max\{0, 1 - w_i b_i\}$. *The subgradient set of* f *at some* $w \in \mathbb{R}^n$ *is the Cartesian product* $\prod_{i=1}^n \mathcal{I}_i$, *where* $\mathcal{I}_i = \{-b_i\}$ *if* $1 - w_i b_i > 0$, $\mathcal{I}_i = \{0\}$ *if* $1 - w_i b_i < 0$ *and* $\mathcal{I}_i = [0, -b_i]$ *if* $1 - w_i b_i = 0$. *The restricted sketched dual program* (55) *only involves the variables* y_i *for the indices* i *such that* $1 = (AS\alpha_\dagger^*)_i b_i$.

For the loss functions of Examples 1, 2 and 3, we compare the empirical performance of the estimator $\widehat{x}^{(1)} = -\lambda^{-1} A^\top y^*$ and an arbitrary estimator $\widehat{x} \in -\lambda^{-1} A^\top \partial f(AQ_S \alpha_\dagger^*)$, and we report results in Figure 3. We observe that $\widehat{x}^{(1)}$ has an increasingly stronger performance compared to $\widehat{x}$ as the sketch size increases.



(a) L_1 regression (b) L_∞ regression (c) Hinge loss

Fig. 3. Recovery error versus sketching dimension $m \in \{2^k \mid 5 \leqslant k \leqslant 10\}$ of the adaptive estimator $\widehat{x}^{(1)} = -\lambda^{-1} A^\top y^*$ versus an estimator $\widehat{x} = -\lambda^{-1} A^\top g$ where $g \in \partial f(AQ_S \alpha_\dagger^*)$ is an arbitrary subgradient (the 'easiest-to-compute'). We use $\lambda = 0.01$, $n = 1000$ and $d = 2000$. We generate a data matrix A with exponential spectral decay $\sigma_j = 0.98^j$, and we use an adaptive Gaussian embedding $S = A^\top \widetilde{S}$. We consider (a) L_1-regression (Example 1), (b) L_∞-regression (Example 2) and (c) SVM classification (Example 3). Results are averaged over 20 trials. Bar plots show (twice) the empirical standard deviations.

2) Solving the plain sketched dual program: When the entire subgradient set $\partial f(AQ_S \alpha_\dagger^*)$ cannot be efficiently computed, one can instead solve directly the plain sketched dual program without restricting y to lie within the subgradient set $\partial f(AQ_S \alpha_\dagger^*)$ and without even computing a low-dimensional solution $\alpha_\dagger^*$. Of practical interest are, for instance, the functions f which, up to a translation, are the support function of a convex set $\mathcal{C}$, i.e.,

$$f(w) = \sup_{z \in \mathcal{C}} z^\top (w - b)\,, \tag{56}$$

for which $\partial f(w) = \operatorname{argmax}_{z \in \mathcal{C}} z^\top (w - b)$. This class of function includes in particular distributionally robust objective functions (e.g., conditional Value-at-Risk) and deterministic robust counterparts [70]–[72] for which computing the entire set of worst-case distributions is, in general, expensive. Hence, one can alternatively obtain y^* by solving the *unrestricted* constrained quadratic program

$$y^* \in \operatorname*{argmin}_{y \in \mathcal{C}} \left\{ y^\top b + \frac{1}{2\lambda} \|Q_S^\top A^\top y\|_2^2 \right\}. \tag{57}$$

This approach corresponds to a *left* sketch of a constrained quadratic program: its computational benefits have been extensively studied in the sketching literature and its statistical performance has been carefully analyzed in the case of oblivious embeddings (see, for instance, [18], [35], [73]). Hence, our analysis extends the range of existing results of the left sketch approach to the case of adaptive embeddings.

VII. Numerical experiments

Datasets. We evaluate Algorithm 1 on the MNIST and CIFAR10 datasets with logistic regression. First, we aim to illustrate that the sketch size can be considerably smaller than the data dimension while recovering a close approximation to the optimal solution which achieves a similar test classification accuracy. Second, we aim to achieve significant computational speed-ups. To solve the primal program (1), we use two standard algorithms for empirical risk minimization, namely, stochastic gradient descent (SGD) with (best) fixed step size and stochastic variance reduction gradient (SVRG) [74] with (best) fixed step size and frequency update of the gradient correction. To solve the adaptive sketched program (2), we use SGD, SVRG and the sub-sampled Newton method [75], [76]: we refer to them as Sketch-SGD, Sketch-SVRG and Sketch-Newton. The latter is well-suited to the sketched program for a relatively small sketch size, as a low-dimensional Newton system can be efficiently solved at each iteration. For both datasets, we use 50000 training and 10000 testing images. We transform each image a using random Fourier features $\psi(a) \in \mathbb{R}^d$, i.e., $\langle \psi(a), \psi(a') \rangle \approx \exp\left(-\gamma \|a - a'\|_2^2\right)$ [2], [77]. For MNIST and CIFAR10, we choose respectively $d = 10000$ and $\gamma = 0.02$, and, $d = 60000$ and $\gamma = 0.002$, so that the primal programs are respectively 10000-dimensional and 60000-dimensional. Then, we train a classifier via a sequence of binary logistic regressions, using a one-versus-all procedure.

Adaptive Gaussian embeddings. We evaluate the test classification error of the first-order estimator $\widehat{x}^{(1)}$. We solve to optimality the primal and sketched programs for values of $\lambda \in \{10^{-4}, 5 \cdot 10^{-5}, 10^{-5}, 5 \cdot 10^{-6}\}$ and sketch sizes $m \in \{128, 256, 512, 1024\}$. In Table II are reported the empirical means and standard deviations, which are averaged over 20 trials. The adaptive sketched program yields a high accuracy classifier for most pairs (λ, m): we match the best primal classifier with values of m as small as 256 for MNIST and 512 for CIFAR10, which respectively corresponds to a dimension reduction by a factor ≈ 40 and ≈ 120. These results also suggest that adaptive sketching induces an implicit regularization effect, which is reminiscent of the benefits of spectral cutoff estimators [78]. For instance, on CIFAR10, using $\lambda = 10^{-5}$ and $m = 512$, we obtain an improvement in test accuracy by more than 2% compared to x^*.

TABLE II
Test classification error of adaptive Gaussian sketching on MNIST and CIFAR10 datasets. The subscript in $\widehat{x}_m^{(1)}$ refers to the sketch size m. Results are averaged over 20 trials, and we report the empirical mean and standard deviation.

λ	x^*_{MNIST}	$\widehat{x}^{(1)}_{128}$	$\widehat{x}^{(1)}_{256}$	$\widehat{x}^{(1)}_{512}$	$\widehat{x}^{(1)}_{1024}$
10^{-4}	5.4	4.8 ± 0.2	5.2 ± 0.1	5.3 ± 0.1	5.4 ± 0.1
$5{\cdot}10^{-5}$	4.6	3.8 ± 0.2	4.0 ± 0.2	4.3 ± 0.1	4.5 ± 0.1
10^{-5}	2.8	3.4 ± 0.8	2.4 ± 0.2	2.5 ± 0.1	2.8 ± 0.1
$5{\cdot}10^{-6}$	2.5	4.9 ± 2.1	2.8 ± 0.3	2.6 ± 0.2	2.4 ± 0.1
λ	x^*_{CIFAR}	$\widehat{x}^{(1)}_{128}$	$\widehat{x}^{(1)}_{256}$	$\widehat{x}^{(1)}_{512}$	$\widehat{x}^{(1)}_{1024}$
$5{\cdot}10^{-5}$	51.6	50.5 ± 0.3	50.6 ± 0.3	50.8 ± 0.2	51.0 ± 0.2
10^{-5}	48.2	54.5 ± 3.2	47.7 ± 0.6	45.9 ± 0.2	46.2 ± 0.2
$5{\cdot}10^{-6}$	47.6	59.8 ± 3.5	51.9 ± 2.1	47.7 ± 0.6	45.8 ± 0.6

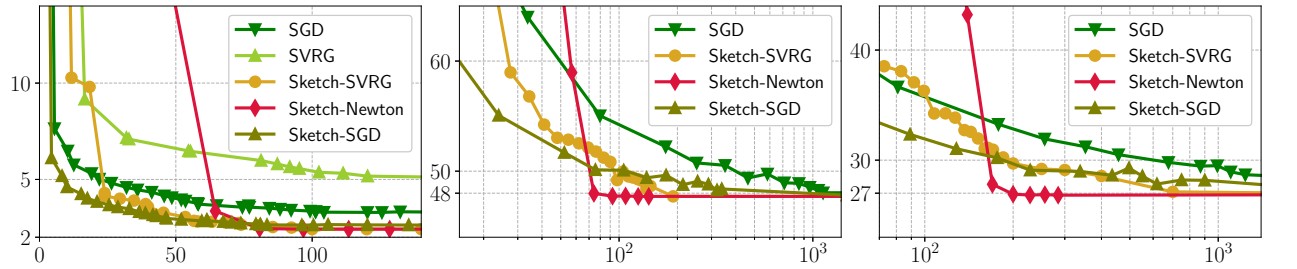
Adaptive versus oblivious Gaussian sketching. We evaluate the test classification error of two baseline estimators, that is, the first-order estimator $\widehat{x}^{(1)}_{\dagger}$ with oblivious Gaussian embeddings as proposed in [31], [32] and described in Section IV-B, and, the first-order estimator $\widehat{x}^N$ with adaptive Nystrom embeddings for which $S = A^\top \widetilde{S}$ with $\widetilde{S}$ a uniformly random column sub-sampling matrix. As reported in Table III, adaptive Gaussian sketching performs better for a wide range of values of sketch size m and regularization parameter λ.

TABLE III
TEST CLASSIFICATION ERROR (IN PERCENTAGE) ON MNIST AND CIFAR10. RESULTS ARE AVERAGED OVER 20 TRIALS AND WE REPORT THE EMPIRICAL MEAN. THE SUBSCRIPT UNDER EACH ESTIMATOR REFERS TO THE SKETCH SIZE m.

λ	x^*_{MNIST}	$\widehat{x}^{(1)}_{256}$	$\widehat{x}^{(1)}_{1024}$	$\widehat{x}^{(1)}_{\dagger,256}$	$\widehat{x}^{(1)}_{\dagger,1024}$	$\widehat{x}^{N}_{256}$	$\widehat{x}^{N}_{1024}$
$5\cdot10^{-5}$	4.6	4.0	4.5	25.2	8.5	5.0	4.6
$5\cdot10^{-6}$	2.0	2.8	2.4	30.1	9.4	3.0	2.7

λ	x^*_{CIFAR}	$\widehat{x}^{(1)}_{256}$	$\widehat{x}^{(1)}_{1024}$	$\widehat{x}^{(1)}_{\dagger,256}$	$\widehat{x}^{(1)}_{\dagger,1024}$	$\widehat{x}^{N}_{256}$	$\widehat{x}^{N}_{1024}$
$5\cdot10^{-5}$	51.6	50.6	51.0	88.2	70.5	55.8	53.1
$5\cdot10^{-6}$	47.6	51.9	45.8	88.9	80.1	57.2	55.8

Wall-clock time speed-ups. For the first-order estimator $\widehat{x}^{(1)}$ with adaptive Gaussian embeddings, we compare the test classification error versus wall-clock time of the aforementioned optimization algorithms. Figure 4 shows results for some values of m and λ. We observe that solving instead the low-dimensional optimization problem offers significant speed-ups on the 10000-dimensional MNIST problem, in particular for Sketch-SGD and for Sketch-SVRG, for which computing the gradient correction is relatively fast. Such speed-ups are even more significant on the 60000-dimensional CIFAR10 problem, especially for Sketch-Newton, and a few iterations suffice to closely reach the approximate solution $\widehat{x}^{(1)}$, with a per-iteration time which is relatively small thanks to dimensionality reduction. Remarkably, it is more than 10 times faster to reach the best test accuracy classifier using the sketched program. In addition to random Fourier features, we carry out another set of experiments with the CIFAR10 dataset, in which we pre-process the images. That is, similarly to [79], [80], we map each image through a random convolutional layer. Then, we kernelize these processed images using a Gaussian kernel with $\gamma = 2 \cdot 10^{-5}$. Using our implementation, the best test accuracy of the kernel primal program (31) we obtained is 73.1%. Sketch-SGD, Sketch-SVRG and Sketch-Newton – applied to the sketched kernel program (32) – match this test accuracy, with significant speed-ups, as reported in Figure 4.



(a) MNIST, $m = 512$, $\lambda = 10^{-5}$. (b) CIFAR, $m = 256$, $\lambda = 10^{-5}$. (c) CIFAR (random layer), $m = 1024$, $\lambda = 5.10^{-6}$.

Fig. 4. Test classification error (percentage) versus wall-clock time (seconds).

Additional numerical details. Experiments were run in Python on a workstation with 512 GB of memory. We use our own implementation of each algorithm for a fair comparison. For SGD, we use a batch size equal to 128. For SVRG, we use a batch size equal to 128 and update the gradient correction every 400 iterations. For Sketch-SGD, we use a batch size equal to 1024. For Sketch-SVRG, we use a batch size equal to 64 and update the gradient correction every 200 iterations. Each iteration of the sub-sampled Newton method (Sketch-Newton) computes a full-batch gradient, and, the Hessian with respect to a batch of size 1500. For SGD and SVRG, we considered step sizes η between 10^{-2} and 10^2. We obtained best performance for $\eta = 10^1$. For the sub-sampled Newton method, we use a step size $\eta = 1$, except for the first 5 iterations, for which we use $\eta = 0.2$. In Figure 4, we did not report results for SVRG for solving the primal (1) on CIFAR10, as the computation time for reaching a satisfying performance was significantly larger than for the other algorithms.

VIII. Information theoretic lower-bounds and optimality for right-sketching

We now consider the fundamental problem of estimating the mean of a random sample $b = Ax_{\rm pl} + w$ where $w \sim \mathcal{N}(0, \frac{\sigma^2}{n} I_n)$, under the assumption that the planted vector $x_{\rm pl}$ satisfies $\|x_{\rm pl}\|_2 \leqslant 1$. Given an estimator $\widehat{x}$, we define its risk as $\mathfrak{R}(\widehat{x}) := \sup_{\|x_{\rm pl}\|_2 \leqslant 1} \mathbb{E}_w \|A(\widehat{x} - x_{\rm pl})\|_2^2$. A critical quantity to characterize the best achievable risk is the *statistical dimension* d_s, defined as $d_s := \min\{k \geqslant 1 \mid \frac{\sigma^2 k}{n} \geqslant \sigma_{k+1}^2\}$. It satisfies the scaling $\frac{\sigma^2 d_s}{n} \asymp \sigma_{d_s+1}^2$. It is well-known (see, for instance, [43], [81], [82]) that the quantity $A^\top b$ is sufficient to construct an optimal estimator.

Information theoretic lower bounds for left-sketching via SA, Sb were developed in [17]. Surprisingly, it was shown that unless $m \asymp d$, left-sketching based estimators are sub-optimal. In turn, an iterative sketching method based on the sketches $\{S_i A, A^T(Ax_i - b)\}_{i=1}^T$ was shown to be statistically optimal for the broader class of constrained least squares problems, including unconstrained least squares, Lasso and nuclear norm constrained problems.

In our context, we are given a sketch AS where $S \in \mathbb{R}^{d \times m}$. Note that right-sketched based optimization problem (2) for the least squares objective is of the form

$$\min_{\alpha} \|AS\alpha - b\|_2^2 + \phi(\alpha) = \min_{\alpha} \|AS\alpha\|_2^2 - 2\alpha^T S^T A^T b + \|b\|_2^2 + \phi(\alpha) \tag{58}$$

where $\phi(\alpha)$ is an arbitrary regularization term. The preceding line shows that right-sketching estimators are functions of $S^T A^T b$. Therefore, bounds on the mutual information between $S^T A^T b$ and $x_{\rm pl}$ can be leveraged to develop information theoretic lower bounds under this observation model. Consequently, we consider right-sketching estimators based on the observation $S^\top A^\top b$, and we aim to characterize the minimax risk $\mathfrak{M}_S := \inf_{\widehat{x}} \mathfrak{R}(\widehat{x})$ where the infimum is taken over estimators $\widehat{x} \equiv \widehat{x}(S^\top A^\top b)$. We aim to show that for an oblivious Gaussian embedding $S \in \mathbb{R}^{d \times m}$, a sketch size $m \asymp d_s$ and polynomial or exponential decays, it holds with high-probability that

$$\mathfrak{M}_S \asymp \frac{\sigma^2 d_s}{n}\,. \tag{59}$$

Moreover, this minimax lower-bound is achieved by the right-sketching estimator. Our proof of the lower-bound is based on the standard local Fano method.

Theorem 9. *Let $m \geqslant 1$ be a sketch size and $S \in \mathbb{R}^{d \times m}$ be a Gaussian embedding, and assume that $d_s \geqslant 4$. Conditional on the event $\|P_{AS}^{\perp} A\|_2^2 \leqslant \frac{\sigma_{d_s+1}^2}{2}$, it holds that*

$$\mathfrak{M}_S \geqslant c_0 \cdot \sigma_{d_s+1}^2 \asymp c_0 \cdot \frac{\sigma^2 d_s}{n}\,, \tag{60}$$

where c_0 is a universal constant such that $c_0 \geqslant \frac{1}{4096}$. Consequently, for a polynomial decay $\sigma_j = j^{-\frac{1+\nu}{2}}$ with $\nu > 0$ for which $d_s \asymp \left(\frac{n}{\sigma^2}\right)^{\frac{1}{2+\nu}}$ and for a sketch size $m \geqslant c_\nu^{poly} \cdot d_s$ where $c_\nu^{poly} \asymp (\nu^{-1} + 1)^{(1+\nu)}$, it holds with probability at least $1 - 6e^{-d_s}$ that

$$\mathfrak{M}_S \geqslant c_0 \cdot \sigma_{d_s+1}^2 \asymp c_0 \cdot \left(\frac{\sigma^2}{n}\right)^{\frac{1+\nu}{2+\nu}}\,. \tag{61}$$

For an exponential decay $\sigma_j = e^{-\frac{\nu j}{2}}$ with $\nu > 0$ such that $n\nu/\sigma^2 \geqslant 5408$ for which $d_s \asymp \frac{1}{\nu}\log(n/\sigma^2)$ and for a sketch size $m \geqslant 4d_s$, it holds with probability at least $1 - 6e^{-d_s}$ that

$$\mathfrak{M}_S \geqslant c_0 \cdot \sigma_{d_s+1}^2 \asymp c_0 \cdot \frac{\sigma^2 \log(n/\sigma^2)}{\nu n}\,. \tag{62}$$

We show next that the risk of the zero-order estimator $\widehat{x}^{(0)}$ with sketch size $m \asymp d_s$ achieves the above lower bound on the minimax rate of estimation (60), as the regularization parameter λ goes to 0.

Corollary 2. *Let $m \geqslant 1$ be a sketch size and $S \in \mathbb{R}^{d\times m}$ be a Gaussian embedding. Assume that $d_s \geqslant 4$. For a polynomial decay $\sigma_j = j^{-\frac{1+\nu}{2}}$ and with a sketch size $m = c_\nu^{poly} \cdot d_s$ where $c_\nu^{poly} \asymp (\nu^{-1}+1)^{(1+\nu)}$, we have with probability at least $1 - 6e^{-d_s}$ that*

$$\lim_{\lambda\to 0} \mathfrak{R}(\widehat{x}^{(0)}) \leqslant \left(c_\nu^{poly} + \frac{1}{2}\right) \cdot \frac{\sigma^2 d_s}{n}\,. \tag{63}$$

For an exponential decay $\sigma_j = e^{-\frac{\nu j}{2}}$ with $\nu > 0$ such that $n\nu/\sigma^2 \geqslant 5408$, and with a sketch size $m = 4d_s$, we have with probability at least $1 - 6e^{-d_s}$ that

$$\lim_{\lambda\to 0} \mathfrak{R}(\widehat{x}^{(0)}) \leqslant 5 \cdot \frac{\sigma^2 d_s}{n}\,. \tag{64}$$

Therefore we conclude that right-sketching is minimax optimal unlike left-sketching under this standard observation model. This result complements the left-sketching lower-bounds from [17] and indicates that right-sketching is more advantageous for problems with small statistical dimension.

IX. Conclusion

Through tighter performance bounds and analytical comparison, we have shown that the dual reconstruction method along with adaptive embeddings yields an estimator $\widehat{x}^{(1)}$ which significant improves over the linear reconstruction map and oblivious sketching in the context of convex smooth optimization, as usually considered in the literature. Furthermore, we have extended this method to non-smooth optimization problems, and our method requires solving an additional dual optimization problem with potentially very few variables: in contrast to optimizing over a random subset of dual variables (e.g., SDNA), our primal low-dimensional approach selects the appropriate subset of dual variables. Most of our results mirror those established for left-sketching methods [35], although they are fundamentally different due to the choice of the adaptive embedding and thus, require a novel analysis technique.

Appendix A
Analysis of adaptive sketches

A. Proof of Lemma 1

A proof of an upper bound on the singular value $\|P_S A^\top\|_2$ is provided in [46], and our analysis is an adaptation of this proof to the restricted case. For two real-valued random variables X and Y, we say that X is stochastically dominated by Y if $\mathbb{P}(X \geqslant \tau) \leqslant \mathbb{P}(Y \geqslant \tau)$ for any $\tau \in \mathbb{R}$, and we write $X \overset{d}{\leqslant} Y$.

Let $S \in \mathbb{R}^{n\times m}$ be a matrix with i.i.d. Gaussian entries $\mathcal{N}(0, 1/m)$. We use the notation $f(A,S) := P^{\perp}_{A^\top S} A^\top$, and we introduce a thin SVD of A, denoted by $A = U\Sigma V^\top$, where $\Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_\rho\}$. Note that

$$P_{A^\top S} = A^\top S(S^\top AA^\top S)^\dagger S^\top A = V\Sigma U^\top S(S^\top U\Sigma \underbrace{V^\top V}_{=I_\rho} \Sigma U^\top S)^\dagger S^\top U\Sigma V^\top = V P_{\Sigma U^\top S} V^\top\,.$$

Consequently, we have that

$$f(A,S) = (I - P_{A^\top S})A^\top = (I - VP_{\Sigma U^\top S}V^\top)V\Sigma U^\top = V\Sigma U^\top - VP_{\Sigma U^\top S}\underbrace{V^\top V}_{=I_\rho}\Sigma U^\top = V\underbrace{(\Sigma - P_{\Sigma U^\top S}\Sigma)}_{=f(\Sigma, U^\top S)}U^\top\,.$$

That is, $f(A,S) = Vf(\Sigma, U^\top S)U^\top$. Let $G \in \mathbb{R}^{\rho\times m}$ be a matrix with i.i.d. Gaussian entries $\mathcal{N}(0,1/m)$. By rotational invariance of the Gaussian distribution, it holds that $G \overset{\mathrm{d}}{=} U^\top S$. Therefore, $f(A,S) \overset{\mathrm{d}}{=} Vf(\Sigma, G)U^\top$. Since V is an isometry (i.e., $\|Vw\|_2 = \|w\|_2$ for any $w \in \mathbb{R}^\rho$), it follows that

$$\mathcal{Z}_f \overset{\mathrm{d}}{=} \sup_{\Delta\in\mathcal{C}_{z^*}} \|f(\Sigma, G)U^\top\Delta\|_2 \,.$$

For $t>0$, we define $M(t) := \begin{bmatrix} tI_k & 0 \\ 0 & \Sigma_{\rho-k}\end{bmatrix}$. Note that $\Sigma \preceq M(t)$ for any $t \geqslant \sigma_1$. According to Lemma 2.5 in [46], it holds that $\|f(\Sigma_1, G)w\|_2 \leqslant \|f(\Sigma_2, G)w\|_2$ for any $w \in \mathbb{R}^\rho$ and any two positive definite diagonal matrices Σ_1 and Σ_2 such that $\Sigma_1 \preceq \Sigma_2$. As a consequence, for any $\Delta \in \mathcal{C}_{z^*}$, it holds almost surely that $\|f(\Sigma, G)U^\top\Delta\|_2 \leqslant \lim_{t\to\infty} \|f(M(t), G)U^\top\Delta\|_2$, and thus,

$$\mathcal{Z}_f \overset{d}{\leqslant} \sup_{\Delta\in\mathcal{C}_{z^*}} \lim_{t\to\infty} \|f(M(t), G)U^\top\Delta\|_2 \,. \tag{65}$$

The following fact has already been proved in [46]: it holds that

$$\lim_{t\to+\infty} f(M(t), G) \overset{\mathrm{d}}{=} \begin{bmatrix} 0 & 0 \\ f(\Sigma_{\rho-k}, X_2)X_1\Lambda^{-1}\Omega & f(\Sigma_{\rho-k}, X_2)\end{bmatrix},$$

where $X_1 \in \mathbb{R}^{(\rho-k)\times k}$, $X_2 \in \mathbb{R}^{(\rho-k)\times k}$, $\Lambda \in \mathbb{R}^{k\times k}$ and $\Omega \in \mathbb{R}^{k\times k}$ are independent random matrices, and, X_1 and X_2 have independent i.i.d. Gaussian entries, Λ is diagonal with entries distributed as the first k singular values of a $k\times m$ Gaussian matrix, and Ω is an orthogonal matrix. Plugging-in this limit in the right-hand side of (65), we obtain

$$\mathcal{Z}_f \overset{\mathrm{d}}{\leqslant} \sup_{\Delta\in\mathcal{C}_{z^*}} \|f(\Sigma_{\rho-k}, X_2)X_1\Lambda^{-1}\Omega U_k^\top\Delta + f(\Sigma_{\rho-k}, X_2)U_{\rho-k}^\top\Delta\|_2 \,.$$

By triangular inequality, it follows that

$$\begin{aligned}
\mathcal{Z}_f &\overset{\mathrm{d}}{\leqslant} \sup_{\Delta\in\mathcal{C}_{z^*}} \|f(\Sigma_{\rho-k}, X_2)X_1\Lambda^{-1}\Omega U_k^\top\Delta\|_2 + \sup_{\Delta\in\mathcal{C}_{z^*}} \|f(\Sigma_{\rho-k}, X_2)U_{\rho-k}^\top\Delta\|_2 \\
&\leqslant \mathrm{rad}\{U_k^\top\mathcal{C}_{z^*}\}\cdot\|f(\Sigma_{\rho-k}, X_2)X_1\Lambda^{-1}\Omega\|_2 + \sup_{\Delta\in\mathcal{C}_{z^*}} \|f(\Sigma_{\rho-k}, X_2)U_{\rho-k}^\top\Delta\|_2 \\
&\leqslant \mathrm{rad}\{U_k^\top\mathcal{C}_{z^*}\}\cdot\|\Sigma_{\rho-k}X_1\Lambda^{-1}\|_2 + \sup_{\Delta\in\mathcal{C}_{z^*}} \|\Sigma_{\rho-k}U_{\rho-k}^\top\Delta\|_2 \,,
\end{aligned}$$

and we used the fact that $\|f(\Sigma_{\rho-k}, X_2)W\|_2 = \|P^\perp_{\Sigma_{\rho-k}X_2}\Sigma_{\rho-k}W\|_2 \leqslant \|\Sigma_{\rho-k}W\|_2$ for any arbitrary matrix W. According to Corollary 10.9 [44], it holds with probability at least $1-6e^{-k}$ that

$$\|\Sigma_{\rho-k}X_1\Lambda^{-1}\|_2 \leqslant 17\sqrt{2}\cdot\sigma_{k+1} + \frac{9}{\sqrt{k}}\cdot\sqrt{\sum_{j>k}\sigma_j^2} \leqslant \underbrace{\max\{17\sqrt{2}, 9\}}_{\leqslant 25}\cdot\left(\sigma_{k+1} + \sqrt{\frac{1}{k}\sum_{j>k}\sigma_j^2}\right) \leqslant 25\cdot R_k(A)\,.$$

In summary, we have shown that with probability at least $1-6e^{-k}$, we have

$$\mathcal{Z}_f \leqslant 25\cdot\mathrm{rad}\{U_k^\top\mathcal{C}_{z^*}\}\cdot R_k(A) + \sup_{\Delta\in\mathcal{C}_{z^*}} \|\Sigma_{\rho-k}U_{\rho-k}^\top\Delta\|_2 \,,$$

which is the first part of the claim. Furthermore, since $\mathcal{C}_{z^*} \subset \mathcal{B}_2^n$, we always have that

$$\mathrm{rad}\{U_k^\top\mathcal{C}_{z^*}\} \leqslant 1\,, \quad \text{and} \quad \sup_{\Delta\in\mathcal{C}_{z^*}} \|\Sigma_{\rho-k}U_{\rho-k}^\top\Delta\|_2 \leqslant \sigma_{k+1}\,.$$

Therefore, $\mathcal{Z}_f \leqslant 25 \cdot R_k(A) + \sigma_{k+1} \leqslant 26 \cdot R_k(A)$ with probability at least $1 - 6e^{-k}$. This holds in particular when $\mathcal{C}_{z^*} = \mathcal{B}_2^n$, which concludes the proof. □

B. Proof of Lemma 2

From Lemma 1, we know that with $m = 2k$ and $S = A^\top \widetilde{S}$ where $\widetilde{S} \in \mathbb{R}^{n \times m}$ has i.i.d. Gaussian entries, it holds with probability at least $1 - 6e^{-k}$ that

$$\mathcal{Z}_f \leqslant 25 \cdot \mathrm{rad}(U_k^\top \mathcal{C}_{z^*}) \cdot R_k(A) + \sup_{\Delta \in \mathcal{C}_{z^*}} \|\Sigma_{\rho-k} U_{\rho-k}^\top \Delta\|_2 \,,$$

From Theorem 7.7.1. in [83] and using that $\mathrm{rad}(\mathcal{C}_{z^*}) \leqslant 1$, we have

$$\mathrm{rad}\big(U_k^\top \mathcal{C}_{z^*}\big) \leqslant c_1' \cdot \frac{\omega(\mathcal{C}_{z^*}) + \sqrt{m}}{\sqrt{n}} \,, \tag{66}$$

with probability at least $1 - 2e^{-k}$, for some universal constant $c_1' > 0$. It remains to control the term $\sup_{\Delta \in \mathcal{C}_{z^*}} \|\Sigma_{\rho-k} U_{\rho-k}^\top \Delta\|_2 = \sup_{\Delta \in \mathcal{C}_{z^*}, t \in \mathcal{E}} \langle t, U_{\rho-k}^\top \Delta \rangle$, where we define the ellipsoid $\mathcal{E} = \{\Sigma_{\rho-k} z \mid \|z\|_2 \leqslant 1\}$. We use again a Chevet-type inequality [84], which yields that

$$\sup_{\Delta \in \mathcal{C}_{z^*}, t \in \mathcal{E}} \langle t, U_{\rho-k}^\top \Delta \rangle \leqslant c_2' \cdot \frac{1}{\sqrt{n}} \left(\omega(\mathcal{C}_{z^*}) \,\mathrm{rad}(\mathcal{E}) + w(\mathcal{E}) \,\mathrm{rad}(\mathcal{C}_{z^*}) \right) \,, \tag{67}$$

with probability at least $1 - c_4 e^{-c_5(\rho-k)}$ for some universal constants $c_4, c_5 > 0$, and where we introduced the ellipsoid $\mathcal{E} = \{\Sigma_{\rho-k} z \mid \|z\|_2 \leqslant 1\}$. Using the facts that $\omega(\mathcal{E}) \leqslant \left(\sum_{j=k+1} \sigma_j^2\right)^{\frac{1}{2}}$, $\mathrm{rad}(\mathcal{E}) = \sigma_{k+1}$ and $\mathrm{rad}(\mathcal{C}_{z^*}) \leqslant 1$, the above inequality becomes

$$\begin{aligned} \sup_{\Delta \in \mathcal{C}_{z^*}} \|\Sigma_{\rho-k} U_{\rho-k}^\top \Delta\|_2 &\leqslant c_2' \left(\sigma_{k+1} \frac{\omega(\mathcal{C}_{z^*})}{\sqrt{n}} + \sqrt{\frac{k}{n}} \cdot \sqrt{\frac{1}{k} \sum_{j=k+1}^{r} \sigma_j^2} \right) \\ &\leqslant c_3' \cdot \frac{\omega(\mathcal{C}_{z^*}) + \sqrt{m}}{\sqrt{n}} \cdot R_k(A) \,. \end{aligned}$$

with probability at least $1 - c_4 e^{-c_5(\rho-k)}$, for some universal constant $c_3' > 0$. By union bound, we obtain the claimed result.

C. Proof of Lemma 3

We use Theorem 2.1 from [45], which states the following. Given a target rank $2 \leqslant k \leqslant \rho$ and a failure probability $0 < \delta < 1$, let $S = A^\top \widetilde{S}$ where $\widetilde{S} \in \mathbb{R}^{n \times m}$ is a SRHT with $n \geqslant m \geqslant 19(\sqrt{k} + \sqrt{8\log(n/\delta)})^2 \log(k/\delta)$. Then, it holds with probability at least $1 - 5\delta$ that $\|P_S^\perp A^\top\|_2 \leqslant \left(4 + \sqrt{\frac{3\log(n/\delta)\log(\rho/\delta)}{m}}\right) \cdot \sigma_{k+1} + \sqrt{\frac{3\log(\rho/\delta)}{m}} \cdot \sqrt{\sum_{j=k+1}^{\rho} \sigma_j^2}$. Picking $\delta = \frac{1}{n}$ and using that $n \geqslant \rho$, we obtain for $m \geqslant 19\left(\sqrt{k} + 4\sqrt{\log n}\right)^2 \log(nk)$ and with probability at least $1 - \frac{5}{n}$ that $\|P_S^\perp A^\top\|_2 \leqslant \left(4 + \sqrt{\frac{12\log^2(n)}{m}}\right) \cdot \sigma_{k+1} + \sqrt{6} \cdot \sqrt{\frac{\log(n)}{m}} \cdot \sqrt{\sum_{j=k+1}^{\rho} \sigma_j^2}$. We conclude by using that $\sqrt{\frac{12\log^2(n)}{m}} \leqslant 1$ and $\sqrt{\frac{\log(n)}{m}} \leqslant \frac{1}{\sqrt{k}}$. □

D. Proof of Theorem 1

From Propositions 1 and 2, we know that there exist $g_{z^*} \in \partial f^*(z^*)$ and $g_{y^*} \in \partial f^*(y^*)$ such that $g_{z^*} + \frac{1}{\lambda} A A^\top z^* = 0$ and $g_{y^*} + \frac{1}{\lambda} A P_S A^\top y^* = 0$. We define the error vector $\Delta := y^* - z^*$, which belongs to the tangent cone $\mathcal{T}_{z^*}$. Subtracting the two

previous equalities, we obtain $g_{y^*} - g_{z^*} + \frac{1}{\lambda} A P_S A^\top \Delta = \frac{1}{\lambda} A P_S^\perp A^\top z^*$. Multiplying by $\lambda \Delta^\top$ and using the fact that $P_S^2 = P_S$, it follows that

$$\lambda \left\langle \Delta, g_{y^*} - g_{z^*} \right\rangle + \|P_S A^\top \Delta\|_2^2 = \left\langle \Delta, A P_S^\perp A^\top z^* \right\rangle . \tag{68}$$

Smoothness of f implies that its Fenchel conjugate f^* is μ^{-1}-strongly convex, i.e., $\langle \Delta, g_{y^*} - g_{z^*} \rangle \geqslant \frac{1}{\mu} \|\Delta\|_2^2$. By definition of $\mathcal{Z}_f$, we have that $\|P_S^\perp A^\top \Delta\|_2 \leqslant \mathcal{Z}_f \|\Delta\|_2$, and consequently,

$$\|P_S A^\top \Delta\|_2^2 = \|A^\top \Delta\|_2^2 - \|P_S^\perp A^\top \Delta\|_2^2 \geqslant \|A^\top \Delta\|_2^2 - \mathcal{Z}_f^2 \|\Delta\|_2^2 .$$

Plugging-in the previous inequalities into (68), we obtain

$$\left(\frac{\lambda}{\mu} - \mathcal{Z}_f^2 \right) \cdot \|\Delta\|_2^2 + \|A^\top \Delta\|_2^2 \leqslant \left\langle \Delta, A P_S^\perp A^\top z^* \right\rangle .$$

Using the assumption $\lambda \geqslant 2\mu \mathcal{Z}_f^2$, it follows that $\frac{\lambda}{\mu} - \mathcal{Z}_f^2 \geqslant \frac{\lambda}{2\mu}$. By Cauchy-Schwarz inequality, we have $|\langle \Delta, A P_S^\perp A^\top z^* \rangle| \leqslant \|P_S^\perp A^\top \Delta\|_2 \|P_S^\perp A^\top z^*\|_2$. Hence, we obtain the inequality

$$\frac{\lambda}{2\mu} \|\Delta\|_2^2 + \|A^\top \Delta\|_2^2 \leqslant \|P_S^\perp A^\top \Delta\|_2 \|P_S^\perp A^\top z^*\|_2 .$$

Using the identity $a^2 + b^2 \geqslant 2ab$ with $a = \sqrt{\frac{\lambda}{2\mu}} \|\Delta\|_2$ and $b = \|A^\top \Delta\|_2$ and the inequality $\|P_S^\perp A^\top \Delta\|_2 \leqslant \mathcal{Z}_f \cdot \|\Delta\|_2$, it follows that

$$\sqrt{\frac{2\lambda}{\mu}} \cdot \|\Delta\|_2 \cdot \|A^\top \Delta\|_2 \leqslant \|\Delta\|_2 \cdot \mathcal{Z}_f \cdot \|P_S^\perp A^\top z^*\|_2 .$$

Using the identities $x^* = -\lambda^{-1} A^\top z^*$ and $\widehat{x}^{(1)} = -\lambda^{-1} A^\top y^*$ and rearranging the above inequality, we obtain

$$\|\widehat{x}^{(1)} - x^*\|_2 \leqslant \sqrt{\frac{\mu}{2\lambda}} \cdot \mathcal{Z}_f \cdot \|P_S^\perp x^*\|_2 . \tag{69}$$

It always holds that $\|P_S^\perp x^*\|_2 \leqslant \|x^*\|_2$, and consequently, we have $\frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2} \leqslant \sqrt{\frac{\mu}{2\lambda}} \cdot \mathcal{Z}_f$. On the other hand, we have $\|\widehat{x}^{(0)} - x^*\|_2^2 = \|P_S(\widehat{x}^{(0)} - x^*)\|_2^2 + \|P_S^\perp x^*\|_2^2$, which further implies $\|P_S^\perp x^*\|_2 \leqslant \|\widehat{x}^{(0)} - x^*\|_2$. Consequently, we also obtain from (69) that $\frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2} \leqslant \sqrt{\frac{\mu}{2\lambda}} \cdot \mathcal{Z}_f \cdot \frac{\|\widehat{x}^{(0)} - x^*\|_2}{\|x^*\|_2}$, and this concludes the proof. □

E. Proof of Theorem 7

Using an induction argument, it suffices to show that for any $t \geqslant 0$ and provided that $\lambda \geqslant 2\mu \mathcal{Z}_f^2$, we have

$$\|\widehat{x}_{t+1}^{(1)} - x^*\|_2 \leqslant \sqrt{\frac{\mu}{2\lambda}} \cdot \mathcal{Z}_f \cdot \|\widehat{x}_t^{(1)} - x^*\|_2 . \tag{70}$$

It should be noted that for $t = 0$, since $\widehat{x}_0^{(1)} = 0$, the latter inequality is exactly the regret bound (17). The proof for any $t \geqslant 0$ follows steps similar to the proof of Theorem 1. Fix $t \geqslant 0$. The Fenchel dual of the sketched program (47) is given by

$$\min_{y \in \mathbb{R}^n} \left\{ f^*(y) - y^\top A P_S^\perp \widehat{x}_t^{(1)} + \frac{1}{2\lambda} \|P_S A^\top y\|_2^2 \right\} .$$

Using arguments similar to the proof of Proposition 2, we obtain that there exists a dual solution $y_t^* \in \mathbb{R}^n$, and that $\alpha_{\dagger,t}^*$ and y_t^* are related through the KKT conditions $y_t^* = \nabla f(A Q_S \alpha_{\dagger,t}^* + A \widehat{x}_t^{(1)})$. Recall that, by definition, $\widehat{x}_{t+1}^{(1)} = -\lambda^{-1} A^\top \nabla f(A Q_S \alpha_{\dagger,t}^* + A \widehat{x}_t^{(1)})$, i.e., $\widehat{x}_{t+1}^{(1)} = -\lambda^{-1} A^\top y_t^*$. We define the error vector $\Delta := y_t^* - z^*$, which belongs to the tangent cone $\mathcal{T}_{z^*}$. By first-order

optimality conditions of y_t^* and z^*, we know that there exist $g_{y_t^*} \in \partial f^*(y_t^*)$ and $g_{z^*} \in \partial f^*(z^*)$ such that $g_{y_t^*} + \frac{1}{\lambda} A P_S A^\top y_t^* - A P_S^\perp \widehat{x}_t^{(1)} = 0$ and $g_{z^*} + \frac{1}{\lambda} A A^\top z^* = 0$. Subtracting both sides of the previous inequalities and multiplying by $\lambda \Delta^\top$, we obtain that

$$\lambda \left\langle \Delta, g_{y_t^*} - g_{z^*} \right\rangle + \|P_S A^\top \Delta\|_2^2 = \langle A P_S^\perp (\lambda \widehat{x}_t^{(1)} + A^\top z^*), \Delta \rangle \,. \tag{71}$$

By definition of $\mathcal{Z}_f$, we have that $\|P_S^\perp A^\top\|_2 \leqslant \mathcal{Z}_f \|\Delta\|_2$, and thus, $\|P_S A^\top \Delta\|_2^2 \geqslant \|A^\top \Delta\|_2^2 - \mathcal{Z}_f^2 \|\Delta\|_2^2$. Plugging-in this inequality into (71) as well as the strong convexity inequality $\langle \Delta, g_{y_t^*} - g_{z^*} \rangle \geqslant \frac{1}{\mu} \|\Delta\|_2^2$, we obtain that

$$\left(\frac{\lambda}{\mu} - \mathcal{Z}_f^2 \right) \|\Delta\|_2^2 + \|A^\top \Delta\|_2^2 \leqslant \langle \Delta, A P_S^\top (\lambda \widehat{x}_t^{(1)} + A^\top z^*) \rangle \,.$$

Using the assumption $\lambda \geqslant 2\mu \mathcal{Z}_f^2$, we get that $\frac{\lambda}{\mu} - \mathcal{Z}_f^2 \geqslant \frac{\lambda}{2\mu}$. Using further the identity $a^2 + b^2 \geqslant 2ab$ with $a = \sqrt{\frac{\lambda}{2\mu}} \|\Delta\|_2$ and $b = \|A^\top \Delta\|_2$, we deduce that

$$\sqrt{\frac{2\lambda}{\mu}} \cdot \|\Delta\|_2 \cdot \|A^\top \Delta\|_2 \leqslant \mathcal{Z}_f \cdot \|\Delta\|_2 \cdot \|\lambda \widehat{x}_t^{(1)} + A^\top z^*\|_2 \,.$$

Using the identities $\widehat{x}_{t+1}^{(1)} = -\lambda^{-1} A^\top y_t^*$, $x^* = -\lambda^{-1} A^\top z^*$ and thus $A^\top \Delta = \lambda (x^* - \widehat{x}_{t+1}^{(1)})$, we finally obtain

$$\|\widehat{x}_{t+1}^{(1)} - x^*\|_2 \leqslant \sqrt{\frac{\mu}{2\lambda}} \cdot \mathcal{Z}_f \cdot \|\widehat{x}_t^{(1)} - x^*\|_2 \,,$$

which is the claimed inequality (70), and this concludes the proof.

F. Proof of Theorem 8

Using the same arguments as in the proofs of Propositions 1 and 2, we obtain that there exist dual solutions z^* and y^* which belong to the image of ∂f. The function f is L-Lipschitz, and this implies that $\|z^*\|_2 \leqslant L$ and $\|y^*\|_2 \leqslant L$. By first-order optimality conditions, there exist subgradients $g_{y^*} \in \partial f^*(y^*)$ and $g_{z^*} \in \partial f^*(z^*)$ such that $g_{y^*} + \frac{1}{\lambda} A P_S A^\top y^* = 0$ and $g_{z^*} + \frac{1}{\lambda} A A^\top z^* = 0$. We define the error vector $\Delta := y^* - z^*$, which belongs to the tangent cone $\mathcal{T}_{z^*}$ and which satisfies $\|\Delta\|_2 \leqslant 2L$. Subtracting the first-order optimality conditions on y^* and z^*, and multiplying by $\Delta^\top$, we obtain that $\langle \Delta, g_{y^*} - g_{z^*} \rangle + \lambda^{-1} \langle \Delta, A P_S A^\top \Delta \rangle = \lambda^{-1} \langle \Delta, A P_S^\perp A^\top z^* \rangle$. By convexity of f, we have $\langle \Delta, g_{y^*} - g_{z^*} \rangle \geqslant 0$. Using $|\langle \Delta, A P_S^\perp A^\top z^* \rangle| \leqslant \|P_S^\perp A^\top \Delta\|_2 \|P_S^\perp A^\top z^*\|_2$, we further obtain that

$$\|A^\top \Delta\|_2^2 \leqslant \|P_S^\perp A^\top \Delta\|_2^2 + \|P_S^\perp A^\top \Delta\|_2 \|P_S^\perp A^\top z^*\|_2 \,. \tag{72}$$

Since $\Delta \in \mathcal{T}_{z^*}$, we have $\|P_S^\perp A^\top \Delta\|_2 \leqslant \|\Delta\|_2 \mathcal{Z}_f \leqslant 2L \mathcal{Z}_f$. Moreover, we have $\|P_S^\perp A^\top z^*\|_2 \leqslant \|z^*\|_2 \|P_S^\perp A^\top\|_2 \leqslant L \|P_S^\perp A^\top\|_2$. Combining the two latter inequalities with (72), we find that $\|A^\top \Delta\|_2^2 \leqslant 4L^2 \mathcal{Z}_f^2 + 2L^2 \mathcal{Z}_f \|P_S^\perp A^\top\|_2$. Dividing by λ^2 and using the identities $x^* = -\lambda^{-1} A^\top z^*$ and $\widehat{x}^{(1)} = -\lambda^{-1} A^\top y^*$, we obtain the claimed inequality

$$\|\widehat{x}^{(1)} - x^*\|_2 \leqslant 2 \cdot \frac{L}{\lambda} \cdot \sqrt{\mathcal{Z}_f^2 + \frac{1}{2} \mathcal{Z}_f \|P_S^\perp A^\top\|_2} \,.$$

On the other hand, under the assumption that $\inf_w f(w) > -\infty$, it holds that $0 \in \mathrm{dom} f^*$. This implies that $-z^* \in \mathcal{T}_{z^*}$, and consequently, $\|P_S^\perp A^\top z^*\|_2 \leqslant L \mathcal{Z}_f$. Since $\Delta \in \mathcal{T}_{z^*}$, we have $\|P_S^\perp A^\top \Delta\|_2 \leqslant 2L \mathcal{Z}_f$. Combining the two latter inequalities

with (72), we obtain the refined inequality

$$\|\widehat{x}^{(1)} - x^*\|_2 \leqslant \sqrt{6} \cdot \frac{L}{\lambda} \cdot \mathcal{Z}_f \,,$$

and this concludes the proof. □

APPENDIX B
ANALYSIS OF OBLIVIOUS SKETCHES

A. Proof of Lemma 4

Let $S \in \mathbb{R}^{d\times m}$ be a matrix with i.i.d. Gaussian entries. By rotational invariance of the Gaussian distribution, there exists a random Haar matrix $Q \in \mathbb{R}^{d\times(d-m)}$ such that $P_S^{\perp} = QQ^\top$. We have $\mathcal{Z}_f = \sup_{\Delta\in\mathcal{C}_{z^*}} \|P_S^{\perp} A^\top \Delta\|_2 = \sup_{\Delta\in\mathcal{C}_{z^*}} \|QQ^\top A^\top \Delta\|_2 = \sup_{\Delta\in\mathcal{C}_{z^*}} \|Q^\top A^\top \Delta\|_2$, where the last equality holds due to the fact that Q is an isometry. According to standard concentration bounds (see, for instance, Theorem 7.7.1 in [83]), it holds with probability at least $1 - 2e^{-(d-m)}$ that

$$\sup_{\Delta,\Delta'\in\mathcal{C}_{z^*}} \|Q^\top A^\top(\Delta - \Delta')\|_2 \leqslant c \cdot \left(\frac{\omega(A^\top \mathcal{C}_{z^*})}{\sqrt{d}} + \sqrt{\frac{d-m}{d}} \cdot \|A\|_2 \right),$$

where $c > 0$ is some universal constant. The claimed result follows from the fact that $0 \in \mathcal{C}_{z^*}$ and thus,

$$\sup_{\Delta\in\mathcal{C}_{z^*}} \|Q^\top A^\top \Delta\|_2 \leqslant \sup_{\Delta,\Delta'\in\mathcal{C}_{z^*}} \|Q^\top A^\top(\Delta - \Delta')\|_2 \,.$$

□

B. Proof of Theorem 5

1) Technical preliminaries:

Lemma 6. *Suppose that $x, y \in \mathbb{R}^n$ such that $\beta\, x^\top y > \|y\|^2$ for some $\beta > 0$. Then, there exists a symmetric matrix H such that $0 \prec H \preceq \beta \cdot I$ and $y = Hx$.*

Proof. Set $H = \beta I + \frac{1}{(y-\beta x)^\top x}(y-\beta x)(y-\beta x)^\top$. The matrix H is well-defined. Indeed, from the assumption $\beta x^\top y > \|y\|^2$, we obtain that $\|y\| < \beta\|x\|$. Consequently, $y^\top x \leqslant \|y\|\|x\| < \beta\|x\|^2$ so that the term $(y - \beta x)^\top x$ in the denominator is negative. This implies in particular that $H \preceq \beta I$. A simple calculation yields that $Hx = y$. It remains to show that $H \succ 0$. This holds provided that $\beta > \|y - \beta x\|^2/(\beta\|x\|^2 - x^\top y)$, i.e., $\|y\|^2 < \beta x^\top y$ which is true by assumption. □

Lemma 7. *Suppose that the function f is γ-strongly convex and μ-strongly smooth. Let z^* be the solution to the dual program, and y^* the solution to the sketched dual program. Let $g_{z^*} \in \partial f^*(z^*)$ and $g_{y^*} \in \partial f^*(y^*)$. Then, there exists a symmetric matrix $0 \prec H \preceq \frac{2}{\gamma} \cdot I$ such that $g_{y^*} - g_{z^*} = H(y^* - z^*)$.*

Proof. From standard convex analysis arguments, the function f^* is $(1/\gamma)$-smooth: this implies that $\|g_{y^*} - g_{z^*}\|_2^2 \leqslant \frac{1}{\gamma}(g_{y^*} - g_{z^*})^\top(y^* - z^*)$. The function f^* is also $(1/\mu)$-strongly convex: this implies that $\frac{1}{\mu}\|y^* - z^*\|_2^2 \leqslant (g_{y^*} - g_{z^*})^\top(y^* - z^*)$. If $y^* = z^*$ then we obtain from the former inequality that $g_{y^*} = g_{z^*}$ and the matrix $H = \frac{2}{\gamma} \cdot I$ trivially satisfies the claim. Hence, we suppose that $y^* \neq z^*$. From the latter inequality, we have that $(g_{y^*} - g_{z^*})^\top(y^* - z^*) > 0$. Combining this with the former inequality, we obtain the strict inequality $\|g_{y^*} - g_{z^*}\|_2^2 < \frac{2}{\gamma}(g_{y^*} - g_{z^*})^\top(y^* - z^*)$. Then, the claim immediately follows from Lemma 6. □

Lemma 8 (Residuals of random projections)**.** *Let $x \in \mathbb{R}^d$ be a given vector such that $\|x\|_2 = 1$. Let $Q \in \mathbb{R}^{d\times m}$ be a partial Haar matrix in $\mathbb{R}^d$. Consider the matrix $P^\perp = I - QQ^\top$. Then, we have the decomposition $\frac{P^\perp x}{\|P^\perp x\|_2} = \alpha\, x + \sqrt{1-\alpha^2}\, XZ$, where $\alpha = \|P^\perp x\|_2$, $X \in \mathbb{R}^{d\times(d-1)}$ is an orthonormal complement to x (i.e., $[x, X]$ is an orthogonal matrix), and Z is a $(d-1)$-dimensional vector uniformly distributed onto the unit sphere $\mathcal{S}^{d-1} := \{w \in \mathbb{R}^d \mid \|w\|_2 = 1\}$ and independent of α.*

Proof. We use the orthogonal decomposition $QQ^\top x = \begin{bmatrix} x & X \end{bmatrix} \cdot \begin{bmatrix} x^\top \\ X^\top \end{bmatrix} QQ^\top x = \|Q^\top x\|_2^2\, x + XX^\top QQ^\top x$, so that $\frac{P^\perp x}{\|P^\perp x\|_2} = \alpha\, x + X\widetilde{Z}$ where $\widetilde{Z} := -\frac{X^\top QQ^\top x}{\|P^\perp x\|_2}$. The vectors αx and $X\widetilde{Z}$ are orthogonal. Taking norms and using that $\|X\widetilde{Z}\|_2 = \|\widetilde{Z}\|_2$, we obtain that $\|\widetilde{Z}\|_2 = \sqrt{1-\alpha^2}$. Setting $Z = \frac{\widetilde{Z}}{\sqrt{1-\alpha^2}}$, we obtain that $\|Z\|_2 = 1$ and the decomposition $\frac{P^\perp x}{\|P^\perp x\|_2} = \alpha\, x + \sqrt{1-\alpha^2}\, XZ$. It remains to show that Z is uniformly distributed on the sphere $\mathcal{S}^{d-1}$ and independent of α. Let $\Omega \in \mathbb{R}^{(d-1)\times(d-1)}$ be a Haar matrix independent of Q. Rotational invariance in distribution of the partial Haar matrix Q implies that $QQ^\top \overset{\mathrm{d}}{=} \begin{bmatrix} x^\top \\ X^\top \end{bmatrix} QQ^\top \begin{bmatrix} x & X \end{bmatrix}$, and furthermore,

$$\begin{bmatrix} 1 & 0 \\ 0 & \Omega \end{bmatrix} \begin{bmatrix} x^\top \\ X^\top \end{bmatrix} QQ^\top \begin{bmatrix} x & X \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \Omega \end{bmatrix} \overset{\mathrm{d}}{=} \begin{bmatrix} x^\top \\ X^\top \end{bmatrix} QQ^\top \begin{bmatrix} x & X \end{bmatrix},$$

i.e.,

$$\begin{bmatrix} x^\top QQ^\top x & x^\top QQ^\top X\Omega \\ \Omega X^\top QQ^\top x & \Omega X^\top QQ^\top X\Omega \end{bmatrix} \overset{\mathrm{d}}{=} \begin{bmatrix} x^\top QQ^\top x & x^\top QQ^\top X \\ X^\top QQ^\top x & X^\top QQ^\top X \end{bmatrix}.$$

Consequently, the joint random variable $(\Omega X^\top QQ^\top x, x^\top QQ^\top x)$ has the same distribution as $(X^\top QQ^\top x, x^\top QQ^\top x)$, i.e., $(\alpha\sqrt{1-\alpha^2}\cdot \Omega Z, 1-\alpha^2) \overset{\mathrm{d}}{=} (\alpha\sqrt{1-\alpha^2}\cdot Z, 1-\alpha^2)$, which further implies that the distribution of ΩZ conditional on α is equal to the distribution of Z conditional on α. The vector ΩZ is uniformly distributed on the unit sphere and independent of α, which implies the same for Z. □

2) Proof of Theorem 5: According to Propositions 1 and 2, there exist $g_{z^*} \in \partial f^*(z^*)$ and $g_{y^*} \in \partial f^*(y^*)$ such that $g_{z^*} + \frac{1}{\lambda} AA^\top z^* = 0$ and $g_{y^*} + \frac{1}{\lambda} AP_S A^\top y^* = 0$. Subtracting the previous two equalities and multiplying by λ, we obtain that $\lambda(g_{y^*} - g_{z^*}) + AP_S A^\top(y^* - z^*) = AP_S^\perp A^\top z^*$. Using the assumption that f is γ-strongly convex, it follows from Lemma 7 that there exists a symmetric matrix H such that $0 \prec H \preceq \frac{2}{\gamma} I$ and $H(y^* - z^*) = g_{y^*} - g_{z^*}$. Substituting the latter equality into the former, left-multiplying both sides of the resulting equation by $\frac{1}{\lambda} A^\top(\lambda H + AP_S A^\top)^{-1}$ and using the identities $x^* = -\lambda^{-1} A^\top z^*$ and $\widehat{x}^{(1)} = -\lambda^{-1} A^\top y^*$, we obtain that $\widehat{x}^{(1)} - x^* = A^\top(\lambda H + AP_S A^\top)^{-1} AP_S^\perp x^*$. Multiplying both sides by ${x^*}^\top P_S^\perp$ and using that $\langle P_S^\perp x^*, \widehat{x}^{(1)} - x^*\rangle \leqslant \|P_S^\perp x^*\|_2 \|\widehat{x}^{(1)} - x^*\|$, it follows that

$$\begin{aligned} \frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2} &\geqslant \frac{\|P_S^\perp x^*\|_2}{\|x^*\|_2} \Big\| \big(\lambda H + AP_S A^\top\big)^{-\frac{1}{2}} A \frac{P_S^\perp x^*}{\|P_S^\perp x^*\|_2} \Big\|_2^2 \\ &\underset{(i)}{\geqslant} \frac{\|P_S^\perp x^*\|_2}{\|x^*\|_2} \Big\| \underbrace{\big(2\lambda\gamma^{-1} I_n + AA^\top\big)^{-\frac{1}{2}} A}_{:= M} \frac{P_S^\perp x^*}{\|P_S^\perp x^*\|_2} \Big\|_2^2. \end{aligned}$$

In inequality (i), we used the fact that $A^\top(\lambda H + AP_S A^\top)^{-1} A \succeq M^\top M$. According to Lemma 8, we have that $\frac{P_S^\perp x^*}{\|P_S^\perp x^*\|_2} \overset{\mathrm{d}}{=} \alpha \frac{x^*}{\|x^*\|_2} + \sqrt{1-\alpha^2}\cdot XZ$, where $\alpha := \|P_S^\perp \frac{x^*}{\|x^*\|_2}\|_2$, $X \in \mathbb{R}^{d\times(d-1)}$ is an orthonormal complement to $\frac{x^*}{\|x^*\|_2}$ and $Z \in \mathbb{R}^{d-1}$ is a

random vector which is uniformly distributed onto the unit sphere in $\mathbb{R}^{d-1}$ and independent of α. Consequently, we have that

$$\begin{aligned}\mathbb{E}_S \frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2} &\geqslant \mathbb{E}_S \|\alpha^{\frac{3}{2}} M \frac{x^*}{\|x^*\|} + \sqrt{\alpha(1-\alpha^2)} MXZ\|_2^2 \\ &\underset{(i)}{=} \mathbb{E}_S\{\alpha^3\} \cdot \frac{\|Mx^*\|_2^2}{\|x^*\|_2^2} + \underbrace{\mathbb{E}_S\{\alpha(1-\alpha^2)\} \cdot \mathbb{E}\|MXZ\|_2^2}_{\geqslant 0} \\ &\underset{(ii)}{\geqslant} \mathbb{E}_S\{\alpha^2\}^{\frac{3}{2}} \cdot \frac{\|Mx^*\|_2^2}{\|x^*\|_2^2}\,.\end{aligned}$$

In inequality (i), we used that the cross-term in the expansion of the square is equal to 0 because of the independence of α and Z and the fact that $\mathbb{E}Z = 0$. In inequality (ii), we used Jensen's inequality to obtain that $\mathbb{E}_S\{\alpha^2\}^{\frac{3}{2}} \leqslant \mathbb{E}_S\{\alpha^3\}$. Using that $\mathbb{E}_S\{\alpha^2\} = 1 - \frac{m}{d}$, $\|M\|_2^2 = \frac{\sigma_1^2}{\sigma_1^2 + \frac{2\lambda}{\gamma}}$ and taking the supremum over $f \in \mathcal{F}_{\gamma,\mu}$, we obtain that

$$\sup_{f \in \mathcal{F}_{\gamma,\mu}} \mathbb{E}_S\left\{\frac{\|\widehat{x}^{(1)} - x^*\|_2^2}{\|x^*\|_2^2}\right\} \underset{(i)}{\geqslant} \sup_{f \in \mathcal{F}_{\gamma,\mu}} \mathbb{E}_S\left\{\frac{\|\widehat{x}^{(1)} - x^*\|_2}{\|x^*\|_2}\right\}^2 \geqslant \left(1 - \frac{m}{d}\right)^3 \cdot \frac{\sigma_1^4}{(\sigma_1^2 + \frac{2\lambda}{\gamma})^2}\,,$$

where inequality (i) is a consequence of Jensen's inequality. This concludes the proof. □

APPENDIX C
RIGHT-SKETCHING AND STATISTICAL OPTIMALITY

A. Technical preliminaries

Given a radius $\delta > 0$, we say that $\{x_1, \ldots, x_M\} \subset \mathbb{R}^d$ is a δ-packing of $\mathcal{B}_2^d$ in the metric $\|P_{AS}A \cdot\|_2$ if $x_j \in \mathcal{B}_2^d$ for any $j \in \{1, \ldots, M\}$, and, $\|P_{AS}A(x_j - x_k)\|_2 > \delta$ for any $j \neq k$. We say that the δ-packing $\{x_1, \ldots, x_M\}$ is maximal if for any $x \in \mathcal{B}_2^d$, there exists $i \in \{1, \ldots, M\}$ such that $\|P_{AS}A(x - x_i)\|_2 \leqslant \delta$. We recall that we denote the rank of the matrix A by ρ.

Lemma 9. *Let $K \in \{1, \ldots, \rho\}$ be any index such that $\sigma_K^2 > \|P_{AS}^{\perp}A\|_2^2$, define $\delta_K := \sqrt{\sigma_K^2 - \|P_{AS}^{\perp}A\|_2^2}$ and let $\delta \in (0, \delta_K)$. Then, there exists a $\frac{\delta}{2}$-packing $\{x_j\}_{j=1}^M$ of $\mathcal{B}_2^d$ in the metric $\|P_{AS}A \cdot\|_2$ such that $\log M \geqslant K \cdot \log 2$, and such that $\|P_{AS}A(x_j - x_k)\|_2 \leqslant 2\delta$ for all $j, k \in \{1, \ldots, M\}$.*

Proof. Let $\widehat{U}\widehat{\Sigma}\widehat{V}^\top$ be a thin SVD of $P_{AS}A$ and denote its rank by $\widehat{\rho}$. Similarly, let $U\Sigma V^\top$ be a thin SVD of A. Let $\widehat{\sigma}_1 \geqslant \ldots \geqslant \widehat{\sigma}_{\widehat{\rho}} > 0$ (resp. $\sigma_1 \geqslant \ldots \geqslant \sigma_\rho > 0$) be the singular values of $P_{AS}A$ (resp. A). It holds that $|\sigma_j^2 - \widehat{\sigma}_j^2| \leqslant \|P_{AS}^{\perp}A\|_2^2$. Consider the Euclidean ball $\mathcal{B}_2^K(\delta^2) := \left\{\theta \in \mathbb{R}^K \mid \sum_{j=1}^K \frac{\theta_j^2}{\delta^2} \leqslant 1\right\}$. For any $\theta \in \mathbb{R}^K$, we have $\sum_{j=1}^K \frac{\theta_j^2}{\delta^2} \underset{(i)}{\geqslant} \sum_{j=1}^K \frac{\theta_j^2}{\sigma_j^2 - \|P_{AS}^{\perp}A\|_2^2} \underset{(ii)}{\geqslant} \sum_{j=1}^K \frac{\theta_j^2}{\widehat{\sigma}_j^2}$, where inequality (i) follows from $\delta^2 \leqslant \sigma_j^2 - \|P_{AS}^{\perp}A\|_2^2$ for all $j = 1, \ldots, K$ and inequality (ii) follows from the fact that $|\sigma_j^2 - \widehat{\sigma}_j^2| \leqslant \|P_{AS}^{\perp}A\|_2^2$ for all $j = 1, \ldots, K$. Therefore, if $\theta \in \mathcal{B}_2^K(\delta^2)$, then $\widetilde{\theta} := [\theta_1, \ldots, \theta_K, 0, \ldots, 0]$ belongs to the ellipsoid $\mathcal{E}_{\widehat{\sigma}}^{\widehat{\rho}} := \left\{\overline{\theta} \in \mathbb{R}^{\widehat{\rho}} \mid \sum_{j=1}^{\widehat{\rho}} \frac{\overline{\theta}_j^2}{\widehat{\sigma}_j^2} \leqslant 1\right\}$. Consequently, if $\{\theta^a\}_{a=1}^M$ is a $\frac{\delta}{2}$-packing of $\mathcal{B}_2^K(\delta^2)$ in the metric $\|\cdot\|_2$, then $\{\widetilde{\theta}^a\}_{a=1}^M$ is a $\frac{\delta}{2}$-packing of $\mathcal{E}_{\widehat{\sigma}}^{\widehat{\rho}}$ in the metric $\|\cdot\|_2$. It is well-known that there exists such a packing $\{\theta^a\}_{a=1}^M$ with cardinality $M \geqslant 2^K$. Hence, there exists a $\frac{\delta}{2}$-packing $\{\widetilde{\theta}^a\}_{a=1}^M$ of $\mathcal{E}_{\widehat{\sigma}}^{\widehat{\rho}}$ in the metric $\|\cdot\|_2$ with cardinality $M \geqslant 2^K$.

We are now ready to construct the claimed packing of $\mathcal{B}_2^d$ in the metric $\|P_{AS}A \cdot\|_2$. For each $a = 1, \ldots, M$, we set $x_a := \widehat{V}\widehat{\Sigma}^{-1}\widetilde{\theta}^a$. Observe that $\|x_a\|_2^2 = \|\widehat{\Sigma}^{-1}\widetilde{\theta}^a\|_2^2 \leqslant 1$, i.e., $x_a \in \mathcal{B}_2^d$. For $a \neq b$, we have $\|P_{AS}A(x_a - x_b)\|_2 = \|\widehat{U}\widehat{\Sigma}\widehat{V}^\top\widehat{V}\widehat{\Sigma}^{-1}(\widetilde{\theta}^a - \widetilde{\theta}^b)\|_2 = \|\widetilde{\theta}^a - \widetilde{\theta}^b\|_2 \geqslant \frac{\delta}{2}$. Furthermore, we have by construction that $\|\widetilde{\theta}^a\|_2 \leqslant \delta$, which implies that $\|P_{AS}A(x_a - x_b)\|_2 \leqslant 2\delta$, and this concludes the proof. □

B. Proof of Theorem 9

Our proof is based on the standard local Fano method. Given an arbitrary vector $\bar{x} \in \mathbb{R}^d$, we denote by $\mathbb{P}_{\bar{x}}$ the probability measure with respect to the distribution $\mathcal{N}(A\bar{x}, \frac{\sigma^2}{n} I_n)$ and by $\mathbb{E}_{\bar{x}}$ the corresponding expectation. Fix a sketch size $m \geqslant 1$, an embedding $S \in \mathbb{R}^{d\times m}$ and let us assume that $\|P_{AS}^{\perp} A\|_2^2 \leqslant \frac{\sigma_{d_s+1}^2}{2}$. Using that $\|P_{AS}\|_2 \leqslant 1$, we have $\mathfrak{M}_S \geqslant \inf_{\widehat{x}} \sup_{\|x_{\mathrm{pl}}\|_2 \leqslant 1} \mathbb{E}_{x_{\mathrm{pl}}} \|P_{AS} A(\widehat{x} - x_{\mathrm{pl}})\|_2^2$. It is thus sufficient to lower bound the latter quantity. We denote $b_S := S^\top A^\top b$, and we let $\widehat{x} \equiv \widehat{x}(b_S)$ be an estimator. We introduce the radius $\delta_m := \sqrt{\sigma_{d_s+1}^2 - \|P_{AS}^{\perp}\|_2^2}$. By assumption, we have $\delta_m \geqslant \frac{\sigma_{d_s+1}}{\sqrt{2}}$. Using Lemma 9 with $K = d_s + 1$ and $\delta = \frac{\delta_m}{8}$, we obtain that there exists a maximal $\frac{\delta_m}{16}$-packing $\{x_1, \ldots, x_M\}$ of $\mathcal{B}_2^d$ in the metric $\|P_{AS} A \cdot\|_2$ such that $\log M \geqslant d_s \cdot \log 2$, and, $\|P_{AS} A(x_j - x_k)\|_2 \leqslant \frac{\delta_m}{4}$ for all j, k. Then, we have

$$\sup_{\|x_{\mathrm{pl}}\|_2 \leqslant 1} \mathbb{E}_{x_{\mathrm{pl}}} \|P_{AS} A(\widehat{x} - x_{\mathrm{pl}})\|_2^2 \geqslant \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}_{x_j} \|P_{AS} A(\widehat{x} - x_j)\|_2^2 \underset{(i)}{\geqslant} \frac{\delta_m^2}{32^2} \cdot \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{x_j} \left(\|P_{AS} A(\widehat{x} - x_j)\|_2 \geqslant \frac{\delta_m}{32} \right),$$

where inequality (i) follows from Markov's inequality. We introduce the test function $\psi(\widehat{x}) := \operatorname{argmin}_{k=1,\ldots,M} \|P_{AS} A(\widehat{x} - x_k)\|_2$. We claim that $\|P_{AS} A(\widehat{x} - x_j)\|_2 < \frac{\delta_m}{32}$ implies that $\psi(\widehat{x}) = j$. Indeed, suppose that $\|P_{AS} A(\widehat{x} - x_j)\|_2 < \frac{\delta_m}{32}$. Then, using the fact that $\{x_i\}_{i=1}^M$ is a $(\delta_m/16)$-packing, we have for any $k \neq j$ that $\|P_{AS} A(\widehat{x} - x_k)\|_2 \geqslant \|P_{AS} A(x_j - x_k)\|_2 - \|P_{AS} A(\widehat{x} - x_j)\|_2 > \delta_m/16 - \delta_m/32 = \delta_m/32$, i.e., $\psi(\widehat{x}) = j$. It follows that

$$\sup_{\|x\|_{\mathrm{pl}} \leqslant 1} \mathbb{E}_{x_{\mathrm{pl}}} \|P_{AS} A(\widehat{x} - x_{\mathrm{pl}})\|_2^2 \geqslant \left(\frac{\delta}{32}\right)^2 \cdot \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{x_j}(\psi(\widehat{x}) \neq j) \geqslant \left(\frac{\delta}{32}\right)^2 \cdot \inf_{\psi} \mathbb{E}_J \big\{ \mathbb{P}_{x_j}(\psi(b_S) \neq j) \mid J = j \big\},$$

where J is a uniformly distributed random variable in $\{1, \ldots, M\}$. Fano's inequality states that the testing error in the latter right-hand side is lower bounded by the quantity $(1 - \frac{I(b_S; J) + \log 2}{\log M})$, where $I(b_S; J)$ denotes the mutual information between the random variables b_S and J. Consequently, we obtain

$$\sup_{\|x\|_{\mathrm{pl}} \leqslant 1} \mathbb{E}_{x_{\mathrm{pl}}} \|P_{AS} A(\widehat{x} - x_{\mathrm{pl}})\|_2^2 \geqslant \left(\frac{\delta}{32}\right)^2 \cdot \left(1 - \frac{I(b_S; J) + \log 2}{\log M}\right). \tag{73}$$

Introducing the mixture distribution $P_{x_J} := \frac{1}{M} \sum_{j=1}^{M} P_{x_j}$, denoting the Kullback-Leibler (KL) divergence between two distributions P and Q by $D_{\mathrm{kl}}(P \,\|\, Q)$ and using the convexity of $Q \mapsto D_{\mathrm{kl}}(P \,\|\, Q)$, we have

$$I(b_S; J) = \frac{1}{M} \sum_{j=1}^{M} D_{\mathrm{kl}}\big(P_{x_j} \,\|\, P_{x_J}\big) \leqslant \frac{1}{M^2} \sum_{j,k=1}^{M} D_{\mathrm{kl}}\big(P_{x_j} \,\|\, P_{x_k}\big).$$

Using that the KL divergence between two Gaussian distributions P_{x_j} and P_{x_k} is equal to $\frac{n}{\sigma^2} \|P_{AS} A(x_j - x_k)\|_2^2$, it follows that

$$I(b_S; J) \leqslant \frac{1}{M^2} \sum_{j,k=1}^{M} \frac{n}{\sigma^2} \|P_{AS} A(x_j - x_k)\|_2^2 \leqslant \frac{n\delta_m^2}{16\sigma^2},$$

where in the last inequality, we used the fact that $\|P_{AS} A(x_j - x_k)\|_2 \leqslant \frac{\delta_m}{4}$ for any j, k. Combining these observations, we obtain from (73) that

$$\sup_{\|x\|_{\mathrm{pl}} \leqslant 1} \mathbb{E}_{x_{\mathrm{pl}}} \|P_{AS} A(\widehat{x} - x_{\mathrm{pl}})\|_2^2 \geqslant \frac{\delta_m^2}{32^2} \left(1 - \frac{n\delta_m^2}{16\sigma^2 d_s \log 2} - \frac{1}{d_s}\right).$$

Using that $\frac{\sigma^2_{d_s+1}}{2} \leqslant \delta_m^2 \leqslant \sigma^2_{d_s+1}$ and $\frac{\sigma^2 d_s}{n} \geqslant \sigma^2_{d_s+1}$, it follows that $\frac{n\delta_m^2}{16\sigma^2 d_s \log 2} \leqslant \frac{n\sigma^2_{d_s+1}}{16\sigma^2 d_s \log 2} \leqslant \frac{1}{16\log 2}$, and thus,

$$\sup_{\|x\|_{\text{pl}}\leqslant 1} \mathbb{E}_{x_{\text{pl}}} \|P_{AS} A(\widehat{x} - x_{\text{pl}})\|_2^2 \geqslant \frac{\sigma^2_{d_s+1}}{2\cdot 32^2} \cdot \underbrace{\left(1 - \frac{1}{16\log 2} - \frac{1}{d_s}\right)}_{\geqslant \frac{1}{2}},$$

and this concludes the proof of the lower bound.

Polynomial decay. Consider a polynomial decay $\sigma_j = j^{-\frac{1+\nu}{2}}$ for some $\nu > 0$. The scaling relation $\frac{\sigma^2 d_s}{n} \asymp \sigma^2_{d_s+1}$ yields that $d_s \asymp \left(\frac{n}{\sigma^2}\right)^{\frac{1}{2+\nu}}$. Fix a target rank $k \geqslant 1$ and a sketch size $m = 2k$. According to Lemma 1 and using the inequality $(a+b)^2 \leqslant 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, we have with probability at least $1 - 6e^{-k}$ that

$$\|P_{AS}^{\perp} A\|_2^2 \leqslant 26^2 \cdot \left(\sigma_{k+1} + \frac{1}{\sqrt{k}} \sqrt{\sum_{j=k+1}^{\rho} \sigma_j^2}\right)^2 \leqslant 2\cdot 26^2 \cdot (\sigma_{k+1}^2 + \frac{1}{k}\sum_{j=k+1}^{\rho} \sigma_j^2)\,.$$

We have $\frac{1}{k}\sum_{j=k+1}^{\rho} \sigma_j^2 \leqslant \frac{1}{k}\int_k^{+\infty} u^{-(1+\nu)} \mathrm{d}u = \frac{1}{\nu k^{1+\nu}}$. It follows that $\|P_{AS}^{\perp} A\|_2^2 \leqslant 2\cdot 26^2 \cdot (\nu^{-1}+1)\cdot k^{-(1+\nu)}$ with probability at least $1 - 6e^{-k}$. Consequently, we have $\|P_{AS}^{\perp} A\|_2^2 \leqslant \frac{\sigma^2_{d_s+1}}{2}$ if $(d_s+1)^{-(1+\nu)} \geqslant 2704 \cdot (\nu^{-1}+1) k^{-(1+\nu)}$, for which it is sufficient to have $m \geqslant \underbrace{(\lceil 2\cdot (2704(\nu^{-1}+1))^{(1+\nu)} \rceil + 1)}_{:= c_\nu^{\text{poly}}} \cdot d_s$.

Exponential decay. Consider an exponential decay $\sigma_j = e^{-\frac{\nu j}{2}}$ for some $\nu > 0$. The scaling relation $\frac{\sigma^2 d_s}{n} \asymp \sigma^2_{d_s+1}$ yields that $d_s \asymp \frac{1}{\nu}\cdot \log(n/\sigma^2)$. Fix a target rank $k \geqslant 1$ and a sketch size $m = 2k$. We have with probability at least $1 - 6e^{-k}$ that $\|P_{AS}^{\perp} A\|_2^2 \leqslant 2\cdot 26^2 \cdot (\sigma_{k+1}^2 + \frac{1}{k}\sum_{j=k+1}^{\rho}\sigma_j^2) \leqslant \frac{2704}{\nu}\cdot e^{-(k+1)\nu}$. Consequently, $\|P_{AS}^{\perp} A\|_2^2 \leqslant \frac{\sigma^2_{d_s+1}}{2}$ if $\frac{5408}{\nu} e^{-(k+1)\nu} \leqslant e^{-(d_s+1)\nu}$, i.e., $m \geqslant 2\cdot d_s + \frac{2}{\nu}\log(5408/\nu)$. Assuming that $\log(n/\sigma^2) \geqslant \log(5408/\nu)$ and using that $d_s \asymp \frac{1}{\nu}\cdot\log(n/\sigma^2)$, we get that $\|P_{AS}^{\perp} A\|_2^2 \leqslant \frac{\sigma^2_{d_s+1}}{2}$ with probability at least $1 - 6e^{-d_s}$ for $m \geqslant 4d_s$.

C. Proof of Corollary 2

Through a simple calculation, we obtain that the risk of $\widehat{x}^{(0)}$, as $\lambda \to 0$, satisfies the bias-variance decomposition

$$\lim_{\lambda\to 0} \mathfrak{R}(\widehat{x}^{(0)}) = \underbrace{\mathbb{E}\|P_{AS} w\|_2^2}_{= m\sigma^2/n} + \sup_{\|x_{\text{pl}}\|_2 \leqslant 1} \|P_{AS}^{\perp} A x_{\text{pl}}\|_2^2 = \frac{\sigma^2 m}{n} + \|P_{AS}^{\perp} A\|_2^2\,. \tag{74}$$

According to (26), the residual error verifies $\|P_{AS}^{\perp} A\|_2^2 \lesssim R^2_{m/2}(A)$, so that

$$\mathfrak{M}_S = \inf_{\widehat{x}} \mathfrak{R}(\widehat{x}) \leqslant \inf_m \lim_{\lambda\to 0} \mathfrak{R}(\widehat{x}^{(0)}) \lesssim \inf_m \left\{ \frac{\sigma^2 m}{n} + R^2_{m/2}(A) \right\}. \tag{75}$$

Consequently, the sketch size m controls this bias-variance trade-off: the larger the sketch size m, the larger the variance term $\frac{\sigma^2 m}{n}$ and the smaller the bias $R^2_{m/2}(A)$, and vice-versa. Furthermore, according to Theorem 9, under the event $\|P_{AS}^{\perp} A\|_2^2 \leqslant \frac{\sigma^2_{d_s+1}}{2}$, it holds that $\mathfrak{M}_S \geqslant c_0 \cdot \sigma^2_{d_s+1}$.

Polynomial decay. Consider a polynomial decay $\sigma_j = j^{-\frac{1+\nu}{2}}$ for some $\nu > 0$. Picking $m = c_\nu^{\text{poly}} \cdot d_s$ and following the same steps as in the proof of Theorem 9, we obtain that $\|P_{AS}^{\perp} A\|_2^2 \leqslant \frac{\sigma^2_{d_s+1}}{2}$ with probability at least $1 - 6e^{-d_s}$. Plugging-in this value of m and this bound on the residual error $\|P_{AS}^{\perp} A\|_2^2$ into (74), it follows that

$$\lim_{\lambda\to 0} \mathfrak{R}(\widehat{x}^{(0)}) \leqslant c_\nu^{\text{poly}} \cdot \frac{\sigma^2 d_s}{n} + \frac{\sigma^2_{d_s+1}}{2} \leqslant (c_\nu^{\text{poly}} + \frac{1}{2}) \cdot \frac{\sigma^2 d_s}{n}\,.$$

with probability at least $1 - 6e^{-d_s}$.

Exponential decay. Consider an exponential decay $\sigma_j = e^{-\frac{\nu j}{2}}$ for some $\nu > 0$. Picking $m = 4d_s$ and following the same steps as in the proof of Theorem 9, we obtain that $\|P_{AS}^{\perp} A\|_2^2 \leqslant \frac{\sigma_{d_s+1}^2}{2}$ with probability at least $1 - 6e^{-d_s}$. Plugging-in this value of m and this bound on the residual error $\|P_{AS}^{\perp} A\|_2^2$ into (74), it follows that

$$\lim_{\lambda \to 0} \mathfrak{R}(\widehat{x}^{(0)}) \leqslant 4 \cdot \frac{\sigma^2 d_s}{n} + \frac{\sigma_{d_s+1}^2}{2} \leqslant 5 \cdot \frac{\sigma^2 d_s}{n} .$$

with probability at least $1 - 6e^{-d_s}$.

APPENDIX D
PROOFS OF FENCHEL DUALITY RESULTS

A. Proof of Proposition 1

The primal objective function is strongly convex, so that it admits a unique minimizer x^*. According to Corollary 31.2.1 in [57] whose assumptions are trivially satisfied, strong duality holds and there exists a dual solution z^*. According to Theorem 31.3 in [57], we have the KKT conditions $x^* = -A^\top z^*/\lambda$ and $z^* = \nabla f(Ax^*)$. In particular, the relation $z^* = \nabla f(Ax^*)$ along with the uniqueness of x^* imply that z^* is unique. If the function f is strictly convex, according to Theorem 26.5 in [57], the mapping ∇f is one-to-one from $\mathbb{R}^n$ to the interior of the domain of f^*. Consequently, $\nabla f(Ax^*) \in$ int dom f^*, i.e., $z^* \in$ int dom f^*. □

B. Proof of Proposition 2

According to Corollary 31.2.1 in [57] whose assumptions are trivially satisfied, there exists a primal solution $\alpha^* \in \mathbb{R}^*$, strong duality holds, and there exists a sketched dual solution $y^* \in$ dom f^*. According to Theorem 31.3 in [57], we have the KKT conditions $S^\top S\alpha^* = -\frac{S^\top A^\top y^*}{\lambda}$ and $y^* = \nabla f(AS\alpha^*)$. If the function f is strictly convex, according to Theorem 26.5 in [57], the mapping ∇f is one-to-one from $\mathbb{R}^n$ to the interior of the domain of f^*. Consequently, $\nabla f(AS\alpha^*) \in$ int dom f^*, i.e., $y^* \in$ int dom f^*. □

ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation under grants IIS-1838179, ECCS-2037304, Facebook Research, Adobe Research and Stanford SystemX Alliance.

REFERENCES

[1] J. Lacotte, M. Pilanci, and M. Pavone, "High-dimensional optimization in adaptive random subspaces," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 10 847–10 857.

[2] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems*, 2008, pp. 1177–1184.

[3] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.

[4] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, "When do neural networks outperform kernel methods?" *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[5] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[6] D. Drusvyatskiy and C. Paquette, "Efficiency of minimizing compositions of convex functions and smooth maps," *Mathematical Programming*, vol. 178, no. 1-2, pp. 503–558, 2019.

[7] M. Pilanci and T. Ergen, "Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks," *International Conference on Machine Learning*, 2020.

[8] T. Ergen and M. Pilanci, "Training convolutional ReLU neural networks in polynomial time: Exact convex optimization formulations," *arXiv preprint arXiv:2006.14798*, 2020.

[9] ——, "Convex geometry of two-layer ReLU networks: Implicit autoencoding and interpretable models," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 4024–4033.

[10] ——, "Convex duality of deep neural networks," *arXiv preprint arXiv:2002.09773*, 2020.

[11] S. S. Vempala, *The random projection method*. American Mathematical Society, 2005, vol. 65.

[12] M. W. Mahoney, "Randomized algorithms for matrices and data," *Foundations and Trends® in Machine Learning*, vol. 3, no. 2, pp. 123–224, 2011.

[13] P. Drineas and M. W. Mahoney, "RandNLA: randomized numerical linear algebra," *Communications of the ACM*, vol. 59, no. 6, pp. 80–90, 2016.

[14] H. Avron, P. Maymounkov, and S. Toledo, "Blendenpik: Supercharging LAPACK's least-squares solver," *SIAM Journal on Scientific Computing*, vol. 32, no. 3, pp. 1217–1236, 2010.

[15] X. Meng, M. A. Saunders, and M. W. Mahoney, "LSRN: A parallel iterative solver for strongly over- or underdetermined systems," *SIAM Journal on Scientific Computing*, vol. 36, no. 2, pp. C95–C118, 2014.

[16] M. Pilanci, "Fast randomized algorithms for convex optimization and statistical estimation," Ph.D. dissertation, UC Berkeley, 2016.

[17] M. Pilanci and M. J. Wainwright, "Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1842–1879, 2016.

[18] J. Lacotte and M. Pilanci, "Faster least squares optimization," *arXiv preprint arXiv:1911.02675*, 2019.

[19] I. K. Ozaslan, M. Pilanci, and O. Arikan, "Iterative hessian sketch with momentum," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7470–7474.

[20] M. Pilanci and M. J. Wainwright, "Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence," *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 205–245, 2017.

[21] B. Bartan and M. Pilanci, "Distributed sketching methods for privacy preserving regression," *arXiv preprint arXiv:2002.06538*, 2020.

[22] ——, "Distributed averaging methods for randomized second order optimization," *arXiv preprint arXiv:2002.06540*, 2020.

[23] M. Derezinski, B. Bartan, M. Pilanci, and M. W. Mahoney, "Debiasing distributed second order optimization with surrogate sketching and scaled regularization," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

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

[25] Z. Qu, P. Richtárik, M. Takác, and O. Fercoq, "SDNA: Stochastic dual Newton ascent for empirical risk minimization," in *International Conference on Machine Learning*, 2016, pp. 1823–1832.

[26] N. Doikov and P. Richtárik, "Randomized block cubic Newton method," *International Conference on Machine Learning*, 2018.

[27] H. Luo, A. Agarwal, N. Cesa-Bianchi, and J. Langford, "Efficient second order online learning by sketching," in *Advances in Neural Information Processing Systems*, 2016, pp. 902–910.

[28] R. Gower, D. Koralev, F. Lieder, and P. Richtárik, "RSN: Randomized subspace newton," in *Advances in Neural Information Processing Systems*, 2019, pp. 616–625.

[29] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362, 2012.

[30] Y. Nesterov and B. T. Polyak, "Cubic regularization of Newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, 2006.

[31] L. Zhang, M. Mahdavi, R. Jin, T. Yang, and S. Zhu, "Recovering the optimal solution by dual random projection," in *Conference on Learning Theory*, 2013, pp. 135–157.

[32] ——, "Random projections for classification: A recovery approach," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 7300–7316, 2014.

[33] Y. Xu, H. Yang, L. Zhang, and T. Yang, "Efficient non-oblivious randomized reduction for risk minimization with improved excess risk guarantee," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 2796–2802.

[34] J. Wang, J. D. Lee, M. Mahdavi, M. Kolar, N. Srebro *et al.*, "Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data," *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 4896–4944, 2017.

[35] M. Pilanci and M. J. Wainwright, "Randomized sketches of convex programs with sharp guarantees," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 5096–5115, 2015.

[36] T. Yang, L. Zhang, R. Jin, and S. Zhu, "Theory of dual-sparse regularized randomized reduction," in *International Conference on Machine Learning*, 2015, pp. 305–314.

[37] D. Yao, P. Zhao, T.-A. N. Pham, and G. Cong, "High-dimensional similarity learning via dual-sparse random projection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 3005–3011.

[38] A. Blum, "Random projection, margins, kernels, and feature-selection," in *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*. Springer, 2005, pp. 52–68.

[39] Q. Shi, C. Shen, R. Hill, and A. van den Hengel, "Is margin preserved after random projection?" in *International Conference on Machine Learning*, 2012.

[40] S. Paul, C. Boutsidis, M. Magdon-Ismail, and P. Drineas, "Random projections for support vector machines," in *International Conference on Artificial Intelligence and Statistics*, 2013, pp. 498–506.

[41] K. Vu, P.-L. Poirion, C. D'Ambrosio, and L. Liberti, "Random projections for trust region subproblems," *arXiv preprint arXiv:1706.02730*, 2017.

[42] R. M. Gower and P. Richtárik, "Randomized iterative methods for linear systems," *SIAM Journal on Matrix Analysis and Applications*, vol. 36, no. 4, pp. 1660–1690, 2015.

[43] Y. Yang, M. Pilanci, M. J. Wainwright *et al.*, "Randomized sketches for kernels: Fast and optimal nonparametric regression," *The Annals of Statistics*, vol. 45, no. 3, pp. 991–1023, 2017.

[44] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.

[45] C. Boutsidis and A. Gittens, "Improved matrix algorithms via the subsampled randomized Hadamard transform," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 3, pp. 1301–1340, 2013.

[46] R. Witten and E. Candes, "Randomized algorithms for low-rank matrix factorizations: Sharp performance bounds," *Algorithmica*, vol. 72, no. 1, pp. 264–281, 2015.

[47] M. Derezinski, F. T. Liang, Z. Liao, and M. W. Mahoney, "Precise expressions for random projections: Low-rank approximation and randomized Newton," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[48] C. K. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Advances in Neural Information Processing Systems*, 2001, pp. 682–688.

[49] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[50] P. Drineas and M. W. Mahoney, "On the Nyström method for approximating a Gram matrix for improved kernel-based learning," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 2153–2175, 2005.

[51] S. Kumar, M. Mohri, and A. Talwalkar, "Sampling methods for the Nyström method," *Journal of Machine Learning Research*, vol. 13, no. Apr, pp. 981–1006, 2012.

[52] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou, "Nyström method vs random Fourier features: A theoretical and empirical comparison," in *Advances in Neural Information Processing Systems*, 2012, pp. 476–484.

[53] F. Bach, "Sharp analysis of low-rank kernel matrix approximations," in *Conference on Learning Theory*, 2013, pp. 185–209.

[54] M. Dereziński and M. K. Warmuth, "Reverse iterative volume sampling for linear regression," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 853–891, 2018.

[55] M. Dereziński and M. W. Mahoney, "Determinantal point processes in randomized numerical linear algebra," *Notices of the American Mathematical Society*, 2020.

[56] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[57] R. T. Rockafellar, *Convex analysis*. Princeton University Press, 2015.

[58] N. Ailon and B. Chazelle, "Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform," in *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, 2006, pp. 557–563.

[59] T. Sarlos, "Improved approximation algorithms for large matrices via random projections," in *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE, 2006, pp. 143–152.

[60] E. Dobriban and S. Liu, "Asymptotics for sketching in least squares regression," in *Advances in Neural Information Processing Systems*, 2019, pp. 3675–3685.

[61] J. Lacotte, S. Liu, E. Dobriban, and M. Pilanci, "Optimal iterative sketching methods with the subsampled randomized Hadamard transform," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[62] A. Alaoui and M. W. Mahoney, "Fast randomized kernel ridge regression with statistical guarantees," in *Advances in Neural Information Processing Systems*, 2015, pp. 775–783.

[63] A. Chowdhury, J. Yang, and P. Drineas, "An iterative, sketching-based framework for ridge regression," in *International Conference on Machine Learning*, 2018, pp. 989–998.

[64] J. Lacotte and M. Pilanci, "Effective dimension adaptive sketching methods for faster regularized least-squares optimization," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[65] C. Gu, *Smoothing spline ANOVA models*. Springer Science & Business Media, 2013, vol. 297.

[66] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics.* Kluwer Adademic, 2004.

[67] G. Wahba, *Spline models for observational data*. SIAM, 1990.

[68] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *Journal of mathematical analysis and applications*, vol. 33, no. 1, pp. 82–95, 1971.

[69] T. Ergen and M. Pilanci, "Convex optimization for shallow neural networks," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2019, pp. 79–83.

[70] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust optimization*. Princeton University Press, 2009, vol. 28.

[71] M. Pilanci, O. Arikan, B. Oguz, and M. Pinar, "Structured least squares with bounded data uncertainties," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 3261–3264.

[72] M. Pilanci, O. Arikan, and M. C. Pinar, "Structured least squares problems and robust estimators," *IEEE transactions on signal processing*, vol. 58, no. 5, pp. 2453–2465, 2010.

[73] J. Lacotte and M. Pilanci, "Optimal randomized first-order methods for least-squares problems," *International Conference on Machine Learning*, 2020.

[74] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems*, 2013, pp. 315–323.

[75] R. Bollapragada, R. H. Byrd, and J. Nocedal, "Exact and inexact subsampled Newton methods for optimization," *IMA Journal of Numerical Analysis*, vol. 39, no. 2, pp. 545–578, 2018.

[76] M. A. Erdogdu and A. Montanari, "Convergence rates of sub-sampled Newton methods," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, 2015, pp. 3052–3060.

[77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[78] D. L. Donoho, I. M. Johnstone *et al.*, "Neo-classical minimax problems, thresholding and adaptive function estimation," *Bernoulli*, vol. 2, no. 1, pp. 39–62, 1996.

[79] S. Tu, R. Roelofs, S. Venkataraman, and B. Recht, "Large scale kernel learning using block coordinate descent," *arXiv preprint arXiv:1602.05310*, 2016.

[80] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do CIFAR-10 classifiers generalize to CIFAR-10?" *arXiv preprint arXiv:1806.00451*, 2018.

[81] P. L. Bartlett, O. Bousquet, S. Mendelson *et al.*, "Local Rademacher complexities," *The Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.

[82] S. Van de Geer, *Empirical Processes in M-estimation*. Cambridge University Press, 2000, vol. 6.

[83] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018, vol. 47.

[84] R. Adamczak, R. Latała, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann, "Chevet type inequality and norms of submatrices," *Studia Mathematica*, vol. 210, no. 1, pp. 35–56, 2012.