# CONVERGENT LEARNING RATES FOR LINEAR SGD

A PREPRINT

**Yaroslav Bulatov**
ML Collective
http://mlcollective.org
San Francisco, CA
yaroslavvb@gmail.com

**Ben Grossman**
Vrije Universiteit Brussel
Benjamin.Grossmann@vub.ac.be

October 14, 2020

**ABSTRACT**

In this note we derive an expression for the largest learning rate which guarantees convergence of SGD in the linear setting. We then specialize to the case of normally distributed observations, which allows us to get a simple linear form of SGD update, and simplifies expressions for largest learning rates.

## 1 Overview

Following setting of [5], let $\mathcal{D}$ be a dataset of observations $\mathcal{D} = (x_1, y_1), \ldots$. We apply Stochastic Gradient Descent (SGD) with learning rate $\alpha$ and least squares loss to find $w$ such for all $i$

$$\langle w x_i \rangle = y_i$$

This is known as the case of consistent equations, or learning in "the interpolation regime." For low enough learning rate $\alpha$, this algorithm produces a sequence of estimates $w_1, \ldots, w_k, \ldots$ converging to the unique point $w^*$ which satisfies the constraint above [1]. How low should $\alpha$ be? Consider the case of batch-size=1 (stochastic gradient descent) and batch-size=infinity (gradient descent)

We are interested in the following thresholds:

- $\alpha < \alpha_a^{\text{gd}}$ is necessary and sufficient for gradient descent to converge to $w^*$
- $\alpha < \alpha_m$ is necessary and sufficient for $k$th step of SGD to decrease expected distance between $w_k$ and $w^*$ (monotonic convergence)
- $\alpha < \alpha_a$ is necessary and sufficient for SGD to converge to $w^*$ (asymptotic convergence)

Consider the case of normally distributed $x$ with covariance $\Sigma$ and mean zero. The Hessian of corresponding optimization problem is $H = E[xx'] = \Sigma$. We show the following

$$\frac{2}{\alpha_a^{\text{gd}}} = \|H\| \tag{1}$$

$$\frac{2}{\alpha_m} = 2\|H\| + \operatorname{Tr} H \tag{2}$$

$$\frac{2}{\alpha_a} = \rho(A) \tag{3}$$

[1] for uniqueness proof in overparameterized setting, see Section 3.3 of [11]

We use $\rho$ to denote spectral radius and $A$ is defined below. Let $s = (s_1, s_2, s_3, \ldots)$ be the vector of eigenvalues of $H$, then

$$A = 2 \begin{pmatrix} s_1 & 0 & 0 & \ldots \\ 0 & s_2 & 0 & \ldots \\ 0 & 0 & s_3 & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix} + \begin{pmatrix} s_1 & s_1 & s_1 & \ldots \\ s_2 & s_2 & s_2 & \ldots \\ s_3 & s_3 & s_3 & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix} \tag{4}$$

We can represent quantities in Eq 1-3 in terms of $s$

$$\frac{2}{\alpha_a^{\text{gd}}} = \|s\|_\infty \tag{5}$$

$$\frac{2}{\alpha_m} = 2\|s\|_\infty + \|s\|_1 \tag{6}$$

$$\frac{2}{\alpha_a} = \lambda_{\max} A \tag{7}$$

$$\tag{8}$$

We can consider learning rate which guarantees a decrease after $k$ steps. For $k \to \infty$, this approaches $\alpha_a$, but for intermediate $k$ we obtain a sequence of increasingly tight bounds. For instance, for $k = 2$ we get

$$\frac{2}{\alpha_a^2} < 2\|s\|_\infty^2 + \|s \odot s\|_1 + \|s\|_\infty \|s\|_1 + \frac{1}{2}\|s\|_1^2 \tag{9}$$

When $x$ is normally distributed, SGD update of error distribution with diagonal covariance matrix produces another distribution with diagonal covariance matrix. This allows us to view update of error covariance after SGD step in $d$ dimensions as multiplication a $d \times d$ matrix $B$. Reusing definition of $s$ from Eq 4 this matrix is below

$$B = I - 2\alpha H + 2\alpha^2 H^2 + \alpha^2 ss'$$

Stationary distribution of SGD corresponds to the top eigenvector of this matrix. Also, this matrix provides an alternative way to derive maximual learning rates for monotonic and asymptotic convergence.

In the rest of this note, we will give an example, summarize previous results, derive expression for $\alpha_m$ (monotonic convergence) in Section 3, expression for $\alpha_a$ (asymptotic convergence) in Section 4, specialize these to Gaussian observations x in sections 3.1 and 4.2, and provide a way to obtain bounds on $\alpha_a$ in Section 4.3. To help understand notation, an example is worked out numerically end-to-end in Appendix B. Linear update form of SGD and the covariance of stationary distribution is derived in Appendix F

### 1.1 Example

Suppose our observations $x$ are centered at zero and normally distributed with covariance matrix $\Sigma$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \tag{10}$$

Applying formulas above gives us the following

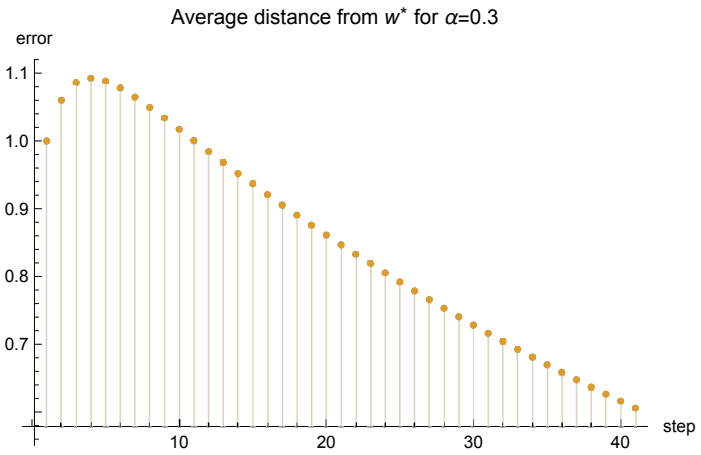
$$\alpha_a^{\text{gd}} = 1 \tag{11}$$

$$\alpha_m = \frac{2}{7} \approx 0.285714 \tag{12}$$

$$\alpha_a = \frac{4}{9 + \sqrt{17}} \approx 0.3048 \tag{13}$$

$$\alpha_a > \frac{2}{\sqrt{47}} \approx 0.29173 \tag{14}$$

We can set $\alpha = 0.3$ and run SGD on this dataset many times to demonstrate the effect of learning rate which guarantees asymptotic but not monotonic convergence:




## 2 Previous results

It can be shown that in the case of 1 dimension and deterministic $x$, the following condition on $\alpha$ is necessary and sufficient for convergence

$$\alpha x^4 < 2x^2 \tag{15}$$

Since we have $h = x^2$ for Hessian $h$, this reduces to the well known bound on convergent learning rate: $\alpha < 2/h$

In the case of stochastic x, the following is necessary and sufficient (Eq 36 of [14])

$$\alpha E[x^4] < 2E[x^2] \tag{16}$$

For the case of $x$ being distributed as standard normal, this gives $2/(3h)$ for the largest learning rate, three times smaller than what's allowed in deterministic case (Eq 15)

For the case of $d$ dimensions, the following is a sufficient condition, with $\prec$ indicating Loewner order[2]

$$\alpha E[xx'xx'] \prec E[xx'] \tag{17}$$

The right-hand side was tightened in [9], section 1.1.2

$$\alpha E[xx'xx'] \prec 2E[xx'] \tag{18}$$

Defossez, Bach showed that the following optimization over symmetric matrices gives sufficient condition for convergence, and conjectured it to also be necessary (Lemma 1 of [5])

$$\frac{1}{\alpha} < \sup_{A \in \mathcal{S}(R^d)} \frac{E[(x'Ax)^2]}{2E[x'A^2x]} \tag{19}$$

In Section 4.1 we show this to be equivalent to the following positive semi-definite constraint

$$\alpha E[xx' \otimes xx'] \prec E[xx' \otimes I] + E[I \otimes xx'] \tag{20}$$

Most recently, [9] generalized Eq 19 to batch sizes beyond 1 and formally showed it to be a necessary condition for convergence.

## 3 Derivation of monotonic convergence

Let $x_k$ denote the example $x$ sampled at $k$th step of SGD, and $w_k$ be the estimate of parameter $w$ at that step. It can be shown that the error vector $\eta_k = w_k - w^*$ obeys the following [3]

[2] assumption A.6 in [1]

[3] Section 1.1 of [5]

$$\eta_{k+1} = (I - \alpha x_k x_k^T)\eta_k \tag{21}$$

Consider error squared $A_k = E[\eta_k \eta_k^T]$, for which the following linear recurrence holds

$$A_{k+1} = E[(I - \alpha xx')A_k(I - \alpha xx')] = T_\alpha(A_k) \tag{22}$$

Equation above serves as the definition of linear operator $T_\alpha$. In Appendix B we will give a matrix representation of $T_\alpha$

Note that the following is true[4]

$$E[\|\eta_k\|^2] = \text{Tr}A_k \tag{23}$$

This is the average distance squared from estimate at step $k$, $w_k$ and target $w^*$. This equivalence means $T_\alpha$ is guaranteed to decrease expected error $E[\|\eta_k\|]$ if and only the induced trace norm of $T_\alpha$ is less than 1, defined as follows

$$\|T\|_1 = \max_{\text{Tr}A=1} \text{Tr}\, T(A) \tag{24}$$

Russo-Dye theorem for induced trace norm [5] tells us that this quantity is maximized by a symmetric rank-1 matrix, hence we can turn Eq 24 into optimization over unit vectors

$$\|T\|_1 = \max_{\|u\|=1} \text{Tr}\, T(uu')$$

Substitute definition of $T$ from Eq 22, and simplify using the following properties: 1) linearity of E 2) linearity of Tr 3) cyclic property of Tr. We get

$$\|T\|_1 = \max_{\|u\|=1} u' E[(I - \alpha xx')^2]u$$

That this is the Raleigh quotient characterization of the largest eigenvalue, equivalent to matrix norm in symmetric case, hence

$$\|T\|_1 = \|E[(I - \alpha xx')^2]\| = \|T(I)\|$$

Alternative way to obtain this expression is to rely on duality of trace and spectral norms, see Appendix D

To find the largest $\alpha$ satisfying $\|T_\alpha(I)\| < 1$, reformulate in terms of positive definite constraints

$$\begin{array}{ll} \text{maximize} & \alpha \\ \text{subject to} & -I \prec T_\alpha(I) \prec I \end{array} \tag{25}$$

Note that $T_\alpha(I)$ can be expanded

$$T_\alpha(I) = E[(I - \alpha xx')^2] = I - 2\alpha E[xx'] + \alpha^2 E[xx'xx'] \tag{26}$$

Since $-I \prec T(I)$ is true by virtue of $T(I)$ being a covariance matrix, we only consider the rightmost inequality. Plug Eq 26 into Eq 25 and simplify to obtain

$$\begin{array}{ll} \text{maximize} & \alpha \\ \text{subject to} & \alpha E[xx'xx'] \prec 2E[xx'] \end{array}$$

If $E[xx']$ is not singular, the solution is obtained at $1/\lambda_{\text{max}}$, where $\lambda_{\text{max}}$ is the largest eigenvalue of the following matrix[6]:

$$E[xx'xx'](2E[xx'])^{-1} \tag{27}$$

[4] $E[\|\eta\|^2] = E[\eta^T\eta] = E[\text{Tr}(\eta^T\eta)] = E[\text{Tr}(\eta\eta^T)] = \text{Tr}E[\eta\eta^T] = \text{Tr}A$

[5] Theorem 3.39 of [16], `https://cs.uwaterloo.ca/~watrous/TQI/TQI.pdf`

[6] See Appendix A

### 3.1 Gaussian case

The following holds for Gaussian-distributed $x$ centered at 0 [7]

$$E[xx'xx'] = 2\Sigma^2 + \Sigma \text{Tr}\Sigma$$

Assuming $\Sigma$ is full-rank and plug into Eq.27, we get

$$E[xx'xx'](2E[xx'])^{-1} = \Sigma + \left(\frac{1}{2}\text{Tr}\Sigma\right) I$$

Largest eigenvalue of this matrix is $\|\Sigma\| + \frac{1}{2}\text{Tr}\Sigma$, hence we have

$$\frac{2}{\alpha_m} = 2\|\Sigma\| + \text{Tr}\Sigma$$

## 4 Derivation of asymptotic convergence

We can derive the necessary and sufficient conditions for asymptotic convergence by examining spectral radius of operator $T$. Consider its matrix form

$$\text{vec}\, C_{k+1} = M \,\text{vec}\, C_k \tag{28}$$

$M$ is a symmetric $d^2$-by-$d^2$ positive semi-definite matrix below

$$\begin{aligned} M &= I \otimes I - \alpha(X^2 \otimes I) - \alpha(I \otimes X^2) + \alpha^2 X^4 = & (29)\\ &= I - \alpha M_p & (30)\end{aligned}$$

We we have defined the following quantities

$$\begin{aligned} X^2 &= E[xx'] & (31)\\ X^4 &= E[xx' \otimes xx'] & (32)\\ M_p &= (I \otimes X^2) + (X^2 \otimes I) - \alpha X^4 & (33)\end{aligned}$$

Since $M$ is symmetric, operator norm $\|M\|$ corresponds to $M$'s spectral radius $\rho$. We know that $M^k$ converges iff $\rho < 1$ [8]

To guarantee this, we need to bound $T(A)$ from two sides, $-I \prec T(A) \prec I$. Because $T(A)$ is a covariance matrix, the leftmost inequality is true independent of $\alpha$. Rearrange and divide rightmost inequality by $\alpha > 0$, to get an equivalent condition: $M_p \succ 0$. Expanding we get

$$(I \otimes X^2) + (X^2 \otimes I) - \alpha X^4 \succ 0 \tag{34}$$

Alternatively

$$\alpha X^4 \prec (I \otimes X^2) + (X^2 \otimes I) \tag{35}$$

The largest rate for which this is true, $\alpha_a$ is determined as the solution of the following optimization problem

$$\begin{array}{ll} \text{maximize} & \alpha \\ \text{subject to} & \alpha X^4 \prec (I \otimes X^2) + (X^2 \otimes I) \end{array} \tag{36}$$

[7] see 20.24 of [13], also Appendix B

[8] Theorem 10.1 of [10] in `http://www.optimization-online.org/DB_FILE/2003/04/640.pdf`

### 4.1 Characterization in terms of moments

Lets show that bound in terms of moments of x in Eq 19 is equivalent to the positive semidefinite bound in Eq 35. Consider that Eq 35 can be represented as follows. For all $a = \text{vec}\, A$, with $A$ a $d$-by-$d$ matrix

$$\alpha a' X^4 a < a'((I \otimes X^2) + (X^2 \otimes I))a \tag{37}$$

Maximizing $\alpha$ subject to this constraint is a generalized eigenvalue problem in disguise, see Appendix B.4.

To show the equivalence of Eq 37 with Eq 19 first express the following quantities as expectations:

$$\begin{aligned} a' X^4 a &= E \sum_{ijkl} x_i x_j x_k x_l A_{ij} A_{kl} = E[(x'Ax)^2] && (38)\\ a(X^2 \otimes I)a' &= a' \,\text{vec}(AX^2) = \text{Tr}\, A'AX^2 = E[x'A'Ax] && (39)\\ a(I \otimes X^2)a' &= E[x'AA'x] = E[x'A'Ax] && (40) \end{aligned}$$

Plugging these into equation 37 we get the following. For all matrices $A$

$$\alpha E[(x'Ax)^2] < 2E[x'A'Ax] \tag{41}$$

Since elements of $A$ only occur as parts of quadratic form, non-symmetric part of $A$ has no effect, hence can restrict attention to symmetric $A$ and simplify further to obtain the intended result

$$\alpha E[(x'Ax)^2] < 2E[x'A^2x] \tag{42}$$

### 4.2 Gaussian case

Lets obtain $\alpha_a$ for the case of $x$ being distributed as Gaussian with covariance matrix $\Sigma = X^2$ and mean 0. Assume eigenvalues of covariance matrix $\Sigma$ are positive and distinct. Then you can show that the following power iteration converges to the top generalized eigenvector of the problem in Eq. 36.

$$\begin{aligned} A_{k+1} &= \text{divide} A_k && (43)\\ A_{k+2} &= \text{multiply} A_{k+1} && (44) \end{aligned}$$

Here "multiply" represents multiplying vec$(A)$ by matrix $X^4$ and "divide" is corresponding division by matrix $(I \otimes X^2) + (X^2 \otimes I)$.

We can extract $\rho$ as the limiting ratio of Frobenius norms

$$\frac{1}{\alpha_a} = \lim_{k \to \infty} \frac{\|A_{k+2}\|_F}{\|A_k\|_F} \tag{45}$$

It's convenient to deal with matrix form rather than vec form, where "multiply" and "divide" can be represented as follows

$$\begin{aligned} \text{divide}(A) &= \text{lyapunov}(X^2, A) && (46)\\ \text{multiply}(A) &= E[xx'Axx'] && (47)\\ & && (48) \end{aligned}$$

Here lyapunov$(X^2, A)$ is an alternative way to represent division of vec$(A)$ by matrix $(I \otimes X^2) + (X^2 \otimes I)$, defined as follows

$$\text{vec}\, \text{lyapunov}(X^2, A) = ((I \otimes X^2) + (X^2 \otimes I))^{-1} \,\text{vec}\, A$$

Equivalently, lyapunov$(X^2, A)$ is a function which produces $Y$ the solution of Lyapunov equation below

$$X^2Y + YX^2 = A$$

Solution is known to exist and be unique when $X^2 = \Sigma$ eigenvalues are all positive[9]

Because of rotational symmetry of the problem, we can assume that $\Sigma$ is diagonal. Starting with diagonal $A_0$, observe that each step of power iteration produces another diagonal matrix – Lyapunov solver in diagonal case reduces to elementwise division. Analogously, multiplication by $X^4$ keeps the argument diagonal.

Now, use Gaussian 4th moment formulas from Appendix B and the fact that $A$ stays diagonal to simplify our updates

$$\begin{aligned} \text{divide}(A) &= A.(2X^2)^{-1} && (49)\\ \text{multiply}(A) &= 2\Sigma A\Sigma + \Sigma \text{Tr}\Sigma && (50)\\ & && (51)\end{aligned}$$

Observe that only diagonal entries of $\Sigma$ and $A$ are non-zero, therefore we can merge the two updates into a single update and represent it in terms of diagonal entries $s$ (of $\Sigma$) and $a$ (of $A$)

$$\begin{aligned} \text{merged}(a) &= a \odot s + 0.5s\|a\|_1 && (52)\\ & && (53)\end{aligned}$$

Correspondingly we can reformulate our Eq 45 in terms of vector $a_k$

$$\frac{1}{\alpha_a} = \lim_{k\to\infty} \frac{\|a_{k+1}\|}{\|a_k\|} \tag{54}$$

Now represent Eq 52 as a linear linear update $a_{k+1} = Ma_k$ where $M$ is defined below

$$M = \begin{pmatrix} s_1 & 0 & 0 & \dots \\ 0 & s_2 & 0 & \dots \\ 0 & 0 & s_3 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix} + \frac{1}{2}\begin{pmatrix} s_1 & s_1 & s_1 & \dots \\ s_2 & s_2 & s_2 & \dots \\ s_3 & s_3 & s_3 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix} \tag{55}$$

Now we can say that $\frac{1}{\alpha_a}$ is equal to to $\rho(M)$, its spectral radius[10]

### 4.3 Additional bounds

Since spectral radius $\rho$ is upper bounded by any matrix norm $\|\cdot\|_*$, we can use the following to produce alternative bounds

$$\rho(M)^n \le \|M^n\|_*$$

If we use $n = 1$ and max-column-norm of $M$, we obtain the following bound

$$\rho \le \|s\|_\infty + 0.5\|s\|_1$$

Note that because $s$ represent eigenvalues of $H$, $\|s\|_1 = \text{Tr}H$ and $\|s\|_\infty = \|H\|$. Also, because $1/\rho$ represents asymptotic critical rate, we can rewrite the result above as follows

$$\frac{2}{\alpha_a} \le 2\|H\| + \text{Tr}(H) = \frac{2}{\alpha_m}$$

[9] Lyapunov stability theorem, Theorem 19.2.1 of [17]

[10] In Equation 4 we used $\frac{2}{\alpha_a}$ instead of $\frac{1}{\alpha_a}$, hence the corresponding bound was $\rho(A) = \rho(2M)$

In other words, this provides a way to derive $\alpha_m$ purely algebraically as a bound on $\alpha_a$. We could now use different matrix norms or different values of $n$ to get tighter bounds. For instance, for max-column-norm and $n = 2$ we obtain the following expression

$$\frac{2}{\alpha_a^2} \leq 2\|H\|^2 + \text{Tr}(H^2) + \|H\|\text{Tr}H + \frac{1}{2}\text{Tr}(H)^2$$

# 5 Appendix

## A Solving semidefinite constraint

We are interested in finding largest $\alpha$ such that $\|T_\alpha\| < 1$. We can formulate this as the following optimization problem

$$\begin{array}{ll} \text{maximize} & \alpha \\ \text{subject to} & \|T_\alpha\| < 1 \end{array}$$

Since our matrices are symmetric, we can can reformulate this problem in terms of semidefinite constraint

$$\begin{array}{ll} \text{maximize} & \alpha \\ \text{subject to} & T_\alpha \prec I \end{array}$$

Expanding and rearranging constraint we get

$$\begin{array}{ll} \text{maximize} & \alpha \\ \text{subject to} & \alpha E[xx'xx'] \prec 2E[xx'] \end{array}$$

Turn this into minimization problem by defining $R = 1/\alpha$ and solving the following

$$\begin{array}{ll} \text{minimize} & R \\ \text{subject to} & E[xx'xx'] \prec 2RE[xx'] \end{array} \tag{56}$$

This is one of the forms of a generalized eigenvalue problem (2.2.3 of [3])

If $E[xx']$ is not singular, we can reduce this to standard eigenvalue problem by multiplying both sides of 56 by inverse of $2E[xx']$ (see [6] for the case of singular matrix)

$$\begin{array}{ll} \text{minimize} & R \\ \text{subject to} & RE[xx'xx'](2E[xx'])^{-1} \succ I \end{array} \tag{57}$$

In full-rank case, solution can be obtained as the solution of the eigenvalue problem below[11]

$$R_{\text{min}} = \lambda_{\text{max}}(E[xx'xx'](2E[xx'])^{-1})$$

## B End-to-end example

As a reminder, operator $T$ represents evolution of error covariance[12] $E[\eta\eta^T]$ where $\eta = w - w^*$, the difference between current estimate of parameter and the target estimate. If $A$ represents error covariance at step $k$, $T(A)$ represents is error covariance at step $k + 1$. In this section we shall derive the explicit form of $T$, and use it to obtain learning rate quantities.

Suppose inputs in our dataset are distributed as a 2-$d$ Gaussian with mean zero and the following covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \tag{58}$$

[11] For rank-deficient case, check [6]

[12] technically it is "error second moment", but "error covariance" has a better ring to it

Corresponding optimization problem has Hessian $H = \Sigma$

The following quantities will be useful later, obtained using Gaussian formulas (see 20.24 of [13], and implementation [13])

$$E[xx'xx'] = 2\Sigma^2 + \Sigma \mathrm{Tr}\Sigma \tag{59}$$
$$E[xx' \otimes xx'] = 2P_d(\Sigma \otimes \Sigma) + \mathrm{vec}\Sigma(\mathrm{vec}\Sigma)' \tag{60}$$
$$\tag{61}$$

Here $P_d$ is the symmetrizer matrix[14], matrix of the linear operator $V(A) = (A + A^T)/2$

Plugging in the values we get the following:

$$E[xx'xx'] = \begin{pmatrix} 5 & 0 \\ 0 & 14 \end{pmatrix} \tag{62}$$
$$E[xx' \otimes xx'] = \begin{pmatrix} 3 & 0 & 0 & 2 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 2 & 0 & 0 & 12 \end{pmatrix} \tag{63}$$
$$\tag{64}$$

### B.1 Explicit form of T

The operator $T$ that governs evolution of error covariance has the following representation in matrix form. Let $a = \mathrm{vec}A$, then

$$\mathrm{vec}T(A) = (I - \alpha E[xx'] \otimes I - I \otimes E[xx'] + \alpha^2 E[xx' \otimes xx'])a \tag{65}$$
$$\mathrm{vec}T(A) = \left[ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} - \alpha \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} + \alpha^2 \begin{pmatrix} 3 & 0 & 0 & 2 \\ 0 & 2 & 2 & 0 \\ 0 & 2 & 2 & 0 \\ 2 & 0 & 0 & 12 \end{pmatrix} \right] a \tag{66}$$

### B.2 Asymptotic convergence

For asymptotic convergence, the matrix $M$ corresponding to operator $T$ needs to have norm less than one

$$M_\alpha = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} - \alpha \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} g + \alpha^2 \begin{pmatrix} 3 & 0 & 0 & 2 \\ 0 & 2 & 2 & 0 \\ 0 & 2 & 2 & 0 \\ 2 & 0 & 0 & 12 \end{pmatrix}$$

$$\text{maximize} \quad \alpha \tag{67}$$
$$\text{subject to} \quad \|M_\alpha\| < 1 \tag{68}$$

This can be turned into minimization problem with semidefinite constraint:

$$\begin{aligned} &\text{minimize} \quad R \\ &\text{subject to} \quad \begin{pmatrix} 3 & 0 & 0 & 2 \\ 0 & 2 & 2 & 0 \\ 0 & 2 & 2 & 0 \\ 2 & 0 & 0 & 12 \end{pmatrix} \prec R \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \end{aligned} \tag{69}$$

Maximum learning rate is extracted from the solution to this problem as $\alpha = 1/R_{\min}$[15]

[13] `https://mathematica.stackexchange.com/questions/230036/speeding-up-gaussian-expectations`

[14] `https://mathematica.stackexchange.com/questions/230167/commutation-symmetrizer-and-duplication-matrices`

[15] see Appendix A for details on semidefinite constraint

Convert to regular eigenvalue problem by dividing both sides by the matrix on the right-hand side, to obtain the answer as largest eigenvalue of the following matrix

$$\begin{pmatrix} \frac{3}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{2}{3} & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & \frac{2}{3} & 0 \\ 1 & 0 & 0 & 3 \end{pmatrix}$$

The largest eigenvalue is $R = \frac{1}{4}\left(\sqrt{17}+9\right)$ hence largest allowable learning rate is $\alpha_m = 1/R \approx 0.304806$

So far we have used the method which applies in a generic setting. Since distribution of $x$ is Gaussian, we could've instead used Eq 3 to obtain equivalent result. Form matrix $A$ out of eigenvalues of $\Sigma$

$$A = 2\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 1 & 6 \end{pmatrix}$$

Eigenvalues of this matrix are $\frac{1}{2}\left(\sqrt{17}+9\right), \frac{1}{2}\left(9-\sqrt{17}\right)$

Hence we have

$$\frac{2}{\alpha_m} = \rho(A) = \frac{1}{2}\left(\sqrt{17}+9\right)$$

Note that generic method required solving eigenvalue problem on $d^2 \times d^2$ matrix, whereas Gaussian-specialized method only needs to consider $d \times d$ matrix.

### B.3 Monotonic convergence

To obtain conditions for monotonic convergence, calculate induced trace norm of $T$, which we know is obtained as $\|T(I)\|$[16]

$$T(I) = E[(I - \alpha xx')^2] = I - 2\alpha E[xx'] + \alpha^2 E[xx'xx']$$

Plug-in numeric values of these moments from Eq 62

$$T(I) = I - 2\alpha\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} + \alpha^2\begin{pmatrix} 5 & 0 \\ 0 & 14 \end{pmatrix}$$

The range of $\alpha$ which guarantee monotonic convergence is the range of $\alpha$ for which $\|T_\alpha(I)\| < 1$. The reciprocal of the largest learning rate for which this is true is the solution of the following generalized eigenvalue problem

$$\begin{array}{ll} \text{minimize} & R \\ \text{subject to} & \begin{pmatrix} 5 & 0 \\ 0 & 14 \end{pmatrix} \prec R\begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} \end{array}$$

Divide both sides by $E[xx']$ to convert this into regular eigenvalue problem

$$R_{\text{min}} = \lambda_{\text{max}}\begin{pmatrix} 5/2 & 0 \\ 0 & 7/2 \end{pmatrix}$$

Since this matrix is diagonal, largest eigenvalue is readily identifiable as 7/2, therefore the largest rate which guarantees monotonic convergence is $\frac{2}{7}$

[16] see Appendix D for proof

### B.4 Asymptotic convergence from the moment bound

We can use equation 19 to obtain equivalent rate. The following expectation identities are useful (20.5.2 of [13])

$$\begin{aligned} E[x'A^2x] &= A^2\Sigma && (70)\\ E[(x'Ax)^2] &= (\text{Tr } A\Sigma)^2 + 2\text{Tr } (A\Sigma)^2 && (71)\\ & && (72)\end{aligned}$$

Therefore to apply equation 19 we need to solve the following maximization problem over symmetric matrices $A$ (using cyclic properties of trace for some rearranging)

$$\frac{1}{\alpha} = \max_{A \in S(R^d)} \frac{(\text{Tr}A\Sigma)^2 + 2\text{Tr}A\Sigma A\Sigma}{2\text{Tr}(A\Sigma A)} \tag{73}$$

We can try to optimize this directly, alternatively, convert this to the generalized eigenvalue problem by noting that for symmetric $A$ and $\Sigma$ the following hold. Let $a = \text{vec}A$, then

$$\begin{aligned} \text{Tr}(A\Sigma A) &= a'(\Sigma \otimes I)a = a'(I \otimes \Sigma)a && (74)\\ \text{Tr}A\Sigma A\Sigma &= a'(\Sigma \otimes \Sigma)a && (75)\\ (\text{Tr}A\Sigma)^2 &= a'\text{vec}\Sigma\text{vec}\Sigma' a && (76)\end{aligned}$$

Substitute this into Equation 73 to get the following optimization over $d \times d$ vectors $a$ [17].

$$\frac{1}{\alpha_{\text{max}}} = \max_a \frac{a'(2\Sigma \otimes \Sigma + \text{vec}\Sigma\text{vec}\Sigma')a}{2a'(\Sigma \otimes I)a}$$

This is readily recognizable as Raleigh quotient of generalized eigenvalue problem $(A, B)$ with

$$\begin{aligned} A &= 2\Sigma \otimes \Sigma + \text{vec}\Sigma\text{vec}\Sigma' && (77)\\ B &= 2\Sigma \otimes I && (78)\end{aligned}$$

. We can divide by $B$ to obtain an equivalent eigenvalue problem

$$\frac{1}{\alpha_{\text{max}}} = \lambda_{\text{max}} \left[ \begin{pmatrix} 3 & 0 & 0 & 2 \\ 0 & 2 & 2 & 0 \\ 0 & 2 & 2 & 0 \\ 2 & 0 & 0 & 12 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}^{-1} \right] = \lambda_{\text{max}} \begin{pmatrix} \frac{3}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} & 0 \\ 0 & 1 & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 3 \end{pmatrix}$$

Find the eigenvalues as roots of characteristic polynomial for this matrix

$$x^4 - 6x^3 + \frac{43x^2}{4} - 6x$$

Largest root is $\frac{1}{4}\left(\sqrt{17} + 9\right)$, hence $\alpha_{\text{max}} \approx 0.304806$

Note that this produced a slightly different eigenvalue problem as in previous section, but the solution was the same.

[17] Normally we would enforce symmetry constraint by using $d(d+1)/2$ sized vector $a$ and a duplication matrix, but in this case it doesn't affect the result

# C Properties of operator T

## C.1 Completely Positive Map

For distribution with finite support, we can write $T$ in the form below[18]

$$T(A) = E_x[B_x A B_x^*] = \sum_x V_x A V_x^*$$

This is the representation used to establish complete positivity in proof of Choi's Theorem [19], meaning $T$ is a completely positive map.

This means it is a positive map, making Russo-Dye theorem applicable. Also it means that $T$ behaves like expectation and various inequalities that hold for $E$ also hold for $T$, see Section 3.3.3 of [2]

A useful representation for such operator is the "operator sum representation" or "Kraus decomposition" of the following form.

$$T(A) = \sum_i V_i A V_i' \tag{79}$$

For instance, if we apply gradient descent with learning rate 1 on our toy problem of (1,2) Gaussian from Appendix B, $T(A)$ can be written in this representation with the following values for $V_i$

$$V_1 = \begin{pmatrix} 0 & 0 \\ 0 & 3 \end{pmatrix}, V_2 = \begin{pmatrix} 0 & \sqrt{2} \\ \sqrt{2} & 0 \end{pmatrix}, V_3 = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix}$$

This is more compact that the eigenvector representation of $T$ which requires 4 vectors. Generally, for Gaussian SGD in $d$ dimensions, the SGD operator needs d(d+1)/2 components in operator sum representation, as opposed to $d^2$ required by rank-1 decomposition. Also, this representation makes convergence of SGD to a fixed distribution evident – if we normalize $T$ to be trace preserving, each entry of the sum in OSR representation in Eq 79 will be a valid covariance matrix, hence we can view operator $T$ as choosing one of the covariance matrices with some probability which makes it a kind of mixing process.

Matrices $V_i$ come up in characterizing properties of the operator, for instance, formula (9.135) of [12] gives a way to characterize information loss incurred by applying $T$ in terms of $\{V_i\}$

## C.2 Self-adjoint

Recall that

$$\operatorname{vec} T(A) = M \operatorname{vec} A$$

Consider form of matrix $M$

$$M = E[I \otimes I - \alpha(xx' \otimes I) - \alpha(I \otimes xx') + \alpha^2 xx' \otimes xx'] \tag{80}$$

Every matrix inside expectation operator is symmetric, therefore $M$ (and hence $T$) is symmetric.

## C.3 Positive definite

We know that operator T with corresponding matrix $M$ is contractive iff the following matrix is positive definite (Proposition 1.3.1 of [2])

$$\begin{pmatrix} I & M \\ M & I \end{pmatrix}$$

This is equivalent to the following statement (Shur component, Theorem 2.3.9 of [17])

[18] for proof in infinite case, use facts from IV.3 of [15]

[19] Theorem 2.22 part 4 of [16], `https://cs.uwaterloo.ca/~watrous/TQI/TQI.pdf`

$$M^2 \prec I$$

Which translates to

$$-I \prec M \prec I$$

Then apply Theorem 2.3.9 of [17], to translate M into form e) in Shur's complement, which is equivalent to form b), which is true because result is a covariance matrix.

#### C.3.1 Alternative proof for $T(I) \succ 0$

Consider the following matrix with $X^4 = E[xx'xx'], X^2 = E[xx']$

$$\begin{pmatrix} I & \alpha X^2 \\ \alpha X^2 & \alpha^2 X^4 \end{pmatrix}$$

By Shur's complement theorem[20], the following two conditions are equivalent:

$$I - 2\alpha X^2 + \alpha^2 X^4 \succ 0$$
$$X^4 \succ X^2 X^2$$

The first condition is equivalent to $T(I) \succ 0$. The latter is true because $X^4 - X^2X^2$ is a covariance matrix of $Y = X \otimes X$.

## D Alternative derivation of maximal rate

- Russo-Dye Theorem gives induced operator norm of $T$: $\|T\| = \|T(I)\|$[21]
- We are instead interested in the induced trace norm $\|T\|_1$ [22]
- Because trace norm and operator norm are dual: $\|T\|_1 = \|T^*\|$ [23]
- Because $T$ is self-adjoint: $T^* = T$[24]
- Hence, $\|T\|_1 = \|T^*\| = \|T\| = \|T(I)\|$

.

For duality of trace and spectral norm, see A.1.6 of [4]. The fact $\|T\|_1 = \|T^*\|$ is also stated without proof as Eq 2.37 in [2]

## E Expected smoothness for linear regression

[7] provides a bound on step size in terms of "expected smoothness". Suppose we are in noise-free realizable regime, in other words we achive zero loss, or $f_i(w^*) = 0$ for all $i$. The following step size is sufficient to guarantee convergence

$$\alpha < \alpha_s = \frac{2}{L} \tag{81}$$

Where $L$ is the expected smoothness constant, defined as follows[25]

$$E\|\nabla f(w)\|^2 \leq \frac{L}{2} E f(w) \tag{82}$$

[20] Theorem 2.3.9 of `https://www.convexoptimization.com/TOOLS/Handbook.pdf`

[21] Russo-Dye theorem for operator norm, Theorem 2.6.4 of [2]

[22] definition and properties, see Section 3.3.2 of [16]

[23] property of induced norms, Theorem 5.6.35 of [8]

[24] Appendix C.2

[25] using $L_{\text{mine}} = 4L_{\text{gowers}}$ for consistency with rest of paper

Let us setup standard linear regression problem and compute its expected smoothness. Let $g(x) = w'x$ be our predictor for observation $x \in \mathbb{R}^d$. We try to minimize the difference between predicted label $y = g(x)$, $y \in \mathbb{R}$ and true label $\hat{y}$ by minimizing the following per-example loss

$$\hat{f}(y) = \frac{1}{2}h(y - \hat{y})^2$$

Here, $h$ is a parameter which is useful for debugging formulas, but can be assumed to be equal to 1.

Our function $f(w)$ represents loss on a single example $x$ evaluated at parameter value $w$, $f = \hat{f} \circ g \circ w$

$$f(w) = \frac{1}{2}h(y - \hat{y})^2 = \frac{1}{2}h(w^T x - w_*^T x)^2 = \frac{1}{2}h(\eta' x)^2$$

We have defined $\eta = w - w_*$, the difference between current estimate $w$ and target parameter $w_*$

Taking expectation over $x$ we compute $Ef(w)$ [26]

$$Ef(w) = \frac{1}{2}h\eta' E[xx']\eta$$

For the left-hand side, note the following

$$\nabla f(w) = h\eta' xx'$$

Hence

$$E[\|\nabla f(w)\|^2] = h^2\eta' E[xx'xx']\eta$$

Plugging these into Eq 82 we get

$$4h\eta' E[xx'xx']\eta \leq L\eta' E[xx']\eta$$

For standard least-squares loss, $h = 1$. Suppose $E[xx']$ is full-rank. Then the optimal value of $L$ is determined as the largest eigenvalue $\rho(A)$ of the following matrix (see Section A)

$$A = 4E[xx'xx'](E[xx'])^{-1}$$

Note that this is identical to Eq 27, except for the extra factor of 4. Therefore, required step size from this analysis is exactly 1/4th of the largest rate which guarantees monotonic convergence rate. In other words

$$\alpha_s = \frac{1}{4}\alpha_m$$

# F Asymptotic distribution of SGD

When inputs are Gaussian, rotational symmetry of SGD allows us to assume that it is diagonal. If we then take diagonal starting covariance matrix $A_0$, SGD updates will keep the result diagonal, so we can reformulate $T(A)$ in terms of diagonal entries of $A$ and $\Sigma$. In other words we can formulate action of SGD operator $T$ by using diagonal extraction operation "diag".

$$\text{diag}T(A) = \widehat{M}\text{diag}A$$

Define the following

$$\begin{aligned} A &= \text{diag}(a) \\ \Sigma &= \text{diag}(h) \end{aligned} \tag{83}$$

[26] Use $\text{var}(Ax) = A\text{var}xA'$ from [13] 20.5 and 20.6

Now a single step of SGD updates $a_i$ as follows

$$a_i \xleftarrow{T} a_i(1 - 2\alpha h_i + 2\alpha^2 h_i^2 + \alpha^2 h_i \langle a, h\rangle)$$

We can write this update in matrix form $a \leftarrow Ma$ where

$$\widehat{M_\alpha} = I - 2\alpha\Sigma + 2\alpha^2\Sigma^2 + \alpha^2 hh'$$

For our toy example with $\text{diag}(1, 2)$ Gaussian, this matrix has the following form

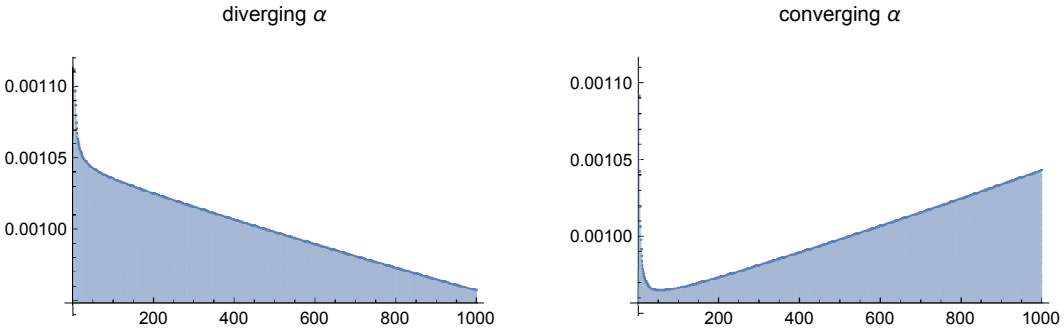
$$\widehat{M_\alpha} = I - 2\alpha \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} + 2\alpha^2 \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} + \alpha^2 \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

This gives an alternative expression for $\alpha_m$ and $\alpha_a$ in terms of spectral and max-row-sum norm of $\widehat{M_\alpha}$

$$\begin{aligned} \alpha_a &= \text{Solve}[\alpha \text{ such that } \|\widehat{M_\alpha}\| = 1]] \\ \alpha_m &= \text{Solve}[\alpha \text{ such that } \|\widehat{M_\alpha}\|_\infty = 1] \end{aligned} \tag{84}$$

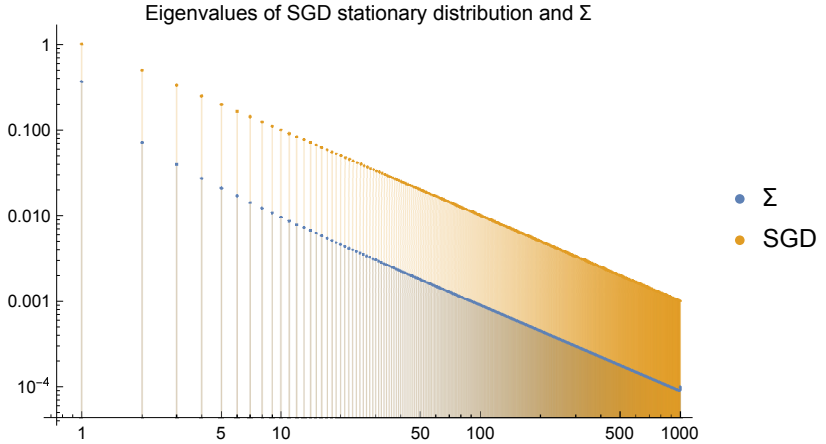
Additionally, this matrix allows us to recover asymptotic distribution of SGD from the top eigenvector of $\widehat{M}$

If we $x$ distributed as 1000 dimensional Gaussian with geometrically decaying eigenvalues, we can look at stationary distribution of SGD errors (normalized by expected magnitude) for learning rates slightly larger and slightly smaller than asymptotic convergence threshold:




You can see that around the value of critical learning rate, asymptotic distribution of SGD flips from being aligned with $\Sigma$ to being aligned with $\Sigma^{-1}$.

We can also examine the decay in eigenvalues of SGD stationary distribution for a convergent learning rate that maximizes decrease in distance to $w_*$ after one SGD step. In the case of $\Sigma$ eigenvalues decaying as $1/k$, SGD stationary distribution exhibits the same $1/k$ behavior throughout the spectrum.



## References


[1] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). June 2013.

[2] Rajendra Bhatia. *Positive Definite Matrices (Princeton Series in Applied Mathematics (24))*.

[3] Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear Matrix Inequalities in System and Control Theory (Studies in Applied and Numerical Mathematics)*.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[5] Alexandre Défossez and Francis Bach. Averaged Least-Mean-Squares: Bias-Variance Trade-offs and Optimal Sampling Distributions. Technical report, 2015.

[6] Benyamin Ghojogh, Fakhri Karray, and Mark Crowley. Eigenvalue and generalized eigenvalue problems: Tutorial. March 2019.

[7] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtarik. SGD: General analysis and improved rates. January 2019.

[8] Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, February 2013.

[9] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. October 2016.

[10] A S Lewis. The mathematics of eigenvalue optimization. *Math. Program.*, 97(1):155–176, July 2003.

[11] Anna Ma, Deanna Needell, and Aaditya Ramdas. Convergence properties of the randomized extended Gauss-Seidel and kaczmarz methods. March 2015.

[12] Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 10 edition, 31 January 2011.

[13] George A F Seber. *A Matrix Handbook for Statisticians*.

[14] D T M Slock. On the convergence behavior of the LMS and the normalized LMS algorithms. *IEEE Trans. Signal Process.*, 41(9):2811–2825, September 1993.

[15] M Takesaki. *Theory of Operator Algebras I*. Springer Science & Business Media, 20 November 2001.

[16] John Watrous. *The Theory of Quantum Information*.

[17] Henry Wolkowicz, Romesh Saigal, and Lieven Vandenberghe, editors. *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications (International Series in Operations Research & Management Science)*.