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 \(\|Ax - b\|^2\) → 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

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

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

\[\|Ax - b\|^2 = \sum_{i=1}^n (\langle a_i, x \rangle - b_i)^2 \rightarrow \min.\]

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

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

The Lipschitz constant of \(f_i\) is \(\|a_i\|_2\), 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).
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

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

where $\|A\|^2_F$ 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 $N(0, 1)$ 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 $\|A x_k - b\|_2$ over the first 10,000 iteration steps (left) and the subsequent 10,000 iteration steps (right).

The picture tells a very interesting story: the error in $\|A x_k - b\|$ 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 $A x_k$ 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 $A x_k \sim b$ 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 $v_\ell$ be a (right) singular vector of $A$ associated to the singular value $\sigma_\ell$. Then, for the sequence $(x_k)_{k=0}^\infty$ obtained in this randomized manner, it holds that

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

Here, $\|A\|^2_F$ 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
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.

![Figure 1.2. Evolution of the normalized error against the leading singular vector $v_1$: fluctuations around the mean.](image)

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 \(\|A^T A(x_k - x)\|_2\): 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 \(\|Ax - b\|_2^2 \to \min\). Abbreviating
\[
\alpha = \max_{1 \leq i \leq n} \frac{||Aa_i||^2}{||a_i||^2},
\]
we have
\[
\mathbb{E} \|Ax_{k+1} - b\|_2^2 \leq \left(1 + \frac{\alpha}{\|A\|_F^2}\right) \|Ax_k - b\|_2^2 - \frac{2}{\|A\|_F^2} \|A^T (Ax_k - b)\|_2^2.
\]
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.
The main point of the inequality is that the last term has an additional matrix multiplication
with $A^T$: we can rewrite it as

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

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

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

are nearly comparable. Thus, this forces $x_k - x$ to be nearly orthogonal to most singular vectors
that correspond 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 $\alpha$. It is clearly important to understand the role of $\alpha$. The size of $\alpha$
provides 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
$\alpha \|v\|_2^2 \geq 2\|A^T v\|_2^2$. We have that $\alpha \sim 2n$, and we have that

$$
\|A^T v\|_2^2 = \langle A^T v, A^T v \rangle = \langle A^T A v, v \rangle.
$$

The eigenvalues of $A^T A$ 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
right-hand side. However, we also have more information: since $L$ discretizes the Laplacian, we expect that

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

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

$$E\|u_{k+1} - u\|_{\dot{H}^1}^2 \leq \left(1 + \frac{\alpha}{\|L\|_F^2}\right)\|u_k - u\|_{\dot{H}^1}^2 - \frac{2}{\|L\|_F^2} \|A\|_F^4 \|u_k - u\|_{\dot{H}^2}^2.$$  

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).  

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

$$\frac{\alpha}{\|A\|_F^4} \|Ax_k - b\|_2^2 - \frac{2}{\|A\|_F^4} \|A^T (Ax_k - b)\|_2^2,$$

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 $\sim 800$ (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 $\sim 696$.

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.
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

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

the estimate

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

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 $\alpha_\ell$ in $\ell$, 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

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

$$= x + r_k - \frac{\langle a_i, r_k \rangle}{\|a_i\|^2} a_i + \frac{b_i - \langle a_i, x \rangle}{\|a_i\|^2} a_i.$$

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 $\|a_i\|^2$, we have

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

This norm can be explicitly squared out as

$$\left\| Ar_k - \frac{\langle a_i, r_k \rangle}{\|a_i\|^2} A a_i \right\|^2 = \|Ar_k\|^2 - 2 \frac{\langle a_i, r_k \rangle}{\|a_i\|^2} \langle Ar_k, A a_i \rangle + \frac{\langle a_i, r_k \rangle^2}{\|a_i\|^4} \|A a_i\|^2.$$
When looking at integer powers, we observe that, by the same reasoning,

This allows us to rewrite the summation as

The second sum can be simplified. We observe that

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

The last sum we bound from above via

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

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

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

\[ \frac{1}{\|A\|_F^2} \sum_{i=1}^{n} \langle a_i, r_k \rangle^2 \| A^\ell a_i \|^2 \|a_i\|^2 \leq \frac{\alpha_\ell}{\|A\|_F^2} \sum_{i=1}^{n} \langle a_i, r_k \rangle^2 \| A^\ell a_i \|^2. \]
It remains to understand the second term: here, we can use the symmetry of the matrix to write
\[
\frac{2}{\|A\|_F^2} \sum_{i=1}^{n} \langle a_i, r_k \rangle \langle A^T r_k, A^T a_i \rangle
\]
\[
= \frac{2}{\|A\|_F^2} \sum_{i=1}^{n} \langle a_i, r_k \rangle \langle A^{2T} r_k, a_i \rangle = \frac{2}{\|A\|_F^2} \langle A r_k, A^{2+1} r_k \rangle
\]
\[
= \frac{2}{\|A\|_F^2} \langle A^{2+1} r_k, A^{2+1} r_k \rangle = \frac{2}{\|A\|_F^2} \|A^{2+1} r_k\|^2.
\]

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 \( N(0, 1) \). Thus, for \( i \) fixed, the distribution of the sum is given by a \( \chi^2 \)-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
\[
P \left( \chi^2_{n-1} - (n-1) \geq 2 \sqrt{(n-1)x + 2x} \right) \leq e^{-x}.
\]
Setting \( x = c \log n \) for some \( c > 1 \), we get
\[
P \left( \chi^2_{n-1} - (n-1) \geq 2 \sqrt{c n \log n + 2 \log n} \right) \leq \frac{1}{n^c},
\]
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.

REFERENCES


