Electronic Transactions on Numerical Analysis.
Volume 54, pp. 610–619, 2021.

ISSN 1068–9613.
DOI: 10.1553/etna_vol54s610

ETNA
Kent State University and
Johann Radon Institute (RICAM)

ON THE REGULARIZATION EFFECT OF STOCHASTIC GRADIENT DESCENT APPLIED TO LEAST-SQUARES*

STEFAN STEINERBERGER†

Abstract. We study the behavior of the stochastic gradient descent method applied to $\|Ax - b\|_2^2 \to \min$ for invertible matrices $A \in \mathbb{R}^{n \times n}$. We show that there is an explicit constant c_A depending (mildly) on A such that

$$\mathbb{E}\,\|Ax_{k+1} - b\|_2^2 \leq \left(1 + \frac{c_A}{\|A\|_F^2}\right) \|Ax_k - b\|_2^2 - \frac{2}{\|A\|_F^2} \left\|A^T A(x_k - x)\right\|_2^2.$$

This is a curious inequality since the last term involves one additional matrix multiplication applied to the error $x_k - x$ compared to the remaining terms: if the projection of $x_k - x$ onto the subspace of singular vectors corresponding to large singular values is large, then the stochastic gradient descent method leads to a fast regularization. For symmetric matrices, this inequality has an extension to higher-order Sobolev spaces. This explains a (known) regularization phenomenon: an energy cascade from large singular values to small singular values acts as a regularizer.

Key words. stochastic gradient descent, Kaczmarz method, least-squares, regularization

AMS subject classifications. 65F10, 65K10, 65K15, 90C06, 93E24

1. Introduction.

1.1. Stochastic gradient descent.

In this paper, we consider the finite-dimensional linear inverse problem

$$Ax = b,$$

where $A \in \mathbb{R}^{n \times n}$ is an invertible matrix, $x \in \mathbb{R}^n$ is the (unknown) signal of interest, and b is a given right-hand side. Throughout this paper, we will use $a_1, \ldots, a_n \in \mathbb{R}^n$ to denote the rows of A. Equivalent to the stated problem, we may try to solve

$$\|Ax - b\|^2 = \sum_{i=1}^{n} \left(\langle a_i, x\rangle - b_i\right)^2 \to \min.$$

Following Needell, Srebro, and Ward [29], we can interpret this as

$$\sum_{i=1}^{n} f_i(x)^2 \to \min, \qquad \text{where} \quad f_i(x) = \langle a_i, x\rangle - b_i.$$

The Lipschitz constant of f_i is $\|a_i\|_{\ell^2}$, which motivates the following basic form of a stochastic gradient descent method: pick one of the n functions with a likelihood proportional to the Lipschitz constant, and then perform a gradient descent for this much simpler function. This results in the Stochastic Gradient Descent (SGD) method,

$$x_{k+1} = x_k + \frac{b_i - \langle a_i, x_k\rangle}{\|a_i\|^2} a_i,$$

which is also known as the *Algebraic Reconstruction Technique* (ART) in computer tomography [11, 14, 15, 25], the *Projection onto Convex Sets Method* [8, 4, 5, 9, 36], or the *Randomized*

*Received December 18, 2020. Accepted August 31, 2021. Published online on October 6, 2021. Recommended by Jianlin Xia. Work supported by the NSF (DMS-2123224) and the Alfred P. Sloan Foundation.
†Department of Mathematics, University of Washington, Seattle, WA 98195, USA (steinerb@uw.edu).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Kaczmarz method [2, 6, 7, 10, 13, 12, 20, 21, 22, 23, 24, 26, 27, 28, 29, 30, 32, 34, 37, 38, 39, 40, 41, 42]. Strohmer and Vershynin [39] showed that

$$\mathbb{E}\|x_k - x\|^2 \leq \left(1 - \frac{1}{\|A^{-1}\|^2 \|A\|_F^2}\right)^k \|x_0 - x\|^2,$$

where $\|A\|_F^2$ is the Frobenius norm. In practice, the algorithm often converges a lot faster initially, and this was studied in [17, 18, 37]. In particular, the authors in [37] obtain an identity in terms of the behavior with regards to the singular values showing that the singular vectors associated to large singular values are expected to undergo a more rapid decay. Motivated by this, we provide rigorous bounds that quantify this energy cascade from large singular values to small singular values by identifying an interesting inequality for the SGD method when applied to the least-squares problem.

1.2. A motivating example. We discuss a simple example that exemplifies the phenomenon that we are interested in. Let us take $A \in \mathbb{R}^{100\times 100}$ by picking each entry independently at random from a standard normal distribution $\mathcal{N}(0,1)$ and then normalizing the rows to $\|a_i\| = 1$. The right-hand side is $b = (1, 1, \ldots, 1)$, and we initialize with $x_0 = 0 \in \mathbb{R}^n$. Figure 1.1 displays the magnitude of $\|Ax_k - b\|_{\ell^2}$ over the first 10.000 iteration steps (left) and the subsequent 10.000 iteration steps (right).

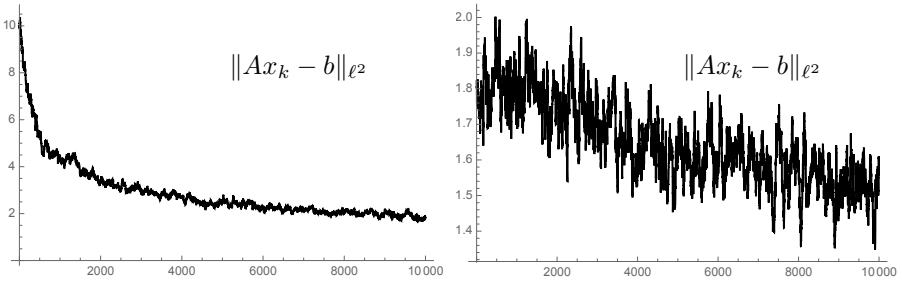

FIG. 1.1. *The size of* $\|Ax_k - b\|_{\ell^2}$*, for* $k = 1, \ldots, 10000$*, (left) followed by the next* 10000 *iterations* $\|Ax_{10000+k} - b\|_{\ell^2}$ *(right). We observe a rapid initial decay, which then slows down.*

The picture tells a very interesting story: the error in $\|Ax_k - b\|$ initially decays quite rapidly before stabilizing in a certain regime: there is still an additional decay but at a much lower rate and with a larger amount of fluctuations. Moreover, for the example in Figure 1.1, we have $\|x_0\| = 0$ and $\|x_{20000}\| \sim 28$, which is not even close to the true solution $\|x\| \sim 128$; nonetheless, the approximation of Ax_k to b is quite good. This leaves us with a curious conundrum: we have a good approximation x_k of the true solution in the sense that $Ax_k \sim b$ even though x_k is not very close to x. One way this can be achieved is if $x_k - x$ is mainly a linear combination of small singular vectors of A. This is related to the following result recently obtained by the author.

THEOREM 1.1 ([37]). *Let* v_ℓ *be a (right) singular vector of* A *associated to the singular value* σ_ℓ*. Then, for the sequence* $(x_k)_{k-0}^{\infty}$ *obtained in this randomized manner, it holds that*

$$\mathbb{E}\,\langle x_k - x, v_\ell\rangle = \left(1 - \frac{\sigma_\ell^2}{\|A\|_F^2}\right)^k \langle x_0 - x, v_\ell\rangle\,.$$

Here, $\|A\|_F$ denotes the Frobenius norm. This shows that we expect $x_k - x$ to be indeed mainly a linear combination of singular vectors associated to small singular values since those

ETNA
Kent State University and
Johann Radon Institute (RICAM)

are the ones undergoing the slowest decay. It also mirrors the bound obtained by Strohmer and Vershynin [39] since $\sigma_\ell \geq \sigma_n = \|A^{-1}\|^{-1}$. While being interesting in itself, this identity does not fully explain the behavior shown above: it provides a bound only in expectation with no control of the variance. Moreover, the inner product does initially undergo some fluctuations. Taking the same type of matrix as above, we see an example of such fluctuations in Figure 1.2.

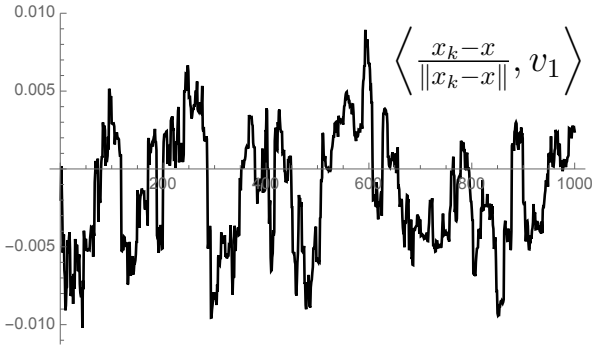

FIG. 1.2. *Evolution of the normalized error against the leading singular vector v_1: fluctuations around the mean.*

1.3. Related results. This type of question is well studied. We refer to Ali, Dobridan, and Tibshirani [1], Defossez and Bach [3], Jain, Kakade, Kidambi, Netrapalli, Pillutla, and Sidford [16], Neu and Rosasco [31], Oymak and Soltanolkotabi [33], Schmidt, Le Roux, and Bach [35] and the references therein. The connection of the SGD method applied to least-squares problems and the Randomized Kaczmarz Method has been pointed out by Needell, Srebro, and Ward [29]. We also mention the papers by Jiao, Jin, and Lu [17] and Jin and Lu [18], who studied a similar question and noted that there is an energy transfer from large singular values to small singular values.

2. Results.

2.1. Main result. The main goal of this note is to provide a simple explanation for the rapid initial regularization: the expected decay of the quantity $\|A(x_{k+1} - x)\|$ under SGD can be bounded from above by a term involving $\left\|A^T A(x_k - x)\right\|_2$: this is the same term except that a matrix has been applied to the existing quantity one more time. This increases the norm of the underlying vector except when $A(x_k - x)$ is mainly the linear combination of singular vectors with small singular values. So as long as this is not the case, we actually inherit strong decay properties, and this leads to the rapid initial regularization.

THEOREM 2.1. *Let $A \in \mathbb{R}^{n \times n}$ be invertible, denote its rows by $a_1, \ldots, a_n$, and apply the Stochastic Gradient Descent method introduced above to $\|Ax - b\|_2^2 \rightarrow \min$. Abbreviating*

$$\alpha = \max_{1 \leq i \leq n} \frac{\|Aa_i\|^2}{\|a_i\|^2},$$

we have

$$\mathbb{E}\,\|Ax_{k+1} - b\|_2^2 \leq \left(1 + \frac{\alpha}{\|A\|_F^2}\right) \|Ax_k - b\|_2^2 - \frac{2}{\|A\|_F^2} \left\|A^T(Ax_k - b)\right\|_2^2.$$

The inequality also holds for $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ as long as $Ax = b$ has a unique solution.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The main point of the inequality is that the last term has an additional matrix multiplication with A^T: we can rewrite it as

$$\left\|A^T(Ax_k - b)\right\|_2^2 = \|A^TA(x_k - x)\|_2^2.$$

This shows that the presence of large singular vectors in $x_k - x$ induces a large decay for $\|A(x_k - x)\|_2^2$. Conversely, once the algorithm has reached the plateau phase (see Figure 1.1), the fact that the decay has slowed down implies that the terms

$$\alpha\|A(x_k - x)\|_2^2 \qquad \text{and} \qquad \|A^TA(x_k - x)\|_2^2$$

are nearly comparable. Thus, this forces $x_k - x$ to be nearly orthogonal to most singular vectors corresponding to large singular values, which, however, shows that it is mainly comprised of small singular vectors and thus explains why $\|A(x_k - x)\| \ll \|x_k - x\|$ is possible in cases where x_k is far away from x. In particular, this suggests why the method could be effective for the problem of finding a vector x such that $Ax \approx b$. One way is to initialize the SGD method with $x_0 = 0$ and run it for a while. Due to the difference in scales and as second-order norms are regularizing first-order norms, we observe that Ax_k converges quite rapidly; whether it converges to something sufficiently close to b for the purpose at hand, is a different question.

2.2. The value of α. It is clearly important to understand the role of α. The size of α essentially governs at what range the algorithm enters the plateau phase. We first note that, since a_i is a row of the matrix A, we have

$$\|Aa_i\|^2 = \sum_{k=1}^{n} \langle a_k, a_i\rangle^2 \geq \|a_i\|^2,$$

and therefore $\alpha \geq 1$. It is easy to see that α is a measure of whether any large singular vectors of A have a large inner product with any of the rows—this may happen but in many settings of interest does not. We can make this statement precise for random matrices.

PROPOSITION 2.2. *Let $A \in \mathbb{R}^{n\times n}$ be comprised of independent Gaussian entries $a_{ij} \sim \mathcal{N}(0,1)$. Then*

$$\alpha \sim (2 + o(1))n$$

with high probability as $n \to \infty$.

The proof of the proposition shows a slightly more precise result, which is not required for the subsequent discussion. We can now analyze precisely what this means in terms of Theorem 2.1. Let us abbreviate $v = Ax_k - b$. Theorem 2.1 guarantees a decay as long as $\alpha\|v\|_2^2 \geq 2\|A^Tv\|_2^2$. We have that $\alpha \sim 2n$, and we have that

$$\|A^Tv\|_2^2 = \left\langle A^Tv, A^Tv\right\rangle = \left\langle A^TAv, v\right\rangle.$$

The eigenvalues of A^TA are distributed in (approximately) $[0, 4n]$. This shows that the Stochastic Gradient Descent method acts as an effective regularizer as long as the projection of v onto the space of eigenvalues in $[n, 4n]$ is large.

2.3. A Sobolev space interpretation. An interesting way to illustrate the result is in terms of partial differential equations. Suppose we try to solve $-\Delta u = f$ on some domain $\Omega \subset \mathbb{R}^n$. After a suitable discretization, this results in a discrete linear system $Lu = f$, where $L \in \mathbb{R}^{n\times n}$ is a discretization of the Laplacian $-\Delta$. By an abuse of notation, u denotes a discrete approximation of the continuous solution and f a discretization of the continuous

ETNA
Kent State University and
Johann Radon Institute (RICAM)

right-hand side. However, we also have more information: since L discretizes the Laplacian, we expect that

$$\langle Lu, u\rangle \sim \int_\Omega |\nabla u|^2 dx \qquad \text{and} \qquad \langle Lu, Lu\rangle \sim \int_\Omega |\Delta u|^2 dx.$$

Here, the first term correspond to the norm of u in the Sobolev space $\dot{H}^1$, while the second term is the norm of u in the Sobolev space $\dot{H}^2$. In fact, this is a common way to define discretized approximations in Sobolev spaces, also known as the spectral definition, since they are defined in terms of the spectrum of L. Suppose now we compute a sequence of approximations u_k via the method outlined above. Then Theorem 2.1 can be rephrased as

$$\mathbb{E}\|u_{k+1} - u\|_{\dot{H}^1}^2 \leq \left(1 + \frac{\alpha}{\|L\|_F^2}\right) \|u_k - u\|_{\dot{H}^1}^2 - \frac{2}{\|L\|_F^2} \|u_k - u\|_{\dot{H}^2}^2 .$$

What is of great interest here is that the decay of the error in $\dot{H}^1$ is driven by the decay of the error in $\dot{H}^2$ (which is usually larger).

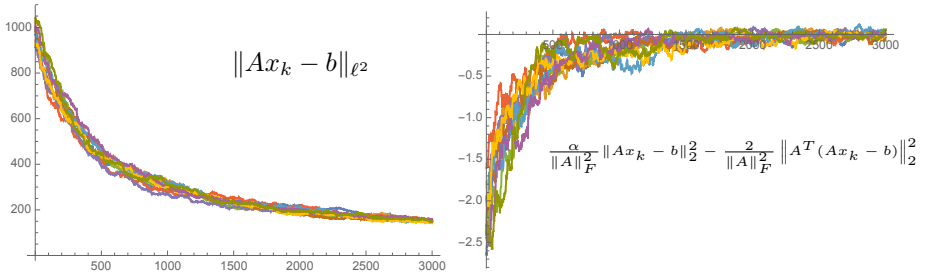

FIG. 2.1. $\|Ax_k - b\|_{\ell^2}$, *for* $k = 1, \ldots, 3000$, *(left) and the decay guaranteed by Theorem 2.1 in expectation (right) over multiple runs.*

2.4. An example. We illustrate this with an example. Choosing $A \in \mathbb{R}^{500\times 500}$ at random (and then, for convenience, normalize the rows to $\|a_i\| = 1$), we solve $Ax = (1, 1, \ldots, 1)$ starting with a random initial vector x_0 where each entry is chosen independently from a standardized $\mathcal{N}(0, 1)$-distribution. We consider both the evolution of $\|Ax_k - b\|_2^2$ across multiple runs as well as the size of

$$\frac{\alpha}{\|A\|_F^2}\|Ax_k - b\|_2^2 - \frac{2}{\|A\|_F^2}\left\|A^T(Ax_k - b)\right\|_2^2,$$

which is the term from our theorem quantifying the expected decay at each step. We see in Figure 2.1 that over 3000 periods, the approximation decays roughly by a factor ~ 800 (with little variation across multiple runs). The bound in the theorem implies an expected decay of -0.23 per time-step, which, over 3000 time steps, leads to a total decay factor of roughly ~ 696.

2.5. Higher powers. If the matrix $A \in \mathbb{R}^{n\times n}$ is symmetric, then we can extend the result to higher powers.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

THEOREM 2.3. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric and invertible. When solving the problem $\|Ax-b\|_2^2 \to \min$ via the Stochastic Gradient Descent method outlined above, we have, for any $\ell \in \mathbb{N}$ with*

$$\alpha_\ell = \max_{1\leq i\leq n} \frac{\|A^\ell a_i\|^2}{\|a_i\|^2},$$

the estimate

$$\mathbb{E}\,\|A^\ell(x_{k+1}-x^*)\|_2^2 \leq \left(1+\frac{\alpha_\ell}{\|A\|_F^2}\right)\|A^\ell(x_k-x^*)\|_2^2 - \frac{2}{\|A\|_F^2}\left\|A^{\ell+1}(x_k-x^*)\right\|_2^2.$$

This shows that the same phenomenon does happen at all scales of 'smoothness'. The applicability of the result is, naturally, depending on the growth of α_ℓ in ℓ, though, generically, one would not expect this to be badly behaved: there is no good reason to expect that the row of a matrix happens to be the linear combination of singular vectors associated to large singular values—though, naturally, this can happen (for example, if A has one very large entry on the diagonal).

2.6. Open problem. The analysis of this particular scheme of gradient descent is somewhat distinguished insofar as picking a descent direction with likelihood proportional to the Lipschitz constant $\|a_i\|_{\ell^2}$ leads to a distinguished scheme undergoing a certain algebraic simplification. One could consider other schemes (see, e.g., [38] for the Random Kaczmarz method), but it is not clear whether they might admit similar inequalities—we believe this to be an interesting problem (both in the Random Kaczmarz setting as well as in the Stochastic Gradient setting).

3. Proofs.

3.1. Proof of Theorem 2.1. *Proof.* To simplify the exposition, we introduce the error

$$r_k = x_k - x.$$

Plugging in, we obtain that if the i-th equation is chosen, then

$$\begin{aligned} x + r_{k+1} = x_{k+1} &= x_k + \frac{b_i - \langle a_i, x_k\rangle}{\|a_i\|^2}a_i = x + r_k + \frac{b_i - \langle a_i, x + r_k\rangle}{\|a_i\|^2}a_i \\ &= x + r_k - \frac{\langle a_i, r_k\rangle}{\|a_i\|^2}a_i + \left(\frac{b_i - \langle a_i, x\rangle}{\|a_i\|^2}a_i\right). \end{aligned}$$

Since x is the exact solution, we have $b_i - \langle a_i, x\rangle = 0$ and

$$r_{k+1} = r_k - \frac{\langle a_i, r_k\rangle}{\|a_i\|^2}a_i.$$

Recalling that the i-th row is chosen with probability proportional to $\|a_i\|^2$,

$$\mathbb{E}\,\|Ar_{k+1}\|^2 = \mathbb{E}\left\|A\left(r_k - \frac{\langle a_i, r_k\rangle}{\|a_i\|^2}a_i\right)\right\|^2 = \sum_{i=1}^n \frac{\|a_i\|^2}{\|A\|_F^2}\left\|Ar_k - \frac{\langle a_i, r_k\rangle}{\|a_i\|^2}Aa_i\right\|^2.$$

This norm can be explicitly squared out as

$$\left\|Ar_k - \frac{\langle a_i, r_k\rangle}{\|a_i\|^2}Aa_i\right\|^2 = \|Ar_k\|^2 - 2\frac{\langle a_i, r_k\rangle}{\|a_i\|^2}\langle Ar_k, Aa_i\rangle + \frac{\langle a_i, r_k\rangle^2}{\|a_i\|^4}\|Aa_i\|^2.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

This allows us to rewrite the summation as

$$\begin{aligned}\mathbb{E}\,\|Ar_{k+1}\|^2 &= \sum_{i=1}^{n} \frac{\|a_i\|^2}{\|A\|_F^2}\left(\|Ar_k\|^2 - 2\frac{\langle a_i, r_k\rangle}{\|a_i\|^2}\langle Ar_k, Aa_i\rangle + \frac{\langle a_i, r_k\rangle^2}{\|a_i\|^4}\|Aa_i\|^2\right)\\ &= \|Ar_k\|^2 - \frac{2}{\|A\|_F^2}\sum_{i=1}^{n}\langle a_i, r_k\rangle\langle Ar_k, Aa_i\rangle + \frac{1}{\|A\|_F^2}\sum_{i=1}^{n}\langle a_i, r_k\rangle^2\frac{\|Aa_i\|^2}{\|a_i\|^2}.\end{aligned}$$

The second sum can be simplified. We observe that $\langle a_i, r_k\rangle$ is the i-th entry of Ar_k. Similarly, $\langle A^TAr_k, a_i\rangle$ is the i-th entry fo AA^TAr_k. Therefore,

$$\begin{aligned}\frac{2}{\|A\|_F^2}\sum_{i=1}^{n}\langle a_i, r_k\rangle\langle Ar_k, Aa_i\rangle &= \frac{2}{\|A\|_F^2}\sum_{i=1}^{n}\langle a_i, r_k\rangle\langle A^TAr_k, a_i\rangle\\ &= \frac{2}{\|A\|_F^2}\langle Ar_k, AA^TAr_k\rangle = \frac{2}{\|A\|_F^2}\|A^TAr_k\|^2.\end{aligned}$$

The last sum we bound from above via

$$\frac{1}{\|A\|_F^2}\sum_{i=1}^{n}\langle a_i, r_k\rangle^2\frac{\|Aa_i\|^2}{\|a_i\|^2} \le \left(\max_i \frac{\|Aa_i\|^2}{\|a_i\|^2}\right)\frac{\|Ar_k\|^2}{\|A\|_F^2}.$$

This results in the desired estimate. ∎

3.2. Proof of Theorem 2.3. *Proof.* We again reduce the problem to that of the study of the error

$$r_{k+1} = r_k - \frac{\langle a_i, r_k\rangle}{\|a_i\|^2}a_i.$$

When looking at integer powers, we observe that, by the same reasoning,

$$\begin{aligned}\mathbb{E}\,\|A^\ell r_{k+1}\|^2 &= \mathbb{E}\left\|A^\ell\left(r_k - \frac{\langle a_i, r_k\rangle}{\|a_i\|^2}a_i\right)\right\|^2\\ &= \sum_{i=1}^{n}\frac{\|a_i\|^2}{\|A\|_F^2}\left\|A^\ell r_k - \frac{\langle a_i, r_k\rangle}{\|a_i\|^2}A^\ell a_i\right\|^2\\ &= \sum_{i=1}^{n}\frac{\|a_i\|^2}{\|A\|_F^2}\left(\|A^\ell r_k\|^2 - 2\frac{\langle a_i, r_k\rangle}{\|a_i\|^2}\langle A^\ell r_k, A^\ell a_i\rangle + \frac{\langle a_i, r_k\rangle^2}{\|a_i\|^4}\|A^\ell a_i\|^2\right)\\ &= \|A^\ell r_k\|^2 - \frac{2}{\|A\|_F^2}\sum_{i=1}^{n}\langle a_i, r_k\rangle\langle A^\ell r_k, A^\ell a_i\rangle + \frac{1}{\|A\|_F^2}\sum_{i=1}^{n}\langle a_i, r_k\rangle^2\frac{\|A^\ell a_i\|^2}{\|a_i\|^2}.\end{aligned}$$

The first term is easy to analyze, and the third term can, as before, be bounded by

$$\frac{1}{\|A\|_F^2}\sum_{i=1}^{n}\langle a_i, r_k\rangle^2\frac{\|A^\ell a_i\|^2}{\|a_i\|^2} \le \frac{\alpha_\ell}{\|A\|_F^2}\sum_{i=1}^{n}\langle a_i, r_k\rangle^2 = \frac{\alpha_\ell}{\|A\|_F^2}\|Ar_k\|^2.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

It remains to understand the second term: here, we can use the symmetry of the matrix to write

$$\begin{aligned}&\frac{2}{\|A\|_F^2}\sum_{i=1}^{n}\langle a_i, r_k\rangle\left\langle A^\ell r_k, A^\ell a_i\right\rangle\\ &\quad=\frac{2}{\|A\|_F^2}\sum_{i=1}^{n}\langle a_i, r_k\rangle\left\langle A^{2\ell} r_k, a_i\right\rangle=\frac{2}{\|A\|_F^2}\left\langle Ar_k, A^{2\ell+1} r_k\right\rangle\\ &\quad=\frac{2}{\|A\|_F^2}\left\langle A^{\ell+1} r_k, A^{\ell+1} r_k\right\rangle=\frac{2}{\|A\|_F^2}\|A^{\ell+1} r_k\|^2. \qquad \square\end{aligned}$$

3.3. Proof of Proposition 2.2. *Proof.* We write

$$\frac{\|Aa_i\|^2}{\|a_i\|^2}=\left\|A\frac{a_i}{\|a_i\|}\right\|^2.$$

The vector $v = a_i/\|a_i\|$ behaves like a randomly chosen vector with respect to the Haar measure. Using independence of the rows, we have, for $v = a_i/\|a_i\|$,

$$\left\|A\frac{a_i}{\|a_i\|}\right\|^2=\|a_i\|^2+\sum_{j\neq i}\langle a_j, v\rangle^2.$$

We note that the second sum is comprised of $n-1$ independent objects. Each of these objects $\langle a_j, v\rangle^2$ has a distribution function that can be determined in closed form: both v and a_j share rotational symmetry. We can thus—to determine the distribution—set $v = (1, 0, \ldots, 0)$ and obtain that $\langle a_j, v\rangle$ is distributed as $\mathcal{N}(0, 1)$. Thus, for i fixed, the distribution of the sum is given by a χ^2-distribution

$$\sum_{j\neq i}\langle a_j, v\rangle^2\sim\chi^2_{n-1}.$$

We use an inequality of Laurent and Massart [19] to argue that

$$\mathbb{P}\left(\chi^2_{n-1}-(n-1)\geq 2\sqrt{(n-1)x}+2x\right)\leq e^{-x}.$$

Setting $x = c\log n$ for some $c > 1$, we get

$$\mathbb{P}\left(\chi^2_{n-1}-(n-1)\geq 2\sqrt{cn\log n}+2\log n\right)\leq\frac{1}{n^c},$$

and, by the union bound, the maximum over n different (possibly dependent) terms does not violate this bound with likelihood $\leq n^{-(c-1)}$. We also see that the value of c only affects the lower-order terms. $\square$

REFERENCES

[1] A. ALI, E. DOBRIDAN, AND R. TIBSHIRANI, *The implicit regularization of stochastic gradient flow for least squares*, Preprint on arXiv, 2020. https://arxiv.org/abs/2003.07802

[2] Z.-Z. BAI AND W.-T. WU, *On convergence rate of the randomized Kaczmarz method*, Linear Algebra Appl., 553 (2018), pp. 252–269.

[3] A. DEFOSSEZ AND F. BACH, *Averaged least-mean-squares: bias-variance trade-offs and optimal sampling distributions*, in Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, G. Lebanon and S. V. N. Vishwanathan, eds., Proceedings of Machine Learning Research 38, San Diego, 2015, pp. 205–213.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[4] F. DEUTSCH, *Rate of convergence of the method of alternating projections*, in Parametric Optimization and Approximation, B. Brosowski and F. Deutsch, eds., Internat. Schriftenreihe Numer. Math., 72, Birkhäuser, Basel, 1985, pp. 96–107

[5] F. DEUTSCH AND H. HUNDAL, *The rate of convergence for the method of alternating projections. II*, J. Math. Anal. Appl., 205 (1997), pp. 381–405.

[6] Y. C. ELDAR AND D. NEEDELL, *Acceleration of randomized Kaczmarz method via the Johnson-Lindenstrauss lemma*, Numer. Algorithms, 58 (2011), pp. 163–177.

[7] T. ELFVING, P. C. HANSEN, AND T. NIKAZAD, *Semi-convergence properties of Kaczmarz's method*, Inverse Problems, 30 (2014), Art. 055007, 16 pages.

[8] H. G. FEICHTINGER, C. CENKER, M. MAYER, H. STEIER, AND T. STROHMER, *New variants of the POCS method using affine subspaces of finite codimension with applications to irregular sampling*, in Visual Communications and Image Processing '92, P. Maragos, ed., Proceedings of SPIE 1818, SPIE, Bellingham, pp. 299–310, 1992.

[9] A. GALANTAI, *On the rate of convergence of the alternating projection method in finite dimensional spaces*, J. Math. Anal. Appl., 310 (2005), pp. 30–44.

[10] D. GORDON, *A derandomization approach to recovering bandlimited signals across a wide range of random sampling rates*, Numer. Algorithms, 77 (2018), pp. 1141–1157.

[11] R. GORDON, R. BENDER, AND G. HERMAN, *Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography*, J. Theoret. Biol., 29 (1970), pp. 471–476.

[12] R. M. GOWER, D. MOLITOR, J. MOORMAN, AND D. NEEDELL, *Adaptive sketch-and-project methods for solving linear systems*, Preprint on arXiv, 2019. https://arxiv.org/abs/1909.03604

[13] R. M. GOWER AND P. RICHTARIK, *Randomized iterative methods for linear systems*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1660–1690.

[14] G. HERMAN, *Image Reconstruction from Projections*, Academic Press, New York, 1980.

[15] G. T. HERMAN AND L. B. MEYER, *Algebraic reconstruction techniques can be made computationally efficient*, IEEE Trans. Medical Imag., 12 (1993), pp. 600–609.

[16] P. JAIN, S. KAKADE, R. KIDAMBI, P. NETRAPALLI, V. PILLUTLA, AND A. SIDFORD, *A Markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares)*, in 37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2017), S. Lokam and R. Ramanujam, eds., LIPIcs, Leibniz-Zentrum für Informatik, Dagstuhl, 2018, pp. 2:1–2:10.

[17] Y. JIAO, B. JIN, AND X. LU, *Preasymptotic convergence of randomized Kaczmarz method*, Inverse Problems, 33 (2017), Art. 125012, 21 pages.

[18] B. JIN AND X. LU, *On the regularizing property of stochastic gradient descent*, Inverse Problems, 35 (2019), Art. 015004, 27 pages.

[19] B. LAURENT AND P. MASSART, *Adaptive estimation of a quadratic functional by model selection*, Ann. Statist., 28 (2000), pp. 1302–1338.

[20] Y. T. LEE AND A. SIDFORD, *Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems*, in 2013 IEEE 54th Annual Symposium on Foundations of Computer Science—FOCS 2013, IEEE Computer Soc., Los Alamitos, 2013, pp. 147–156.

[21] D. LEVENTHAL AND A. S. LEWIS, *Randomized methods for linear constraints: convergence rates and conditioning*, Math. Oper. Res., 35 (2010), pp. 641–654.

[22] J. LIU AND S. J. WRIGHT, *An accelerated randomized Kaczmarz algorithm*, Math. Comp., 85 (2016), pp. 153–178.

[23] A. MA, D. NEEDELL, AND A. RAMDAS, *Convergence properties of the randomized extended Gauss-Seidel and Kaczmarz methods*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1590–1604.

[24] J. D. MOORMAN, T. K. TU, D. MOLITOR, AND D. NEEDELL, *Randomized Kaczmarz with averaging*, BIT, 61 (2021), pp. 337–359.

[25] F. NATTERER, *The Mathematics of Computerized Tomography*, Wiley, New York, 1986.

[26] D. NEEDELL, *Randomized Kaczmarz solver for noisy linear systems*, BIT, 50 (2010), pp. 395–403.

[27] D. NEEDELL AND J. A. TROPP, *Paved with good intentions: analysis of a randomized block Kaczmarz method*, Linear Algebra Appl., 441 (2014), pp. 199–221.

[28] D. NEEDELL AND R. WARD, *Two-subspace projection method for coherent overdetermined systems*, J. Fourier Anal. Appl., 19 (2013), pp. 256–269.

[29] D. NEEDELL, R. WARD, AND N. SREBRO, *Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm*, in Advances in Neural Information Processing Systems 27 (NIPS 2014), Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, eds., Curran Assoc., Red Hook, 2014, pp. 1017–1025.

[30] D. NEEDELL, R. ZHAO, AND A. ZOUZIAS, *Randomized block Kaczmarz method with projection for solving least squares*, Linear Algebra Appl., 484 (2015), pp. 322–343.

[31] G. NEU AND L. ROSASCO, *Iterate averaging as regularization for stochastic gradient descent*, Proceedings of the 31st Conference On Learning Theory, S. Bubeck, V. Perchet, and P. Rigollet, eds., Proceedings of

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Machine Learning Research 75, San Diego, 2018, pp. 3222–3242.

[32] J. NUTINI, B. SEPEHRY, I. LARADJI, M. SCHMIDT, H. KOEPKE, AND A. VIRANI, *Convergence rates for greedy Kaczmarz algorithms, and faster randomized Kaczmarz rules using the orthogonality graph*, in Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence, A. Ihler and D. Janzing, eds., AUAI Press, Arlington, 2016, pp. 547–556.

[33] S. OYMAK AND M. SOLTANOLKOTABI, *Overparameterized nonlinear learning: Gradient descent takes the shortest path?*, in Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and R. Salakhutdinov, eds., Proceedings of Machine Learning Research 97, San Diego, 2019, pp. 4951–4960.

[34] C. POPA, *Convergence rates for Kaczmarz-type algorithms*, Numer. Algorithms, 79 (2018), pp. 1–17.

[35] M. SCHMIDT, N. LE ROUX, AND F. BACH, *Minimizing finite sums with the stochastic average gradient*, Math. Program. Ser. A, 162 (2017), pp. 83–112.

[36] K. SEZAN AND H. STARK, *Applications of convex projection theory to image recovery in tomography and related areas*, in Image Recovery: Theory and Application, H. Stark, ed., Academic Press, New York, 1987, pp. 415–462.

[37] S. STEINERBERGER, *Randomized Kaczmarz converges along small singular vectors*, SIAM J. Matrix Anal. Appl., 42 (2021), pp. 608–615.

[38] ———, *A weighted randomized Kaczmarz method for solving linear systems*, Math. Comp., 90 (2021), pp. 2815–2826.

[39] T. STROHMER AND R. VERSHYNIN, *A randomized Kaczmarz algorithm for linear systems with exponential convergence*, J. Fourier Anal. Appl., 15 (2009), pp. 262–278.

[40] Y. S. TAN AND R. VERSHYNIN, *Phase retrieval via randomized Kaczmarz: theoretical guarantees*, Inf. Inference, 8 (2019), pp. 97–123.

[41] J.-J. ZHANG, *A new greedy Kaczmarz algorithm for the solution of very large linear systems*, Appl. Math. Lett., 91 (2019), pp. 207–212.

[42] A. ZOUZIAS AND N. M. FRERIS, *Randomized extended Kaczmarz for solving least squares*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 773–793.