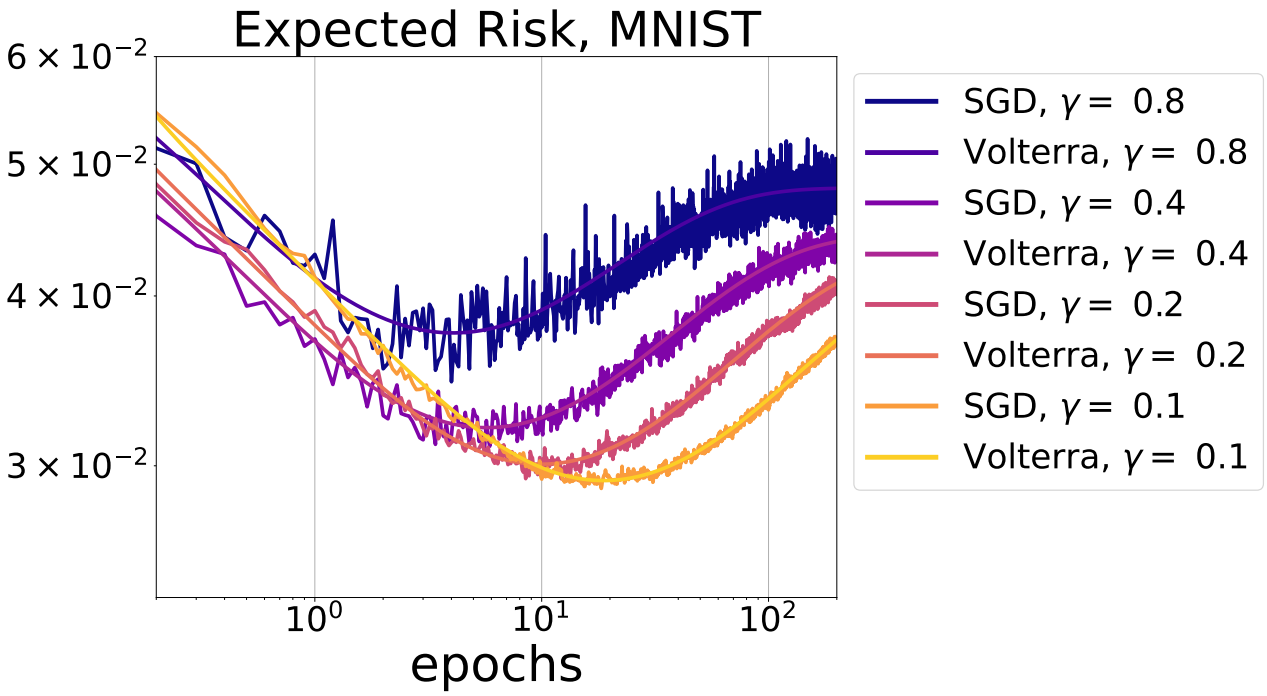
High-dimensional limits of stochastic gradient descent

Elliot Paquette

Version: July 19, 2023



Course Content. Stochastic gradient descent, random matrix theory

Summary of Contents

Foreword and Acknowledgements

These notes were developed for the "Stochastic methods and computation" summer school, organized by Si Tang at Lehigh University in July 2023. These notes develop the probabilistic analysis stochastic gradient descent on idealized high-dimensional objective functions. Moreover, the goal of this analysis is to reveal properties of the algorithm itself, in how it responds to different choices of hyperparameters and how different problem-geometries interact with those choices. Furthermore, the mathematical analysis is performed by approximating problems in the large-dimensional limit, and in showing that the resulting problem simplifies.

Some of the work presented here is my own, together with the input of many extraordinary coauthors. I would especially like to acknowledge Courtney Paquette, to whom I am indebted for everything I've learned about optimization theory. She is furthermore responsible for much of the technical and mathematical developments displayed here. Many figures were created with her. ♡

Besides Courtney, I am indebted to my colleagues and students:

1. Kiwon Lee and Fabian Pedregosa, who were instrumental in the first version construction of the Volterra model [Paq+21].
2. Ben Adlam and Jeffrey Pennington, who guided the development of this work to have more machine learning implications [Paq+22a] and [Paq+22b].
3. Elizabeth Collins-Woodfin who developed the analysis for streaming models [CP23a], and Elizabeth and Inbar Seroussi, whose work on generalized linear models has pushed this past the least squares context to GLMs (forhcoming at time of writing).

Besides those who have worked directly on projects that are presented here, Andrew Cheng deserves mention for his work on proportional batch size SGD limits.

I'd like to add special thanks to Si Tang and ByeongHo Bahn for feedback and corrections.
Elliot Paquette

July 4, 2023

1 Background

In this section, we collect various probability and linear algebra background which will be helpful for working with all of the theory here.

1.1 Tensors and calculus

We suppose that $\mathcal{V}_j$ for $j = 1, 2, 3$ are some finite-dimensional Hilbert spaces. Recall that as a vector space $\mathcal{V}_1 \otimes \mathcal{V}_2$ is all (finite) linear combinations of *simple* tensors, i.e., those of the form $a \otimes b$ where $a \in \mathcal{V}_1$ and $b \in \mathcal{V}_2$. This becomes an algebra, allowing scalars to commute, i.e., for $c \in \mathbb{R}$

$$c(a \otimes b) = (ca) \otimes b = a \otimes (cb),$$

and by allowing $\otimes$ to distribute over addition,

$$(a + b) \otimes c = (a \otimes c) + (b \otimes c) \quad \text{and} \quad a \otimes (b + c) = (a \otimes b) + (a \otimes c). \tag{1}$$

In what follows, we will need to contract along various tensors. To facilitate this, we introduce a generalization of the inner product. Each $\mathcal{V}_1$ and $\mathcal{V}_2$ carries with it an inner product which we denote by $\langle \cdot, \cdot \rangle_{\mathcal{V}_1}$ and $\langle \cdot, \cdot \rangle_{\mathcal{V}_2}$ respectively. This induces a natural inner produce on $\mathcal{V}_1 \otimes \mathcal{V}_2$, which for simple tensors is defined by

$$\langle a \otimes b, c \otimes d \rangle_{\mathcal{V}_1 \otimes \mathcal{V}_2} = \langle a, c \rangle_{\mathcal{V}_1} \langle b, d \rangle_{\mathcal{V}_2}. \tag{2}$$

This is extended to the full space $\mathcal{V}_1 \otimes \mathcal{V}_2$ by bilinearity.

This, for example, can be connected to the Frobenius inner product. If we represent an element $A \in \mathcal{V}_1 \otimes \mathcal{V}_2$ in the orthonormal basis $\{e_i \otimes f_j\}$ as

$$A = \sum_{i,j} A_{ij} e_i \otimes f_j, \tag{3}$$

then we have the identification

$$\langle A, B \rangle_{\mathcal{V}_1 \otimes \mathcal{V}_2} = \sum_{i,j} A_{ij} B_{ij} = \mathrm{Tr}(AB^T).$$

This gives a convenient representation for quadratic forms as well:

$$\langle A, x \otimes x \rangle_{\mathbb{R}^d \otimes \mathbb{R}^d} = \mathrm{Tr}(Axx^T) = x^T A x. \tag{4}$$

Higher tensor powers. For higher tensor powers, the dot products written above extend naturally to

$$\mathcal{V}_1 \otimes \mathcal{V}_2 \otimes \mathcal{V}_3. \tag{5}$$

Namely for $a_i, b_i \in \mathcal{V}_i$ for $i = 1, 2, 3$,

$$\langle a_1 \otimes a_2 \otimes a_3, b_1 \otimes b_2 \otimes b_3 \rangle_{\mathcal{V}_1 \otimes \mathcal{V}_2 \otimes \mathcal{V}_3} = \langle a_1, b_1 \rangle \langle a_2, b_2 \rangle \langle a_3, b_3 \rangle.$$

This is once more extended by multi-linearity, and we further extend it to higher tensor powers.

Partial contractions. For tensors it is also helpful to consider contractions over partial directions. Once more, for simple tensors, $t_i = (a_i \otimes b_i) \in \mathcal{V}_1 \otimes \mathcal{V}_2$ for $i = 1, 2$,

$$\langle t_1, t_2 \rangle_{\mathcal{V}_1} := \langle a_1, a_2 \rangle_{\mathcal{V}_1} (b_1 \otimes b_2) \in \mathcal{V}_2^{\otimes 2}. \tag{6}$$

This is also extended as a bilinear map $(\mathcal{V}_1 \otimes \mathcal{V}_2)^{\otimes 2} \to \mathcal{V}_2^{\otimes 2}$. This extends to higher tensor powers analogously, and also to the more general situation of products of $\mathcal{V}_1 \otimes \mathcal{V}_2$ with $\mathcal{V}_1 \otimes \mathcal{V}_3$ as a bilinear mapping:

$$\langle \cdot, \cdot \rangle_{\mathcal{V}_1} : (\mathcal{V}_1 \otimes \mathcal{V}_2) \otimes (\mathcal{V}_1 \otimes \mathcal{V}_3) \to \mathcal{V}_2 \otimes \mathcal{V}_3 \tag{7}$$

by the formula for simple tensors in (6). This includes the case where one of $\mathcal{V}_2$ or $\mathcal{V}_3$ may be 1-dimensional.

We shall reserve the notation $\langle \cdot, \cdot \rangle$ for the complete contraction between two tensors, in whichever space they reside, and we shall add the subscript whenever a partial contraction is needed. We note that having done the partial contraction, it may be helpful to complete the contraction to a full contraction. This is performed by the *trace* operation, which on the Hilbert space $\mathcal{V} \otimes \mathcal{V}$, is defined for simple tensors by

$$\mathrm{Tr}(v \otimes w) = \langle v, w \rangle_{\mathcal{V}}, \tag{8}$$

and which extends to all $\mathcal{V} \otimes \mathcal{V}$ by linearity. In the context of (6), we can then write

$$\mathrm{Tr}(\langle t_1, t_2 \rangle_{\mathcal{V}_1}) = \langle a_1, a_2 \rangle_{\mathcal{V}_1} \langle b_1, b_2 \rangle_{\mathcal{V}_2} = \langle t_1, t_2 \rangle,$$

which by linearity therefore identifies $\mathrm{Tr}(\langle \cdot, \cdot \rangle_{\mathcal{V}_1})$ as the full contraction.

Norms. Recall that for a matrix A, there are three traditional matrix norms, beginning with the Frobenius (or Hilbert-Schmidt) norm $\| \cdot \|$, operator norm $\| \cdot \|_\sigma$ and trace norm $\| \cdot \|_*$

$$\|A\| = \sqrt{\mathrm{Tr}(A^T A)}, \quad \|A\|_\sigma = \sup_{x,y \neq 0} \frac{(x^t A y)}{\|x\| \|y\|}, \quad \|A\|_* = \sup_{\|B\|_\sigma = 1} \mathrm{Tr}(B^T A).$$

These generalize to 2-tensors and higher tensors in an analgous fashion. For 2-tensors $A \in \mathcal{V}_1 \otimes \mathcal{V}_2$, the induced norm on the Hilbert space generalizes the Hilbert-Schmidt norm, through

$$\|A\|^2 = \langle A, A \rangle_{\mathcal{V}_1 \otimes \mathcal{V}_2}.$$

More generally, for higher tensor products, the induced Hilbert space is the natural generalization. Note that by Cauchy-Schwarz this also admits a variational representation

$$\|A\| = \sup_{B, \|B\| = 1} \langle A, B \rangle.$$

As for the operator norm, we take the above definition which defined the operator norm as supremum over simple unit tensors. We will call these the σ-norm and denote it by $\|\cdot\|_\sigma$. This norm is also commonly known as the *injective tensor norm*. Explicitly, if $\varphi \in \mathcal{V}_1 \otimes \mathcal{V}_2 \otimes \ldots \otimes \mathcal{V}_k$, then we define its σ-norm by

$$\|A\|_\sigma := \sup_{\substack{\|y_i\|_{\mathcal{V}_i}=1 \\ i=1,2,\ldots,k}} \langle A, y_1 \otimes y_2 \otimes \ldots \otimes y_k \rangle,$$

where $y_1 \otimes y_2 \otimes \ldots \otimes y_k \in \mathcal{V}_1 \otimes \mathcal{V}_2 \otimes \ldots \otimes \mathcal{V}_k$ is a simple tensor. Note the norm

$$\|y_1 \otimes y_2 \otimes \ldots \otimes y_k\|^2 = \langle y_1, y_1 \rangle \langle y_2, y_2 \rangle \cdots \langle y_k, y_k \rangle = 1,$$

and hence we have by the variational representation $\|A\|_\sigma \leq \|A\|$.

Finally for the nuclear norm, we just generalize it as the dual norm of the injective norm, setting

$$\|A\|_* := \sup_{B, \|B\|_\sigma = 1} \langle A, B \rangle.$$

Using the variational representations we observe

$$\|A\|_\sigma \leq \|A\| \leq \|A\|_*. \tag{9}$$

Calculus for tensors. The functions given above are compositions of smooth functions f with linear functions, and we would like to perform many Taylor approximations of these functions. We recall briefly how differential calculus works here and connect it with the tensor notation above.

For a (smooth) function $f : \mathcal{V}_1 \to \mathcal{V}_2$ on (finite dimensional) Hilbert spaces $\mathcal{V}_1, \mathcal{V}_2$, its (Fréchet) derivative $\mathrm{D}f$ can be identified as a mapping from $\mathcal{V}_1 \to \mathcal{L}(\mathcal{V}_1, \mathcal{V}_2)$, the space of linear operators from $\mathcal{V}_1 \to \mathcal{V}_2$ so that for all $x, h \in \mathcal{V}_1$

$$\lim_{t \downarrow 0} \frac{f(x+th) - f(x)}{t} = (\mathrm{D}f)(x)[h].$$

The space $\mathcal{L}(\mathcal{V}_1, \mathcal{V}_2)$ can be represented as elements of the tensor product $\mathcal{V}_2 \otimes \mathcal{V}_1$, by picking an orthonormal basis $\{e_j\}$ for $\mathcal{V}_1$ and then identifying

$$(\mathrm{D}f)(x) \leftrightarrow \sum_j (\mathrm{D}f)(x)[e_j] \otimes e_j,$$

which is (in effect) its Jacobian matrix representation. This procedure can now be iterated, as $\mathrm{D}f$ is a mapping between $\mathcal{V}_1$ and a new vector space $\mathcal{L}(\mathcal{V}_1, \mathcal{V}_2) \cong \mathcal{V}_2 \otimes \mathcal{V}_1$, and hence

$$\mathrm{D}^2 f : \mathcal{V}_1 \to \mathcal{L}(\mathcal{V}_1, \mathcal{L}(\mathcal{V}_1, \mathcal{V}_2)) \cong \mathcal{V}_2 \otimes \mathcal{V}_1 \otimes \mathcal{V}_1.$$

In the case that the output of f is 1-dimensional (so that $\mathcal{V}_2 \cong \mathbb{R}$) we may furthermore identify the second derivative $(\mathrm{D}^2 f)(x)$ with an element of $\mathcal{V}_1 \otimes \mathcal{V}_1$. A parallel approach identifies the third derivative as

$$\mathrm{D}^3 f : \mathcal{V}_1 \to \mathcal{L}(\mathcal{V}_1, \mathcal{L}(\mathcal{V}_1, \mathcal{L}(\mathcal{V}_1, \mathcal{V}_2))) \cong \mathcal{V}_2 \otimes \mathcal{V}_1^{\otimes 3}.$$

In this way, we have that

$$\mathrm{D}^k f : \mathcal{V}_1 \to \mathcal{V}_2 \otimes \mathcal{V}_1^{\otimes k}.$$

Similarly, when $\mathcal{V}_2 \cong \mathbb{R}$, we can identify $\mathcal{V}_2 \otimes \mathcal{V}_1^{\otimes k} \cong \mathcal{V}_1^{\otimes k}$.

Chain rule with tensors. The class of statistics (and losses) we consider are compositions of smooth maps. In this section, we show how one can use the tensor notation to simplify the chain rule for higher order derivatives. Supposing one has two smooth maps f, g with $f : \mathcal{V}_1 \to \mathcal{V}_2$ and $g : \mathcal{V}_2 \to \mathcal{V}_3$, the chain rule states that $g \circ f$ is a smooth map from $\mathcal{V}_1 \to \mathcal{V}_3$ and its derivative is a map from $\mathcal{V}_1$ to $\mathcal{L}(\mathcal{V}_1, \mathcal{V}_3)$. Moreover it's derivative is given by

$$\mathrm{D}(g \circ f)(x)[h] = (\mathrm{D}g)(f(x))[(\mathrm{D}f)(x)[h]].$$

If we represent these as tensors, then $(\mathrm{D}g)(f(x))$ is in $\mathcal{V}_3 \otimes \mathcal{V}_2$ and $(\mathrm{D}f)(x)$ is in $\mathcal{V}_2 \otimes \mathcal{V}_1$, and hence we can as well represent the chain rule by

$$\mathrm{D}(g \circ f)(x) = \langle (\mathrm{D}g)(f(x)), (\mathrm{D}f)(x) \rangle_{\mathcal{V}_2} \in \mathcal{V}_3 \otimes \mathcal{V}_1, \tag{10}$$

showing along which axis the contraction is taken. We note the ordering is important here. The input space is always taken to be on the right.

Applying this in the case of a directional derivative, suppose we take a smooth function $\varphi : \mathcal{V} \to \mathbb{R}$. Then for any fixed $x, \Delta \in \mathcal{V}$, the map $\psi : t \mapsto \varphi(x + t\Delta)$ is a smooth function of $\mathbb{R}$, and we may compute its Taylor approximation. In particular, we are interested in approximating $\varphi(x + \Delta)$ or equivalently $\psi(1)$. If we approximate $\varphi(x + \Delta)$ by the third order Taylor expansion at x with remainder, we have

$$\varphi(x + \Delta) = \psi(1) = \psi(0) + \psi'(0) + \tfrac{1}{2}\psi''(0) + \frac{1}{2}\int_0^1 (1 - t)^2 \psi^{(3)}(t)\, \mathrm{d}t.$$

Applying the chain rule, if we set $x(t) = x + t\Delta$, then $(\mathrm{D}x)(t)$ is constant and equal to Δ. Therefore, we deduce that

$$\begin{aligned}
\psi'(0) &= \langle (\mathrm{D}\varphi)(x), \Delta \rangle, \\
\psi''(0) &= \langle (\mathrm{D}^2\varphi)(x), \Delta^{\otimes 2} \rangle, \\
\psi^{(3)}(t) &= \langle (\mathrm{D}^3\varphi)(x(t)), \Delta^{\otimes 3} \rangle.
\end{aligned}$$

To derive this, in particular, the 2nd and 3rd derivatives, we used linearity to conclude

$$\begin{aligned}\psi''(t) = \mathrm{D}(\langle(\mathrm{D}\varphi)(x(t)), \Delta\rangle_{\mathcal{V}}) &= \langle \mathrm{D}((\mathrm{D}\varphi)(x(t))), \Delta\rangle_{\mathcal{V}} \\ &= \langle\langle(D^2\varphi)(x(t)), \Delta\rangle_{\mathcal{V}}, \Delta\rangle_{\mathcal{V}} \\ &= \langle(D^2\varphi)(x(t)), \Delta^{\otimes 2}\rangle_{\mathcal{V}\otimes\mathcal{V}}.\end{aligned}$$

We note that in the second line, there is in principle an ambiguity $\langle(D^2\varphi)(x(t)), \Delta\rangle_{\mathcal{V}}$, in that $(D^2\varphi)(x(t))$ is an element of $\mathcal{V}\otimes\mathcal{V}$. However, as the second derivative is symmetric (as φ is smooth and so mixed partials can be interchanged), contraction along either axis works. We summarize with the following generic directional derivative expansion for scalar C^3-smooth functions $\varphi : \mathcal{V} \to \mathbb{R}$

$$\begin{aligned}\varphi(x+\Delta) = \varphi(x) &+ \langle(\mathrm{D}\varphi)(x), \Delta\rangle \\ &+ \frac{1}{2}\langle(\mathrm{D}^2\varphi)(x), \Delta^{\otimes 2}\rangle \\ &+ \frac{1}{2}\int_0^1 (1-t)^2\langle(\mathrm{D}^3\varphi)(x+t\Delta), \Delta^{\otimes 3}\rangle \, \mathrm{d}t.\end{aligned} \tag{11}$$

1.2 *Resolvents*

Resolvents are a powerful tool for the manipulation of high-dimensional matrices and for doing random matrix theory.

Definition 1 (Resolvent): For a matrix $A \in \mathbb{M}(n,n)$, its resolvent $R(z;A)$ is the matrix valued function $z \mapsto (A - z\,\mathrm{Id}_n)^{-1}$, defined on the subset of the complex plane where $\mathbb{C} \setminus \mathrm{Spec}(A)$. We will usually abbreviate this by writing $(A-z)^{-1}$.

We use $\mathrm{Spec}(A)$ to denote the set of eigenvalues of A and Id_n to denote the $n \times n$ identity matrix.

The resolvent is a well-behaved complex function, in the following sense:

Definition 2 (Meromorphic): A function $f : \mathbb{C} \to \mathbb{C} \cup \{\infty\}$ is meromorphic if it is analytic except at isolated points where (at λ) it has a pole, i.e., it diverges no faster than $|z-\lambda|^{-k}$ as $z \to \lambda$ for some $k \in \mathbb{N}$.

Example 1: Rational functions

$\{p(z)/q(z) \in \mathbb{C}(z)\}$ for polynomials p and q are meromorphic functions.

This extends to matrices by asking that each entry has this property:

Definition 3 (Matrix meromorphic functions): A matrix valued function is meromorphic if every entry is meromorphic.

Theorem 1: Resolvents are meromorphic

Let $A \in \mathbb{M}(n,n)$. Then $R(z;A)$ is a meromorphic function, and its poles are precisely $\mathrm{Spec}(A)$.

Proof. Let $A = SJS^{-1}$ be a Jordan decomposition of A, so that J has a block diagonal representation as $J = \mathrm{diag}(J_1, J_2, \cdots)$, ($J_i$'s are Jordan blocks). Now $(J - z)$ is again block diagonal, and observing $(A-z)^{-1} = S(z-J)^{-1}S^{-1}$

$$R(z;A) = S \left(\begin{array}{c|c|c} R(z;J_1) & 0 & \cdots \\ \hline 0 & R(z;J_2) & \cdots \\ \hline 0 & 0 & \ddots \end{array}\right) S^{-1}.$$

Thus it suffices to evaluate the resolvent of a single Jordan block and to show it has a pole precisely at the eigenvalue of the block. Suppose J is a Jordan block

$$J = \begin{pmatrix} \lambda & 1 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda \end{pmatrix}.$$

Then by an explicit computation, we can verify

$$(J_\lambda - z)^{-1} = \begin{pmatrix} y & y^2 & \cdots & y^{n-1} \\ 0 & y & \cdots & y^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y \end{pmatrix}, \text{where } y = (\lambda - z)^{-1}.$$

□

Corollary 1 (Diagonalizble case): A matrix $A \in \mathbb{M}(n,n)$ is diagonalizable (i.e., comprised of all size 1 Jordan blocks) if and only if $R(z;A)$ only has simple poles (the largest inverse power of $\lambda - z$ that appears is 1). Moreover, if we let $\{\lambda_j\}$ be the eigenvalues of A and $\{(u_j, v_j)\}$ be corresponding left and right eigenvectors normalized so that $\langle u_j, v_j \rangle = 1$

$$R(z;A) = \sum_{j=1}^{n} \frac{v_j u_j^T}{\lambda_j - z}. \tag{12}$$

If we furthermore have that A is symmetric, we have the following elementary estimate:

Corollary 2 (Resolvent-norm): If $A \in \mathbb{M}(n,n)$ is symmetric, then it terms of orthonormal eigenvectors u_j and eigenvalues λ_j

$$R(z;A) = \sum_{j=1}^{n} \frac{u_j u_j^T}{\lambda_j - z}. \tag{13}$$

Moreover, we have the operator norm estimate

$$\|R(z;A)\|_\sigma \leq \frac{1}{d(z,\mathrm{Spec}(A))} \leq \frac{1}{|\Im z|}.$$

Proof. Equation (13) is (12) in the case that A is unitarily diagonalizable, and so has $v_j = u_j$. In the $\{u_j\}$-basis (which is an orthonormal change of basis), the resolvent is therefore diagonal, and so its spectral norm is given by its largest entry in modulus. As all these eigenvalues are real, this gives the second estimate. □

1.3 *Perturbation Formulas*

Theorem 2: Perturbation formulas

For $A, B \in \mathbb{M}(n,n)$ and $y, z \in \mathbb{C}$, we have

1. $R(z;A) - R(y;A) = (z-y)R(z;A)R(y;A)$,
2. $R(z;A) - R(z;B) = R(z;A)(B-A)R(z;B)$.

Proof. It suffices to establish the equations at points z where both A and B are invertible. Then for the first equality, multiply $(A-z)$ and $(A-y)$ on left and right, respectively, on both sides. □

Using the theorem above, if $A = B + E$ for E sufficiently small,

$$\begin{aligned} R(z;B+E) &= R(z;B) - R(z;B+E)ER(z;B) \\ &= R(z;B) - R(z;B)ER(z;B) + R(z;B)ER(z;B)ER(z;B) + \cdots. \end{aligned}$$

This can stop at a finite point, or if the spectral radius of $ER(z;B) < 1$, we can develop it as a convergent series. Similarly, for z sufficiently close to y,

$$R(z;A) = R(y;A) + (z-y)R(y;A)^2 + (z-y)^2 R(y;A)^3 + \cdots.$$

As a corollary, we have all derivatives in the resolvent:

Corollary 3 (Resolvent Derivatives): The derivatives of the resol-

vent are given by, for any $k \in \mathbb{N}$ and at all $z \in \mathbb{C} \setminus \mathrm{Spec}(A)$

$$\frac{\mathrm{d}^k R(z;A)}{(\mathrm{d}z)^k} = k! R(z;A)^{k+1}.$$

In the special case that A and B differ by a low-rank matrix, there is another formula which can be more fruitful:

Corollary 4 (Woodbury identity): For $U, V \in \mathbb{M}(n,k)$ and $C \in \mathbb{M}(k,k)$ which is invertible

$$R(z;A+UCV^T) = R(z;A) - R(z;A)U(C^{-1}+UR(z;A)V^T)^{-1}V^T R(z;A),$$

for all z for which $(C^{-1}+UR(z;A)V^T)^{-1}$ exists. In particular, when $k = 1$ and without loss of generality when $C = 1$, we have

$$R(z;A+UV^T) - R(z;A) = \frac{R(z;A)UV^T R(z;A)}{1+UR(z;A)V^T}.$$

1.4 *Spectral mapping*

Theorem 3: The Residue Formula

If $U \subset \mathbb{C}$ is a connected, simply connected open set, $f : U \to \mathbb{C}$ is meromorphic, and γ is a smooth chain in U disjoint from the poles of f, we have

$$\frac{1}{2\pi\imath}\oint_\gamma f(z)dz = \sum_{poles\ \lambda\in U} \mathrm{Res}(f;\lambda)\,\mathrm{Ind}(\gamma;\lambda), \tag{14}$$

where

- $\mathrm{Res}(f;z) = r_{-1}$ if $f(z) = \sum_{k\in\mathbb{Z}} r_k(z-\lambda)^k$ is a series converging in a sufficiently small neighborhood of λ, and
- $\mathrm{Ind}(\gamma;\lambda)$ is the number of times γ winds counterclockwise around λ.

A chain is a sum of curves. Integration with respect to a chain is the sum of integrals overa ll the cuves in the chain.

Definition 4 (Holomorphic Functional Calculus): If $f : U \to \mathbb{C}$ is analytic and $U \supseteq \mathrm{Spec}(A)$, then for smooth simple γ enclosing $\mathrm{Spec}(A)$ with index 1,

$$f(A) := \frac{-1}{2\pi i}\oint_\gamma f(z)R(z;A)dz. \tag{15}$$

The main point of this definition is that it recovers composition.

Theorem 4: Holomorphic functional calculus

If $U \supseteq \mathrm{Spec}(A)$ and given analytic functions $f : U \to \mathbb{C}$ and $g : U \to U$, we have $f(g(A)) = (f \circ g)(A)$.

Example 2: Exponentials

If f is entire and $f(z) = \sum_{k=0}^{\infty} a_k z^k$, then we could also define $f(A) = \sum_{k=0}^{\infty} a_k A^k$. This coincides with (14). Now we also have

$$\exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!} = \frac{-1}{2\pi i} \oint_\gamma e^z R(z; A) dz.$$

Conversely, if $\Re \mathrm{Spec}(A) > 0$, $\log A = \frac{-1}{2\pi i} \oint_\gamma \log(z) R(z; A) dz$ where we take $\log z$ the principal branch. Moreover, $\log(\exp(A)) = A$.

Finally we note that for symmetric A, we can give a simple spectral representation.

Definition 5 (Symmetric spectral mapping): If $A \in \mathbb{M}(n, n)$ is symmetric and f is a real-valued function defined in a neighborhood of $\mathrm{Spec}(A)$ then in terms of orthonormal eigenvectors u_j and eigenvalues λ_j

$$f(A) := \sum_{j=1}^{n} f(\lambda_j) u_j u_j^T. \tag{16}$$

This agrees with the holmorphic functional calculus when f is analytic in a neighborhood $\mathrm{Spec}(A)$ using Corollary 2.

1.5 Martingales and concentration

(Discrete time) Martingales are processes satisfying the two following properties:

Definition 6 (Martingale): A Martingale $(M_n : n \geq 0)$ adapated to a filtration $(\mathscr{F}_n : n \geq 0)$ is a real-valued stochastic process satisfying:

1. $\mathbb{E}|M_n| < \infty$ for all $n \geq 0$.
2. $\mathbb{E}(M_{n+1} \mid \mathscr{F}_n) = M_n$.

On replacing the second equality by $\geq$ we get a submartingale and likewise $\leq$ leads to a supermartingale.

Martingales are essential tools for the analysis of stochastic processes. They generally allow the analysis of many different processes.

A typical application of martingales is the following:

Lemma 1 (Doob Maximal inequality): For any non-negative submartingale $(M_n : n \geq 0)$ and any $a > 0$ and all $n \geq 1$

$$\Pr(\max_{0 \leq k \leq n} M_k \geq a) \leq \frac{\mathbb{E}M_n}{a}.$$

Submartingales can be manufactured from martingales by applying a convex function:

Exercise 1 (convex): Suppose that $\phi : \mathbb{R} \to \mathbb{R}$ is convex and that $(M_n : n \geq 0)$ is a martingale. Show that if $(\phi(M_n) : n \geq 0)$ has finite expectation, then it is a submartingale. If further ϕ is nondecreasing, then the same holds if $(M_n : n \geq 0)$ is a submartingale

Martingales moreover can be manufactured from other process by taking their *Doob decomposition*.

Definition 7 (Predictable): A stochastic process $(X_n : n \geq 0)$ is *predictable* if X_0 is deterministic and X_n is $\mathscr{F}_{n-1}$–measurable for all $n \in \mathbb{N}$.

(Note)[1]

Using this, any adapted process can be decomposed into a martingale and predictable part.

Theorem 5: Doob decomposition

Any real-valued process $(X_n : n \geq 0)$ having $\mathbb{E}|X_n| < \infty$ for all n and adapted to a filtration $(\mathscr{F}_n : n \geq 0)$ can be uniquely decomposed as $X_n = M_n + A_n$ where $M_0 = 0$, $(M_n : n \geq 0)$ is a martingale and $(A_n : n \geq 0)$ is predictable. Moreover

$$A_n = \mathbb{E}X_0 + \sum_{j=1}^{n} \mathbb{E}(X_j - X_{j-1} \mid \mathscr{F}_{j-1}).$$

The process $(A_n : n \geq 0)$ is called the *compensator* of $(X_n : n \geq 0)$.

The bracket process is an important is an important special case. [2]

Define

[1] This implies adaptedness, but moreover, it means that at the n-th step, you could have determined the process available in the $(n-1)$-st.

[2] This is going to intuitively represent the accumulated ammount of "randomess" of a martingale. This measure can be skeweed to be larger than in some sense it should be if the second moments of increments of the martingale barely exist (or do not exist at all!) in which case this is not really useful. So it is almost always appears paired with the condition that $|M_j - M_{j-1}| \leq 1$ almost surely, which is more helpful.

Definition 8 (Bracket process): For a martingale $(M_n : n \geq 0)$, the *bracket process* $[M_n]$ is the compensator of M_n^2, i.e.

$$\begin{aligned}[M_n] &= \mathbb{E}M_0^2 + \sum_{j=1}^{n} \mathbb{E}(M_j^2 - M_{j-1}^2 \mid \mathscr{F}_{j-1}) \\ &= \mathbb{E}M_0^2 + \sum_{j=1}^{n} \mathbb{E}((M_j - M_{j-1})^2 \mid \mathscr{F}_{j-1}).\end{aligned}$$

One of the simplest criteria for convergence of a stochastic process can be given in terms of this bracket process.

Theorem 6: Bracket process & convergence

Suppose that $(M_n : n \geq 0)$ is a martingale. By monotonicity, $[M]_\infty := \lim_{n\to\infty}[M]_n$ exists almost surely (but may be infinite). On the event $[M]_\infty < \infty$, $M_n \xrightarrow[n\to\infty]{\text{a.s.}} M_\infty$, which exists and is finite almost surely.

1.6 Subgaussian Martingale concentration

When the increments of a martingale are sufficiently bounded, it is possible to make much stronger estimates of the maximum value of a martingale, and this leads to some of the most important applications of martingales: tail bounds for random variables.

Definition 9 (Subgaussian): A centered random variable X is V-subgaussian if

$$\mathbb{E}e^{\lambda X} \leq e^{\lambda^2 V/2} \quad \text{for all } \lambda \in \mathbb{R}.$$

If the random variable is not centered, there are competing definitions of what V-subgaussian should mean. The clearest alternative definition would be an estimate of its ψ_2–norm defined in Definition 10.

This also leads to a definition of a norm which is convenient for quick tail bounds.

Definition 10 (Orlicz-norms): For any $p \geq 1$ and any real-valued random variable X, define the ψ_p–Orlicz norm

$$\|X\|_{\psi_p} = \inf\{t \geq 0 : \mathbb{E}\exp(|X|^p/t^p) \leq 2\}.$$

This connects to the previous definition through the following estimate:

Lemma 2 (Orlicz characterization of Subgaussian): There are absolute constants C_1 and C_2 so that:

1. If a centered random variable X is V-subgaussian, then
$$\|X\|_{\psi_2} \leq C_1\sqrt{V}.$$
2. Conversely, if X is centered and $\|X\|_{\psi_2} < \infty$ then X is $C_2\|X\|_{\psi_2}^2$ subgaussian.

Besides the subgaussian case, this has another extremely important special case:

Definition 11 (Subexponential): A random variable X is V-subexponential if
$$\|X\|_{\psi_1} \leq V.$$

See [Ver18, Chapter 2] for an elaboration on various equivalent formulations of subgaussian and subexponential processes.

For a martingale, we can define an upgraded bracket process, replaces a sum of conditional variances by the sum of conditional subgaussian increments.

Definition 12 (Subgaussian Bracket): A martingale $(M_n : n \geq 0)$ is (V_n)–conditionally subgaussian for an adapted process $(V_n : n \geq 1)$ if for all $n \geq 1$ and all $\lambda \in \mathbb{R}$
$$\mathbb{E}[e^{\lambda(M_n - M_{n-1})} \mid \mathscr{F}_{n-1}] \leq e^{\lambda^2 V_n} \quad \text{a. s.}$$

Define the subgaussian bracket $[\![M_n]\!]$ as the smallest, non-negative, non-decreasing adapted process so that $(M_n : n \geq 0)$ is conditionally subgaussian with process $([\![M_n]\!] - [\![M_{n-1}]\!] : n \geq 1)$.
Say that a martingale $(M_n)_{n=1}^N$ is subgaussian if $[\![M_N]\!] < \infty$ a. s.

This leads immediately to a tail bound for a martingale which enjoys this conditional subgaussian property.

Theorem 7: Subgaussian Azuma

Suppose that $(M_n : n \geq 0)$ that is a subgaussian with martingale. Then for any $n, t, S \geq 0$,
$$\Pr(\{\sup_{0\leq k\leq n}(M_k - M_0) \geq t\} \cap \{[\![M_n]\!] \leq S\}) \leq \exp\left(-\frac{t^2}{2S}\right).$$

Proof. By subtracting M_0 from the martingale, we may assume M_0 is

0. Define a new process, for any $\lambda \in \mathbb{R}$,

$$\mathcal{E}_n := \exp\big(\lambda M_n - \lambda^2 [\![M_n]\!]/2\big).$$

Then by the conditional subgaussian assumption $(\mathcal{E}_n : n \geq 0)$ is a supermartingale. Let T be the first time k that $M_k \geq t$ or that $[\![M_k]\!] > S$. Then by optional stopping, for $\lambda \geq 0$

$$1 \geq \mathbb{E}(\mathcal{E}_{T\wedge n}).$$

On the event $\{T \leq n\} \cap \{[\![M_n]\!] \leq S\}$, we have

$$\mathcal{E}_{T\wedge n} \geq \exp\big(\lambda t - \lambda^2 [\![M_T]\!]/2\big) \geq \exp\big(\lambda t - \lambda^2 S/2\big).$$

Thus

$$1 \geq \Pr(\{T \leq n\} \cap \{[\![M_n]\!] \leq S\}) \exp\big(\lambda t - \lambda^2 S/2\big).$$

Rearranging we have shown that for any $\lambda \geq 0$,

$$\Pr(\{\sup_{0\leq k\leq n} M_k \geq t\} \cap \{[\![M_n]\!] \leq S\}) \leq \exp\big(-\lambda t + \lambda^2 S/2\big).$$

Optimzing over $\lambda \geq 0$, we select $\lambda = t/S$ which shows the bound. □

A simple special case is for increments that are bounded.

> **Lemma 3 (Bounded implies subgaussian):** Suppose that X is mean 0 and $X \in (a, b)$ for $a, b \in \mathbb{R}$. Then
>
> $$\mathbb{E}\exp(\lambda X) \leq \exp((b-a)^2\lambda^2/8).$$
>
> Or simply, X is $(b-a)^2/4$–subgaussian.

Proof. Suppose without loss of generality that $b \leq a$. We can represent X as a convex combination, by

$$X = b\frac{X-a}{b-a} + a\frac{b-X}{b-a}.$$

Then by convexity for all $\lambda \in \mathbb{R}$

$$\mathbb{E}\exp(\lambda X) \leq \mathbb{E}\left(\exp(\lambda b)\frac{X-a}{b-a} + \exp(\lambda a)\frac{b-X}{b-a}\right).$$

Using that X has mean 0,

$$\mathbb{E}\exp(\lambda X) \leq \exp(\lambda b)\frac{-a}{b-a} + \exp(\lambda a)\frac{b}{b-a} =: f(\lambda).$$

Taking the log-derivative

$$\frac{d}{d\lambda}\log f(\lambda) = \frac{-ab\exp(\lambda b) + ab\exp(\lambda a)}{-a\exp(\lambda b) + b\exp(\lambda a)}.$$

With courage, we take another derivative, and then bound it above by $(b-a)^2/4$, uniformly in $\lambda \in \mathbb{R}$. Then, integrating twice,

$$\log f(\lambda) \leq \frac{\lambda^2}{2}\frac{(b-a)^2}{4}.$$

□

As a corollary, we derive the classical Azuma inequalities.

Corollary 5 (Azuma): Suppose that $(M_n : n \geq 0)$ is a martingale and $(A_n : n \geq 1)$ is a predictable process such that for all $1 \leq k \leq n$, $|M_k - M_{k-1}| \leq A_k$, then for all $t \geq 0$

$$\Pr\big(\{\max_{0\leq k\leq n}(M_k - M_0) \geq t\} \cap \{\sum_1^n A_k \leq A\}\big) \leq \exp\left(-\frac{t^2}{2A}\right).$$

If A_k are in fact deterministic, then we derive the convential Azuma inequality

$$\Pr\big(\max_{0\leq k\leq n}(M_k - M_0) \geq t\big) \leq \exp\left(-\frac{t^2}{2\sum_{k=1}^n A_k^2}\right).$$

1.7 *Subexponential Martingale concentration*

Martingales whose increments are only subexponential still retain a strong tail bound which is not quite Gaussian, but is generally Gaussian on a large enough range to recover most of what one needs from such a tail bound. The following is an adaptation of *Bernstein's inequality* to the martingale case (c.f. [Ver18, Theorem 2.8.1], where the nonmartingale bound is proven. The adaptation to the martingale case is a small extension):

Lemma 4 (Martingale Bernstein inequality): If $(M_n)_1^N$ is a martingale on the filtered probability space $(\Omega, (\mathscr{F}_n)_1^N, \Pr))$ and we define

$$\sigma_n := \left\| \inf\{t \geq 0 : \mathbb{E}\left(e^{|M_n - M_{n-1}|/t} | \mathscr{F}_{n-1}\right) \leq 2\} \right\|_{L^\infty(\Pr)}, \tag{17}$$

then there is an absolute constant $C > 0$ so that, for all $t > 0$,

$$\Pr\left(\sup_{1\leq n\leq N} |M_n - M_0| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{t}{C\|\sigma\|_\infty}, \frac{t^2}{C\|\sigma\|_2^2}\right\}\right), \tag{18}$$

where the norms $\|\sigma\|_p$ are the ℓ^p vector norms of $(\sigma_n : 1 \leq n \leq N)$.

Another, related inequality is Freedman's inequality, which trades stronger *a priori* control on the increments for simple control on the bracket.

Lemma 5 (Freedman inequality): Suppose $(M_n)_1^N$ is a martingale on the filtered probability space $(\Omega, (\mathscr{F}_n)_1^N, \Pr))$ and suppose its increments are all bounded by 1 almost surely Then there is an absolute constant $C > 0$ so that, for all $S, t > 0$,

$$\begin{aligned} &\Pr\left(\left\{\sup_{1\leq n\leq N} |M_n - M_0| \geq t\right\} \cap \{[M_N] \leq S\}\right) \\ &\leq 2\exp\left(-\min\left\{\frac{t}{C}, \frac{t^2}{CS}\right\}\right). \end{aligned} \tag{19}$$

1.8 Itô calculus

We will use simple multivariable Itô calculus for continous semi-martingales. An introduction to this type of theory can be found, for example in [Oks13] or in [KS91]. We will not attempt to develop this theory entirely here, but in this text we will use the simplest theory of (strong) solutions of stochastic differential equations. Furthermore, we shall show how this interacts with the tensor formalism introduced earlier.

Recall that:

Definition 13 (Brownian motion): A Brownian motion $(B_t : t \geq 0)$ is a continuous function (almost surely) with the poperty that $B_0 = 0$ and for any finite collection $0 = t_0 < t_1 < t_2 < \cdots < t_k$ the collection $(B_{t_j} - B_{t_{j-1}} : 1 \leq j \leq k)$ are indepedent, mean 0, Gaussian and have variances $(|t_j - t_{j-1}| : 1 \leq j \leq k)$. A standard d-dimensional Brownian motion is a vector of independent Brownian motions.

We suppose that $(\Omega, (\mathscr{F}_t : t \geq 0), \Pr)$ is a filtered probability space with a d-dimensional Brownian motion $(B_t : t \geq 0)$ so that B_t is $\mathscr{F}_t$ measurable for all $t \geq 0$ (i.e. it is adapted).

The filtration we take to be right-continuous.

In continuous time, we again define continuous martingales:

Definition 14 (Continuous Martingale): A continuous martingale $(M_t : t \geq 0)$ adapated to filtration $(\mathscr{F}_t : t \geq 0)$ is a real-valued stochastic process satisfying:

1. $\mathbb{E}|M_t| < \infty$ for all $t \geq 0$.
2. $\mathbb{E}(M_t \mid \mathscr{F}_s) = M_s$ for all $t > s \geq 0$

Replacing the second equality by $\geq$ we get a submartingale and likewise $\leq$ leads to a supermartingale.

Continuous martingales and stochastic processes are slightly incomplete in that it is helpful to enlarge this class slightly. So we define:

Definition 15 (Local martingale): A local martingale $(X_t : t \geq 0)$ is a continuous adapted process to $(\mathscr{F}_t : t \geq 0)$ with the property that there is a sequence of stopping times T_k with $T_k \xrightarrow[k\to\infty]{\text{a.s.}} \infty$ and so that the stopped process $X_t^{T_k} := X_{t \wedge T_k}$ are martingales.

Definition 16 (Itô integral): For an adapted continuous process V in the space of matrices $\mathbb{M}(p,d)$ having $\|V\|_\sigma$ bounded by 1 almost surely, the Itô integral can be given by the in-probability limit

$$\int_0^t V_s \, \mathrm{d}B_s = \Pr \cdot \lim_{k\to\infty} \sum_{j=1}^{k} V_{t_{j-1}}(B_{t_j} - B_{t_{j-1}}),$$

where the maximal spacing in the mesh $0 = t_0 < t_1 < t_2 < \cdots < t_k = t$ tends to 0 with k.

This can be seen to be independent of the choice of mesh and be subsequently extended to unbounded integrands V by approximation by bounded ones.

Definition 17 (Itô process): An Itô process $(X_t : t \geq 0)$ in $\mathbb{R}^o$ is one for which we can represent

$$X_t = X_0 + \int_0^t u_s \, \mathrm{d}s + \int_0^t V_s \, \mathrm{d}B_s,$$

with the latter integral given by the Itô integral and where u and V are continuous adapted processes satisfying that almost surely, for each $t \geq 0$,

$$\int_0^t (\|u_s\| + \|V_s\|) \, \mathrm{d}s < \infty.$$

This is often represented in differential form by

$$\mathrm{d}X_t = u_t \, \mathrm{d}t + V_t \, \mathrm{d}B_t.$$

An Itô process is *finite variation* if and only if $V_t \equiv 0$.

A key result connects Itô processes and martingales

Theorem 8: Martingale representation

An Itô process is *local martingale* if and only if $u_t \equiv 0$. If furthermore $\mathbb{E}|X_t| < \infty$ for all $t > 0$ then it is a martingale. Conversely, if a martingale (M_t) adapted to $(\mathscr{F}_t)$ satisfies $\mathbb{E}|M_t|^2 < \infty$ for any t, then it is an Itô process.

For Itô processes, we have Itô's formula.

Theorem 9: Itô's formula

Suppose $g : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ is C^2 and suppose that $(X_t : t \geq 0)$ is an Itô process. Then $g(t, X_t)$ is again an Itô process and

moreover

$$\mathrm{d}g(t,X_t) = (\partial_t g(t,X_t) + \langle \nabla_x g(t,X_t), u_t\rangle + \tfrac{1}{2}\langle \nabla_x^2 g(t,X_t), V_t\rangle)\,\mathrm{d}t + \langle \nabla_x g(t,X_t), V_t\,\mathrm{d}B_t\rangle.$$

For a continuous-time local martingale, $(M_t : t \geq 0)$, in $\mathbb{R}$, we define its bracket process by:

Definition 18 (Bracket process): The bracket process $[M]_t$ is the unique finite variation process with $[M]_0 = 0$ so that $M_t^2 - [M]_t$ is a local martingale. If $\mathrm{d}M_t = V_t\,\mathrm{d}B_t$ then

$$\mathrm{d}[M]_t = \|V_t\|^2\,\mathrm{d}t.$$

The bracket process gives a quick way to produce tail bounds for local martingales.

Exercise 2 (Exponential martingale): Use Itô's formula (applied to $X_t = (M_t, [M]_t)$) to show that $\exp(M_t - \frac{1}{2}[M]_t)$ is a local martingale.

Lemma 6 (Concentration for Brownian martingales): Suppose that $(M_t : t \geq 0)$ is a local martingale. For any $T, S, x > 0$

$$\Pr(\{\max_{0\leq t\leq T}(M_t - M_0) \geq x\} \cap \{[M]_T \leq S\}) \leq \exp\left(-\frac{x^2}{2S}\right).$$

Note that this is precisely the analogue of the Discrete Freedman's inequality Lemma 5.

Proof. By subtracting M_0 from M we may assume $M_0 = 0$. Using that $Y_t := \exp(\lambda M_t - \frac{\lambda^2}{2}[M]_t)$ is a local martingale, there are stopping times T_k so that $Y_t^{T_k}$ are martingales. Let $\vartheta = \min\{t : [M]_t \geq S \text{ or } |M_t| > x\}$. As $Y^{T_k\wedge\vartheta}$ is a continuous martingale,

$$\mathbb{E}[Y_T^{T_k\wedge\vartheta}] = \mathbb{E}[Y_0^{T_k\wedge\vartheta}] = 1.$$

By Fatou's Lemma, we may take $k \to \infty$ and conclude

$$\mathbb{E}[Y_T^{\vartheta}] \leq \liminf_{k\to\infty} \mathbb{E}[Y_T^{T_k\wedge\vartheta}] = 1.$$

On the event $\{[M]_T \leq S\} \cap \{\max_{0\leq t\leq T} M_t \geq x\}$ we have for $\lambda \geq 0$

$$Y_T^{\vartheta} \geq \exp(\lambda x - \tfrac{\lambda^2}{2}S).$$

Hence

$$\Pr(\{[M]_T \leq S\} \cap \{\max_{0\leq t\leq T} M_t \geq x\}) \leq \exp(-\lambda x + \tfrac{\lambda^2}{2}S).$$

Setting $\lambda = x/S$ gives

$$\Pr(\{[M]_T \leq S\} \cap \{\max_{0 \leq t \leq T} M_t \geq x\}) \leq e^{-\lambda^2/(2S)}.$$

□

2 SGD and optimization theory

SGD (stochastic gradient descent) [3] has raised to prominence as a multipurpose, simple algorithm for the optimization of many random functions. There are probably lots of reasons for its success, first and foremost being that it is a gradient-based algorithm; gradients, especially in high-dimensions are hugely important in that the optimal search directions tend to evade any fixed choices. [4] Another reason for its success, or more precisely the success of a larger umbrella of stochastic gradient methods, is that the algorithm is quite extensible: minibatch SGD, momentum SGD [Sut+13], Adagrad [DHS11], RMSProp [HSS12], and most prominently Adam [KB14] all extend SGD by fusing it with other optimization techniques. This list covers the lion's share of optimization algorithms used for machine learning as of today.

To introduce SGD, we will consider the *finite-sum framework*. This is an example of a *structure*[5] we impose on the objective function f to be optimized.

Definition 19 (Finite sum): An optimization problem is *finite sum* if its objective function f can be given by

$$\min_{x\in\mathbb{R}^d}\{f(x)\} \quad \text{where} \quad f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad x \in \mathbb{R}^d. \tag{20}$$

The parameter d represents the dimensionality of the parameter space, and n represents the number of functions.

In the typical *empirical risk minimization* framework (discussed below), the n would represent the cardinality of the training data-set and each f_i would represent the *risk* associated to the i-th datapoint.

We shall also assume that the functions f_i have amount of smoothness. For exploiting any form of gradient method, we need to have a derivative. Further, this derivative almost always needs some amount of tameness.

Definition 20 (Lipschitz gradients): The objective function $f : \mathbb{R}^d \to \mathbb{R}$ has Lipchitz gradients with constant L if ∇f exists and

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

for all $x, y \in \mathbb{R}^d$. For the finite sum problem, we say its summands have Lipschitz gradients if there is a constant L such that for all $x, y \in \mathbb{R}^d$

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|.$$

[3] It has been argued that the "descent" should be dropped from the name of this algorithm, owing to the fact that the algorithm need not always descend (and hence does not fit into the larger class of descent algorithms [BCN18a].

[4] One may wonder about why does one only use gradients? Higher-order optimization that takes advantage of Hessian information can be faster, but the computational costs of even computing the Hessian (or of approximating it) grow with dimension. See the discussion in [Bot10].

[5] Structure, in the context of optimization theory, is the set of assumptions one puts on the objective function f which allows it to be meaningfully manipulated or optimized. Standard examples include convexity or smoothness.

Exercise 3 (Quadratic Upper Bound): Suppose that f has Lipschitz gradients. Show that for any $x, y \in \mathbb{R}^d$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{L}{2}\|y - x\|^2$$

by setting $g(t) = f(x + t(y - x))$ and using $f(y) - f(x) = \int_0^1 g'(t)\,\mathrm{d}t$.

Remark 1 (A little less smooth): A sufficient condition for Lipschitz gradients is that f is twice differentiable with a second-derivative matrix (Hessian matrix) bounded in norm. While Lipschitz gradients is a little weaker than this, it is not by much. A weaker structure which is common is just that f itself is Lipschitz or even α-pseudo-Lipschitz, meaning

$$|f(x) - f(y)| \leq L\|x - y\|(1 + \|x\|^\alpha + \|y\|^\alpha).$$

A final very common structure to consider is *convexity*. Convexity makes lots of problems simpler to analyze. So when one has convexity, it is a shame not to use it. However, in contrast to smoothness assumptions (which are essentially necessary, in some form, to being able to run SGD), convexity is not necessary.

Definition 21 (Strong convexity): The objective function $f : \mathbb{R}^d \to \mathbb{R}$ is strongly convex with constant $\mu > 0$ if for any $x, y \in \mathbb{R}^d$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{\mu}{2}\|y - x\|^2.$$

If this holds with $\mu = 0$, then the function is convex. If one has the above with $\mu = 0$ but with a *strict* inequality, then the function is strictly convex.

Exercise 4 (Function value growth): Suppose that f is continuously differentiable and x^* is a stationary point of f, i.e. $\nabla f(x) = 0$. Show that if f is strongly convex then x^* is a global minimizer, and moreover for all x

$$f(x) - f(x^*) \geq \tfrac{\mu}{2}\|x - x^*\|^2.$$

Hence x^* is a global minimizer.

This shows that any local minimizer is a global minimizer, and hence the global minimizer is unique. For strictly convex functions, these conclusions remain true, but we can only conclude $f(x) - f(x^*) > 0$ for all $x \neq x^*$.

2.1 *SGD on the finite-sum*

The finite-sum framework allows us to pose a very general version of SGD:

Definition 22 (SGD): Stochastic gradient descent, with step-size schedule γ_k, has the iterates

$$x_{k+1} = x_k - \gamma_k \nabla f_{i_k}(x_k),$$

where i_k is a (usually random) choice of function. We let $(\mathscr{F}_k : k \in \mathbb{N}_0)$ be a filtration with respect to which the sequence $\{(i_k, x_k)\}$ are adapated. Some of the most important examples are given by:

1. (Random-sample/multi-pass) The $i_k \overset{\text{law}}{=} \text{Unif}(\{1, 2, \cdots, n\})$ are chosen iid.
2. (Single-shuffle) A single permutation $\pi : \{1, 2, \cdots, n\} \to \{1, 2, \cdots, n\}$ is drawn uniformly at random, and then we set $i_{nr+k} = \pi(k)$ for all non-negative integers r.
3. (Random-shuffle) After each *epoch* (6[o]) , we draw a new permutation $\pi_r : \{1, 2, \cdots, n\} \to \{1, 2, \cdots, n\}$ uniformly at random and set $i_{nr+k} = \pi_r(k)$.
4. (One-pass) Here one runs the single-shuffle algorithm but simply stops after (or before) one full pass over the dataset.

[6] An epoch is one full pass over the dataset. We will use it to mean n here, even in the multi-pass case.

The goal of these notes is to establish the algorithmic performance implications of choices such as these and the step-size schedule $\{\gamma_k\}$. How large should they be chosen? In cases where there are many solutions, which solutions are selected and how does it depend on the choices of step-size or shuffling scheme?

Remark 2 (Minibatch SGD): A natural extension of SGD processes multiple gradient estimates in parallel. In this case, one forms updates

$$x_{k+1} = x_k - \gamma_k \sum_{i \in B_k} \nabla f_i(x_k),$$

for a random subset $B_k \subseteq \{1, 2, \cdots, n\}$. In practice, SGD on the finite sum is essentially always run in batches, which can be chosen analogously to all the methods in Definition 22. Part of the reason is architectural: working in batches ensures allows one to perform fewer gradient queries and/or the hardware itself (especially GPUs) parallelize multi-dimensional tensor contraction (up to some bounds on dimensionality) to be the same wall-clock speed as a single dot product. Hence in a given update-loop one may want to increase the batch size to

take advantage of this.

However, from the mathematical point of view, this begs the question if there is any difference in the behavior of the algorithm as a consequence of the batch-size. Lots of work has focused on the idea of "variance reduction", which is to say that minibatch SGD updates are smaller-variance updates of the underlying gradient.

But this intuition ignores dimensionality effects – if all the gradient estimators are orthogonal, there is no averaging effect occuring within the sum. So there is only a 'reduction of variance' once the batch-size starts to exceed the effective dimensionality of the gradients. In many of the setups here, that means that $|B_k|$ needs to be proportional to d, or moreover if $|B_k|/d \to 0$ one reproduces small-batch limits.

In a setup where batch-size grows proportionally to dimension, one can recover an entire theory that runs in parallel to what is presented here. In particular one sees that there is a saturation effect once batch is sufficiently large (first observed in [MBB18]; see also for a sharper analysis in [Lee+22] using assumptions similar to the ones here). Proportional batch methods have also been analyzed in [Ger+22] using similar machinery.

Remark 3 (Momentum methods): Momentum methods are another direction of generalization, in which one keeps a running average of gradient estimates and then uses this running average to update the function. This provides another axis along which to consider the behavior of stochastic gradient methods This was popularized in machine learning possibly by [Sut+13]; there remains a relatively healthy controversy over whether or not momentum matters for stochastic optimization, but this may be partly because of the precise form of the momentum (see especially [MY18], [Kid+18] and [PP21] which give versions which appear to correctly capture some of this momentum effect in small-batch settings) or because of interactions of batch size and momentum [BCW22] [Lee+22].

2.2 *Risks*

To measure the performance of SGD, it is helpful to adopt the language of *risk*. This gives us a precise way of describing the algo-

rithmic performance of SGD. Suppose that we have a distribution $\mathcal{D}$ on $\mathbb{R}^m \times \mathbb{R}^p$, where m represents an ambient data dimensionality and the second $\mathbb{R}^p$ represents a (p-dimensional) output or label. The basic statistical learning theory challenge is to find a function $M : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^p$ which for a given choice of parameters $x \in \mathbb{R}^d$ and a data-point (a, b) sampled from $\mathcal{D}$, minimizes a loss $\ell : \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^p \to [0, \infty)$.

Definition 23 (Statistical risk): The *statistical risk* (or *population risk* or *expected risk*) is the function

$$\mathscr{P} : x \mapsto \mathbb{E}(\ell(x, M(x, a), b)) \quad \text{where} \quad (a, b) \overset{\text{law}}{=} \mathcal{D}.$$

In other words, having selected our parameters x and our chosen loss, how much do our modeling mistakes cost?

In practice, having a finite dataset, we might reserve some of these data for the "test" dataset and keep the remainder (often the vast majority) for the training data set. Having estimated the parameters $x \in \mathbb{R}^d$, we could measure how well we did by statistically estimating $\mathscr{P}$ using the test data set. For the estimation of the parameters x, we use instead:

Definition 24 (Empirical risk): The *empirical risk* (or *training loss*) for n samples $(a_i, b_i)_{i=1}^n$ drawn iid from $\mathcal{D}$ is

$$\mathscr{L} : x \mapsto \frac{1}{n} \sum_{i=1}^{n} (\ell(x, M(x, a_i), b_i)).$$

Note that mathematically, this is nothing but the statistical risk, but with the distribution $\mathcal{D}$ replaced by the *empirical* distribution of samples. Moreover, this leads naturally to the empirical risk minimization problem:

Definition 25 (Empirical risk minimization): The *empirical risk minimization* problem for n samples $(a_i, b_i)_{i=1}^n$ drawn iid from $\mathcal{D}$ and model M is the finite-sum problem

$$\min_{x \in \mathbb{R}^d} \left\{ \mathscr{L}(x) = \frac{1}{n} \sum_{i=1}^{n} \ell(x, M(x, a_i), b_i) \right\}.$$

To make this concrete, we illustrate a few canonical examples.

Example 3: Linear regression

The most important, basic example is 1-dimensional linear regression (i.e. $p = 1$). Here we take the model M to just be a

linear function, so that $M(x,a) := \langle x,a\rangle$ and so the parameter dimensionality d matches the data dimensionality m. Further, we suppose the data-distribution $\mathcal{D}$ comes from a linear model, which is to say that

$$b = \langle a,\beta\rangle + \epsilon$$

for a ground truth $\beta \in \mathbb{R}^d$; a noise random variable ϵ which is independent of a, mean 0, finite variance; and a random vector a from some distribution on $\mathbb{R}^d$. Often, we take $a \overset{\text{law}}{=} \text{Normal}(0,\Sigma)$ for a $d \times d$ covariance matrix Σ.

The conventional loss to take in this setting is the mean-squared error, so that $\ell(x,a,b) = \frac{1}{2}(b-a)^2$. In this case, we have the following explicit formula for the population risk

$$\mathscr{R}(x) = \tfrac{1}{2}\mathbb{E}(\epsilon + \langle a,\beta\rangle - \langle a,x\rangle)^2 = \tfrac{1}{2}\mathbb{E}\epsilon^2 + \tfrac{1}{2}\langle \beta - x, \Sigma(\beta - x)\rangle. \tag{21}$$

Note that when $\Sigma \succ 0$, this has a unique minimizer at $\beta = x$, and moreover the loss is strictly convex.

The empirical risk also can also be represented simply. Suppose we have n data-target pairs (a_i,b_i) for $1 \le i \le n$. If we let A be a matrix whose n rows are given by $\{a_i\}$ and b be the column vector of $\{b_i\}$ then

$$\mathscr{L}(x) = \tfrac{1}{2n}\sum_{i=1}^{n}(\langle a_i,x\rangle - b_i)^2 = \tfrac{1}{2n}\|Ax-b\|^2. \tag{22}$$

Example 4: Ridge regression and penalties

A small generalization of this problem is to add a *regularizer*, or effectively to modify the loss to penalize large weights. The ℓ_2–regularized loss is $\ell(x,a,b) = \frac{1}{2}((b-a)^2 + \lambda\|x\|^2)$. This makes the empirical risk

$$\mathscr{L}(x) = \tfrac{1}{2n}\|Ax-b\|^2 + \tfrac{\lambda}{2}\|x\|^2. \tag{23}$$

It should be noted that in context, one may wish to consider either the regularized population risk, or alternatively the unregularized population risk (21).

For any positive $\lambda > 0$, the empirical risk has a unique minimizer, as a consequence of the strong convexity of $\mathscr{L}$. The minimizer of the regularized empirical risk always exists, and is called the *ridge estimator*.

Other penalty terms may also be added; especially, adding a $\|x\|_1$–norm penalty leads to (one form of) the Lasso problem.

Exercise 5 (Ridge regression): how that the ℓ_2–regularized risk $\mathscr{L}$ is strongly convex (with constant λ) (and therefore $\nabla\mathscr{L}(X) = 0$ is uniquely solvable) and find its solution.

Example 5: Generalized linear models

One step more complicated than the linear models are generalized linear models (GLMs). With $p = 1$, one supposes the model M is a composition of a linear model and a nonlinearity; so

$$M(x, a) := \phi(\langle x, a\rangle).$$

Two notable cases are that of *phase retrieval*, in which case

$$\phi(x) = |x|. \tag{24}$$

This is one of the simplest nonconvex problems that can be formulated in high dimensions.

Another is *binary logistic regression* in which case

$$\phi(x) = \frac{e^x}{1 + e^x}. \tag{25}$$

In this case, the model $M(x, a)$ gains the extra interpretation as a probability. In particular, it may represent the probability that a data-point a has membership in some class, and so this is well-suited to a classification problem.

The data distribution $\mathcal{D}$ might be many things, but one natural choice is that the data follows the model we are trying to fit. If we do so, then the data distribution $\mathcal{D}$ are assumed to be given by

$$b = \phi(\langle a, \beta\rangle),$$

where a follows some distribution. Noise may also be added, but the precise location of the noise in the model differs from case to case.

In the case of binary logistic regression, one may instead suppose that

$$b = X \cdot \chi + (1 - \chi)\mathbf{1}_{\langle a,\beta\rangle > 0},$$

where $\chi \overset{\text{law}}{=} \text{Bernoulli}(\epsilon)$ and $X \overset{\text{law}}{=} \text{Bernoulli}(\frac{1}{2})$ are independent of a. This represents a data distribution in which class

membership is given, but with some amount of mislabeling error.

Finally for the losses, it is common with phase retrieval to simply choose the mean-squared error. If one takes this, without regularization and without noise, then

$$\mathscr{P}(x) := \tfrac{1}{2}\mathbb{E}(|\langle a, x\rangle| - |\langle a, \beta\rangle|)^2 \tag{26}$$

In some cases (especially the case of Gaussian a), it is possible to produce explicit expressions for the risk $\mathscr{P}$, but generally this is impossible. For logistic regression, it is common to use the KL–divergence

$$\ell(x, M, b) = b\log(\tfrac{b}{M}) + (1-b)\log(\tfrac{1-b}{1-M})$$

or the closely related cross-entropy loss. (7[o])

[7] The cross-entropy differs from the KL–divergence by addition of the entropy $b\log b + (1-b)\log(1-b)$. This additional term does not affect the gradients of the loss with respect to M, and hence it induces the same SGD dynamics.

Example 6: Generalized linear models II

Generalized linear models also naturally can take $p > 1$. This allows natural generalizations of the Example 5 such as multiclass logistic regression, for classifying multiple classes. Here we now suppose x is a 2-tensor, living in $\mathbb{R}^m \otimes \mathbb{R}^p$. The inner product $\mathbb{R}^m \ni a \mapsto \langle x, a\rangle_m$ contracts the m-dimensional part of x with that of a. (See partial contractions (6)). A generalized linear model is now one in which for some $g : \mathbb{R}^p \to \mathbb{R}$, $M(x, a) = g(\langle x, a\rangle)$.

For a concrete example, we introduce multi-class logistic regression. For functions $\phi : \mathbb{R} \to \mathbb{R}$ we extend them functions from $\mathbb{R}^p \to \mathbb{R}$ by applying ϕ coordinate-wise. The model is given by:

$$M(x, a) := \frac{e^{\langle x,a\rangle_m}}{\langle e^{\langle x,a\rangle_m}, \mathbb{1}\rangle}, \tag{27}$$

where $\mathbb{1}$ is the all-1 vector. We may assume the data distribution $\mathcal{D}$ is given as (a, b) where b is a one-hot (8[o]) class vector and a is an element of $\mathbb{R}^m$. For the loss, one may take once more the KL–divergence

$$\ell(x, a, b) = \sum_{i=1}^{p} b_i \log \tfrac{b_i}{a_i}.$$

[8] The one-hot representation of a class is the vector of all 0 save for in the entry given by the class, in which it is one.

Example 7: The two layer neural network

Neural networks generalize Example 6 further by effectively composing generalized linear models. The *multilayer perceptron* or MLP is the simplest example of this, and can be considered as compositions, in a sense, of generalized linear models. In a *two-layer neural network* (or *one-hidden layer neural network*), one composes two of these. In the notation of Example 6, if we set (see 9[o])

$$N(x,a) := (\langle x,a\rangle)_+$$

where x is an $\mathbb{R}^m \otimes \mathbb{R}^h$–dimensional parameter tensor (and so the output is $\mathbb{R}^h$–dimensional) then with M as in (27),

$$(\mathbb{R}^m \otimes \mathbb{R}^h) \times (\mathbb{R}^h \otimes \mathbb{R}^p) \times \mathbb{R}^m \ni ((x_1,x_2),a) \mapsto M(x_2, N(x_1,a))$$

is a relatively common construction of a neural network used for classification purposes. Typically, further layers and more purpose-built layers would be added to improve the performance (see for example [LeC+98], which was one of the first instances of "deep learning", and which has 6 hidden layers). Once more, one could use KL–divergence for the training purposes.

[9] The function $x \mapsto (x)_+$, meaning positive part, is the ReLU activation function, which is a popular choice.

2.3 *Streaming/Online stochastic gradient descent*

In the case of running streaming SGD for the problem of empirical risk minimization, at every step $k \leq n$ of the algorithm, one draws a new datapoint (a_{k+1}, b_{k+1}) and then performs an SGD update:

Definition 26 (Streaming SGD): Streaming (aka online) SGD is the algorithm with updates given by (28).

$$X_{k+1} = X_k - \gamma_k \nabla_{X_k} \ell(X_k, M(X_k, a_{k+1}), b_{k+1}). \tag{28}$$

Thus at the n-th step, the algorithm has used precisely n datapoints, and moreover, one may naturally view n as a free parameter, representing the number of datapoints used. This means that the algorithm is adapted to the filtration $(\mathscr{F}_k : k \geq 0)$ generated by the sequence of datapoints $((a_k, b_k) : k \geq 0)$. In the case of empirical risk minimization, there is effectively no difference between this and one-pass SGD, except for how the size of the data-set is discussed.

Streaming can be viewed as a form of stochastic gradient descent for directly minimizing the population risk $\mathscr{P}$. Namely commuting

expectation and the gradient, one has for streaming SGD

$$X_{k+1} = X_k - \gamma_k \nabla_{X_k} \mathscr{P}(X_k) - \xi_{k+1},$$

where ξ_{k+1} is the martingale increment

$$\xi_{k+1} = \gamma_k \nabla_{X_k} \ell(X_k, M(X_k, a_{k+1}), b_{k+1}) - \mathbb{E}[\gamma_k \nabla_{X_k} \ell(X_k, M(X_k, a_{k+1}), b_{k+1}) \mid \mathscr{F}_k].$$

> **Remark 4 (Streaming is an idealization):** While it is attractive to consider an algorithm which directly minimizes population risk, this is almost invariably a data-inefficient procedure. There can be circumstances where compute time, rather than data is the limiting feature (see the discussion in e.g. [Bot10] or [NNS20]), in which case one may wish to use something like streaming SGD.
> Regardless, as a theoretical exercise, it is definitely true that streaming is a simpler algorithm to mathematically understand, owing to the underlying independence of the updates.

2.4 *Classical convergence of stochastic gradient descent*

One of the traditional methods of analysis of stochastic gradient descent is as a stochastic process, establishing its almost sure convergence properties. Consider a stochastic algorithm defined by

$$X_{k+1} := X_k - \gamma_k(\nabla F(X_k) + \xi_{k+1}) \tag{29}$$

for some random vectors ξ_k with $\mathbb{E}(\xi_{k+1} \mid \mathscr{F}_k) = 0$. This is satisfied by both the multi-pass and one-pass versions of SGD for the finite-sum problem, provided all the f_i have bounded first derivatives.

Theorem 10: Mean convergence of SGD

Suppose that $F \geq 0$ satisifes:

1. F has Lipschitz gradients with constant L and F is μ-strongly convex.
2. The noise ξ_{k+1} satisfies $\mathbb{E}(\|\xi_{k+1}\|^2 \mid \mathscr{F}_k) \leq M\|\nabla F(X_k)\|^2$.
3. The step-size γ is constant and satisfies

$$0 < \gamma < \tfrac{2}{(1+M)L}.$$

Then with x_* the global minimizer of F

$$\mathbb{E}(F(X_k) - F(x_*)) \leq e^{-2\mu\alpha k} \times \mathbb{E}(F(X_0) - F(x_*)),$$

where $\alpha = \gamma(1 - \gamma\frac{L(1+M)}{2})$.

Proof. We look at an increment under SGD. Using the Lipschitz gradient property (or more precisely Exercise 3)

$$F(X_{k+1}) - F(X_k) \leq \langle \nabla F(X_k), X_{k+1} - X_k \rangle + \tfrac{L}{2}\|X_{k+1} - X_k\|^2.$$

Substituting the definition of the iterates,

$$F(X_{k+1}) - F(X_k) \leq -\gamma \langle \nabla F(X_k), \nabla F(X_k) + \xi_{k+1} \rangle + \tfrac{L\gamma^2}{2}\|\nabla F(X_k) + \xi_{k+1}\|^2.$$

Hence if we take conditional expectations on both sides

$$\mathbb{E}(F(X_{k+1}) - F(X_k) \mid \mathscr{F}_k) \leq -\gamma \|\nabla F(X_k)\|^2 + \tfrac{L\gamma^2(1+M)}{2}(\|\nabla F(X_k)\|^2).$$

Hence by how the step-size is chosen

$$\mathbb{E}(F(X_{k+1}) - F(X_k) \mid \mathscr{F}_k) \leq -\alpha \|\nabla F(X_k)\|^2.$$

Now we need the following conclusion of strong convexity: for x^* the global minimizer [10]

$$(F(x) - F(x^*)) \leq \frac{1}{2\mu}\|\nabla F(x)\|^2.$$

[10] This, while not totally obvious just follows from rearranging the definition of Strong convexity applied to the points x and $x - \frac{1}{\mu}\nabla F(x)$, and then using that $F(x^*)$ is a global minimizer.

Thus we conclude

$$\mathbb{E}(F(X_{k+1}) - F(x^*) \mid \mathscr{F}_k) \leq (1 - 2\alpha\mu)\mathbb{E}(F(X_k) - F(x^*) \mid \mathscr{F}_k),$$

which by induction proves the theorem. □

> **Remark 5 (Bibliographic note):** This was adapted from the excellent notes of [BCN18a].

This additional randomness could also come from a lot of sources: it may be added artificially to improve the behavior of the algorithm, such as *data augmentation strategies* (see for example [SK19] for a survey of the technique and see [HS21] for some related optimization considerations as discussed below), but frequently it is the result of using a computationally efficient stochastic estimator for the true gradient (which is generally the reason for minibatch SGD).

To analyze the algorithm, a good starting point is the Taylor expansion

$$F(X_{k+1}) = F(X_k) - \gamma_k \langle \nabla F(X_k), \nabla F(X_k) + \xi_{k+1} \rangle + R_{k+1}. \tag{30}$$

Theorem 11: Robbins-Monro convergence of SGD

1. Suppose that $F \geq 0$, F has Lipschitz gradients, $\|\nabla F\|^2$ is bounded, and $\mathbb{E}(\|\xi_{k+1}\|^2 \mid \mathscr{F}_k) \leq K$.

2. Suppose that F has compact sublevel sets, so that for all $t > 0$, $\{x \in \mathbb{R}^d : F(x) \leq t\}$ is compact.

3. Suppose that γ_k satisfies the Robbins-Monro condition

$$\sum_{k=1}^{\infty} \gamma_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \gamma_k^2 < \infty.$$

Let $\mathcal{S}$ be the set of stationary points of F, i.e. those $x \in \mathbb{R}^d$ for which $\nabla F(x) = 0$. Then (29) converges in that it satisfies $X_k \xrightarrow[k\to\infty]{\text{a.s.}} \mathcal{S}$, which is to say its distance from the set $\mathcal{S}$ tends to 0.

The first assumptions give control over the errors in the Taylor approximation. The R_{k+1} carries a factor of γ_k^2 and so it will be absolutely summable.

Exercise 6 (Convergence of R): Show that if F has Lipschitz gradients, $\|\nabla F\|^2$ is bounded, and $\mathbb{E}(\|\xi_{k+1}\|^2 \mid \mathcal{F}_k) \leq K$. Suppose $\sum_{j=1}^{\infty} \gamma_k^2 < \infty$, then $\sum_{k=1}^{\infty} R_k < \infty$.

Remark 6 (Notions of convergence): The above convergence shows that X_k converges to a stationary point, meaning a point of $\mathcal{S}$. It does not necessarily show convergence of X_k to a local, let alone a global, minimizer. Under further hypotheses on F, one can characterize the stationary points: most significantly, if F is strictly convex then there is unique minimizer X^* of the problem $\min F(X)$ and moreover it is the unique stationary point.

Remark 7 (Other convergence approaches): There are many versions of convergence of SGD that are proven throughout the literature. In [Bot98], multiple criteria are given for almost sure convergence. See also [BCN18b] for more versions of convergence in mean, more in the direction of Theorem 10.

Proof. We define the martingale M_k for $k > 0$ by

$$M_k = \sum_{j=1}^{k} \gamma_j \langle \nabla F(X_j), \xi_{j+1} \rangle.$$

This is a martingale which satisfies

$$\begin{aligned}\mathbb{E}M_k^2 &= \sum_{j=1}^{k} \gamma_j^2 \mathbb{E}\langle \nabla F(X_j), \xi_{j+1}\rangle^2 \\ &\leq \sum_{j=1}^{k} \gamma_j^2 \big(\sup_x \|\nabla F(x)\|^2\big) K.\end{aligned}$$

By assumption this is therefore bounded independently of k, and so we have by martingale convergence that there is a randopm variable M_∞ almost surely finite so that

$$M_k \xrightarrow[k\to\infty]{\text{a.s.}} M_\infty \quad \text{and} \quad \sup_k |M_k| < \infty.$$

From (30), this implies that

$$F(X_k) - F(X_0) \leq M_k + \sum_{j=1}^{k} R_j$$

Then the martingale and finite variation parts are both bounded, in that

$$\sup_k F(X_k) < \infty \quad \text{a. s.}.$$

We also have that

$$\Delta_k \coloneqq \sum_{j=1}^{k} -\gamma_j \langle \nabla F(X_j), \nabla F(X_j)\rangle$$

is non-increasing, and so either it tends to $-\infty$ or converges. If it tends to $-\infty$, we would contradict that $F(X_k) \geq 0$, since we have

$$F(X_k) - F(X_0) = \Delta_k + M_k + \sum_{j=1}^{k} R_j,$$

and the other terms are bounded. It furthermore follows that in fact $F_\infty \coloneqq \lim_{k\to\infty} F(X_k)$ exists almosts surely.

So we introduce the event $\mathcal{E}_P \coloneqq \{\sup_k F(X_k) < P\}$ and note that $\cup_{P=1}^{\infty} \mathcal{E}_P$ has probability 1, it suffices to show that on every $\mathcal{E}_P$ we have $\|\nabla F(X_k)\| \xrightarrow[k\to\infty]{\text{a.s.}} 0$ or in other words for every $\epsilon > 0, P > 1$

$$\Pr(\mathcal{E}_P \cap \{\|\nabla F(X_k)\|^2 > \epsilon \text{ for infinitely many } k\}) = 0.$$

Now on the event $\mathcal{E}_P$, SGD remains in the set $K = \{x : F(x) \leq P\}$ for all time, which by assumption is compact. So we may work inside of K with the subspace topology. Recall that $\mathcal{S}$ is the set of stationary points, and let $\mathcal{U}_\epsilon$ be the set $\{x \in K : \|\nabla F(x)\|^2 \geq \epsilon\}$. Then this is disjoint from $\mathcal{S}$, and by compactness of $K \cap \mathcal{S}$ we can find a $\delta > 0$ sufficiently small that the closed δ-neighborhood V_δ of $\mathcal{U}_\epsilon$ is disjoint

from $\mathcal{S}$. Also by compactness there is an $\eta > 0$ so that $\|\nabla F(x)\|^2 > \eta$ uniformly on $K \cap \mathcal{S}$.

Now we show that X_k cannot visit $\mathcal{U}_\epsilon$ infinitely often. If we wait long enough, the contributions of the noise M_k and the Taylor error terms R_k will be uniformly small. In particular, we can find a T sufficiently large (and random) such that

$$\max_{k \geq T} \big(|M_k - M_T| + \sum_T^k R_j \big) \leq \eta\delta/(4\|\nabla F\|_\infty).$$

Likewise, if we perform a martingale decomposition of X_k, we can write

$$X_k - X_T = \sum_{j=T}^{k-1} -\gamma_j(\nabla F(X_j) + \xi_j) =: \sum_{j=T}^{k-1} -\gamma_j \nabla F(X_j) + (Z_k - Z_T),$$

for a martingale $(Z_k : k)$. By martingale convergence, we can also ensure T is long enough that $\max_k \|(Z_k - Z_T)\| \leq \delta/8$. If $\tau > T$ is a time at which X_k is in $\mathcal{U}_\epsilon$, then

$$|X_k - X_\tau| \leq \sum_{\tau+1}^{k} \gamma_j \|\nabla F\|_\infty + \delta/4.$$

Let σ be the first time after τ that the process leaves V_δ. Then

$$\sum_{\tau+1}^{\sigma} \gamma_j \|\nabla F\|_\infty \geq |X_\sigma - X_\tau| - \delta/4 \geq 3\delta/4.$$

Now on this time window, we have

$$F(X_\sigma) - F(X_\tau) \leq -\sum_{\tau+1}^{k} \gamma_j \eta + \eta\delta/(2\|\nabla F\|_\infty) \leq -\eta\delta/(4\|\nabla F\|_\infty).$$

Thus every time that X_k enters $\mathcal{U}_\epsilon$, the objective function $F(X_k)$ must subsequently drop by a fixed amount. As $F(X_k) \xrightarrow{\text{a.s.}} F_\infty$, this is impossible, and hence we have $\|\nabla F(X_k)\|^2 \xrightarrow[k\to\infty]{\text{a.s.}} 0$ on $\mathcal{E}_P$. By compactness of K, we also have that X_k converges to $\mathcal{S}$. □

> **Remark 8 (ODE interpolation):** Another way to argue the convergence above is to show that the iterates asympototically approximate a solution to an ordinary differential equation. The classical Robbins and Monro argument actually uses this. It shows that the path of the algorithm asymptotically almost surely converges to gradient flow, $\frac{d}{dt}\mathcal{X}(t) = -\nabla F(\mathcal{X}(t))$ where we identify $X_k \approx \mathcal{X}(t_k)$ with $t_k = \sum_1^k \gamma_j$. This can only true in

the sense that

$$\lim_{k\to\infty} \max_{n\geq k} \|X_n - \mathcal{X}^{(k)}(t_n)\| = 0$$

where $\mathcal{X}^{(k)}$ is gradient flow with initial condition $\mathcal{X}^{(k)}(t_k) = X_k$.
Refinements of this argument further show SDE approximations, for $X_n - \mathcal{X}^{(k)}(t_n)$. See [KY].

Exercise 7 (Recurrence to Convergence): Suppose that $F \geq 0$ but that ∇F and $\nabla^2 F$ are only continuous (instead of bounded). Suppose however that with γ_k satisfying the Robbins-Monro condition, the process X_k returns infinitely often to some compact set K, with probability 1. Show that F converges to a stationary point of $\mathcal{S} \cap K$.

2.5 *The pessimism of almost sure convergence*

The good part about the Robbins-Monro type condition on $\{\gamma_k\}$ in Theorem 11 is that it does not depend at all on the problem – one is guaranteed convergence with decreasing step-sizes such as $\gamma_k = 1/k$. But convergence is an asymptotic statement, and practically speaking, one must decide at which finite time to stop the algorithm. So rates of convergence, which necessarily depend on the problem setup, are important.

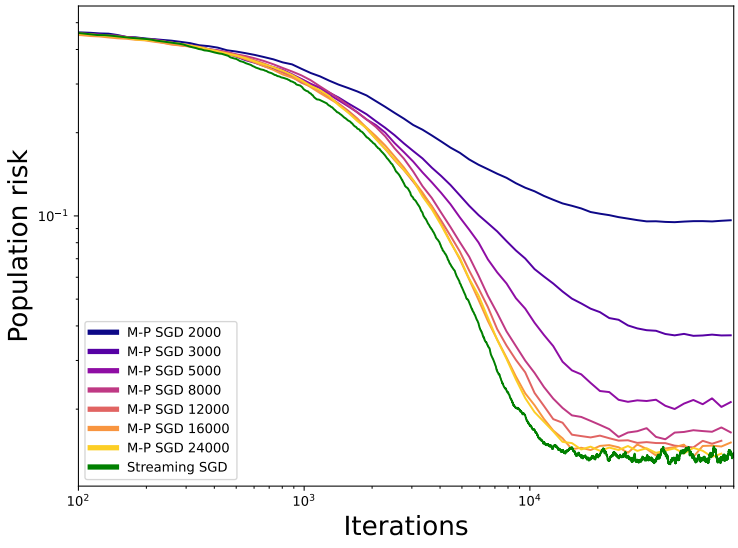


Figure 1: Linear regression (see Example 3). Constant step-size SGD with step-size within a factor of 3 of the largest stable step-size. Fixed dimension $d = 2000$, identity data covariance. Increasing numbers of samples, with multipass SGD. Streaming is the "infinite data" version.

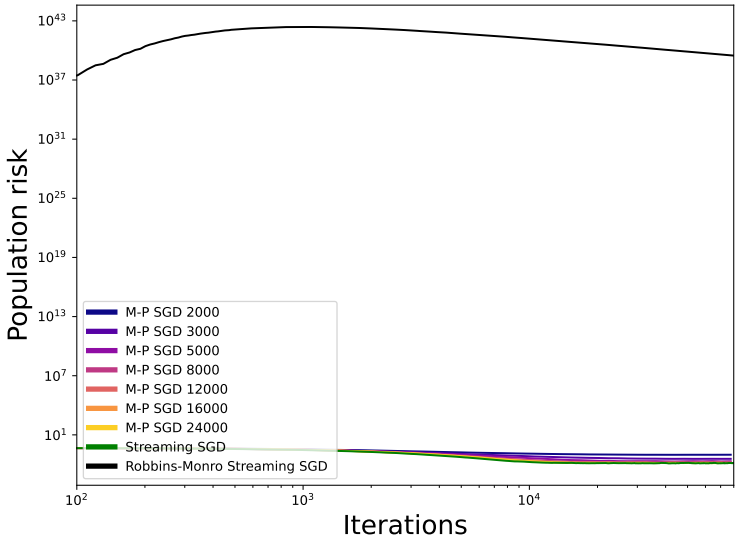


Figure 2: Linear regression (see Example 3). Same setup as Figure 1 with one additional curve, the Robbins-Monro stepsize $\gamma_k = 1/k$. By Theorem 11, the black curve converges. Rescaling the step-size (for example dividing by d) gives a curve which is effectively constant over the same time-scale.

Furthermore, in high-dimensional settings such as those displayed in Figures 1, it is important to account for the magnitude of the gradients. Moreover, in dimension-independent terms, the additional errors incurred from simply picking a constant step-size may be small, as measured by the risk. On the other hand, constant step-size SGD *may* not converge, as if the noise generated by SGD does not vanish, one may have a non-degenerate stationary distribution.

The overarching goal of these notes are to develop the mathematics behind the figures presented here, and in particular to formulate an algorithmic analysis which is accurate in high dimensions.

3 *High-dimensional limits: streaming SGD in the case autonomous order parameters*

In the previous section, we saw an example of a simple high-dimensional (high measured in the thousands) linear regression problem where the Robbins-Monro step-size schedule performed poorly and a constant step-size performed better. The Robbins-Monro schedule paid no heed to the underlying problem parameters, and indeed for a fair comparison one could add problem dependent constants (for example see [KNS16]). However, in the example given, the unavoidable conclusion is the step-size is just too slow, and one possible explanation is that the strategy paid no attention to the dimensionality of the problem.

To conceptualize what it means for the dimension to be large, however, we need to change the dimension and understand its effects. Doing this reveals some important lessons: The most significant

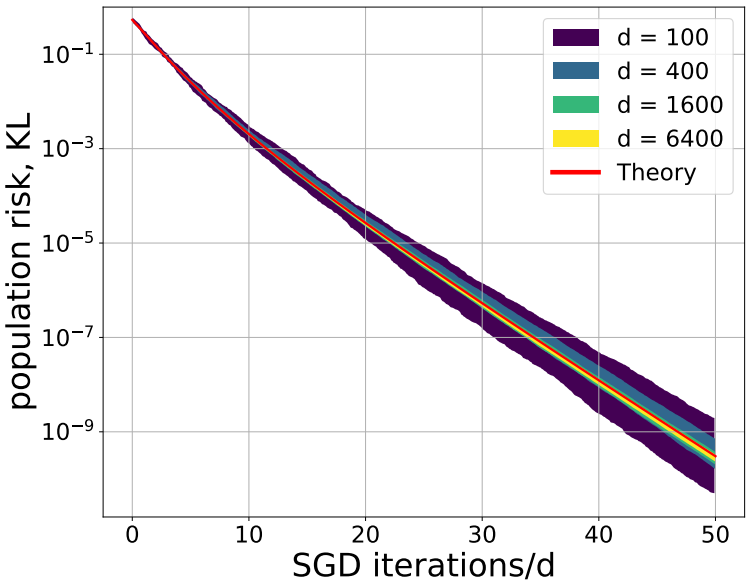


Figure 3: Population risk of logistic regression (see Example 5). In each dimension, 10 runs of streaming SGD for logistic regression are performed. We then display 80% confidence intervals over time (i.e. we discard the largest and smallest at error at each point in time). The curves concentrate around a high-dimensional limit value. Note that time is scaled by dimension. In the isotropic case, this risk curve follows an autonomous ODE. (This in fact is non-isotropic, for which there is a Volterra curve, similar to those discussed in the next section.)

observation is that the risk curve concentrates around a dimension-independent limit. Moreover, this curve depends in a nontrivial way on the stepsize.

In other words, there is some dynamical system hiding in plain sight, such that on sending dimension to infinity, the risks are described by this dynamical system. We begin by illustrating this with a simple example.

Example 8: Isotropic Gaussian Linear Regression

We follow Example 3, and run streaming SGD on it. We suppose the data distribution $\mathcal{D}$ is such that $(a, b) \sim \mathcal{D}$ means

$$a \overset{\text{law}}{=} N(0, \mathrm{Id}), \quad \epsilon \overset{\text{law}}{=} N(0, \eta^2 \mathrm{Id}), \quad b = \langle a, \beta \rangle + \epsilon,$$

where $\beta = \beta$ will be a vector in $\mathbb{R}^d$ of norm 1. The loss is $\ell(x, u, v) = \frac{1}{2}(u - v)^2$ and the risk $\mathscr{P}$ is given by

$$\mathscr{P}(x) = \tfrac{1}{2}(\eta^2 + \|\beta - x\|^2).$$

Streaming SGD on this problem is given by

$$x_{k+1} = x_k - \gamma_k(\langle x_k - \beta, a_{k+1} \rangle - \epsilon_{k+1})a_{k+1}.$$

Now to perform an analysis of this, we will look for a way to describe the limit as dimension tends to infinity of the risk of SGD over time. The good starting point for this type of analysis is to compute the evolution in time of the expected risk of SGD. It will turn out to be enough to compute the mean and covariance matrix of the updates SGD.

Remark 9 (Tensor formalism): It will be helpful when working with high-dimensional limits to use tensor representations, as even for algorithms which only involve matrix-vector products, one is forced to consider higher tensors. We can naturally identify matrices $M = M_{i,j}$ with 2-tensors (see Section 1.1). The covariance matrix of a random vector a is then identified with

$$\mathbb{E} a \otimes a.$$

The norm-squared of a vector x can be alternatively represented, using the contraction operator as

$$\|x\|^2 = \langle x, x \rangle = \langle x \otimes x, \mathrm{Id} \rangle = \mathrm{Tr}(x \otimes x).$$

A quadratic form of x and a matrix A can be represented by

$$x^t A x = \mathrm{Tr}(A x x^T) = \langle A, x \otimes x \rangle.$$

Example 9: Dynamical analysis of Isotropic Gaussian Linear Regression

Let $\mathscr{F}_k$ be the σ-algebra generated by $((a_j, b_j) : 0 \leq j \leq k)$. The conditional mean and conditional variance of this update

are given by

$$\mathbb{E}[(\langle x_k - \beta, a_{k+1}\rangle - \epsilon_{k+1})a_{k+1} \mid \mathscr{F}_k] = x_k - \beta = \nabla \mathscr{P}(x_k),$$

and the covariance matrix (see Exercise 9 below) is given by

$$\begin{aligned}
&\mathbb{E}[(\langle x_k - \beta, a_{k+1}\rangle - \epsilon_{k+1})^2 a_{k+1} \otimes a_{k+1} \mid \mathscr{F}_k] \\
&= \left(\mathbb{E}[(\langle x_k - \beta, a_{k+1}\rangle)^2 \mid \mathscr{F}_k] + \eta^2\right) \mathrm{Id} + 2((x_k - \beta) \otimes (x_k - \beta)) \\
&= 2\mathscr{P}(x_k)\,\mathrm{Id} + 2((x_k - \beta) \otimes (x_k - \beta)).
\end{aligned}$$

It will turn out the correct way to view this in high-dimensions is as a principal term (the first one) and a lower order correction, i.e.

$$\mathbb{E}[(\langle x_k - \beta, a_{k+1}\rangle - \epsilon_{k+1})^2 a_{k+1} \otimes a_{k+1} \mid \mathscr{F}_k] \approx 2\mathscr{P}(x_k)\,\mathrm{Id}\,.$$

Suppose that we consider the evolution of the risk itself under SGD, which is to say we consider the update

$$\mathscr{P}(x_{k+1}) - \mathscr{P}(x_k) = \tfrac{1}{2}(\|x_{k+1} - x_k\|^2 + 2\langle x_{k+1} - x_k, x_k - \beta\rangle).$$

If we set $\mathscr{R}(x) := \frac{1}{2}\|\beta - x\|^2$, then computing the conditional expectation, we arrive at

$$\begin{aligned}
&\mathbb{E}[\mathscr{R}(x_{k+1}) - \mathscr{R}(x_k) \mid \mathscr{F}_k] \\
&= \tfrac{\gamma_k^2}{2}\,\mathrm{Tr}\big(2\mathscr{P}(x_k)\,\mathrm{Id} + 2((x_k - \beta) \otimes (x_k - \beta))\big) \\
&- \gamma_k \langle x_k - \beta, x_k - \beta\rangle \\
&= -2\gamma_k \mathscr{R}(x_k) + \gamma_k^2(d\mathscr{R}(x_k) + d\eta^2/2 + \mathscr{R}(x_k)).
\end{aligned}$$

The major factor to consider in this recurrence is the d which appears in the γ_k^2 term.

For both first and second order terms to survive in a limit, we must take $\gamma_k \asymp \gamma_k^2 d$ (meaning as order of magnitudes in d), which implies that $\gamma_k \asymp 1/d$. Moreover to achieve a non-degenerate limit, we should set $\gamma_k = \gamma(k/d)/d$ for a continuous function $\gamma(\cdot)$. If we set $\rho(t) = \lim_{d\to\infty} \mathbb{E}[\mathscr{R}(x_{[td]})]$ then the above equation becomes an Euler approximation for the ordinary differential equation

$$\dot{\rho} = -2\gamma(t)\rho + \gamma^2(t)(\rho + \eta^2/2).$$

Remark 10 (Risk curve and stability): This can be solved explicitly. In the case of $\eta \equiv 0$, it is

$$\rho(t) = \rho(0) \exp\left(- \int_0^t (2\gamma(s) - \gamma^2(s)) \, \mathrm{d}s \right).$$

Note that for constant $\gamma(t) \equiv \gamma$, the curve is convergent if and only if $\gamma < 2$ and bounded if and only if $\gamma \leq 2$, (which can also be reasoned just from the ODE). Note further the risk tends to 0. In the case $\eta > 0$ and constant $\gamma < 2$ the risk does not tend to 0, but we can further solve for the limiting risk as the stationary point of ρ by setting $\dot{\rho} = 0$:

$$\rho(\infty) = \frac{\gamma^2 \eta^2}{4\gamma - 2\gamma^2}.$$

Exercise 8 (Concentration): Show using martingale concentration that the difference of $\mathscr{R}(x_{[td]})$ and $\mathbb{E}\mathscr{R}(x_{[td]})$ tends to 0 in probability as $d \to \infty$ for any fixed $t > 0$.

Exercise 9 (Wick rule computations): The *Wick rule* gives a quick way to compute expectations of tensors formed from Gaussians. Suppose $a \overset{\text{law}}{=} N(0, K)$. For a simple tensors f_i for $1 \leq i \leq 4$,

$$\begin{aligned}
&\mathbb{E}\langle a^{\otimes 4}, f_1 \otimes f_2 \otimes f_3 \otimes f_4 \rangle \\
&= \langle K, f_1 \otimes f_2 \rangle \langle K, f_3 \otimes f_4 \rangle \\
&+ \langle K, f_1 \otimes f_3 \rangle \langle K, f_2 \otimes f_4 \rangle \\
&+ \langle K, f_1 \otimes f_4 \rangle \langle K, f_2 \otimes f_3 \rangle.
\end{aligned}$$

This extends by multilinearity to 4-tensors B by

$$\begin{aligned}
&\mathbb{E}\langle a^{\otimes 4}, B \rangle \\
&= \langle K, \langle K, B \rangle_{1,2} \rangle_{3,4} \\
&+ \langle K, \langle K, B \rangle_{1,3} \rangle_{2,4} \\
&+ \langle K, \langle K, B \rangle_{1,4} \rangle_{2,3},
\end{aligned}$$

where $\langle \cdot, \cdot \rangle_{a,b}$ refers to contracation along axes a and b. Show that

$$\begin{aligned}
&\mathbb{E}(\langle a, y \rangle^2 \langle a^{\otimes 2}, \mathrm{Id}_m \rangle) \\
&= \mathbb{E}\langle a^{\otimes 4}, y \otimes y \otimes \mathrm{Id}_m \rangle \\
&= \langle K, y \otimes y \rangle \langle K, \mathrm{Id}_m \rangle + 2 \langle \langle K, y \rangle \otimes \langle K, y \rangle, \mathrm{Id}_m \rangle \\
&= y^t K y \, \mathrm{Tr}(K) + 2 y^t K^2 y.
\end{aligned}$$

3.1 *Hidden finite dimensional risk manifold*

The key to the previous example (Example 8) was that the equation for the dynamics of the risk was autonomous: the evolution of the risk depends only on the current value of the risk. This generalizes the situation seen in Example 9, in which the risk itself describes an autonomous evolution.

Definition 27 (Hidden risk manifold): Say that family of empirical risk minimization problems, indexed by model dimensionality d, lie on a *hidden risk manifold* of dimension k if for any d there are C^2 functions $u^{(d)} : \mathbb{R}^d \to \mathbb{R}^k$ and $F_1, F_2 : \mathbb{R}^k \to \mathbb{R}^k$ with:

1. $u_1^{(d)}(x) = \mathscr{P}(x)$ and $\mathscr{P}$ is uniformly coercive:

$$\lim_{\|x\| \to \infty} \liminf_{d \to \infty} u_1^{(d)}(x) = \infty.$$

2. F_1, F_2 are continuous functions;
3. with $\gamma_k \equiv \gamma / d$

$$\| d\mathbb{E}[u(x_1) - u(x_0) \mid \mathscr{F}_0] + \gamma F_1(u(x_0)) - \gamma^2 F_2(u(x_0)) \| \to 0$$

uniformly on compact sets of $\|x_0\|$ as $d \to \infty$ for fixed γ;

4. with $\gamma_k \equiv \gamma / d$

$$d\mathbb{E}[\|u(x_1) - u(x_0)\|^2 \mid \mathscr{F}_0] \to 0$$

uniformly on compact sets of $\|x_0\|$ as $d \to \infty$ for fixed γ.

Remark 11 (Origins of the hidden risk manifold): This is an adaptation of formulation of [BAGJ22], which contains, in addition, some notable worked examples and further theoretical elaborations. While not formalized in this way, some of the ideas of this limit appear in the earlier in the work of [SS95]. This idea has also appeared [Vei+22], [Arn+23], and [AGJ21].

The notion has appeared in the physics literature, where it is described as the closure of the equations of motion for the order parameters [Gol+20]. In situations where the data covariance is non-identity, this procedure usually has trouble, [Gol+20; YO19]. In Section 4 we we show one way to handle this.

As a first central example, the Isotropic Gaussian satisfies these assumptions.

Example 10: Isotropic Gaussians satisfy HRM

We only need a single observable:

$$u(x) = \mathscr{P}(x) = \tfrac{1}{2}\|\beta - x\|^2 + \tfrac{1}{2}\eta^2.$$

This risk is clearly uniformly coercive. The computations in Example 9 show that these ERMs satisfy Part 2 and 3 of Definition 27 with

$$F_1(u) = 2u - \eta^2 \quad \text{and} \quad F_2(u) = u.$$

Finally it can be checked that for some constant $C(\|x_0\|)$

$$\mathbb{E}[\|u(x_1) - u(x_0)\|^2 \mid \mathscr{F}_0] \le C(\|x_0\|)\gamma^2/d^2.$$

Theorem 12: Hidden risk manifold

Suppose a family of empirical risk minimization problems have a hidden risk manifold and $\{x_k^{(d)}\}$ is streaming SGD on these problems. Suppose the initialization satisfies $u(x_0) \to \mu_0$ as $d \to \infty$. Let μ be the solution of the initial value problem on $\mathbb{R}^k$

$$\dot{\mu} = -\gamma F_1(\mu) + \gamma^2 F_2(\mu), \quad \mu(0) = \mu_0.$$

Suppose that the solution of this IVP exists for all time. Uniformly on compact sets of t

$$u(x_{[td]}) \xrightarrow[d\to\infty]{\Pr} \mu(t).$$

Proof. Fix an $R > 0$ and let τ_R be the first time k the norm of x_k exceeds R in norm, i.e.

$$\tau_R = \inf\{k : \|u(x_k)\| > R\}.$$

It suffices to show that uniformly on compact sets of time [11]

$$u(x_{[td]}^{\tau_R}) \xrightarrow[d\to\infty]{\Pr} \mu^{\sigma_R}(t),$$

where σ_R is the first time t that $\|\mu(t)\| > R$. Having shown this, and since $\sigma_R \to \infty$ as $R \to \infty$, it follows that $\tau_R \to \infty$ in probability as $d \to \infty$ followed by R, i.e. for any M

$$\lim_{R\to\infty} \limsup_{d\to\infty} \Pr(\tau_R \le M) = 0.$$

Thus, we will have shown the claimed convergence of u to μ without stopping.

[11] The process x_k^τ refers to the *stopped process*, given by $x_k^\tau = x_{k\wedge\tau}$. Likewise μ^τ is run to the first time $t > \tau$ at which point it is frozen.

Now for a given R from the Part 1 of Definition 27 that $\mathscr{P}$ is uniformly coercive, we have that there is an M sufficiently for all d sufficiently large $\|x_k^{\tau_R}\| \leq M$ for all $k < \tau_R$. From Part 4 of Definition 27, we have that $\|x_k^{\tau_R}\| \leq M+1$ even at the final time (where the process x_k can jump outside the ball, but by the moment bound given can only jump a little with probability going to 1) with probability going to 1 as $d \to \infty$. For any t, we perform a Doob decomposition of u up to time $\ell \leq [td]$, which gives

$$u(x_\ell) = u(x_0) + \sum_{k=0}^{\ell-1} \mathbb{E}[u(x_{k+1}) - u(x_k) \mid \mathscr{F}_k] + M_\ell.$$

From Part 4 of Definition 27, we have from Doob's inequality and Doob's L^2–maximal inequality

$$\max_{0\leq k\leq \ell} \|M_k\| \xrightarrow[d\to\infty]{\Pr} 0.$$

Hence From Part 3 of Definition 27

$$\max_{0\leq \ell\leq td} \| - u(x_\ell) + u(x_0) + \frac{1}{d}\sum_{k=0}^{\ell-1}\{-\gamma F_1(u(x_k)) + \gamma^2 F_2(u(x_k))\}\| \xrightarrow[d\to\infty]{\Pr} 0.$$

This is now a uniform approximation of the IVP in the statement of the theorem. The theorem follows from Gronwall's inequality and Part 2 of Definition 27. □

Example 11: GLMs with isotropic features

We suppose that we have GLM (Example 5) in a student-teacher format. That is, suppose that we have have $M(x,a) = \phi(\langle x, a\rangle)$ and suppose $\beta \in \mathbb{R}^d$ is given and has unit norm. Suppose we have a data distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$ where $(a,b) \sim \mathcal{D}$ means

$$b = M(\beta, a) \quad \text{and} \quad a \overset{\text{law}}{=} N(0, \mathrm{Id}_d).$$

Now we suppose the loss $\ell(x,u,v) = \ell(u,v)$ is given and is C^1 with derivative bounded uniformly in norm.

The population risk is given by

$$\mathscr{P}(x) := \mathbb{E}\ell(M(x,a), M(\beta,a)),$$

and constant step-size streaming SGD is given by

$$\begin{aligned} x_{k+1} &= x_k - \tfrac{\gamma}{d}\nabla_x \ell(M(x_k, a_{k+1}), M(\beta, a_{k+1})) \\ &= x_k - \tfrac{\gamma}{d} a_{k+1}\phi'(\langle x_k, a_{k+1}\rangle)\ell_u(\phi(\langle x_k, a_{k+1}\rangle), \phi(\langle \beta, a_{k+1}\rangle)). \end{aligned}$$

We observe that

$$\nabla_x \mathscr{P}(x) = \mathbb{E}\nabla_x \ell(M(x,a), M(\beta,a)).$$

This is an expectation over a two-dimensional Gaussian distribution $(\langle x,a\rangle, \langle \beta, a\rangle)$. Thus, this can be computed from two covariances

$$\langle x,\beta\rangle = \mathbb{E}(\langle x,a\rangle\langle\beta,a\rangle) \quad \text{and} \quad \langle x,x\rangle = \mathbb{E}(\langle x,a\rangle\langle x,a\rangle).$$

Now under suitable assumptions ($\mathscr{P}$ being coercive, ℓ, ϕ being sufficiently bounded), it can be verified that this pair of observables determines the entire evolution of the system, i.e.

$$u(x) = \big(\mathscr{P}(x), \langle x,x\rangle, \langle x,\beta\rangle\big)$$

is a hidden risk manifold.

Exercise 10 (Smooth phase retrieval): In the case that $\phi(x) = x^2$ and $\ell(u,v) = \frac{1}{2}(u-v)^2$, find F_1 and F_2.

Example 12: The Saad-Solla neural network

Following [SS95], consider a setup in which for a 2-tensor $x \in \mathbb{R}^m \otimes \mathbb{R}^p$

$$M_p(x,a) := g(\langle x,a\rangle_m) \quad \text{where} \quad g(x) = \operatorname{erf}(x/\sqrt{2})$$

and suppose one considers the student-teacher setup in which $a \overset{\text{law}}{=} N(0, \mathrm{Id}_m)$ and mean-squared error loss $\ell(u,v) = \frac{1}{2}(u-v)^2$. The number of hidden units in the student and teacher layers are different and given by p, q respectively. Hence the risk is given by

$$\mathscr{P}(x) = \mathbb{E}\ell(M_p(x,a), M_q(\beta,a)).$$

This can be evaluated explicitly in terms of the correlation matrices

$$Q = \langle x,x\rangle_m, \quad T = \langle \beta,\beta\rangle_m, \quad R = \langle x,\beta\rangle_m.$$

These correlation matrices. For a 2-tensor $A \in \mathcal{V}^{\otimes 2}$, setting $\mathfrak{D}(A)$ to be the vector $(1/\sqrt{1+A_{ii}} : 1 \le i \le \dim(\mathcal{V}))$,

$$\begin{aligned}\mathscr{P}(x) = \tfrac{1}{\pi}\big(&\operatorname{Tr}\arcsin(Q \otimes (\mathfrak{D}(Q) \otimes \mathfrak{D}(Q)))\\ &+ \operatorname{Tr}\arcsin(T \otimes (\mathfrak{D}(T) \otimes \mathfrak{D}(T)))\\ &- 2\operatorname{Tr}\arcsin(R \otimes (\mathfrak{D}(Q) \otimes \mathfrak{D}(T)))\big),\end{aligned}$$

with the arcsin applied entrywise. Moreover, the triple $(\mathscr{P}, Q, T, R)$ form a hidden risk manifold. See [SS95] for a qualitative discussion of the resulting ODEs.

4 *High dimensional analysis of streaming SGD on the correlated least squares problem*

This is adapted from the article [CP23b]. Portions adapted from [Paq+22b], [Paq+22a], and [Paq+21].

A unifying theme of the examples in the previous section were that (1) the data were isotropic Gaussian $N(0, \mathrm{Id}_m)$ and (2) the risks could be described by a family of order parameters related to correlations $\mathbb{E}\langle a, x\rangle_m \otimes \langle a, x\rangle_m$ (and when relevant $\mathbb{E}\langle a, x\rangle_m \otimes \langle a, \beta\rangle_m$). The reliance on isotropic Gaussian data calls into question to what extent this theory could ever apply to more involved setups. So we may wish to generalize it, which brings us to point (2): if we look at the case of correlated data, are there still hidden variables which describe the evolution of risk?

Example 13: Nonisotropic linear regression requires unboundedly many statistics

We now suppose the data distribution $\mathcal{D}$ is such that $(a, b) \sim \mathcal{D}$ means

$$a \overset{\text{law}}{=} N(0, K), \quad \epsilon \overset{\text{law}}{=} N(0, \eta^2 \mathrm{Id}), \quad b = \langle a, \beta\rangle + \epsilon,$$

where $\beta = \beta$ will be a vector in $\mathbb{R}^d$ of norm 1. The loss is $\ell(x, u, v) = \frac{1}{2}(u - v)^2$ and the risk $\mathscr{P}$ is given by

$$\mathscr{P}(x) = \tfrac{1}{2}(\eta^2 + \langle K, (\beta - x)^{\otimes 2}\rangle).$$

The risk now satisfies a 1-step update given by

$$\begin{aligned}&\mathbb{E}[\mathscr{P}(x_{k+1}) - \mathscr{P}(x_k) \mid \mathscr{F}_k] \\ &= \tfrac{\gamma_k^2}{2} \operatorname{Tr}\big(2\mathscr{P}(x_k)K + 2(K(x_k - \beta) \otimes L(x_k - \beta))\big) \\ &- \gamma_k \langle K^2, (x_k - \beta)^{\otimes 2}\rangle)\end{aligned}$$

Now, unfortunately, the gradient descent term (meaning that which is linear in γ_k is no longer just the risk $\mathscr{P}$, owing to the presence of the K^2. One may attempt to add $u_2(x) := \langle K^2, (x_k - \beta)^{\otimes 2}\rangle)$, but when considering its evolution under SGD, this just leads to a gradient descent term $\langle K^3, (x_k - \beta)^{\otimes 2}\rangle$. So there is not a finite family of statistics that can be used to autonomously describe the evolution.

So we need another framework for describing the high-dimensional limit dynamics, beyond what has already been presented. We shall put the following assumptions on the data. (See Section 1.5).

Assumption 1 (Data assumptions): A sample (a, b) from the distribution $\mathcal{D}$ satisfies the following:

1. That data a is centered and has covariance matrix $K := \mathbb{E}a \otimes a$ which has operator-norm bounded independent of d.
2. The data satisfies a Hanson-Wright type inequality: for all $t \geq 0$ and for any deterministic matrix B
$$\Pr\left(\left|a^T Ba - \mathbb{E}a^T Ba\right| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{t^2 d^{-4\varepsilon}}{\|B\|^2}, \frac{td^{-2\varepsilon}}{\|B\|_\sigma}\right\}\right).$$
3. Conditionally on a, the distribution of b is given by $\langle a, \beta\rangle + \eta w$ where w is mean 0, variance 1 and is subgaussian with $\|w\|_{\psi_2} \leq d^\varepsilon$.
4. The ground truth β is assumed to have norm at most d^ε.

Throughout this section we shall only discuss streaming SGD for the least squares problem with constant step-size $\gamma_k \equiv \gamma/d$. Hence the the iterates are given, in terms of a stream of data $(a_j, b_j)_1^\infty$, by

$$x_k - \beta = (\mathrm{Id}_d - \tfrac{\gamma}{d} a_k a_k^T)(x_{k-1} - \beta) + \tfrac{\gamma}{\sqrt{d}}\eta w_k, \tag{31}$$

where $(w_j)_1^\infty$ are the standardized noises in the targets.

Homogenized SGD for streaming linear regression. To accomplish this task, we introduce an idealized process which captures the large-dimensional behvior. Homogenized SGD is defined to be a continuous time process with initial condition $\mathbf{X}_0 = x_0$ that solves the stochastic differential equation

$$\mathrm{d}\mathbf{X}_t = -\gamma\nabla\mathscr{P}(\mathbf{X}_t)\,\mathrm{d}t + \gamma\sqrt{\tfrac{2}{d}\mathscr{P}(\mathbf{X}_t)K}\,\mathrm{d}B_t \tag{32}$$

where B_t is standard Brownian motion in dimension d, where we rcall $\mathscr{P}$ is the population risk:

$$\mathscr{P}(x) := \tfrac{1}{2}\mathbb{E}_{(a,b)}(\langle a, x\rangle - b)^2, \quad (a, b) \sim \mathcal{D}. \tag{33}$$

We will formulate a comparison theorem between $\mathbf{X}_t$ and x_k. To do so, we use the following probabilistic notion:

Definition 28 (Overwhelming probability): We use the probabilistic modifier *with overwhelming probability* to mean a statement holds except on an event of probability at most $e^{-\omega(\log d)}$ where $\omega(\log d)$ tends to ∞ faster than $\log d$ as $d \to \infty$.

To quantify the growth of functions, we use the following:

Definition 29 (C2 norm): Define $\|\cdot\|_{C^2}$ on functions $q : \mathbb{R}^d \to \mathbb{C}$

$$\|q\|_{C^2} := \sup_x \|\nabla^2 q(x)\|_\sigma + \|\nabla q(0)\| + |q(0)|,$$

with the norms on the right hand side being given by the operator and Euclidean norm respectively.

Our main theorem is given by the following:

Theorem 13: Streaming SGD limit

Suppose the data satisfies Assumption 1. For any quadratic $q : \mathbb{R}^d \to \mathbb{R}$, and for any deterministic initialization x_0 with $\|x_0\| \leq 1$, there is a constant $C(\|K\|_\sigma)$ so that the processes $\{x_k\}_{k=0}^n$ and $\{\mathbf{X}_t\}_{t=0}^{n/d}$ satisfy for any n satisfying $n \leq d \log d / C(\|K\|_\sigma)$

$$\sup_{0 \leq k \leq n} \left| q(x_k) - q(\mathbf{X}_{k/d}) \right| < \|q\|_{C^2} \cdot e^{C(\|K\|_\sigma)\frac{n}{d}} \cdot d^{-\frac{1}{2}+9\varepsilon} \tag{34}$$

with overwhelming probability.

The processes x_k and $\mathbf{X}_t$ are independent, and hence this is also a statement about concentration. In particular, the statement is also true if we replace $q(\mathbf{X}_{k/d})$ by $\mathbb{E}q(\mathbf{X}_{k/d})$.

4.1 *Explicit risk curves*

Unlike results from the previous section, this is not quite a complete solution to describing the limiting risk curves in the high-dimensional limit. Indeed, this process still exists in a d-dimensional space and not in a space of dimension independent of d. So to find the risk curves, we still have an argument to do.

The main idea we use here is to consider the complex curve, with $R(z; K)$ given by the resolvent (see Section 1.2 for a discussion of resolvent properties that we use)

$$Q_t(z) := \tfrac{1}{2}\langle R(z; K), (\mathbf{X}_t - \beta) \otimes (\mathbf{X}_t - \beta)\rangle \quad z \in \mathbb{C}. \tag{35}$$

It will suffice to consider $Q_t(z)$ on a curve $\Gamma \subset \mathbb{C}$ that encloses the spectrum of K. As we have supposed that K has an operator norm independent of d, we can suppose that this curve is independent of d and encloses a 1-neighborhood of all eigenvalues of K.

Now from Cauchy's integral formula (see the spectral mapping theorem), we have

$$\mathscr{P}(\mathbf{X}_t) := \tfrac{1}{2}\langle K, (\mathbf{X}_t - \beta) \otimes (\mathbf{X}_t - \beta)\rangle = \frac{-1}{2\pi i} \oint_\Gamma z Q_t(z)\, \mathrm{d}z. \tag{36}$$

Hence, the risk can be extracted from $Q_t(z)$. Now on the other hand, applying Itô's formula

$$\begin{aligned} \mathrm{d}Q_t(z) &= -\gamma\langle KR(z;K), (\mathbf{X}_t-\beta)\otimes(\mathbf{X}_t-\beta)\rangle\,\mathrm{d}t \\ &\quad + \gamma\langle R(z;K), (\mathbf{X}_t-\beta)\otimes\sqrt{\tfrac{2K\mathscr{P}(\mathbf{X}_t)}{d}}dB_t\rangle \\ &\quad + \tfrac{\gamma^2\mathscr{P}(\mathbf{X}_t)}{d}\langle R(z;K), \sqrt{K}dB_t\otimes\sqrt{K}dB_t\rangle. \\ &= -\gamma\langle zR(z;K)+\mathrm{Id}_d, (\mathbf{X}_t-\beta)\otimes(\mathbf{X}_t-\beta)\rangle\,\mathrm{d}t + \mathrm{d}M_t(z) \\ &\quad + \tfrac{\gamma^2\mathscr{P}(\mathbf{X}_t)}{d}\langle R(z;K), K\,\mathrm{Id}_d\rangle\,\mathrm{d}t. \end{aligned} \tag{37}$$

The process $dM_t(z)$ is the martingale term, i.e. all those terms linear in dB_t. In summary

$$\mathrm{d}Q_t(z) = -2\gamma z Q_t(z)\,\mathrm{d}t + \tfrac{\gamma^2\mathscr{P}(\mathbf{X}_t)}{d}\,\mathrm{Tr}(KR(z;K))\,\mathrm{d}t - \gamma\|\mathbf{X}_t-\beta\|^2\,\mathrm{d}t + \mathrm{d}M_t(z).$$

Hence using an integrating factor, we have

$$\mathrm{d}(e^{2\gamma z t}Q_t(z)) = e^{2\gamma z t}\tfrac{\gamma^2\mathscr{P}(\mathbf{X}_t)}{d}\,\mathrm{Tr}(KR(z;K))\,\mathrm{d}t - \gamma e^{2\gamma z t}(\|\mathbf{X}_t-\beta\|^2\,\mathrm{d}t + \mathrm{d}M_t(z)).$$

This can be solved explicitly to give

$$Q_t(z) = Q_0(z)e^{-2\gamma z t} + \int_0^t e^{-2\gamma z(t-s)}\tfrac{\gamma^2\mathscr{P}(\mathbf{X}_s)}{d}\,\mathrm{Tr}(KR(z;K))\,\mathrm{d}s + \mathrm{E}_t(z).$$

The term $\mathrm{E}_t(z)$ is an error term containing both terms which will vanish in subsequent steps and a martingale term which we must show vanishes (owing to the extra factor of $\sqrt{d}$ that it carries). From this, we can extract the risk $\mathscr{P}(\mathbf{X}_t)$ by integrating over Γ. Specifically using (36)

$$\mathscr{P}(\mathbf{X}_t) = \frac{-1}{2\pi i}\oint_\Gamma z\left(Q_0(z)e^{-2\gamma z t} + \int_0^t e^{-2\gamma z(t-s)}\tfrac{\gamma^2\mathscr{P}(\mathbf{X}_s)}{d}\,\mathrm{Tr}(KR(z;K))\,\mathrm{d}s + \mathrm{E}_t(z)\right)\mathrm{d}z. \tag{38}$$

Each of these terms we integrate separately.

Gradient flow term. For the first term, $zQ_0(z)e^{-2\gamma z t}$, we can identify it as a function of gradient flow.

> **Definition 30 (Gradient Flow):** Gradient flow $(\mathscr{X}_t : t \geq 0)$ on the objective function $\mathscr{P}$ with initialization X is the solution of the ODE
>
> $$\dot{\mathscr{X}}_t = -\nabla\mathscr{P}(\mathscr{X}_t)$$
>
> with initial state $\mathscr{X}_0 = X$.

In the least-squares problem, this can be explicitly solved, which yields:

Lemma 7 (Least squares gradient flow): If $\mathscr{P}(x) = \frac{1}{2}\langle K, (x-\beta)^{\otimes 2}\rangle + \eta^2/2$ and initial state of gradient flow of SGD is X, then

$$\mathscr{X}_t - \beta = e^{-tK}(X-\beta).$$

Proof. From uniqueness of the gradient flow ODE, it suffices to simply verify that

$$\dot{\mathscr{X}}_t = -K(\mathscr{X}_t - \beta) = \nabla\mathscr{P}(\mathscr{X}_t)$$

and that at initialization $\mathscr{X}_0 = X$. □

From spectral mapping, we have

$$\begin{aligned}\frac{-1}{2\pi i}\oint_\Gamma zQ_0(z)e^{-2\gamma zt}\,\mathrm{d}z &= \frac{1}{2}\left\langle \frac{-1}{2\pi i}\oint_\Gamma ze^{-2\gamma zt}R(z;K)\,\mathrm{d}z, (\mathbf{X}_0-\beta)^{\otimes 2}\right\rangle \\ &= \frac{1}{2}\langle Ke^{-2\gamma Kt}, (\mathbf{X}_0-\beta)^{\otimes 2}\rangle \\ &= \frac{1}{2}\langle K, (\mathscr{X}_{\gamma t}-\beta)^{\otimes 2}\rangle.\end{aligned}$$

Thus the first terms is precisely the risk of gradient flow run from the same initialization. The noise term (which is quadratic in γ) can again by spectral mapping can be identified, from which (38) can be expressed as

$$\mathscr{P}(\mathbf{X}_t) = \mathscr{P}(\mathscr{X}_{\gamma t}) + \int_0^t \mathrm{Tr}(K^2e^{-2\gamma K(t-s)})\frac{\gamma^2\mathscr{P}(\mathbf{X}_s)}{d}\,\mathrm{d}s - \frac{1}{2\pi i}\oint_\Gamma z\mathrm{E}_t(z)\,\mathrm{d}z.$$

Hence we introduce the *Volterra model* for the risk by

Definition 31 ((Finite-dimensional) Volterra risk model): Let $\mathscr{X}_t$ be the path of gradient flow started from initialization $\mathbf{X}_0$. Let $\mathcal{K}_\gamma$ be the function from $[0,\infty) \to [0,\infty)$ given by

$$\mathcal{K}_\gamma(t) := \gamma^2\frac{\mathrm{Tr}(K^2e^{-2\gamma Kt})}{d}.$$

Then the Volterra risk model is the solution of the convolution-type Volterra equation

$$\Psi(t) := \mathscr{P}(\mathscr{X}_{\gamma t}) + \int_0^t \mathcal{K}_\gamma(t-s)\Psi(s)\,\mathrm{d}s.$$

After establishing control on the error terms above, we will have shown the following

Theorem 14: Homogenized SGD risk curve

For any $\varepsilon > 0$, any $T > 0$

$$\sup_{0 \le t \le T} |\mathscr{P}(\mathbf{X}_t) - \Psi(t)| < C(T, \|K\|_\sigma) d^{-1/2+\varepsilon}$$

with overwhelming probability.

This is a similar Gronwall inequality argument and uses concentration of Brownian martingales. See [Paq+22a, Theorem 1.1] for details.

4.2 *Optimization implications of the Volterra risk model.*

From here, we can already make some optimization conclusions. We first note that while the comparison between the true risk $\mathscr{P}(x_{td})$ and $\Psi(t)$ only holds in the limit as $d \to \infty$, the curve $\Psi(t)$ exists at each finite d.

The first observation is that $F(\gamma t) := \mathscr{P}(\mathscr{X}_{\gamma t})$ is always decreasing, for all γ and moreover decreases as $t \to \infty$. The limit risk is given by

Lemma 8 (Gradient flow risk): The risk under gradient flow converges $F(\infty) = \eta^2/2$ and moreover converges like

$$F(\gamma t) \le F(\infty) + e^{-2\gamma\lambda(K)t}(F(0) - F(\infty))$$

where $\lambda(K)$ is the smallest positive eigenvalue of K. This is asymptotically correct in that

$$(F(\gamma t) - F(\infty))^{1/t} \to e^{-2\gamma\lambda(K)}.$$

Proof. From Lemma 7, we have the explicit integral curve $\mathscr{X}_t - \beta = e^{-tK}(X - \beta)$, with X the initialization of gradient flow. It follows that

$$F(\gamma t) = \tfrac{1}{2}\langle K e^{-2K\gamma t}, (X-\beta)^{\otimes 2}\rangle + \tfrac{\eta^2}{2}.$$

On taking $t \to \infty$ this inner product converges to 0. The rate of convergence can be quantified in terms of the smallest positive eigenvalue of K, which is given by

$$(F(t) - F(\infty))^{1/t} \to e^{-2\lambda(K)}.$$

We also have the non-asympototic guarantee

$$(F(t) - F(\infty)) \le e^{-2\lambda(K)t}(F(0) - F(\infty)).$$

□

Since the function F always is bounded, the boundedness of the solution of the Volterra model can be stated entirely in terms of the kernel K. One also can deduce rates of convergence, which we give in terms of the *Malthusian exponent*.

Definition 32 (Malthusian exponent): For a convolution Volterra equation, the Malthusian exponent λ^* is given by

$$\lambda^*(\mathcal{K}_\gamma) = \inf\left\{\lambda > 0 : \int_0^\infty e^{2\gamma\lambda t}\mathcal{K}_\gamma(t)\ \mathrm{d}t = 1\right\}$$

if it exists.

Exercise 11 (Malthusian exponent exists at finite d): In finite dimensions (i.e. with $\mathcal{K}_\gamma(t)$ given as in the Volterra model with finite dimensional K), the Malthusian exponent always exists and is always less than the smallest eigenvalue of $\lambda(K)$.

Theorem 15: Volterra model optimization properties

The Volterra risk model Ψ satisfies the following.

1. The risk Ψ remains bounded if and only if $\gamma \leq \frac{2\operatorname{Tr} K}{d}$, and the limiting risk $\Psi(\infty) = F(\infty)(1 - \frac{\gamma d}{2\operatorname{Tr} K})^{-1}$.
2. If for $\gamma < \frac{2\operatorname{Tr} K}{d}$, then $\Psi(t)^{1/t} \to e^{-2\gamma\lambda^*}$.

While the Malthusian exponent is always larger than $\lambda(K)$, this leaves open whether or not $\lambda^*(K)$ is vanishingly close to $\lambda(K)$. To answer this, it is simplest to pass to an infinite dimensional setting.

4.3 *Infinite dimensional Volterra equation*

The Volterra model in Definition 31 still depends on the dimensionality of the underlying problem; it also can be derived without further modelling considerations of the covariances structure d or initialization. It can also be advantageous to derive a true dimension-independent limit, which for example clarifies those γ at which the Malthusian exponent plays a dimension-independent role. To derive a dimension-independent limit, it is enough to suppose that the empirical measure of eigenvalues of K converges to a limit.

Definition 33 (Empirical spectral measure): The empirical spectral measure μ of K is the d-point atomic measure

$$\mu_K(dx) = \frac{1}{d}\sum_{j=1}^{d}\delta_{\lambda_j(K)}(dx)$$

where $\{\lambda_j\}$ are the eigenvalues of K.

If the empirical spectral measure μ_K converges weakly to some compactly supported limit measure μ, and the driving curve $F(t;d) \to F(t;\infty)$ uniformly on compact sets of t, then uniformly on compact sets of time

$$\Psi(t;d) \to \Psi(t;\infty),$$

where the infinite-dimensional Volterra model satisfies the following.

Definition 34 (Infinite Dimensional Volterra Model): The finite dimensional Volterra model with gradient flow risk curve F and spectral measure μ is the solution of

$$\Psi(t) = F(\gamma t) + \int_0^t \mathcal{K}_\gamma(t-s)\Psi(s)\,\mathrm{d}s,$$

where

$$\mathcal{K}_\gamma(t) = \gamma^2 \int_0^\infty x^2 e^{-2\gamma x t}\mu(dx).$$

This generalizes the finite-dimensional model by taking μ to be the empirical spectral measure of K and F given by $\mathscr{P}(\mathscr{X}_t)$.

The convergence analysis Theorem 15 remains true, with $2\operatorname{Tr} K/d = 2\int_0^\infty x\mu(dx)$. In the infinite-dimensional case, the Malthusian exponent of the convolution-Volterra equation may cease to exist. If $\lambda(\mu)$ is the left-edge of the support of μ, then λ^* does not exist for γ such that

$$\frac{\gamma}{2}\int_{\lambda(\mu)}^\infty \frac{x^2}{x-\lambda(\mu)}\mu(x) = \int_0^\infty e^{2t\gamma\lambda(\mu)}K_\gamma(t)\,\mathrm{d}t < 1.$$

This leads to the following infinite-dimensional version of Theorem 15.

Theorem 16: Volterra limit

For a limiting spectral measure μ, the infinite dimensional Volterra risk model Ψ satisfies:

1. The risk Ψ remains bounded if and only if $\gamma \le 2\int_0^\infty x\mu(dx)$, and the limiting risk $\Psi(\infty)$ is given by $F(\infty)(1-\frac{\gamma}{2\int_0^\infty x\mu(dx)})^{-1}$.
2. If for $\gamma < 2\int_0^\infty x\mu(dx)$, the Malthusian exponent exists then $(\Psi(t)-\Psi(\infty))^{1/t} \to e^{-2\gamma\lambda^*}$.
3. If for $\gamma < 2\int_0^\infty x\mu(dx)$, the Malthusian exponent does not exist, the convergence rate (at exponential scale) is the

same as $F(t)$.

Remark 12 (Precise rates): Under the further assumption that $\mu([\lambda(K), \lambda(K) + t])t^{-\alpha} \to c > 0$ as $t \to 0$ it is possible to give more precise statements for the behavior of the rates (such as $\Phi(t)e^{\rho t}t^{\beta} \to c$ as $t \to \infty$). Under the assumption x_0 is isotropic subgaussian, it is possible to give more precise particular asymptotic equivalences (i.e. without the $1/t$ exponent).

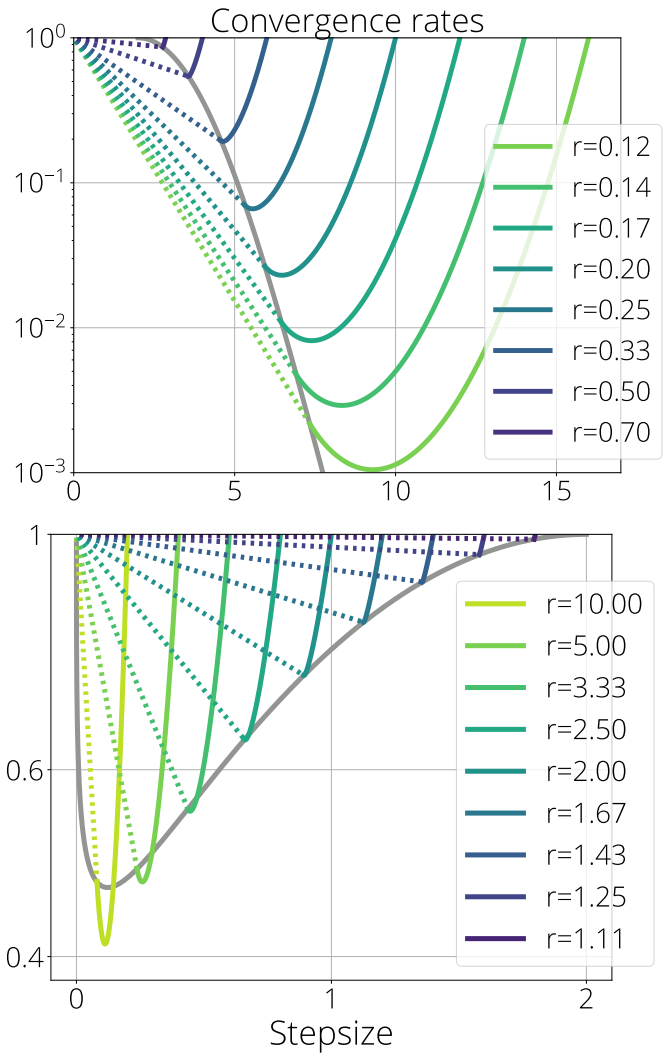


Figure 4: **Phase transition of the convergence rate** (y-axis) as a function of the stepsize (x-axis, γ) for the isotropic features model at infinite dimensions. Thus μ is Marchenko-Pastur (depending on aspect ratio r) and gradient flow is given by an isotropic starting vector. Smaller stepsizes (dotted) yield convergence rates which depend linearly on γ with a slope that is always frozen on $\lambda(\mu)$ – this coincides with the convergence rate of the underlying gradient flow. The convergence rate changes behavior once it hits the *critical stepsize* (solid gray, γ_*), becoming a non-linear function of γ (a discontinuity occurs in the second derivative of the convergence rate with respect to γ). The critical stepsize appears to be a good predictor for the optimal stepsize. In addition, the more overparameterized the data matrix ($r \to 0$) is, the smaller the window of convergent stepsizes and as its Hessian becomes ill-conditioned ($r \to 1$), the linear rate degenerates and the high temperature phase disappears.

4.4 Proof sketch of the homogenized SGD comparison

We give a reduced version of the proof of Theorem 13. In effect we show that $q(x_k)$ nearly satisfies the conclusion of Itô's lemma. Further, we show the martingale terms in both of the Doob decompositions are small, and hence it suffices to show the predictable parts of $q(x_k)$ and $q(\mathbf{X}_t)$ are close.

To advance the discussion, we compute this Doob decomposition. To take advantage of the simpler structure afforded by removing β, introduce

$$v_k := x_k - \beta \quad \text{and} \quad V_t := \mathbf{X}_t - \beta. \tag{39}$$

We shall extend the first integer indexed function to real-valued indices by setting $v_t = v_{\lfloor t \rfloor}$. We also let $(\mathscr{F}_t : t \geq 0)$ be the filtration generated by $(v_t : t \geq 0)$ and $(V_{t/d} : t \geq 0)$. Hence for all $k \in \mathbb{N}$, v_k is measurable with respect to $\mathscr{F}_k$. Recalling the recurrence (31), for a quadratic q

$$\begin{aligned} q(v_k) - q(v_{k-1}) &= -\gamma(\nabla q(v_{k-1}))^T(\Delta_k) + \tfrac{\gamma^2}{2}(\Delta_k)^T(\nabla^2 q)(\Delta_k), \\ \text{where} \quad m_k &= a_k/\sqrt{d} \\ \Delta_k &= m_k(m_k^T v_{k-1} - \eta w_k) \end{aligned} \tag{40}$$

The equation above can each be decomposed as a predictable part and two martingale increments

$$\begin{aligned} q(v_k) - q(v_{k-1}) =& -\gamma(\nabla q(v_{k-1}))^T\left(\tfrac{1}{d}Kv_{k-1}\right) + \Delta\mathcal{M}_k^{\text{lin}} \\ &+ \tfrac{\gamma^2}{2}\operatorname{Tr}(\tfrac{1}{d}K(\nabla^2 q))\left(\tfrac{1}{d}v_{k-1}^T K v_{k-1} + \mathbb{E}[\eta_k^2]\right) \\ &+ \Delta\mathcal{E}_k^{\text{quad}} + \Delta\mathcal{M}_k^{\text{quad}}, \\ \text{where} \quad \Delta\mathcal{M}_k^{\text{quad}} :=& \Delta_k^T\nabla^2 q\Delta_k - \mathbb{E}[\Delta_k^T\nabla^2 q\Delta_k \mid \mathcal{F}_{k-1}]. \end{aligned} \tag{41}$$

The remainder of the martingale increments are given by $\Delta\mathcal{M}_k^{\text{lin}}$ and are all linear in Δ_k. The predictable parts have been further decomposed into the leading order terms and an error term $\Delta\mathcal{E}_k^{\text{quad}}$.

These predictable parts, in turn, depend on different statistics $q_1(v_{k-1})$. It turns out to approximately describe the risk, we can work on a manifold indexed by a curve in $\mathbb{C}^2$ which approximately closes. Specifically, we let

$$\begin{aligned} &\mathcal{Q}_n(q) := \mathcal{Q}_n(q, K) = \\ &\Big\{q(x), \quad (\nabla q(x))^T R(z;K)x, \quad x^T R(y;K)(\nabla^2 q)R(z;K)x, \\ &\qquad (\nabla q(x))^T R(z;K)\beta, \quad x^T R(y;K)(\nabla^2 q)R(z;K)\beta, \quad \forall\, z, y \in \Gamma\Big\}. \end{aligned} \tag{42}$$

In order to control the martingales, it is convenient to impose a

stopping time

$$\tau := \inf\{k : ||v_k|| > d^\varepsilon\} \cup \{td : ||V_t|| > d^\varepsilon\}, \tag{43}$$

and we introduce the corresponding stopped processes

$$v_k^\tau = v_{k\wedge\tau}, \quad V_t^\tau = V_{t\wedge(\tau/d)}. \tag{44}$$

We prove a version of our theorem for the stopped processes and then show that the stopping time is greater than n with overwhelming probability.

Our key tool for comparing v_{td} and V_t is the following lemma.

Lemma 9 (Comparison of SGD to HSGD): Given a quadratic q with $\|q\|_{C^2} \le 1$, with $\mathcal{Q} = \mathcal{Q}_n(q) \cup \mathcal{Q}_n(\mathscr{P}) \cup \mathcal{Q}_n(\|\cdot\|^2)$ as above,

$$\begin{aligned}&\max_{0\le t\le \frac{n}{d}} |q(v_{td}^\tau) - q(V_t^\tau)| \le \\ &\sup_{0\le t\le n/d} \left(|\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{lin},\tau}| + |\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{quad},\tau}| + |\mathcal{E}_{\lfloor td\rfloor}^{\mathrm{quad},\tau}| + |\mathcal{M}_t^{HSGD,\tau}|\right) \\ &+ C(\|K\|_\sigma)\cdot \sup_{g\in\mathcal{Q}} \int_0^{n/d} |g(v_{sd}^\tau) - g(V_s^\tau)|ds.\end{aligned} \tag{45}$$

Here $\mathcal{M}_t^{HSGD,\tau}$ is the martingale part in the semimartingale decomposition of $q(V_t^\tau)$.

Proof. Owing to the similarities of this claim with the proof in [Paq+22a, Proposition 4.1], we just illustrate the main idea. The idea is that if we take a $g \in \mathcal{Q}$, and we apply (41), then in the predictable part of $g(v_t)$ we have

$$I_1 := \int_0^t \nabla g(v_{sd})^T K v_{sd}\, \mathrm{d}s, \quad I_2 := \int_0^t \nabla g(v_{sd})^T \beta\, \mathrm{d}s, \quad I_3 := \int_0^t v_{sd}^T K v_{sd}\, \mathrm{d}s.$$

These also appear with coefficients that can be bounded solely using $\|g\|_{C^2}$ and $\|K\|_\sigma$. We get the same, applying Itô's lemma to $g(V_t)$, albeit with the replacement $v_t \to V_t$. We wish to bound for example $I_1(v_t) - I_1(V_t)$. We do this by expressing its integrand as $p(v_t) - p(V_t)$ for polynomial p. If g is linear (the final row of (42)), then p is again linear. For example, if it is $g(x) = \nabla q(x)^T R(z;K)\beta$, then p is again linear and is given by

$$p(x) = x^T K R(z;K)\beta = +x^T R(z;K)\beta - z x^T \beta,$$

where we have used the resolvent identity $(K - z)R(z;K) = I$. Note the function $x^T R(z;K)\beta$ is contained in $\mathcal{Q}$ by virtue of being in $\mathcal{Q}_n(\|\cdot$

$\|^2$). Moreover, by Cauchy's integral formula, we can represent $x^T\beta$ by averaging $\frac{-1}{2\pi i}x^T R(y;K)\beta$ over $y \in \Gamma$. Hence

$$|p(v_{td}) - p(V_t)| \leq \|\Gamma\| \max_{g\in\mathcal{Q}} |g(v_{td}) - g(V_t)|$$

with $\|\Gamma\|$ the length of the curve (which can be bounded in terms of $\|K\|_\sigma$). The same manipulations lead finally to showing every term included in $\mathcal{Q}$ can be controlled in a similar manner, using the other elements of the class $\mathcal{Q}$. □

The second important idea is to discretize the set $\mathcal{Q}$.

Lemma 10 (Discretize the spectral curve): There exists $\bar{Q} \subseteq \mathcal{Q}$ with $|\bar{Q}| \leq C(\|K\|_\sigma)d^{4m}$ such that, for every $q \in \mathcal{Q}$, there is some $\bar{q} \in \bar{Q}$ satisfying $||q - \bar{q}||_{C^2} \leq d^{-2m}$.

Proof. On the spectral curve Γ, we can bound the norm of the resolvent. Since

$$\frac{\mathrm{d}}{\mathrm{d}z}R(z;K) = (K - zI)^{-2},$$

we have it is norm bounded by an absolute constant. The arc length of the curve is at most $C(\|K\|_\sigma)$, and so by choosing a minimal net $d^{-2\varepsilon}$ of the manifold $\Gamma \times \Gamma$, the lemma follows. □

Now the main technical part of the argument is to control the martingales and errors. As we work with the stopped process v_k^τ we introduce the stopped proccesses $\mathcal{M}_k^{\mathrm{lin},\tau}, \mathcal{M}_k^{\mathrm{quad},\tau}, \mathcal{E}_k^{\mathrm{quad},\tau}$, which are defined analogously to (44).

Lemma 11 (Martingale bounds): For any quadratic q with $\|q\|_{C^2} \leq 1$, the terms $\mathcal{M}_k^{\mathrm{lin},\tau}, \mathcal{M}_k^{\mathrm{quad},\tau}, \mathcal{E}_k^{\mathrm{quad},\tau}$ satisfy the following bounds with overwhelming probability (with a bound which is uniform in q) for $n \leq d\log d$

i $\sup_{1\leq k\leq n} |\mathcal{M}_k^{\mathrm{lin},\tau}| \leq d^{-\frac{1}{2}+5\varepsilon}$,

ii $\sup_{1\leq k\leq n} |\mathcal{M}_k^{\mathrm{quad},\tau}| \leq d^{-\frac{1}{2}+9\varepsilon}$,

iii $\sup_{1\leq k\leq n} |\mathcal{E}_k^{\mathrm{quad},\tau}| \leq d^{-1+9\varepsilon}$.

Combining Lemmas 9 and 10, along with the above, we conclude that, for any $\bar{q} \in \bar{Q}$ with $\|q\|_{C^2} = 1$,

$$|\bar{q}(v_{td}^\tau) - \bar{q}(V_t^\tau)| \leq 4d^{-\frac{1}{2}+9\varepsilon} + C(\|K\|_\sigma) \max_{g\in\mathcal{Q}} \int_0^t |g(v_{sd}^\tau) - g(V_s^\tau)|ds. \tag{46}$$

Hence by Lemma 10 and by bounding $\|g\|_{C^2}$ over all Q,

$$\max_{g\in Q} |q(v_{td}^\tau) - q(V_t^\tau)| \le C(\|K\|_\sigma)\left(d^{-2} + d^{-\frac{1}{2}+9\varepsilon} + \int_0^t \max_{g\in Q} |g(v_{sd}^\tau) - g(V_s^\tau)|ds\right). \tag{47}$$

By Gronwall's inequality, this gives us that with overwhelming probability

$$\max_{g\in Q} \max_{0\le t\le n/d} |g(v_{td}^\tau) - g(V_t^\tau)| \le C(\|K\|_\sigma)(d^{-2} + 4d^{-\frac{1}{2}+9\varepsilon})e^{C(\|K\|_\sigma)n/d}. \tag{48}$$

Now we note that the norm function $x \mapsto \|x\|^2$ is one of the quadratics included in Q. Hence if we let $\mathcal{G}$ be the event in the above display, and we let $\mathcal{E} = \{\max_{0\le s\le n/d} \|V_s\| \le d^{\varepsilon/2}\}$, then we have

$$\mathcal{G} \cap \mathcal{E} \cap \{\tau \le n/d\} \subseteq \{\|v_\tau\| - \|v_{\tau-1}\| \ge d^{\varepsilon/2}\} \cap \{\tau \le n/d\}.$$

This is because on the event $\{\tau \le n/d\} \cap \mathcal{E}$ we must have had $\|v_\tau\| > d^\varepsilon$, but in the step before τ, we had $v_{\tau-1}$ could be compared to $V_{\tau-1}$ (due to $\mathcal{G}$, and we had the norm of $V_{\tau-1}$ was small. Now it is easily seen that with overwhelming probability, no increment of SGD between time 0 and n/d can increase the norm by a power of d. So to complete the proof it suffices to show $\mathcal{E}$ holds with overwhelming probability.

Thus the proof is completed by the following:

Lemma 12 (Non-explosiveness of HSGD): For any $\delta > 0$ and any $t > 0$ with overwhelming probability

$$\max_{0\le s\le t} \|\mathbf{X}_s\|^2 \le e^{C(\|K\|_\sigma)t} d^\delta.$$

Proof. We apply Itô's formula to $\phi(\mathbf{X}_t) := \log(1 + \|\mathbf{X}_t\|^2)$, from which we have

$$\begin{aligned} \mathrm{d}\phi(\mathbf{X}_t) = &-2\gamma \frac{\mathbf{X}_t \cdot \nabla \mathscr{R}(\mathbf{X}_t)}{1+\|\mathbf{X}_t\|^2}\,\mathrm{d}t + \frac{\mathbf{X}_t \cdot \gamma\sqrt{\frac{2}{d}\mathscr{R}(\mathbf{X}_t)K}\,\mathrm{d}B_t}{1+\|\mathbf{X}_t\|^2} \\ &+ \Big(\frac{\mathscr{R}(\mathbf{X}_t)}{1+\|\mathbf{X}_t\|^2}\frac{2\gamma^2}{d}\operatorname{Tr}(K) - \frac{2\gamma^2 \mathscr{R}(\mathbf{X}_t)\mathbf{X}_t^T K \mathbf{X}_t}{d}\Big)\,\mathrm{d}t \end{aligned}$$

The drift terms and the quadratic variation terms can be bounded by some $C(\|K\|_\sigma)$. Hence with this constant, for all $r \ge 0$,

$$\Pr(\max_{0\le s\le t} \phi(\mathbf{X}_s) \ge C(\|K\|_\sigma)(t + r\sqrt{t})) \le 2\exp(-r^2/2).$$

Taking $r = \sqrt{\log d \log\log d}$, we conclude that with overwhelming probability

$$\max_{0\le s\le t} \phi(\mathbf{X}_s) \le C(\|K\|_\sigma)(t + \sqrt{t \log d \log\log d}).$$

□

4.5 Controlling the errors

The main goal of this section is to control the martingale terms and error terms; in particular we prove Lemma 11. We will also record for future use an estimate on ∇q that follows from $\|\cdot\|_{C^2}$ control.

$$||\nabla q(x)|| \leq \|\nabla^2 q\|_\sigma \cdot ||x|| + ||\nabla q(0)|| \leq \|q\|_{C^2} \cdot (||x|| + 1). \tag{49}$$

Martingale for gradient part of recurrence.

Proof. Comparing (40) and (41), we see that for $k \leq \tau$

$$\begin{aligned} \Delta\mathcal{M}_k^{\text{lin},\tau} &= \left[\left(w_{k-1}^T m_k \right) \left(m_k^T v_{k-1}^\tau - \eta_k \right) - \tfrac{1}{d} w_{k-1}^T K v_{k-1}^\tau \right] \\ &=: [\Delta\mathcal{M}_k^{\text{lin}\,1,\tau} - \Delta\mathcal{M}_k^{\text{lin}\,2,\tau}], \\ &\text{where} \quad w_{k-1} := -\gamma \nabla q(v_{k-1}^\tau) + \tfrac{\gamma^2 \mathrm{d}}{d}(v_{k-1}^\tau + \beta). \end{aligned} \tag{50}$$

Note for $k > \tau$, the stopped martingale increment is 0. Using (49), $\|w_{k-1}\| \leq C(\gamma, \mathrm{d}) d^\varepsilon$. We will separately bound the contributions from $\Delta\mathcal{M}_k^{\text{lin}\,1,\tau}$ and $\Delta\mathcal{M}_k^{\text{lin}\,2,\tau}$ in terms of their Orlicz norms. For the first part, for any fixed k, we condition on $\mathcal{F}_{k-1}$ and Assumption 1, we conclude

$$||\Delta\mathcal{M}_k^{\text{lin}\,1,\tau}||_{\psi_1} \leq \left\| w_{k-1}^T m_k \right\|_{\psi_2} \left\| m_k^T v_{k-1}^\tau - \eta_k \right\|_{\psi_2} \leq C d^{-\frac{1}{2}+2\varepsilon} \cdot d^{-\frac{1}{2}+2\varepsilon} \tag{51}$$

where C is some absolute constant. For the second part, we have

$$|\Delta\mathcal{M}_k^{\text{lin}\,2,\tau}| = |\tfrac{1}{d} w_{k-1}^T K v_{k-1}^\tau| \leq C d^{-1+2\varepsilon}. \tag{52}$$

Combining these, we see that, for every k,

$$\sigma_{k,1} := \inf\{t > 0 : \mathbb{E}[\exp(|\Delta\mathcal{M}_k^{\text{lin}\,1,\tau} - \Delta\mathcal{M}_k^{\text{lin}\,2,\tau}|/t)|\mathcal{F}_{k-1}] \leq 2\} \leq C d^{-1+4\varepsilon} \tag{53}$$

and, by the martingale Bernstein inequality,

$$\begin{aligned} &\Pr\left(\sup_{1\leq k\leq n} |\mathcal{M}_k^{\text{lin},\tau} - \mathbb{E}\mathcal{M}_0^{\text{lin}}| \geq t \right) \\ &\leq 2\exp\left(-\min\left\{ \frac{t}{c \max \sigma_{k,1}}, \frac{t^2}{c\sum_{k=1}^n \sigma_{k,1}} \right\} \right) \\ &\leq 2\exp\left(-\min\left\{ C t d^{1-4\varepsilon}, C t^2 d^{2-8\varepsilon} n^{-1} \right\} \right). \end{aligned} \tag{54}$$

As we assume that $n \leq d \log d$ then this gives us

$$\sup_{1\leq k\leq n} |\mathcal{M}_k^{\text{lin},\tau}| \leq d^{-\frac{1}{2}+5\varepsilon} \tag{55}$$

with overwhelming probability. □

Martingale for Hessian part of recurrence.

Proof. Next we consider the contribution from the Hessian part of the recurrence. We write

$$\begin{aligned}&\frac{\gamma^2}{2}(m_k m_k^T v_{k-1}^\tau - m_k\eta_k)^T(\nabla^2 q)(m_k m_k^T v_{k-1}^\tau - m_k\eta_k)\\&= \mathbb{E}\left[\frac{\gamma^2}{2}(m_k m_k^T v_{k-1}^\tau - m_k\eta_k)^T(\nabla^2 q)(m_k m_k^T v_{k-1}^\tau - m_k\eta_k)|\mathcal{F}_{k-1}\right] + \Delta\mathcal{M}_k^{\text{quad}}.\end{aligned} \tag{56}$$

Rearranging the terms, we get

$$\Delta\mathcal{M}_k^{\text{quad}} = A_k B_k - \mathbb{E}[A_k B_k|\mathcal{F}_{k-1}] \tag{57}$$

where

$$A_k := m_k^T(\nabla^2 q)m_k, \quad B_k := (m_k^T v_{k-1}^\tau - \eta_k)^2. \tag{58}$$

This can be expanded as

$$\begin{aligned}\Delta\mathcal{M}_k^{\text{quad}} =&(A_k - \mathbb{E}[A_k])(B_k - \mathbb{E}[B_k]) + \mathbb{E}[A_k]\mathbb{E}[B_k] - \mathbb{E}[A_k B_k]\\&+ (A_k - \mathbb{E}[A_k])\mathbb{E}[B_k] + (B_k - \mathbb{E}[B_k])\mathbb{E}[A_k],\end{aligned} \tag{59}$$

so we focus first on obtaining subexponential bounds for the quantities $A_k - \mathbb{E}[A_k]$ and $B_k - \mathbb{E}[B_k]$ using the Hanson-Wright inequality. For A_k, we have

$$\begin{aligned}&\Pr(|A_k - \mathbb{E}A_k| \geq t)\\&\leq 2\exp\left[-c\min\left(\frac{t^2}{d^{-2+4\varepsilon}||\nabla^2 q||_{HS}^2}, \frac{t}{d^{-1+2\varepsilon}||\nabla^2 q||}\right)\right]\\&\leq 2\exp[-c'\min(t^2 d^{1-4\varepsilon}, td^{1-2\varepsilon})] \leq 2\exp[-c'' td^{\frac{1}{2}-2\varepsilon}]\end{aligned} \tag{60}$$

and thus we have the subexponential bound

$$||A_k - \mathbb{E}[A_k]||_{\psi_1} < Cd^{-\frac{1}{2}+2\varepsilon}. \tag{61}$$

Next we obtain a subexponential bound for B_k. For the part of B_k not involving η_k, we use Hanson-Wright to get

$$\begin{aligned}&\Pr\left(\left|m_k^T v_{k-1}^\tau (v_{k-1}^\tau)^T m_k - \mathbb{E}m_k^T v_{k-1}^\tau (v_{k-1}^\tau)^T m_k\right| \geq t\right)\\&\leq 2\exp\left[-c\min\left(\frac{t^2}{d^{-2+4\varepsilon}||v_{k-1}^\tau (v_{k-1}^\tau)^T||_{HS}^2}, \frac{t}{d^{-1+2\varepsilon}||v_{k-1}^\tau (v_{k-1}^\tau)^T||}\right)\right]\\&\leq 2\exp[-c\min(t^2 d^{2-8\varepsilon}, td^{1-4\varepsilon})].\end{aligned} \tag{62}$$

For the terms involving η_k, we use the Orlicz bounds from the assumptions in the set-up to obtain

$$\begin{aligned}||m_k^T v_{k-1}^\tau \eta_k||_{\psi_1} &\leq ||m_k^T v_{k-1}^\tau||_{\psi_2} \cdot ||\eta_k||_{\psi_2} = d^{-\frac{1}{2}+2\varepsilon} d^{-\frac{1}{2}+\varepsilon}\\&= d^{-1+3\varepsilon}.\end{aligned} \tag{63}$$

Since also $||\eta_k^2||_{\psi_1} = d^{-1+2\varepsilon}$ combining the bounds (62) and (63), we have

$$||B_k - \mathbb{E}[B_k]||_{\psi_1} < Cd^{-1+4\varepsilon}. \tag{64}$$

Furthermore, we have

$$\mathbb{E}[A_k] = O(1), \qquad \mathbb{E}[B_k] = O(d^{-1}), \tag{65}$$

uniformly for all k based on the assumptions on η_k and m_k. We now use (61), (64), (65) to bound each term of (59) in turn.

To bound the contribution from $(A_k - \mathbb{E}[A_k])(B_k - \mathbb{E}[B_k])$, we observe that, for each k, with overwhelming probability, $|A_k - \mathbb{E}[A_k]| < d^{-\frac{1}{2}+3\varepsilon}$ and $|B_k - \mathbb{E}[B_k]| < d^{-1+5\varepsilon}$, so we can conclude that, with overwhelming probability,

$$\sum_{k=1}^{n} \Big|(A_k - \mathbb{E}[A_k])(B_k - \mathbb{E}[B_k])\Big| < nd^{-\frac{3}{2}+8\varepsilon} < d^{-\frac{1}{2}+9\varepsilon}. \tag{66}$$

For the second term of (59) we have

$$\Big|\mathbb{E}[A_k]\mathbb{E}[B_k] - \mathbb{E}[A_kB_k]\Big| = \Big|\mathbb{E}\big[(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\big]\Big| \le \mathbb{E}\Big|(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\Big|. \tag{67}$$

We can bound this quantity using

$$\begin{aligned} &\Pr\left(\left|(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\right| \ge t\right) \\ &\le \Pr(|A_k - \mathbb{E}A_k| \ge \sqrt{t}) + \Pr(|B_k - \mathbb{E}B_k| \ge \sqrt{t}) \\ &\le 4\exp\left[-c\min(td^{1-4\varepsilon}, \sqrt{t}d^{1-4\varepsilon})\right] \end{aligned} \tag{68}$$

where the bound in the last line comes from combining (60) and (64). Using this bound, we obtain

$$\begin{aligned} \Big|\mathbb{E}[A_k]\mathbb{E}[B_k] - \mathbb{E}[A_kB_k]\Big| &\le \int_0^\infty x\Pr\left(\left|(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\right| \ge x\right)dx \\ &\le \int_0^1 4x\exp(-cxd^{1-4\varepsilon})dx + \int_1^\infty 4x\exp(-c\sqrt{x}d^{1-4\varepsilon})dx \end{aligned} \tag{69}$$

Making the change of variables $y = xd^{1-4\varepsilon}$ in the first integral and $z = \sqrt{x}d^{1-4\varepsilon}$ in the second integral, this becomes

$$4d^{-2+8\varepsilon}\int_0^{d^{1-4\varepsilon}} y\exp(-cy)dy + 4d^{-4+8\varepsilon}\int_{d^{1-4\varepsilon}}^\infty z^2\exp(-cz)dz = O(d^{-2+8\varepsilon}). \tag{70}$$

Thus,

$$\sum_{k=1}^{n}\Big|\mathbb{E}[A_k]\mathbb{E}[B_k] - \mathbb{E}[A_kB_k]\Big| = O(nd^{-2+8\varepsilon}). \tag{71}$$

Finally, we note that the remaining terms of (59), namely $(A_k - \mathbb{E}[A_k])\mathbb{E}[B_k]$ and $(B_k - \mathbb{E}[B_k])\mathbb{E}[A_k]$, are martingale increments with

$$||(A_k - \mathbb{E}[A_k])\mathbb{E}[B_k]||_{\psi_1} \le Cd^{-\frac{3}{2}+2\varepsilon}, \qquad ||(B_k - \mathbb{E}[B_k])\mathbb{E}[A_k]||_{\psi_1} \le Cd^{-1+4\varepsilon}. \tag{72}$$

Applying the Martingale Bernstein inequality, we conclude

$$\begin{aligned}
&\Pr\left(\sup_{1\le k\le n}\left|\sum_{j=1}^{k}(A_j-\mathbb{E}[A_j])\mathbb{E}[B_j]+(B_j-\mathbb{E}[B_j])\mathbb{E}[A_j]\right|\ge t\right)\\
&\quad\le 2\exp\left(-\min\left\{\frac{t}{c\max\sigma_{k,1}},\frac{t^2}{c\sum_{k=1}^{n}\sigma_{k,1}}\right\}\right)\\
&\quad\le 2\exp\left(-\min\left\{Ctd^{1-4\varepsilon},Ct^2d^{2-8\varepsilon}n^{-1}\right\}\right).
\end{aligned}\tag{73}$$

Thus, for $n\le d\log d$, we get

$$\sup_{1\le k\le n}\left|\sum_{j=1}^{k}(A_j-\mathbb{E}[A_j])\mathbb{E}[B_j]+(B_j-\mathbb{E}[B_j])\mathbb{E}[A_j]\right|\le d^{-\frac{1}{2}+5\varepsilon}\tag{74}$$

with overwhelming probability. Finally, combining the bounds from (66), (71), (74), we conclude that, for $n\le d\log d$,

$$\sup_{1\le k\le n}|\mathcal{M}_k^{\mathrm{quad},\tau}|\le d^{-\frac{1}{2}+8\varepsilon}\tag{75}$$

with overwhelming probability. This completes the proof of part (ii) of the lemma.

For part (iii), we observe that $\Delta\mathcal{E}_k^{\mathrm{quad},\tau}=\mathbb{E}[A_kB_k]-\mathbb{E}[A_k]\mathbb{E}[B_k]+O(d^{-2+4\epsilon})$, the error terms arising from u_k cross terms, so that the bound of $\mathcal{E}_k^{\mathrm{quad},\tau}$ follows immediately from (71). □

5 *Homogenization of Multipass SGD on the least squares*

This is adapted from [Paq+22a], building on earlier work in [Paq+21].

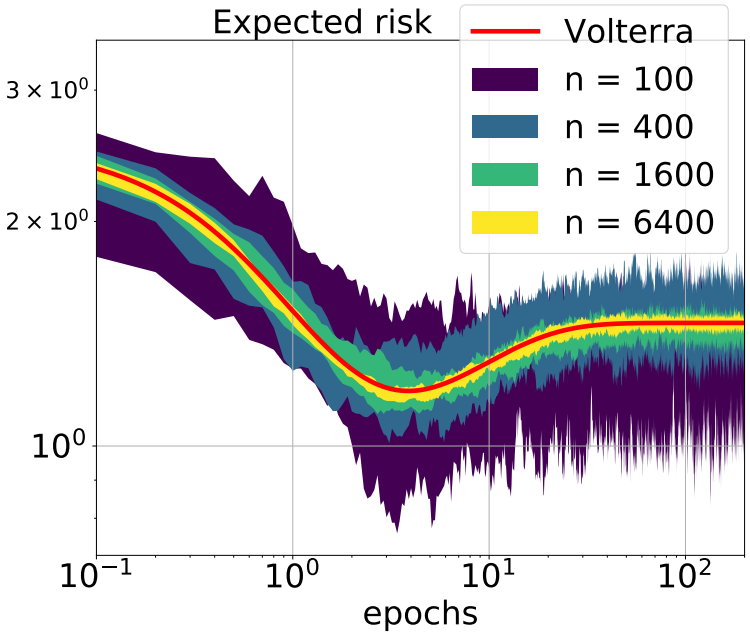


Figure 5: Risk curves of SGD across different dimensions. In each dimension, 10 runs of multi-pass constant step-size SGD are performed on a least squares problem, and the test error is computed over time. We then display 80% confidence intervals over time (i.e. we discard the largest and smallest at error at each point in time). The curves concentrate around a high-dimensional limit value. Note that time is scaled in epochs. The Volterra curve is the limiting risk curve.

In this section, we will deal exclusively with multi-pass SGD on the least squares problem. Strictly speaking, this will no longer purely concern the problem of linear regression (although this remains the main motivating application). Suppose that we are given an $n \times d$ matrix A and a target vector b. We look at the least squares problem

$$\min_{x \in \mathbb{R}^d} \left\{ \mathscr{L}(x) := \tfrac{1}{2n} \| \langle A, x \rangle_d - b \|^2 = \tfrac{1}{2n} \sum_{i=1}^{n} (\langle a_i, x \rangle - b_i)^2 \right\}.$$

The SGD we now consider is

$$x_{k+1} = x_k - \gamma_k (\langle a_{i_{k+1}}, x \rangle - b_{i_{k+1}}) a_{i_{k+1}}, \quad \{i_k\} \text{ iid Unif}(\{1, 2, \cdots, n\}). \tag{76}$$

This is multi-pass SGD.

Now a fruitful point of view in this case is to actually recast this as streaming SGD, which is possible if we view $\mathcal{D}$ as the empirical distribution of the pairs $((a_i, b_i) : 1 \le i \le n)$ so that samples from $\mathcal{D}$ are given by

$$(a, b) \overset{\text{law}}{=} (a_i, b_i), \quad i \overset{\text{law}}{=} \text{Unif}(\{1, 2, \cdots, n\}).$$

The expected risk, considered this way, would be the empirical risk $\mathscr{L}$. For clarity, we shall still refer to it as the empirical risk, as in an

ERM context, it may be helpful to still consider a population risk. However, this does give a clear guess for how to approximate the resulting SGD in high-dimensions. Define the sample covariance matrices

$$\hat{K} := \frac{1}{n} A^T A \quad \text{and} \quad \check{K} := \frac{1}{n} A A^T, \tag{77}$$

where the first is the (usual) feature-feature covariance and the second is (up to scaling) an empirical estimator of the covariance between the samples. If we use (32) as a guide, then with $\gamma_k = \gamma(k/d)/d$

$$\mathrm{d}\mathbf{X}_t = -\gamma(t)\left(\nabla \mathscr{L}(\mathbf{X}_t) + \sqrt{\tfrac{2}{d}\mathscr{L}(\mathbf{X}_t)\hat{K}}dB_t\right). \tag{78}$$

On the other hand, what is clear is that this distribution cannot satisfy Assumption 1 in two important ways. First the data absolutely cannot generically Part 2 (the Hanson–Wright inequality) uniformly in B, as the case of B being given by an outer product $a_1 \otimes a_1$, which will cause large non-concentration issues. Second, there is no underlying model for the targets b, and no clear candidate for a target β.

Assumption 2 (Empirical data assumptions): Suppose that the norm of $\hat{K}$ (and hence $\check{K}$) is bounded above independent of n and d. Suppose Γ is the contour enclosing $[0, \|\hat{K}\|]$ at distance 1. Suppose there is a $\theta \in (0, \frac{1}{4})$ for which

1. $\max_{z\in\Gamma} \max_{1\le i\le n} |e_i^T R(z;\check{K}) b| \le n^{\theta-1/2}$.
2. $\max_{z\in\Gamma} \max_{1\le i\ne j\le n} |e_i^T R(z;\check{K}) e_j^T| \le n^{\theta-1/2}$.
3. $\max_{z\in\Gamma} \max_{1\le i\le n} |e_i^T R(z;\check{K}) e_i - \frac{1}{n}\operatorname{Tr} R(z;\check{K})| \le n^{\theta-1/2}$.

In a random matrix theory context, such types of results are standard. That is, under quite general assumptions, if we suppose that the rows of A are given by independent samples from a high-dimensional distribution, one gets that the off-diagonal resolvent entries of $\check{K}$ are small and the on-diagonal entries approximate the trace. See for example [KY17].

We also need that the initialization does not pick out a part of the feature covariance matrix which is unusally dense.

Assumption 3 (Non-spectral Init): Let Γ be the same contour as in Assumption 2 and let $\theta \in (0, \frac{1}{2})$. Then

$$\max_{z\in\Gamma} \max_{1\le i\le d} |e_i^T R(z;\hat{K}) x_0| \le d^{\theta-1/2}.$$

Exercise 12 (Initialization): Show that if $\sqrt{d}x_0$ has iid mean 0, subgaussian entries, $\hat{K}$ has bounded norm then Assumption 3 holds with overwhelming probability.

Finally for comparison of SGD to its homogenized counter-part, we need that the the risk we consider is well-behaved.

Assumption 4 (Quadratic statistics): Suppose $\mathscr{R} : \mathbb{R}^d \to \mathbb{R}$ is quadratic, i.e. there is a symmetric matrix $T \in \mathbb{R}^{d\times d}$, a vector $u \in \mathbb{R}^d$, and a constant $c \in \mathbb{R}$ so that

$$\mathscr{R}(x) = \tfrac{1}{2}x^T Tx + u^T x + c. \tag{79}$$

We assume that $\mathscr{R}$ satisfies $\|\mathscr{R}\|_{C^2} \le C$ for some C independent of n and d. Moreover, we assume the following (for the same Γ and θ) as in Assumption 2:

$$\begin{aligned} &\max_{z,y\in\Gamma}\max_{1\le i\le n} \tfrac{1}{n}|e_i^T A\widehat{T}A^T e_i - \mathrm{Tr}(\hat{K}\widehat{T})| \le \|T\|n^{-\theta}, \quad \text{where} \\ &\widehat{T} = R(z)TR(y) + R(y)R(z), \quad R(z) = R(z;\hat{K}). \end{aligned} \tag{80}$$

Then under all these assumptions, we can compare this risk as it evolves under SGD to the same under homogenized SGD.

Theorem 17: Homogenization of multi-pass SGD

Suppose $n \ge d^{\tilde{\varepsilon}}$ and $n \le d^C$ and suppose that Assumptions 2, 3 and 4 are in force. There is a $\varepsilon > 0$ depending only on θ and $\tilde{\varepsilon}$ so that for any deterministic $T > 0$

$$\sup_{0\le t\le T} |\mathscr{R}(x_{td}) - \mathscr{R}(\mathbf{X}_t)| \le d^{-\tilde{\varepsilon}/2}$$

with overwhelming probability.

Example 14: SGD for Linear regression

As a principle example suppose one takes a linear regression setup where for a fixed $d \times d$ covariance matrix $\Sigma \succ 0$ of bounded norm, we set a sample $(a, b) \overset{\text{law}}{=} \mathcal{D}$ to be constructed by

$$a = \sqrt{\Sigma}z, \quad b = \langle a, \beta\rangle + \eta w,$$

where z is an iid 1-subgaussian vector and w is mean 0 1-subgaussian. We set $\mathscr{P}(x) = \frac{1}{2}\mathbb{E}(\langle a, x\rangle - b)^2$.

Let $((a_i, b_i) : 1 \le i \le n)$ be n samples from this distribution, and form a matrix (A, b) by setting the rows of A to be given

by the samples $\{a_i\}$. Then provided n/d is bounded below independent of d, Assumption 2 holds for any $\theta > 0$ with overwhelming probability. Suppose x_0 is as in Exercise 12, so that Assumption 3 holds. Finally both $\mathscr{P}$ and $\mathscr{L}$ satisfy Assumption 4 with overwhelming probability. Hence for $\mathscr{R}$ given by either of $\mathscr{P}$ or $\mathscr{L}$,

$$\sup_{0\le t\le T} |\mathscr{R}(x_{td}) - \mathscr{R}(\mathbf{X}_t)| \le d^{-\tilde{\varepsilon}/2}$$

with overwhelming probability.

Remark 13 (Random (fully connected) feature regression): In random features regression, suppose that one has an underlying data distribution $\mathcal{D}_0$ on $\mathbb{R}^m \otimes \mathbb{R}^p$. Motivated by neural networks (and especially by *wide* neural networks), one considers an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and one introduces a weight matrix $W \in \mathbb{R}^d \otimes \mathbb{R}^m$. Then one transforms the data to make a new distribution $\mathcal{D}$ by setting a sample from $(a, b) \overset{\text{law}}{=} \mathcal{D}$ to be given by

$$\textsc{Output} : (\sigma(\langle W, a\rangle_m), b) \quad \text{where} \quad (a, b) \sim \mathcal{D}_0.$$

If W is drawn simply from $N(0, \mathrm{Id}_m \otimes \mathrm{Id}_p)$, then this is a random fully-connected feature model. More general, structured covariances can be used to produce more elaborate and interesting models: see [RR08].

Random features models can also be seen to satisfy the assumptions of Theorem 17; see [Paq+22a] for details.

This means we have a Volterra risk model for the training loss:

Definition 35 ((Empirical) Volterra model for training loss): Let $\mathscr{X}_t$ be the path of gradient flow started from initialization $\mathbf{X}_0$ for minimizing the empirical risk, i.e.

$$\dot{\mathscr{X}_t} = -\nabla \mathscr{L}(\mathscr{X}_t).$$

Let $\mathcal{K}_\gamma$ be the function from $[0, \infty) \to [0, \infty)$ given by

$$\mathcal{K}_\gamma(t) := \gamma^2 \frac{\mathrm{Tr}(\hat{K}^2 e^{-2\gamma \hat{K} t})}{d}.$$

Then the Volterra risk model is the solution of the convolution-

type Volterra equation

$$\Psi(t) := \mathscr{L}(\mathscr{X}_{\gamma t}) + \int_0^t \mathcal{K}_\gamma(t-s)\Psi(s)\ \mathrm{d}s.$$

Now to give the population risk $\mathscr{P}$, it is helpful to return to the behaviour of homogenized SGD.

Often in this context, this is also referred to as the generalization error, meaning how well the estimator performs on a new sample from the distribution.

The empirical risk curve concentrates around the Volterra risk model, as in Theorem 14. It follows that for homgenized SGD, we actually have the following approximation

$$\mathrm{d}\mathbf{X}_t \approx -\gamma(t)\big(\nabla\mathscr{L}(\mathbf{X}_t) + \sqrt{\tfrac{2}{d}\Psi(t)K}dB_t\big),$$

which we shall see actually describes a Gaussian centered around gradient flow.

To describe gradient flow, we need a surrogate for β. The correct relacement comes from properly projecting b Namely, we decompose

$$\|Ax-b\|^2 = \|Ax - A(A^TA)^{-1}A^Tb + \eta\|^2 = \|A(x-\beta^*)\|^2 + \eta^2,$$

where η is a vector orthogonal to the rows of A, i.e. $A^T\eta = 0$. Here we have set $\beta^* = (A^TA)^{-1}A^Tb$. It follows that we have

$$\nabla\mathscr{L}(x) = \hat{K}(x-\beta^*)$$

Then β^* is the appropriate generalizer of β in the sense that gradient flow on $\mathscr{L}$ acts by

$$\mathscr{X}_t - \beta^* = e^{-t\hat{K}}(\mathscr{X}_0 - \beta^*),$$

and moreover, homogenized SGD can be expressed as

$$\mathrm{d}\mathbf{X}_t \approx -\gamma(t)\big(\hat{K}(\mathbf{X}_t-\beta^*) + \sqrt{\tfrac{2}{d}\mathscr{L}(\mathbf{X}_t)\hat{K}}dB_t\big).$$

Hence, working for simplicity in the case $\gamma(t)\equiv\gamma$,

$$\mathrm{d}(e^{\gamma t\hat{K}}(\mathbf{X}_t-\beta^*)) = e^{\gamma t\hat{K}}\sqrt{\tfrac{2}{d}\mathscr{L}(\mathbf{X}_t)\hat{K}}dB_t \approx e^{\gamma t\hat{K}}\sqrt{\tfrac{2}{d}\Psi(t)\hat{K}}dB_t.$$

This leads to the following approximation

Lemma 13 (Gaussian approximation): With $\gamma(t)\equiv\gamma$, we have that for any T and any $\epsilon>0$, with overwhelming probability

$$\sup_{0\le t\le T}\left|\mathbf{X}_t - \mathscr{X}_{\gamma t} + \int_0^t \gamma e^{-\gamma(t-s)\hat{K}}\sqrt{\tfrac{2}{d}\Psi(s)\hat{K}}dB_s\right| \le d^{-1/2+\epsilon}.$$

Thus for evaluation against another statistic, such as $\mathscr{P}(\mathbf{X}_t)$, we have:

Corollary 6 (Generalization error model): The generalization error $\mathscr{P}(\mathbf{X}_t)$ evolves according to the risk curve

$$\mathscr{P}(\mathbf{X}_t) = \mathscr{P}(\mathscr{X}_{\gamma t}) + \int_0^t \tfrac{\gamma^2}{d} \operatorname{Tr}(e^{-2\gamma(t-s)\hat{K}} K\hat{K})\Psi(s)\ \mathrm{d}s + \mathrm{E}_t$$

The error E_t tends to 0 like $d^{-1/2+\epsilon}$ with overwhelming probability uniformly on compact sets.

5.1 *Comparison of single and multi-pass case*

A few major qualitative points can be made here. In an empirical risk minimization framework, as multi-pass SGD minimizes the empirical risk $\mathscr{L}$, it has the ability to *overfit*. Thus, running longer in multi-pass SGD can in fact actually degrade test loss performance.

On the other hand, the *excess risk* of using multi-pass SGD over gradient flow, which is the term

$$\text{Excess-risk}(t) := \int_0^t \tfrac{\gamma^2}{d} \operatorname{Tr}(e^{-2\gamma(t-s)\hat{K}} K\hat{K})\Psi(s)\ \mathrm{d}s,$$

depends qualitatively on two main features, the size of γ and the behavior of the training loss Ψ. In particular, when $\Psi(t) \to 0$ (which in particular implies that γ is less than the convergence threshold $2\operatorname{Tr}(\hat{K})/d$) then the excess risk of SGD tends to 0.

In situations where $\Psi(t) \to \Psi(\infty) > 0$, then the excess risk incurred tends to

$$\text{Excess-risk}(\infty) := \tfrac{\gamma}{2d} \operatorname{Tr}(K\Pi(\hat{K}))\Psi(\infty),$$

where $\Pi(\hat{K})$ is the projection onto the span of $\hat{K}$. Note that in the streaming case, there is also excess risk caused by SGD over gradient flow, and it follows a similar recipe.

From a risk minimization point of view, one can ask whether the danger of overfitting using multi-pass SGD outweighs the cost of using one-pass SGD, which is limited in the number of steps by the number of data points? The answer is complicated and depends greatly on the problem, see as an illutration Figure 6 and 7.

5.2 *Proof strategy for homogenized SGD*

The general plan of the proof follows that of streaming SGD, with a few important differences. We give an overview of the strategy here. As there, we look to evaluate the updates of a quadratic test statistic over time.

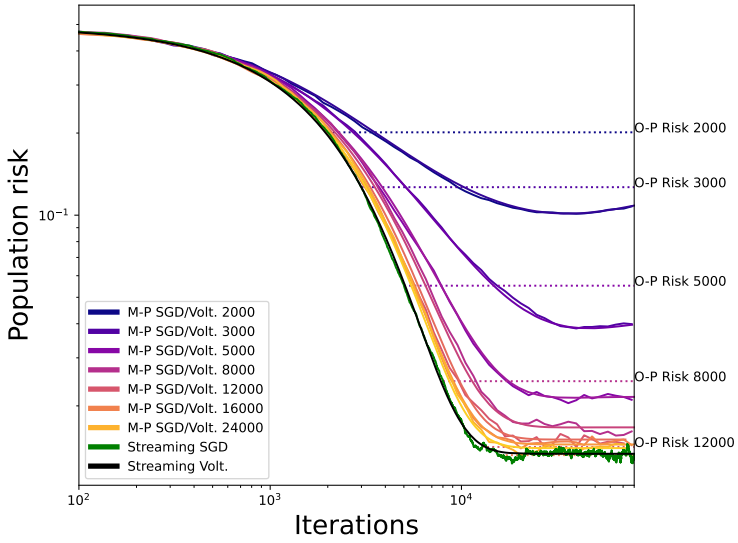


Figure 6: Risk curves for a simple linear regression problem. Multi-pass SGD, its high dimensional equivalent (i.e. "Volterra"), Streaming SGD (i.e. one-pass with varying dataset size), and the expected risk of homogenized SGD ("Streaming Volterra") are all plotted. Risk levels for streaming SGD at various levels n are plotted for comparison against the corresponding multi-pass version. Note that at smaller dataset sizes, multi-pass SGD improves greatly over one-pass SGD. At higher dataset sizes, they are similar and in fact multi-pass SGD always underperforms.

Now for an update, we have from (76)

$$\begin{aligned} q(x_{k+1}) - q(x_k) &= -\gamma_k \langle \nabla q(x_k), a_{i_{k+1}} \rangle (\langle a_{i_{k+1}}, x_k \rangle - b_{i_{k+1}}) \\ &\quad + \tfrac{\gamma_k^2}{2} \langle \nabla^2 q(x_k), a_{i_{k+1}}^{\otimes 2} \rangle (\langle a_{i_{k+1}}, x_k \rangle - b_{i_{k+1}})^2. \end{aligned} \tag{81}$$

We then proceed to compute the conditional means of both of these terms.

The first term we connect to the empirical risk, via

$$\begin{aligned} &\mathbb{E}[\langle \nabla q(x_k), a_{i_{k+1}} \rangle (\langle a_{i_{k+1}}, x_k \rangle - b_{i_{k+1}}) \mid \mathscr{F}_k] \\ &= \tfrac{1}{n} \langle \nabla q(x_k), A^T (Ax - b) \rangle \\ &= \langle \nabla q(x_k), \nabla \mathscr{L}(x_k) \rangle. \end{aligned}$$

As for the second term, we define $f_i(x) = \frac{1}{2}(\langle a_i, x_k \rangle - b_i)^2$ and observe this allows us to express it as

$$\tfrac{1}{2} \langle \nabla^2 q(x_k), a_{i_{k+1}}^{\otimes 2} \rangle (\langle a_{i_{k+1}}, x_k \rangle - b_{i_{k+1}})^2 = \langle \nabla^2 q(x_k), a_{i_{k+1}}^{\otimes 2} \rangle f_{i_{k+1}}(x_k).$$

If $a_{i_{k+1}}$ *were* independent of $\nabla^2 q(x_k) = \nabla^2 q$ (recall that q is quadratic), then we could approximate $\langle \nabla^2 q(x_k), a_{i_{k+1}}^{\otimes 2} \rangle$ (using something like the Hanson-Wright inequality) by

$$\langle \nabla^2 q(x_k), a_{i_{k+1}}^{\otimes 2} \rangle \approx \langle \nabla^2 q(x_k), \hat{K} \rangle.$$

Hence, it would suffice to work on the event $\mathcal{E}^q$ on which

$$\max_i |\langle \nabla^2 q, a_i^{\otimes 2} - \hat{K} \rangle| \leq \|\nabla q\|^2 n^{-1/2+\theta}.$$

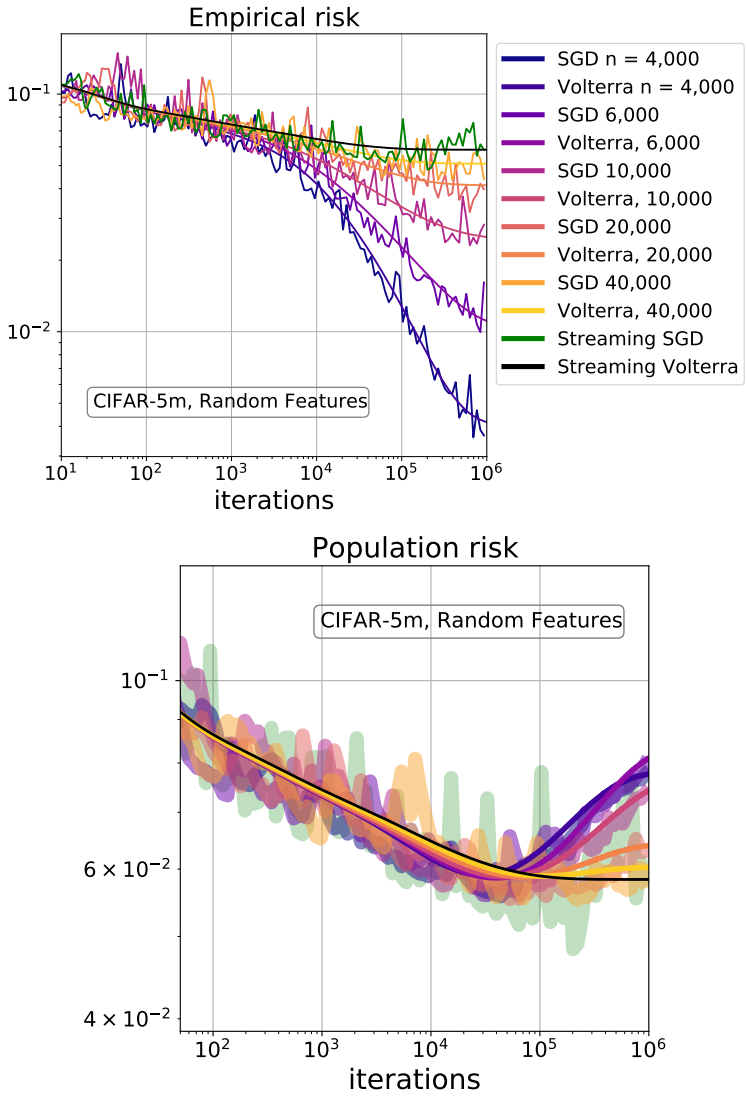


Figure 7: Risk curves for fully-connected random features model (with $d = 6000$) built on CIFAR-5m, empirical (top) and test-loss (bottom). The CIFAR-5m dataset [NNS21] is a synthetically generated 5 million data-point set of images, with the same class structure and image geometry as CIFAR-10. We compare running SGD on these curves as we vary the size of the subset used in each run. Note that in generalization performance, multi-pass SGD continues to improve generalization performance up to around $2 \times 10^4 = 20,000$ iterations (for all n displayed), which is about 5 epochs in the $n = 4000$ case. Achieving the same performance with streaming requires about $10^5 = 100,000$ iterations.

Now suppose we introduce the stopping time τ given by

$$\inf\{k : \max_i f_i(x_k) \geq n^\theta\},$$

then for $k > \tau$ we have

$$|\mathbb{E}[\langle \nabla^2 q(x_k), a_{i_{k+1}}^{\otimes 2} \rangle f_{i_{k+1}}(x_k) \mid \mathscr{F}_k] - \langle \nabla^2 q, \hat{K} \rangle \mathscr{L}(x_k)| \leq n^{-1/2+2\theta}.$$

Thus, we have a martingale decomposition

$$\begin{aligned} q(x_{k+1}^\tau) - q(x_k^\tau) = &-\gamma_k \langle \nabla q(x_k^\tau), \nabla \mathscr{L}(x_k^\tau) \rangle + \gamma_k \Delta M_k^{\text{lin}} \\ &+ \tfrac{\gamma_k^2}{2} \langle \nabla^2 q, \hat{K} \rangle \mathscr{L}(x_k^\tau) + \gamma_k \text{KL}_k + \gamma_k^2 \Delta M_k^{\text{quad}}, \end{aligned} \tag{82}$$

where KL_k is a deterministic error controlled by $n^{-1/2+2\theta}$ on $\mathcal{E}^q$. The martingale increments ΔM_k^{lin} and ΔM_k^{quad} can be controlled using martingale concentration techniques and the implied control from the stopping time τ.

Now as in (42), we perform this analysis over a class of functions. This function class only need to be modified slightly, to account for the change of β. So we define:

$$\begin{aligned} &\mathcal{Q}_n(q) := \mathcal{Q}_n(q, \hat{K}) = \\ &\Big\{ q(x), \quad (\nabla q(x))^T R(z; \hat{K}) x, \quad x^T R(y; \hat{K}) (\nabla^2 q) R(z; K) x, \\ &\qquad (\nabla q(x))^T R(z; \hat{K}) A^T b, \quad x^T R(y; \hat{K}) (\nabla^2 q) R(z; K) A^T b, \quad \forall\, z, y \in \Gamma \Big\}, \end{aligned} \tag{83}$$

and as in the streaming setting, we use a function class $\mathcal{Q} = \mathcal{Q}_n(\mathscr{L}, \hat{K}) \cup \mathcal{Q}_n(\|\cdot\|^2, \hat{K}) \cup \mathcal{Q}_n(\mathscr{R}, \hat{K})$, where $\mathscr{R}$ is the additional risk that we look to use.

Now the remainder of the proof proceeds as follows.

1. We need to show we can work on the event $\mathcal{E}^q$ for all $q \in \mathcal{Q}$. This is where Assumption 4 plays its role (specifically for $\mathscr{R}$). For the norm $\|\cdot\|^2$ and for $\mathscr{L}$, we get this control from Assumption 2.

2. Let σ be the first time k that $\mathscr{L}(x_k) > C$ (for a large but unimportant C independent of d, n). Now show that τ does not occur before σ with overwhelming probability. This uses a *bootstrap* argument, which shows that under the assumption the max-coordinate has some initial control, it can be improved with high probability.

3. The martingale terms are controlled with overwhelming probability using Freedman-inequality type bounds.

References

[AGJ21] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. "Online stochastic gradient descent on non-convex losses from high-dimensional inference". In: *The Journal of Machine Learning Research* 22.1 (2021), pp. 4788–4838.

[Arn+23] Luca Arnaboldi et al. "Escaping mediocrity: how two-layer networks learn hard single-index models with SGD". In: *arXiv preprint arXiv:2305.18502* (2023).

[BAGJ22] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. "High-dimensional limit theorems for sgd: Effective dynamics and critical scaling". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 25349–25362.

[BCN18a] Léon Bottou, Frank E Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning". In: *Siam Review* 60.2 (2018), pp. 223–311.

[BCN18b] Léon Bottou, Frank E Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning". In: *SIAM review* 60.2 (2018), pp. 223–311.

[BCW22] Raghu Bollapragada, Tyler Chen, and Rachel Ward. "On the fast convergence of minibatch heavy ball momentum". In: *arXiv preprint arXiv:2206.07553* (2022).

[Bot10] Léon Bottou. "Large-scale machine learning with stochastic gradient descent". In: *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*. Springer. 2010, pp. 177–186.

[Bot98] Leon Bottou. "Online learning and stochastic approximations". In: *On-linelearning in neural networks* 17.9 (1998), p. 142.

[CP23a] Elizabeth Collins-Woodfin and Elliot Paquette. "High-dimensional limit of one-pass SGD on least squares". In: *arXiv e-prints*, arXiv:2304.06847 (Apr. 2023), arXiv:2304.06847. DOI: 10.48550/arXiv.2304.06847. arXiv: 2304.06847 [math.PR].

[CP23b] Elizabeth Collins-Woodfin and Elliot Paquette. "High-dimensional limit of one-pass SGD on least squares". In: *arXiv e-prints*, arXiv:2304.06847 (Apr. 2023), arXiv:2304.06847. DOI: 10.48550/arXiv.2304.06847. arXiv: 2304.06847 [math.PR].

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7 (2011).

[Ger+22] Cedric Gerbelot et al. "Rigorous dynamical mean field theory for stochastic gradient descent methods". In: *arXiv preprint arXiv:2210.06591* (2022).

[Gol+20] Sebastian Goldt et al. "Modeling the influence of data structure on learning in neural networks: The hidden manifold model". In: *Physical Review X* 10.4 (2020), p. 041044.

[HS21] Boris Hanin and Yi Sun. "How Data Augmentation affects Optimization for Linear Regression". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 8095–8105. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/442b548e816f05640dec68f497ca38ac-Paper.pdf.

[HSS12] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent". In: *Cited on* 14.8 (2012), p. 2.

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[Kid+18] Rahul Kidambi et al. "On the insufficiency of existing momentum schemes for stochastic optimization". In: *2018 Information Theory and Applications Workshop (ITA)*. IEEE. 2018, pp. 1–9.

[KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*. Springer. 2016, pp. 795–811.

[KS91] Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*. Second. Vol. 113. Graduate Texts in Mathematics. Springer-Verlag, New York, 1991, pp. xxiv+470. DOI: 10.1007/978-1-4612-0949-2. URL: https://doi.org/10.1007/978-1-4612-0949-2.

[KY] H. Kushner and G.G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability.

[KY17] A. Knowles and J. Yin. "Anisotropic local laws for random matrices". In: *Probab. Theory Related Fields* 169.1-2 (2017), pp. 257–352.

[LeC+98] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[Lee+22] Kiwon Lee et al. "Trajectory of Mini-Batch Momentum: Batch Size Saturation and Convergence in High Dimensions". In: *To Appear in NeurIPS 2022*, arXiv:2206.01029 (June 2022), 38pp. arXiv: 2206.01029 [math.OC].

[MBB18] Siyuan Ma, Raef Bassily, and Mikhail Belkin. "The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 3325–3334.

[MY18] Jerry Ma and Denis Yarats. "Quasi-hyperbolic momentum and Adam for deep learning". In: *International Conference on Learning Representations*. 2018.

[NNS20] Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. "The deep bootstrap framework: Good online learners are good offline generalizers". In: *arXiv preprint arXiv:2010.08127* (2020).

[NNS21] P. Nakkiran, B. Neyshabur, and H. Sedghi. "The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers". In: *International Conference on Learning Representations (ICLR)*. 2021.

[Oks13] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[Paq+21] Courtney Paquette et al. "SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 3548–3626. arXiv: 2102.04396 [math.OC].

[Paq+22a] Courtney Paquette et al. "Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties". In: *arXiv e-prints*, arXiv:2205.07069 (May 2022), 64pp. arXiv: 2205.07069 [math.ST].

[Paq+22b] Courtney Paquette et al. "Implicit Regularization or Implicit Conditioning? Exact Risk Trajectories of SGD in High Dimensions". In: *To Appear in NeurIPS 2022*, arXiv:2206.07252 (June 2022), 33pp. arXiv: 2206.07252 [stat.ML].

[PP21] Courtney Paquette and Elliot Paquette. "Dynamics of Stochastic Momentum Methods on Large-scale, Quadratic Models". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 9229–9240. arXiv: 2106.03696 [math.OC]. URL: https://proceedings.neurips.cc/paper/2021/file/4cf0ed8641cfcbbf46784e620a0316fb-Paper.pdf.

[RM51] Herbert Robbins and Sutton Monro. "A stochastic approximation method". In: *The annals of mathematical statistics* (1951), pp. 400–407.

[RR08] A. Rahimi and B. Recht. "Random features for large-scale kernel machines". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 20. 2008, pp. 1177–1184.

[SK19] Connor Shorten and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *Journal of big data* 6.1 (2019), pp. 1–48.

[SS95] David Saad and Sara Solla. "Dynamics of on-line gradient descent learning for multilayer neural networks". In: *Advances in neural information processing systems* 8 (1995).

[Sut+13] Ilya Sutskever et al. "On the importance of initialization and momentum in deep learning". In: *International conference on machine learning*. PMLR. 2013, pp. 1139–1147.

[Vei+22] Rodrigo Veiga et al. "Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 23244–23255.

[Ver18] Roman Vershynin. *High-dimensional probability*. Vol. 47. Cambridge Series in Statistical and Probabilistic Mathematics. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018, pp. xiv+284. DOI: 10.1017/9781108231596. URL: https://doi.org/10.1017/9781108231596.

[YO19] Yuki Yoshida and Masato Okada. "Data-dependence of plateau phenomenon in learning with neural network—Statistical mechanical analysis". In: *Advances in Neural Information Processing Systems* 32 (2019).