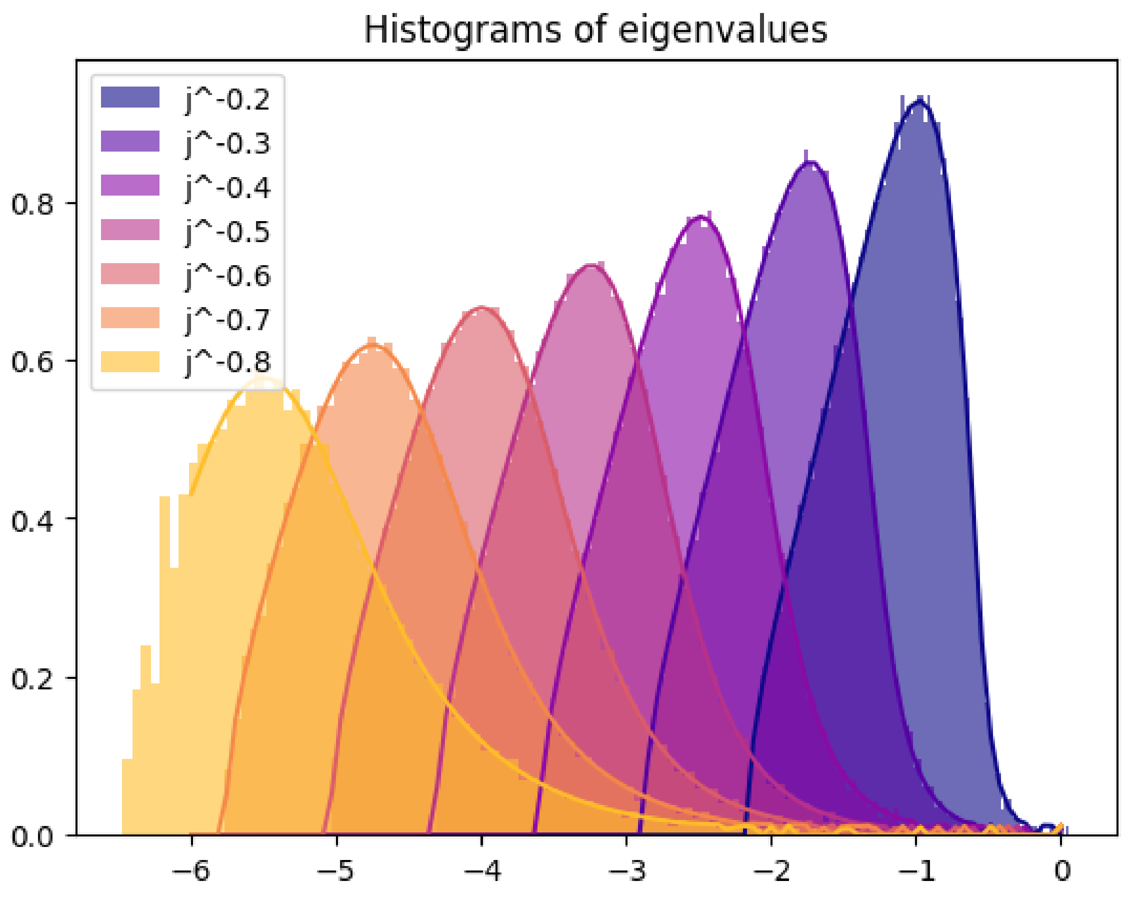
Random matrix theory of high-dimensional optimization

Elliot Paquette

Version: July 1, 2024



Course Content. Random matrix theory, Dyson equations, Deformed Marchenko-Pastur laws, Optimization theory of the high-dimensional random least squares problem, stochastic gradient descent

Summary of Contents

Foreword and Acknowledgements

These notes were developed to complement two summer schools from 2024: the random matrix theory summer school at the University of Michigan, and the CRM-PIMS probability summer school in Montreal.

They are intended to random matrix theory and theory for analyzing the gradient based optimization on random problems. The random matrix theory was highly influenced by the Dyson equation method, which provides a systematic way to analyze lots of structured random matrix problems. The Dyson method for Wigner-like (or weakly correlated) random matrices has relatively recently been deployed to great extent in systematizing many difficult estimates; see [Erd19] for an account. In particular, it is well-tuned to giving deterministic expressions for statistics of resolvents $R(z; A^T A)$ where A is a random matrix, and many random matrix problems admit convenient representations in terms of resolvents.

One idiosyncracy (or potentially, one innovation) of these notes is the *Newton flow*, for comparing the resolvent $R(z; A^T A)$ to the solution of the Dyson equation. This was done to give a self-contained way to show how estimates made for the resolvent $R(z; A^T A)$ propagate to error estimates between $R(z; A^T A)$ and the associated deterministic equivalent. These notes also implement this Newton flow in the canonical semicircle law case, to show how this method can be used.

The second component of these notes is the analysis of gradient based optimization algorithms on random least squares problems (especially gradient descent and stochastic gradient descent), in particular showing how the loss curves of the these algorithms can be represented in terms of various random matrix statistics developed in the first part.

At the time of writing this, the notes are still works in progress, and they should be updated over the course of the CRM-PIMS summer school!

These notes would not be possible without all the wonderful feedback from the participants of the Michigan summer school in random matrix theory. Thanks for all your efforts! I've reinforced parts of the document which should hopefully address some sources of confusion on my part that you all uncovered.

Thanks also to Yizhe Zhu and Ben Robinson who gave excellent feedback on various components of these notes. Thanks also to Lucas Benigni for discussions that helped develop some sections of these notes. Thanks to Mert Vural for catching typos and providing feed-

back on the first edition of these notes.
Elliot Paquette

July 1, 2024

Acknowledgements for the 2023 version of these notes "high dimensional limits of stochastic gradient descent"

These notes integrate an earlier version of these notes. 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 everyhing 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

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

This gives a convenient representation for quadratic forms as well:

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

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

In context, it may be simpler to use a numerical subscript (denoting which axis or axes are constracted) or the unique label of the space.

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

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

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

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 resolvent in the spectral parameter are given by, for any $k \in \mathbb{N}$ and at all $z \in \mathbb{C} \setminus \operatorname{Spec}(A)$

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

The resolvent is also differentiable in the matrix, and we have, whenever $z \in \mathbb{C} \setminus \operatorname{Spec}(A)$

$$D(R(z;A))[B] := \lim_{\epsilon \to 0} \frac{R(z;A+\epsilon B) - R(z;A)}{\epsilon} = -R(z;A)BR(z;A),$$

which is to say the directional derivative of the resolvent in A in the B direction is as reported.

In the special case that A and B differ by a low-rank matrix, there is another formula which can be more fruitful:[1]

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

[1] While it is a corollary, it is easier to simply check the identity directly, multiplying through by $(A - z\,\mathrm{Id})$ on both sides and turning the crank.

$\mathbb{M}(k,k)$ which is invertible

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

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

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

In the same vein, for working with resolvents, it is frequently helpful to be able to compute entries or blocks of the resolvent in terms of other entries. The following is a direct consequence of the using row-reduction to invert a matrix:[2]

[2] This differs slightly from other versions, such as Wikipedia, and is advantageous when both A and D have easy and/or interpretable inverses.

Lemma 1 (Schur complement formula): Suppose that the matrix M is written in blocks as

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

Then provided all the inverses are well-defined M^{-1} can be given by

$$M^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}.$$

Proof. Multiply MM^{-1} and verify it is Id. □

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

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

where

- $\operatorname{Res}(f; z) = r_{-1}$ 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 over all the cuves in the chain.

When integrating a matrix-valued function, it can be defined entry-by-entry. From the linearity of the mapping of matrix-to-entry, this must commute with integration in a parameter.

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 $\mathrm{Re}\,\mathrm{Spec}(A) > 0$, $\log A = \frac{-1}{2\pi i}\oint_\gamma \log(z)R(z;A)dz$ 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 *Stieltjes transforms*

The Stieltjes transform is an essential tool for working with random matrices, where it largely plays the role of the Fourier transform of traditional random matrix theory.

Definition 6 (Stieltjes): For a finite measure μ on $\mathbb{R}$, the *Stieltjes transform* of μ is the function

$$s_\mu(z) := \int_{\mathbb{R}} \frac{\mu(\mathrm{d}x)}{x - z},$$

which maps the upper-half plane $\mathbb{H}$ to itself.

The Stieltjes transform encodes properties of the measure analytically. In particular, the mapping of $\mu \to s_\mu$ is injective, and so to know the Stieltjes transform is to know the measure. Recovering the measure μ from s_μ can be done by Stieltjes inversion.

Theorem 5: Stieltjes Inversion

For a Stieltjes transform s_μ, the measure μ can be recoverd by taking the limit

$$\frac{1}{\pi} \operatorname{Im} s_\mu(x + \tfrac{i}{t})\,\mathrm{d}x \xrightarrow[t\to\infty]{\text{law}} \mu(\mathrm{d}x).$$

Proof. By explicit computation, we have the imaginary party of s_μ is given by

$$s_\mu(x + \tfrac{i}{t}) = \int_{\mathbb{R}} \frac{t\mu(\mathrm{d}y)}{((y-x)t)^2 + 1},$$

which is the density of the convolution of μ with a $X\frac{1}{t}$ for a Cauchy random variable X. This law converges to μ on taking $t \to \infty$. □

This leads to an equivalent formulation of weak convergence of probability measures (and weak-in-probability convergence of random measures)[3]:

Lemma 2 (Stieltjes characterization of weak convergence): A sequence of (Borel) probability measures $\{\mu_n\}$ converges weak-* to μ_∞ if and only if for all $z \in \mathbb{H}$, $s_{\mu_n}(z) \to s_{\mu_\infty}(z)$.

We also can use functions that look like Stieltjes transforms to construct measures. A key piece of complex analysis is the Herglotz representation theorem:

[3] A sequence of random finite measures μ_n on $\mathbb{R}$ converges weakly in probability if there is a (possibly random) measure μ_∞ so that for any bounded continuous ϕ $\int \phi\mu_n \xrightarrow[n\to\infty]{\Pr} \int \phi\mu$.

Theorem 6: Herglotz

Suppose that F is an analytic function from $\mathbb{H} \to \mathbb{H}$. Then there exists real numbers a, b with $a \geq 0$ and a measure μ on $\mathbb{R}$ so that $\int \frac{\mu(\mathrm{d}x)}{1+x^2} < \infty$ so that for all $z \in \mathbb{H}$

$$F(z) = az + b + \int_{\mathbb{R}} \left\{ \frac{1}{x - z} - \frac{x}{1 + x^2} \right\} \mu(\mathrm{d}x).$$

See [DK05, Theorem 1.4.2] for a proof.

In particular, for the purpose of identifying Stieltjes transforms of finite measures, it suffices to further know the behavior of F for z having large imaginary part.

Corollary 5 (Characterization of Stieltjes transforms): Suppose that F is an analytic function from $\mathbb{H} \to \mathbb{H}$, and suppose further that

$$F(it)t \xrightarrow[t\to\infty]{} ic \quad \text{where} \quad c \in [0, \infty).$$

Then there exists μ on $\mathbb{R}$ with $\mu(\mathbb{R}) = c$ so that

$$F(z) = \int_{\mathbb{R}} \frac{1}{x - z} \mu(\mathrm{d}x).$$

Proof. As $F : \mathbb{H} \to \mathbb{H}$ is analytic, we may apply Theorem 6. Taking the imaginary part of $F(it)t$, we have

$$\operatorname{Im} F(it)t = at^2 + \int_{\mathbb{R}} \frac{t^2}{x^2 + t^2} \mu(\mathrm{d}x).$$

As this converges to c we must have, from dominated oncvergence that $a = 0$ and $\mu(\mathbb{R}) = c$. We therefore have

$$F(z) = b + \int_{\mathbb{R}} \left\{ \frac{1}{x - z} - \frac{x}{1 + x^2} \right\} \mu(\mathrm{d}x),$$

and since μ is finite, we may expand the terms of the integral, to conclude for some other real constant b'

$$F(z) = b' + \int_{\mathbb{R}} \frac{1}{x - z} \mu(\mathrm{d}x).$$

Taking the real part

$$\operatorname{Re} F(it) = b' + \int_{\mathbb{R}} \frac{x}{x^2 + t^2} \mu(\mathrm{d}x),$$

and hence as $\operatorname{Re} F(it) \to 0$, we have $b' = 0$ as well, which concludes the proof. □

1.6 Martingales and concentration

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

Definition 7 (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 3 (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 8 (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)[4]

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

[4] 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.

Theorem 7: 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. [5] Define

[5] 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 9 (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 8: 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.

1.7 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 10 (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 11.

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

Definition 11 (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 4 (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 12 (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 13 (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 $\lambda \in \mathbb{R}$

$$\mathbb{E}[e^{\lambda(M_n - M_{n-1})} \mid \mathscr{F}_{n-1}] \leq e^{\lambda^2 V_n/2} \quad \text{a.s.}$$

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 9: 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 5 (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 $\lambda \in \mathbb{R}$

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

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 6 (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 $t \geq 0$

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

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

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

1.8 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 6 (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 7 (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 $S, t > 0$,

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

1.9 *Moment concentration of multilinear forms of independent random variables*

While Orlicz-type bounds are efficient at bounding probabilities with exponential rates, we often do not need such precision, and simpler polynomial rates of high order will suffice. (As we often have to take union bounds over families of events which are a polynomial in the system size).

These are most easily formulated in terms of the simpler Hölder norms. So for a $p \geq 1$ we use the simpler:

Definition 14 (Hölder-norms): For any $p \geq 1$ and any real-valued random variable X, define

$$\|X\|_p = (\mathbb{E}(|X|^p))^{1/p}.$$

The subgaussian and subexponential random variables can be characterized in terms of their moments. In particular, we can define norms equivalent to the Orlicz-norms in terms of the usual Hölder norms:

Definition 15 (Moment-norms): For any $p \geq 1$ and any real-valued random variable X, define the m_p norm

$$\|X\|_{m_p} = \inf\{q \geq 1 : q^{-1/p} \times \|X\|_q\},$$

and a dyadic counterpart

$$\|X\|_{md_p} = \inf\{k \in \mathbb{N} : 2^{-k/p} \times \|X\|_{2^k}\}.$$

These are equivalent to the Orlicz norms up to constants that depend only on p:

Theorem 10: Moment-Characterization

For any $p \geq 1$ there are positive constants a_p, b_p, c_p so that for all real-valued random variables X,

$$a_p\|X\|_{\psi_p} \leq b_p\|X\|_{md_p} \leq \|X\|_{m_p} \leq \|X\|_{md_p} \leq c_p\|X\|_{\psi_p}.$$

The first multilinear form to bound is the sum of independent real-valued random variables. A key idea in this direction is *symmetrization*. In particular, suppose that $\{X_j\}_1^n$ are independent real-valued random variables. Enlarge the probability space to include an independent copy $\{Y_j\}_1^n$ of these random vectors and independent collection $\{\epsilon_j\}_1^n$ of iid Rademacher random variables[6]. The distribu-

[6] Unif$\{1, -1\}$

tions of $X_j - Y_j$ are all symmetric and independent, and so

$$(X_j - Y_j : 1 \leq j \leq n) \overset{\text{law}}{=} (\epsilon_j(X_j - Y_j) : 1 \leq j \leq n).$$

Then by Jensen's inequality for $q \geq 1$

$$\begin{aligned}
\left\| \sum_j (X_j - \mathbb{E}X_j) \right\|_q^q &= \mathbb{E}\left| \sum_j (X_j - \mathbb{E}Y_j) \right|^q \\
&\leq \mathbb{E}\left| \sum_j (X_j - Y_j) \right|^q \\
&= \left\| \sum_j (\epsilon_j(X_j - Y_j)) \right\|_q^q \\
&\leq 2^q \left\| \sum_j (X_j - \mathbb{E}X_j) \right\|_q^q .
\end{aligned}$$

In the last step, we have applied the triangle inequality for the q-norms and again used the symmetries in law.

Now the signs ϵ_j are iid subgaussian, and so we may apply subgaussian concentration *conditionally* on $(X - Y)$. Writing $\| \cdot | \mathscr{F} \|_{(\cdot)}$ for the norms, conditioned on $(X - Y)$,

$$\begin{aligned}
\left\| \sum_j (\epsilon_j(X_j - Y_j)) | \mathscr{F} \right\|_q &\leq q^{1/2} \left\| \sum_j (\epsilon_j(X_j - Y_j)) | \mathscr{F} \right\|_{m_2} \\
&\leq c_2 q^{1/2} \left(\sum_j ((X_j - Y_j)^2) \right)^{1/2} \\
&\leq c_2 (2q)^{1/2} \left(\sum_j ((X_j - \mathbb{E}X_j)^2) \right)^{1/2}
\end{aligned}$$

for c_2 an absolute constant relating to subgaussian norms. Taking the expected q-th power on both sides

$$\left\| \sum_j (X_j - \mathbb{E}X_j) \right\|_q^q \leq c_2^q (2q)^{q/2} \left\| \sum_j (X_j - \mathbb{E}X_j)^2 \right\|_{q/2}^{q/2}. \tag{20}$$

Thus we have halved the order of the Hölder norm at the cost of squaring the summands. In principle, this could be developed further (for example recursively applying this identity), but it will be enough to get a bound on sums. [7]

[7] If

Lemma 8 (Linear form): Suppose $\{X_j\}_1^n$ are independent and suppose that $a = \{a_j\}_1^n$ are complex constants. Then there is an absolute constant c so that for all $q \geq 2$

$$\left\| \sum_j a_j (X_j - \mathbb{E}X_j) \right\|_q \leq c\sqrt{q} \|a\|_2 \max_j \|X_j - \mathbb{E}X_j\|_q.$$

Proof. By taking real and imaginary parts, it suffices to prove the claim for real scalars. For real scalars, by (20)

$$\left\| \sum_j a_j (X_j - \mathbb{E}X_j) \right\|_q^2 \leq c_2^2 (2q) \left\| \sum_j a_j^2 (X_j - \mathbb{E}X_j)^2 \right\|_{q/2}. \tag{21}$$

Applying the triangle inequality to the right hand side,

$$\left\|\sum_j a_j(X_j - \mathbb{E}X_j)\right\|_q^2 \le c_2^2(2q)\sum_j a_j^2 \left\|(X_j - \mathbb{E}X_j)^2\right\|_{q/2},$$

and the claimed bound follows quickly. □

This strategy generalizes to higher forms as well, with one crucial observation.[8] The random signs $\{\epsilon_j\}$ can also be selected independent of the σ-algebra of invariant functions

[8] This idea is from [Whi60].

$$\mathscr{F} := \sigma\left(F(X,Y) : F : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, F(A,B) = F(B,A)\right),$$

as the signs correpond to swapping entries of X and Y. Hence for a symmetric matrix A

$$\begin{aligned}\langle X \otimes X, A\rangle - \langle Y \otimes Y, A\rangle &= \langle (X-Y)\otimes(X+Y), A\rangle \\ &\overset{\text{law}}{=} \sum_j \epsilon_j (X_j - Y_j)(A(X+Y))_j.\end{aligned}$$

Thus, conditioning on $\mathscr{F}$

$$\|\langle X\otimes X - Y \otimes Y, A\rangle | \mathscr{F}\|_q \le c_2 q^{1/2}\left(\sum_j (X_j - Y_j)^2\left((A(X+Y))_j\right)^2\right)^{1/2}.$$

For $q \ge 2$, we can take the q-th power on both sides and bound

$$\begin{aligned}\mathbb{E}\left(\langle X\otimes X - Y\otimes Y, A\rangle\right)^q &\le (2c_2^2 q)^{q/2}\sum_j \mathbb{E}|X_j - Y_j|^q\left|(A(X+Y))_j\right|^q \\ &\le (8c_2^2 q)^{q/2}\sum_{i,j}|A_{i,j}|^q \mathbb{E}|X_j - Y_j|^q |X_i + Y_i|^q \\ &\le 4(16\cdot 8c_2^2 q)^{q/2}\left(\sum_{i,j}|A_{i,j}|^q\right)\max_j \mathbb{E}|X_j|^{2q}.\end{aligned}$$

Note that the sum of q-th powers of A we can bound by

$$\sum_{i,j}|A_{i,j}|^q \le \left(\sum_{i,j}|A_{i,j}|^2\right)\|A\|_{op}^{q-2} \le \|A\|_{HS}^q,$$

from which we conclude:

Lemma 9 (Quadratic form): Suppose $\{X_j\}_1^n$ are independent, centered and have variance 1. Suppose that A is a complex matrix. Then there is an absolute constant C so that for all $q \ge 2$

$$\|\langle X\otimes X, A\rangle - \operatorname{Tr}(A)\|_q \le C\sqrt{q}\|A\|_{HS}\max_j \|X_j\|_{2q}.$$

Proof. For general complex A, the quadratic form $\langle X \otimes X, A \rangle$ can be written as a sum of quadratic forms in real symmetric matrices, each of which has Hilbert-Schmidt norm bounded (up to absolute constants) by the Hilbert-Schmidt norm of A. This can now be bounded by the method preceding the lemma. □

This generalizes to any order quadratic form the same way:

Lemma 10 (General form): Suppose $\{X_j\}_1^n$ are independent, centered and have variance 1. Suppose that A is a complex k-tensor for $k \geq 3$. Then there is an absolute constant C_k so that for all $q \geq 2$

$$\left\| \langle X^{\otimes k}, A \rangle - \mathbb{E} \langle X^{\otimes k}, A \rangle \right\|_q \leq C_k \sqrt{q} \|A\| \max_j \|X_j\|_{kq}.$$

Here $\| \cdot \|$ is the induced Hilbert-space norm, which is the sum of absolute squares of all entries of A.

1.10 *Gaussian interpolation and concentration*

Gaussian measure satisfies a wide-ranging, far stronger form of concentration, often just going by the name *Gaussian concentration*. This makes many (otherwise tricky) concentration of measure estimates practically trivial. The starting point for this is Gaussian integration by parts.

> **Lemma 11 (Stein's Lemma):** If $Z \overset{\text{law}}{=} N(0, \Sigma)$, then provided $f : \mathbb{R}^d \to \mathbb{R}$ is absolutely continuous and $\mathbb{E}\|\nabla f(Z)\| < \infty$
>
> $$\mathbb{E} Z f(Z) = \Sigma \mathbb{E} \nabla f(Z).$$

Proof. We may assume by approximation that f is compactly supported and C^1. In the one-dimensional case, with $Z \overset{\text{law}}{=} N(0,1)$

$$\mathbb{E} Z f(Z) = \mathbb{E} f'(Z).$$

This follows directly from integration-by-parts applied to the Gaussian density. Hence if we repeat this coordinate-by-coordinate, then we conclude that for $f : \mathbb{R}^d \to \mathbb{R}$ and for $Z \overset{\text{law}}{=} N(0, \mathrm{Id})$

$$\mathbb{E} Z f(Z) = \mathbb{E} \nabla f(Z).$$

So, for the claim, we now represent $Z = \sqrt{\Sigma} Y$ as an iid Gassian vector Y. We may also represent $f(Z) = (f \circ \sqrt{\Sigma})(Y) =: g(Y)$, which remains compactly supported and C^1. Then

$$\mathbb{E} Z f(Z) = \sqrt{\Sigma} \mathbb{E} Y g(Y) = \sqrt{\Sigma} \mathbb{E} \nabla g(Y) = \Sigma \mathbb{E} \nabla f(Z).$$

□

The next key idea is *Gaussian interpolation*.

> **Definition 16 (Gaussian interpolation):** For a general, mean-0 random vector X with finite second moments and covariance matrix Σ, we can define a flow from X to an independent Gaussian vector Y with matching covariance, given by
>
> $$Y^\alpha = \alpha X + \sqrt{(1-\alpha^2)} Y \quad \text{where} \quad \alpha \in [0,1].$$

This flow can be used in multiple ways: first, if we differentiate expectations along this flow and control the change, we can show that some expectations are close to the same as those of matching Gaussian moments, which is an instance of *universality*. We can also use it with two independent Gaussians X and Y of matching covariance, in which case we can use it to show concentration.[9]

[9] If we can flow from $f(Y^0)$ to $f(Y^1)$ without changing very much, then $f(Y^0)$ must be close to deterministic.

Lemma 12 (Dynamical covariance representation): Suppose X, Y are iid $N(0,\Sigma)$ and f, g are absolutely continuous and satisfy

$$\mathbb{E}\left(f(X)^2 + \|\Sigma\nabla f(X)\|^2 + g(X)^2 + \|\Sigma\nabla g(X)\|^2\right) < \infty.$$

Then

$$\operatorname{Cov}(f(X), g(X)) = \int_0^1 \mathbb{E}\langle \Sigma\nabla f(X), \nabla g(Y^\alpha)\rangle \,\mathrm{d}\alpha.$$

Proof. By approximation, we can assume that f and g are C^2 and have compact support. Then by independence of Y^1 from Y^0

$$\begin{aligned}\operatorname{Cov}(f(X), g(X)) &= \int_0^1 \mathbb{E} f(X)\frac{\mathrm{d}}{\mathrm{d}\alpha} g(Y^\alpha)\,\mathrm{d}\alpha \\ &= \int_0^1 \mathbb{E} f(X)\langle \nabla g(Y^\alpha), X - \tfrac{\alpha}{\sqrt{1-\alpha^2}} Y\rangle \,\mathrm{d}\alpha. \\ &= \int_0^1 \mathbb{E}\operatorname{Tr}\left(f(X)\nabla g(Y^\alpha)\otimes (X - \tfrac{\alpha}{\sqrt{1-\alpha^2}}Y)\right)\mathrm{d}\alpha.\end{aligned}$$

Now we apply Gaussian integration by parts once more. We note that taking the X gradient and Y gradients of $\nabla g(Y^\alpha)$ produce

$$\nabla_X \nabla g(Y^\alpha) = \alpha\nabla^2 g(Y^\alpha) \quad \text{and} \quad \nabla_Y \nabla g(Y^\alpha) = \sqrt{1-\alpha^2}\nabla^2 g(Y^\alpha),$$

with $\nabla^2 g$ the Hessian matrix of g. On the other hand the Y gradient of f is 0, and so we have

$$\begin{aligned}\operatorname{Cov}(f(X), g(X)) = \int_0^1 &\mathbb{E}\operatorname{Tr}(\Sigma\nabla f(X)\otimes\nabla g(Y^\alpha) + \alpha f(X)\Sigma\nabla^2 g(Y^\alpha) \\ &- \sqrt{1-\alpha^2}\tfrac{\alpha}{\sqrt{1-\alpha^2}}\alpha f(X)\Sigma\nabla^2 g(Y^\alpha))\,\mathrm{d}\alpha \\ = \int_0^1 &\mathbb{E}\operatorname{Tr}(\Sigma\nabla f(X)\otimes\nabla g(Y^\alpha))\,\mathrm{d}\alpha \\ = \int_0^1 &\mathbb{E}\langle\Sigma\nabla f(X), \nabla g(Y^\alpha)\rangle\,\mathrm{d}\alpha.\end{aligned}$$

□

A direct consequence of this representation is Gaussian concentration.[10]

[10] This is adapted from [AT07, Chapter 2] who further ascribes it to [CY12].

Theorem 11: Gaussian concentration

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz and $X \overset{\text{law}}{=} N(0,\Sigma)$. Suppose further that for some deterministic finite L

$$\langle\Sigma\nabla f(X), \nabla f(X)\rangle \le L^2$$

almost surely. Then $\mathbb{E}|f(X)|$ is finite and for all $t \geq 0$

$$\Pr(f(X) - \mathbb{E}f(X) \geq t) \leq \exp\left(-t^2/(2L^2)\right).$$

Proof. The integrability of $\mathbb{E}e^{C|f(X)|}$ for any C can be checked by expressing $X = \mu + \sqrt{\Sigma}Z$ for iid normal Z and mean μ and bounding on shells of $\|Z\|$. Furthermore, by shifting f, we may without loss of generality assume $\mathbb{E}f(X) = 0$.

Now we apply Lemma 12 with $g = e^{\lambda f}$ for fixed λ (or in fact we will reverse the roles of f and g), and we conclude for any $\lambda \geq 0$

$$\mathbb{E}f(X)e^{\lambda f(X)} = \int_0^1 \mathbb{E}\left(\langle \Sigma\nabla f(X), \lambda\nabla f(Y^\alpha)\rangle e^{\lambda f(X)}\right) \mathrm{d}\alpha.$$

Then giving an almost sure bound on the inner product,

$$\mathbb{E}f(X)e^{\lambda f(X)} \leq \lambda L^2 \mathbb{E}e^{\lambda f(X)}.$$

Hence with $M(\lambda) = \mathbb{E}e^{\lambda f(X)}$,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}M(\lambda) \leq \lambda L^2 M(\lambda).$$

Thus from Gronwall's inequality, for all $\lambda \geq 0$

$$M(\lambda) \leq M(0)e^{\lambda^2L^2/2} = e^{\lambda^2L^2/2}.$$

The tail bound now follows the same way as the usual Gaussian tail bounds (c.f. Theorem 9). □

1.11 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 17 (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 18 (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 19 (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 20 (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

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

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 21 (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 $t \geq 0$,

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

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 12: 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 13: 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 22 (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 13 (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 7.

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

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

Setting $\lambda = x/S$ gives

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

□

2 *Dyson equations and deterministic equivalents*

One of the central tools in the analysis of random matrices is the *resolvent* $R(z; A)$.[11] When random matrices are defined by their entries, resolvents allow the efficient computation of spectral data about the matrix. The key combination of ideas is (1): the smooth dependence of the resolvent on the entries of the matrix (and in particular the Woodbury identity Corollary 4) and (2): the ability to efficiently extract from the resolvent eigenvalue and eigenvector information.

[11] If unfamiliar, look at the background section on resolvents Section 1.2!

The smooth dependence of the resolvent on its entries is especially effective, when considering random matrices with independent entries of near equal variance. In this case, we would rightly expect the resolvent (which is a smooth function of many random variables) to behave as though it were deterministic; this is to say, we can find a deterministic matrix around which the random resolvent concentrates.[12]

[12] It is quite natural to jump towards the *expected* resolvent here. In some cases (many cases), this causes no problem, and we can and should look at the expected resolvent. On the other hand, if the random matrix A can have eigenvalues in some open set $\mathcal{U}$ of the plane with positive probability, the expected resolvent (which has a notion of spectrum) will put spectrum in this set $\mathcal{U}$. This can be undesirable as the deterministic equivalent will lose some qualitative similarities of the random matrix, such as the notion of 'spectral edge.'

The construction of this deterministic equivalent is most easily motivated by the case of a Gaussian symmetric random matrix A. Gaussians are determined by their covariance structure, and so this model is fully specified by a mean matrix $\mathbb{E}A$ and a 4-tensor that describes the covariances of $A - \mathbb{E}A$.[13] A slightly simpler setup is to assume that the matrix A arises by letting $A = P + \Sigma G \Sigma$ where P is deterministic (mean) matrix, $\Sigma \succeq 0$ is a positive semidefinite covariance matrix, and G is GOE:

[13] A full account of what can happen with a general Gaussian 4-tensor is beyond the scope of these notes. But some parts we will leave at this level of generality.

> **Definition 23 (GOE):** A matrix $G \in \mathbb{M}(n)$ has the n-dimensional GOE (Gaussian orthogonal ensemble) distribution if $\{G_{ij} : i \geq j\}$ are normally distributed, mean 0, and have the normalization $\mathbb{E}G_{ij}^2 = (1 + \delta_{ij})$. Call the shifted matrix $P + SGS$ affine-GOE if $P, S \in \mathbb{M}(n)$ are symmetric and $S \succeq 0$.

The normalization of diagonal vs. off-diagonal entries of the G is by convention and is chosen to give G the rotation-invariance property, so that $OGO^\intercal \overset{\text{law}}{=} G$ for all orthogonal matrices O. As a consequence, there is no loss of generality in taking $S \succeq 0$, as we can apply a polar decomposition to S.

The *Dyson-equation* for the resolvent provides a concise way to find the matrix around which this resolvent concentrates. The Dyson equation is derived, for an n-dimensional affine-GOE (or indeed for any Gaussian random matrix) starting from the tautological identity

$$R(z; A)(A - z\,\mathrm{Id}) = \mathrm{Id}\,. \tag{22}$$

To this equation, we will take expectations on both sides, assuming for the moment that $z \in \mathbb{H}$, the upper half-plane.

We will need a fundamental identity about the Gaussian distribution, which goes by *Stein's lemma* or *Gaussian integration-by-parts*. To formulate this, we'll need the following function class:

Definition 24 (Pseudo-Lipschitz): A function f between Banach spaces $(X, \|\cdot\|_X) \to (Y, \|\cdot\|_Y)$ is α-pseudo-Lipschitz with constant L (for $\alpha \geq 0$) if

$$\|f(x_1) - f(x_2)\|_Y \leq L\|x_1 - x_2\|_X(1 + \|x_1\|_X + \|x_2\|_X)^{\alpha}.$$

As a consequence of Rademacher's Theorem, pseudo-Lipschitz functions on $\mathbb{R}^n$ are almost-everywhere differentiable, and their derivative grows no faster than the α-power of the $\|x\|$ (in any norm) as $x \to \infty$.

Stein's Lemma 11 now tells us that for these functions that if $Z \stackrel{\text{law}}{=} N(0, \Sigma)$, then provided f is pseudo-Lipschitz

$$\mathbb{E}Zf(Z) = \Sigma\mathbb{E}\nabla f(Z).$$

With this in hand, we can apply it to (22). Write $P = \mathbb{E}A$, and consider the expectation of $R(z;A)_{ij}(A_{jk} - P_{jk})$. Then following Lemma 11, we have for fixed entries i, j, k [14]

$$\mathbb{E}R(z;A)_{ij}(A_{jk} - P_{jk}) = \mathbb{E}\langle C^{(jk)}, \nabla R(z;A)_{ij}\rangle$$

where $C^{(jk)} \in \mathbb{M}(n)$ is a slice of the covariance tensor $\mathbb{E}((A_{ab} - P_{ab})(A_{jk} - P_{jk}))$, where $\nabla R(z;A)_{ij}$ is the Jacobian matrix of R with respect to A, and which gives (from Corollary 3)

$$(\nabla R(z;A)_{ij})_{ab} = -R(z;A)_{ia}R(z;A)_{bj}.$$

Hence we can represent this as an outer product,

$$(\nabla R(z;A)_{ij}) = -\left(\mathbf{e}_i^T R(z;A)\right) \otimes \left(R(z;A)\mathbf{e}_j\right).$$

Now if we let $\tilde{A}$ be an iid copy of $A - P$, we can represent

$$\begin{aligned}\mathbb{E}R(z;A)_{ij}(A_{jk} - P_{jk}) &= -\mathbb{E}\langle \tilde{A}\tilde{A}_{jk}\left(\mathbf{e}_i^T R(z;A)\right) \otimes \left(R(z;A)\mathbf{e}_j\right)\rangle \\ &= -\mathbb{E}\left(\tilde{A}_{jk}\mathbf{e}_i^T R(z;A)\tilde{A}R(z;A)\mathbf{e}_j\right).\end{aligned}$$

Hence summing over j, we arrive at

$$\mathbb{E}\left(R(z;A)(A-P)\right)_{ik} = -\mathbb{E}\left(R(z;A)\tilde{A}R(z;A)\tilde{A}\right)_{ik}.$$

We conclude the following fundamental identity:

Lemma 14 (Proto-Dyson equation): Suppose that A is a Gaussian random matrix and suppose that $z \in \mathbb{C}$ is such that $A \mapsto R(z;A)$ is pseudo-Lipschitz, then with $P = \mathbb{E}A$ and $\tilde{A}$

[14] We use $\langle , \rangle$ as a generic inner-product, which will vary depending on the objects to which it is applied. For vectors it is the standard $\ell - 2$ inner-product. For matrices it is the Hilbert-Schdmit inner product. See the discussion in Section 1.1

A simple condition that guarantees pseudo-Lipschitzness is just that A is symmetric and $z \in \mathbb{H}$.

We write Proto-Dyson equation, as in every situation in this text, we will want to simplify this equation further. There a handful of situations (for example the Dyson equation for the Gaussian *Unitary* Ensemble) in which this equation is exact.

an iid copy of $A - P$,

$$\mathbb{E}\left(R(z;A)(A-P)\right) = -\mathbb{E}\left(R(z;A)\tilde{A}R(z;A)\tilde{A}\right).$$

For a fully general covariance tensor, we may be unable to simplify further. But for affine-GOE, we can evaluate the covariance tensor

$$\begin{aligned}
C_{ab}^{(jk)} &= \mathbb{E}((SGS)_{ab}(SGS)_{jk}) \\
&= \sum_{qrst} \mathbb{E}(S_{aq}G_{qr}S_{rb}S_{js}G_{st}S_{tk}) \\
&= \sum_{qr} \left(S_{aq}S_{rb}S_{jq}S_{rk} + S_{aq}S_{rb}S_{jr}S_{qk}\right) \\
&= \left((S^2)_{aj}(S^2)_{bk} + (S^2)_{ak}(S^2)_{bj}\right).
\end{aligned}$$

Hence, applying this to the Gaussian integration-by-parts

$$\begin{aligned}
&\mathbb{E}R(z;A)_{ij}(A_{jk} - P_{jk}) \\
&= \mathbb{E}\langle C^{(jk)}, \nabla R(z;A)_{ij}\rangle \\
&= -\mathbb{E}\sum_{ab}\left((S^2)_{aj}(S^2)_{bk} + (S^2)_{ak}(S^2)_{bj}\right) R(z;A)_{ia}R(z;A)_{bj} \\
&= -\mathbb{E}\left((R(z;A)S^2)_{ij}(R(z;A)S^2)_{jk} + (R(z;A)S^2)_{ik}(R(z;A)S^2)_{jj}\right).
\end{aligned}$$

Finally, summing in j, we conclude

$$\begin{aligned}
&\mathbb{E}R(z;A)(A-P) \\
&= -\mathbb{E}\left((R(z;A)S^2R(z;A)S^2) + (R(z;A)S^2)\operatorname{Tr}(R(z;A)S^2)\right).
\end{aligned} \tag{23}$$

Now it will transpire that the first term is basically negligible with respect to the second.[15]

Based upon this, we define the Dyson equation for the affine-GOE.

Definition 25 (Dyson equation for the affine GOE): Let $\mathbb{M}^+(n)$ be the set of all symmetric, complex, $n \times n$ matrices with positive-definite imaginary part. Define $\mathcal{S}$ as the linear map on $\mathbb{M}(n)$ given by

$$\mathcal{S}(M) := S^2 \operatorname{Tr}(MS^2),$$

which restricts to a self-map of $\mathbb{M}^+(n)$ when $S \succ 0$ (and to the closure of $\mathbb{M}^+(n)$, otherwise). The Dyson equation on $\mathbb{M}^+(n)$ is the matrix equation

$$M(P - \mathcal{S}(M) - z\,\mathrm{Id}) = \mathrm{Id}.$$

This leads to a general method of analysis, for finding the leading order behavior of many random matrix theory questions.

[15] If we imageine that $R(z;A)$ and S^2 are matrices of operator-norm bounded independent of dimension, then the entirety of the first term is bounded in operator-norm, while the second term will be (generically) an order of n larger.

Definition 26 (The Dyson equation method in RMT):

1. Find a Dyson equation (apply Gaussian integration by parts to (22)) for a Gaussian random matrix A, and finding leading terms.
2. Show the Dyson equation (Definition 25) is uniquely solvable.
3. Show that a solution of the perturbed Dyson equation

$$M(P - \mathcal{S}(M) - z\,\mathrm{Id}) = \mathrm{Id} + \xi$$

   for an error term ξ is close to the solution of the unperturbed Dyson equation (stability).
4. Show that the resolvent of a random matrix (Gaussian or otherwise) solves a perturbed Dyson equation, and hence is close to the unperturbed equation.

The exact metrics in which these comparisons are made depend on the desired result (in which sense the solution of M should be close to $R(z; A)$). If all we care about is the bulk spectral properties, then this sense is the normalized trace.

This method was greatly championed by [Erd19], [AEK20], [AEK19]. In more limited forms, it underpins all Stieltjes-transform based analyses of the spectral distribution, going back to [MP67].

It turns out that the second point in this list (for the affine GOE Dyson equation), holds in complete generality.

Theorem 14: Uniqueness of the solution of the Dyson equation

There is a unique solution of the Dyson equation for the affine GOE.

The theorem in generality is due to [HFS07, Theorem 2.1], which relies on the contractivity of the mapping $M \mapsto (P - \mathcal{S}(M) - z\,\mathrm{Id})^{-1}$ in an appropriate metric (the Carathéodory metric, which generalizes the hyperbolic metric). In many simple cases, it can be checked directly by using the Schwarz Lemma.

The general Dyson equation for the affine GOE is given by

$$M(P - S^2 \operatorname{Tr}(MS^2) - z\,\mathrm{Id}) = \mathrm{Id}\,.$$

Rearranging,

$$M = (P - S^2 \operatorname{Tr}(MS^2) - z\,\mathrm{Id})^{-1}.$$

Hence if we introduce $m(z) := \operatorname{Tr}(MS^2)$,

$$m(z) = \operatorname{Tr}\left(S^2(P - S^2 m(z) - z\,\mathrm{Id})^{-1}\right).$$

So, if we solve for the scalar m, we can express M in terms of m by

$$M = (P - S^2 m(z) - z\,\mathrm{Id})^{-1}.$$

Thus there is a single scalar equation that should be solved to find M. We highlight a few important cases below.

Example 3: Wigner matrices and the semicircle

Suppose that $P = 0$ and $S^2 = \frac{1}{n}\operatorname{Id}$. Then the Dyson equation becomes

$$M(-\tfrac{1}{n}\operatorname{Tr}(M)\operatorname{Id} - z\operatorname{Id}) = \operatorname{Id}.$$

Taking the normalized trace of both sides, and setting $s(z) = \frac{1}{n}\operatorname{Tr}(M)$, we arrive at

$$s(z)(-s(z) - z) = 1.$$

As this is a quadratic equation, we can solve it explicitly, which leads to

$$s(z) = \frac{-z \pm \sqrt{z^2 - 4}}{2}$$

The choice of root is determined by the condition $\operatorname{Im} s(z) > 0$. Using the principal branch $\sqrt{\cdot}$ this turns out to be

$$s(z) = \frac{-z + \sqrt{z-2}\sqrt{z+2}}{2}.$$

Having solved for s, the matrix M is therefore the diagonal matrix

$$M = (-s(z) - z)^{-1}\operatorname{Id} = s(z)\operatorname{Id},$$

where we have applied the quadratic equation to both sides.

Example 4: The free additive convolution with the semicircle

In the more general case that $S^2 = \frac{1}{n}\operatorname{Id}$, we have the more complicated Dyson equation

$$M(P - \tfrac{1}{n}\operatorname{Tr}(M)\operatorname{Id} - z\operatorname{Id}) = \operatorname{Id}.$$

Hence, setting $s(z) = \frac{1}{n}\operatorname{Tr}(M)$, we conclude

$$s(z) = \frac{1}{n}\operatorname{Tr}\left(P - (s(z) + z)\operatorname{Id}\right)^{-1}.$$

This equation has no general closed-form solution. But it represents an important operation, referred to as the *free additive convolution* (in this case, more specifically, the free additive convolution of the spectrum of P with the semicircle law), and lots of properties of it are known.
The solution of the Dyson equation is then

$$M = \left(P - (s(z) + z)\operatorname{Id}\right)^{-1}.$$

Example 5: The free multiplicative convolution with the semicircle

In contrast, in the case that $P = 0$, we instead get

$$M(-S^2\operatorname{Tr}(MS^2) - z\,\mathrm{Id}) = \mathrm{Id}\,.$$

If we set $m(z) = \operatorname{Tr}(MS^2)$, then this solves

$$m(z) = \operatorname{Tr}\left(\frac{S^2}{-S^2 m(z) - z\,\mathrm{Id}}\right).$$

Hence the Stieltjes transform $s(z) = \frac{1}{n}\operatorname{Tr}(M)$ satisfies

$$s(z) = \frac{1}{n}\operatorname{Tr}\left(\frac{1}{-S^2 m(z) - z\,\mathrm{Id}}\right).$$

This gives the *free multiplicative convolution* of the spectrum of S^2 with the semicircle.
The solution of the Dyson equation is in this case

$$M = \left(-S^2 - (s(z) + z)\,\mathrm{Id}\right)^{-1}.$$

2.1 *Stability and the Newton flow*

Suppose we let $F(M; z)$ be the left-hand-side of the Dyson equation

$$F(M; z) := M(P - S^2\operatorname{Tr}(MS^2) - z\,\mathrm{Id}). \tag{24}$$

The resolvent will be an approximate solution of the Dyson equation, which is to say

$$F(R(z; A); z) - \mathrm{Id} = \xi \approx 0,$$

in a way which will need to be quantified, and we would like to deduce from this that

$$R(z; A) \approx M, \quad \text{where} \quad F(M; z) = \mathrm{Id},$$

which uniquely defines M.

To do this, we need a version of something like the inverse function therem. Since the equations are potentially dimension dependent, we would also like it to be somewhat quantitative.

So we define the following ordinary differential equation:

This is a continuous limit of the 'damped Newton's method' for solving an equation. This is not the unique flow that one could consider for the task that we use. For example, one could change the flow by inserting a positive operator on the right hand side.

Definition 27 (Newton flow): Define the Newton flow, an ordinary differential equation for $\mathcal{M}(t) \in \mathbb{M}(n)$ given by

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mathcal{M}(t); z) = -(F(\mathcal{M}(t); z) - \mathrm{Id}),$$

where we set $\xi = F(\mathcal{M}(0); z) - \mathrm{Id}$.

Provided the flow is well-posed, meaning that we can solve for $\frac{d}{dt}\mathcal{M}(t)$,[16] we get that

$$(F(\mathcal{M}(t);z) - \mathrm{Id}) = e^{-t}\xi,$$

and hence the Newton ODE is equivalent to

$$\frac{d}{dt}F(\mathcal{M}(t);z) = -e^{-t}\xi. \tag{25}$$

Thus, as input to the analysis of the Newton ODE, it suffices to estimate the error $((F(\mathcal{M}(0);z)) - \mathrm{Id})$ at initialization (which will be the probabilistic input for the analysis of the resolvent), as well as showing the flow is well-posed.

For the affine-GOE, the equation that must be solved for $\mathcal{M}(t)$ is

$$\frac{d}{dt}F(\mathcal{M}(t);z) = \dot{\mathcal{M}}(P - S^2\,\mathrm{Tr}(XS^2) - z\,\mathrm{Id}) + \mathcal{M}(-S^2\,\mathrm{Tr}(\dot{\mathcal{M}}S^2)). \tag{26}$$

[16] Existence-uniqueness theory for ODEs say that we can extend a solution of an initial value-problem with locally Lipschitz coefficients (as the Newton flow has) to a maximal interval of time, either forward or backward from the inital time. Say we are going forward in time from 0, then either the solution exists for all time, or else it exists on an interval $[0, T)$ for finite T. If the solution extends continuously to T, then we could further extend the solution in time. So well-posedness means that we can solve for $\frac{d}{dt}\mathcal{M}(t)$ at all times in terms of dynamical variables that remain continuous (in particular they do not blow up along the flow).

The case of GOE.

We start with the case of GOE, for which the equation becomes

$$\frac{d}{dt}F(\mathcal{M}(t);z) = \dot{\mathcal{M}}(-\mathfrak{s} - z) + \mathcal{M}(-\dot{\mathfrak{s}}),$$

where $\mathfrak{s}(t) = \frac{1}{n}\,\mathrm{Tr}(\mathcal{M}(t))$. Hence if we apply the normalized trace to both sides, we arrive at

$$\dot{\mathfrak{s}}(2\mathfrak{s} + z) = \frac{d}{dt}(\mathfrak{s}(\mathfrak{s}+z)) = e^{-t}\frac{1}{n}\,\mathrm{Tr}(\xi).$$

Thus provided $2\mathfrak{s} + z$ remains away from 0, we can solve for $\dot{\mathfrak{s}}$. If we solve for $\dot{\mathfrak{s}}$, we can then solve for the $\dot{\mathcal{M}}$ provided $\mathfrak{s} + z$ does not hit 0, and the flow $\dot{\mathcal{M}}$ is well-posed.

The simplest formal consequence of this is the following stability:

Lemma 15 (Semicircle-law-stability): Suppose that $|2s(z) + z|^2 > 2|\frac{1}{n}\,\mathrm{Tr}(\xi)|$ where $s(z)$ is the Stieltjes transform of the semicircle. Then the ODE for $\mathfrak{s}$ exists for all time and

$$|\mathfrak{s}(0) - \mathfrak{s}(\infty)| \le \frac{|\frac{1}{n}\,\mathrm{Tr}(\xi)|}{\sqrt{|2s(z)+z|^2 - 2|\frac{1}{n}\,\mathrm{Tr}(\xi)|}}.$$

If furthermore, the right-hand side of this equation is strictly less than 1, then the solution of $\mathcal{M}$ exists for all-time, and moreover,

$$\mathcal{M}(0) = \frac{1}{\mathfrak{s}(0) + z}\left(\mathcal{M}(\infty)(\mathfrak{s}(\infty) + z) - \xi\right).$$

Note that when $\mathrm{Im}\, z > 0$, since the semicircle law has non-negative imaginary part, $|2s(z) + z|$ is bounded below by $\mathrm{Im}\, z$. In fact, using the formula for $s(z)$ we have $2s + z = \sqrt{z-2}\sqrt{z+2}$ which vanishes at the spectral edges. This reflects that near the spectral edges, the spectrum of the GOE has a different scaling.

Proof. First, we need $(2\mathfrak{s}+z)$ to remain away from 0. The flow is reversible, and so we can either start it from 0 or we can run it back from ∞. At time infinity, the formula is explicit and is given by the formula for the Stieltjes transform of the semicircle (which in particular has positive imaginary part everywhere in the upper half plane). Thus provided that

$$\sqrt{2|\tfrac{1}{n}\operatorname{Tr}(\xi)|} < |2\mathfrak{s}(\infty)+z| = |2s(z)+z|,$$

we shall conclude (in what follows) that solution of $\mathfrak{s}$ exists for all time. This is because

$$\frac{\mathrm{d}}{\mathrm{d}t}(2\mathfrak{s}(t)+z)^2 = 2e^{-t}\tfrac{1}{n}\operatorname{Tr}(\xi),$$

and hence for any neighborhood of t around ∞

$$(2\mathfrak{s}(t)+z)^2 = (2s(z)+z)^2 - 2(1-e^{-t})\tfrac{1}{n}\operatorname{Tr}(\xi),$$

as we are led to the lower bound (over all t)

$$|2\mathfrak{s}(t)+z|^2 \geq |2s(z)+z|^2 - 2|\tfrac{1}{n}\operatorname{Tr}(\xi)|.$$

Therefore, integrating the flow for $\dot{\mathfrak{s}}$, we conclude

$$|\mathfrak{s}(0)-\mathfrak{s}(\infty)| \leq \frac{|\tfrac{1}{n}\operatorname{Tr}(\xi)|}{\sqrt{|2s(z)+z|^2 - 2|\tfrac{1}{n}\operatorname{Tr}(\xi)|}}.$$

Now as for the whole Newton flow, we have

$$\dot{\mathcal{M}}(-\mathfrak{s}-z) + \mathcal{M}(-\dot{\mathfrak{s}}) = -e^{-t}\xi,$$

and hence integrating in a neighborhood of $t = \infty$ in which the solution exists[17]

$$\mathcal{M}(\infty)(-\mathfrak{s}(\infty)-z) - \mathcal{M}(t)(-\mathfrak{s}(t)-z) = -e^{-t}\xi.$$

Then we can solve for $\mathcal{M}(t)$ and produce

$$\mathcal{M}(t) = \frac{1}{\mathfrak{s}(t)+z}\left(\mathcal{M}(\infty)(\mathfrak{s}(\infty)+z) - e^{-t}\xi\right),$$

and so this flow exists and is continuous provided $\mathfrak{s}(t)+z$ avoids 0.

At time ∞, we have

$$\frac{1}{s(z)+z} = -s(z),$$

which is bounded above uniformly in z in modulus by 1. Hence

$$|(\mathfrak{s}(t)+z)| \geq 1 - |\mathfrak{s}(\infty)-\mathfrak{s}(t)|,$$

and so provided that

$$\frac{|\tfrac{1}{n}\operatorname{Tr}(\xi)|}{\sqrt{|2s(z)+z|^2 - 2|\tfrac{1}{n}\operatorname{Tr}(\xi)|}} < 1,$$

we have the Newton flow exists for all time. □

[17] By changing time by $\tau = \log(1/(1-t))$, we can bring the infinite interval of time $(0,\infty)$ to $(0,1)$. Due to the e^{-t}, this makes the equation well-behaved in the τ variables at $\tau = 1$, and so there is an interval of time around $\tau = 1$ in which the solution exists.

2.2 *Proofs of the semicircle law*

We now complete the proof of the semicircle law using the machinery developed.

Theorem 15: (Easiest) Wigner semicircle

Suppose that X is a real Wigner matrix, so that $\{X_{ij}, i \geq j\}$ are independent and the first two moments match the GOE. Suppose further that $\mathbb{E}X_{ij}^8$ is bounded uniformly. Then the empirical spectral measure of $X/\sqrt{n}$ converges weakly in probability to the semicircle law.

This is often referred to as the 'leave-one-out' method of proof of the semicircle. Much weaker moment conditions are needed for the semicircle law to hold. For example, if the entries are iid, then a finite second moment suffices [Arn71].

Proof. Set $W = X/\sqrt{n}$. The statement is implied by showing that for any fixed z with $\operatorname{Im} z > 0$,

$$\tfrac{1}{n}\operatorname{Tr}(W - z\operatorname{Id})^{-1} - s(z) \xrightarrow[n\to\infty]{\Pr} 0,$$

with $s(z)$ the Stieltjes transform of the semicircle law.

To do this, by the Stability lemma, Lemma 15 it suffices to show that $\frac{1}{n}\operatorname{Tr}(\xi) \xrightarrow[n\to\infty]{\Pr} 0$ where for fixed z with $\operatorname{Im} z > 0$

$$\xi = (W - z\operatorname{Id})^{-1}(-\tfrac{1}{n}\operatorname{Tr}((W - z\operatorname{Id})^{-1}) - z\operatorname{Id}) - \operatorname{Id}.$$

Set $s_n(z) := \frac{1}{n}\operatorname{Tr}((W - z\operatorname{Id})^{-1})$. Then this satisfies

$$\tfrac{1}{n}\operatorname{Tr}(\xi) = s_n(z)\,(-s_n(z) - z) - 1.$$

We set $W^{[1]}$ to be the matrix in which we remove the first row and column of W. From the Schur complement formula,

$$\Big((W-z)^{-1}\Big)_{11} = (X_{11}/\sqrt{n} - z - \langle Y, (W^{[1]} - z)^{-1}Y\rangle/n)^{-1}.$$

Here $Y = Y^{[1]}$ is the first column of W, after removing its first entry.

The quadratic form concentrates, conditionally on $(W^{[1]} - z)^{-1}$, as we can evaluate its conditional variance

$$\begin{aligned}&\mathbb{E}\left(|\langle Y, (W^{[1]} - z)^{-1}Y\rangle - \operatorname{Tr}((W^{[1]} - z)^{-1})|^4 \;\middle|\; X^{[1]}\right)\\ &\leq C\|(W^{[1]} - z)^{-1}\|_F^4\\ &\leq Cn^2|\operatorname{Im} z|^{-4}.\end{aligned}$$

Hence by Markov's inequality, we can bound (with $W^{[j]}$ and $Y^{[j]}$ analogously defined in terms of removing the j-th row and column from W)

$$\max_{1\leq j\leq n} \tfrac{1}{n}|\langle Y^{[j]}, (W^{[j]} - z)^{-1}Y^{[j]}\rangle - \operatorname{Tr}((W^{[j]} - z)^{-1})| \xrightarrow[n\to\infty]{\Pr} 0.$$

The difference in the resolvents of W and $W^{[1]}$ (where we include a row and column of zeros in place of the first row/column) can be expressed by

$$R(z;W) = R(z;W^{[1]}) + R(z;W)(W^{[1]} - W)R(z;W^{[1]}).$$

Hence we can bound

$$|s_n(z) - \tfrac{1}{n}(R(z;W^{[1]}))| \le \tfrac{1}{n}|\operatorname{Im} z|^{-2}\|(W^{[1]} - W)\|_*,$$

where we have bounded the resolvents in norm and $\|\cdot\|_*$ is the nuclear norm. As this is a rank 2 matrix, we can bound the nuclear norm by a (dimension-independent) constant multiple of Hilbert-Schmidt norm and hence by the Euclidean norm of the first column of W.

Now as we can bound the fourth moment of this norm,

$$\max_{1\le j\le n} \tfrac{1}{n}\|(W^{[1]} - W)\|_* \xrightarrow[n\to\infty]{\Pr} 0$$

Putting everything together, we have that

$$\max_{1\le j\le n} \left| \left((W-z)^{-1}\right)_{jj} - (-z - s_n(z))^{-1} \right| \xrightarrow[n\to\infty]{\Pr} 0,$$

and hence

$$\tfrac{1}{n}|\operatorname{Tr}(\xi)| \xrightarrow[n\to\infty]{\Pr} 0.$$

□

This proof suffices for a global semicircle law. It does not generalize well to other statistics of the resolvent. One convenient class of statistics are *generalized entries*.

> **Definition 28 (Generalized entry):** A generalized entry of a matrix M is the trace $\operatorname{Tr}(AM)$ for a matrix A of nuclear norm 1.

The proof above does not extend easily to bounding generalized entries (in particular, controlling $\operatorname{Tr}(\xi A)$ requires exploiting cancellations between different resolvents). For the Gaussian case, this is a trivial consequence of powerful general Gaussian concentration of measure estimates.

Theorem 16: (Easy) (Gaussian) Wigner semicircle

Let G be GOE. Then there is a constant $C > 0$ so for any $\|A\|_* = 1$, any $t \ge 1$ and any $z \in \mathbb{H}$,

$$|\operatorname{Tr}(AR(z;G/\sqrt{n})) - \operatorname{Tr}(A)s(z)| \le Ct|\operatorname{Im} z|^{-4}n^{-1/2}$$

with probability $1 - e^{-t^2}$, provided the right hand side of the display is less than 1.

Proof. Let $Y = G/\sqrt{n}$. Applying (23) (with $S^2 = \frac{1}{n}\,\mathrm{Id}$)

$$\mathrm{Id} = -\frac{1}{n^2}\mathbb{E}[R(z;Y)^2] + \mathbb{E}[(R(z;Y))(-s_n(z) - z)],$$

where $s_n(z) = \frac{1}{n}\,\mathrm{Tr}(R(z;Y))$. Rearranging, we have

$$\mathbb{E}\xi = \frac{1}{n^2}\mathbb{E}[R(z;Y)^2].$$

Then for a generalized entry A

$$|\mathbb{E}\,\mathrm{Tr}(\xi A)| \leq \tfrac{1}{n^2}|\operatorname{Im} z|^{-2},$$

where we have bounded the operator norm of $\|R(z;Y)^2\|_{op} \leq |\operatorname{Im} z|^{-2}$.

In preparation to apply Gaussian concentration, we can now estimate the fluctuations of $\mathrm{Tr}(\xi A)$

$$\mathrm{Tr}(\xi A) = \mathrm{Tr}(R(z;Y)A)(-s_n(z) - z) - \mathrm{Id}\,.$$

Computing a Y_{ij} partial of this we get

$$\begin{aligned}\partial_{Y_{ij}}\mathrm{Tr}(\xi A) &= \partial_{Y_{ij}}\left(\mathrm{Tr}(R(z;Y)A)(-s_n(z) - z) - \mathrm{Id}\right)\\ &= (R(z;Y)AR(z;Y))_{ij}(s_n(z)+z) + \tfrac{1}{n}\,\mathrm{Tr}(R(z;Y)A)((R(z;Y))^2)_{ij}.\end{aligned}$$

Thus on summing the squares in preparation to use Gaussian concentration, Theorem 11, for an absolute constant $C > 0$

$$\begin{aligned}\sum_{ij}|\partial_{Y_{ij}}\mathrm{Tr}(\xi A)|^2 &\leq C\left(|s_n(z)+z|^2\|R(z;Y)AR(z;Y)\|^2 + \tfrac{1}{n^2}|\,\mathrm{Tr}(R(z;Y)A)|^2\|(R(z;Y))^2\|^2\right)\\ &\leq C\left(|\operatorname{Im} z|^{-6}\|A\|^2 + \tfrac{1}{n}|\operatorname{Im} z|^{-6}\|A\|_*^2\right).\end{aligned}$$

Hence we have a tail bound for all $t \geq 0$ and some absolute constant $C > 0$,[18]

$$\Pr(|\,\mathrm{Tr}(\xi A) - \mathbb{E}\,\mathrm{Tr}(\xi A)| > t) \leq 2\exp\left(\frac{-nt^2|\operatorname{Im} z|^6}{C\|A\|_*^2}\right).$$

□

The first proof of the Wigner semicircle (Theorem 15) in fact bounded on-diagonal entries of the resolvent. This can be extended to off-diagonal entries of the resolvent as well, but this stops short of giving an estimate for generalized entries. So perhaps it should not be surprising that for GOE, which is invariant to orthogonal changes of bases, we could easily give a proof of the semicircle law in terms of generalized entries.

[18] Here we have bounded $\|A\| \leq \|A\|_*$. This is costly for some A, in particular when $A = \frac{1}{n}\,\mathrm{Id}$, in which case we gain for free a factor of n in this tail bound. The z dependence can also be improved (albeit this requires more work) by leveraging that the Stieltjes transform $s_n(z)$ and Frobenius norm of R has better $\operatorname{Im} z$ dependence than one expects a priori, at least with high probability.

If we are to claim the Dyson equation also determines the law of the eigenvalues of non-Gaussian matrices, then we are fundamentally claiming that only the mean and covariance of the random matrix X is enough to make ξ small in the sense of generalized entries. Or said otherwise, if we were to replace X by a Gaussian matrix, we should not change the moments of ξ by much.

Now suppose ξ is a polynomial in generalized entries of $R(z;W)$ with $W = X/\sqrt{n}$ and X Wigner. Then we can look at moments of generalized entries of resolvents. For the Dyson equation to hold, we are claiming that these moments should be the same as with moments of generalized entries of $Y = G/\sqrt{n}$, normalized GOE. So we can try to do an exchange argument, in which we swap rows of W out for matching Gaussian rows one at a time.

Let W be the normalized Wigner matrix $X/\sqrt{n}$ and let $G, \tilde{G}$ be two additional independent GOE matrices.

Define W^0 be $G/\sqrt{n}$ and define $\{W^j\}$ to be a sequence of matrices in which we swap in rows and columns of W one at a time, i.e.

$$W^j - W^{j-1} = (\mathbf{e}_j \otimes \Delta_j + \Delta_j \otimes \mathbf{e}_j), \quad \Delta_{j,i} = \begin{cases} \frac{(X-\tilde{G})_{j,i}}{\sqrt{n}}, & 1 \leq i < j, \\ \frac{(X-G)_{j,j}}{2\sqrt{n}}, & i = j, \\ \frac{(\tilde{G}-G)_{j,i}}{\sqrt{n}}, & i > j. \end{cases}$$

Let Δ_j^+ and Δ_j^- be analogous, where we take the random variables in the definition of Δ with the $\pm$ sign respectively (so that $\Delta_j = \Delta_j^+ - \Delta_j -$)

Set $G^j = R(z;W^j)$, and set $\hat{G}^{j-1}$ to be the resolvent of the matrix where we have set all entries in the support of Δ_j to 0. Note that we have the representation for any $f \in \mathbb{R}^d$

$$\mathbf{e}_j \otimes f + f \otimes \mathbf{e}_j = UCU^T, \quad U = [\mathbf{e}_j, f], \quad C = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right).$$

Using the Woodbury formula, Corollary 4, for off-diagonal updates

$$G^j = \hat{G}^{j-1} - \hat{G}^{j-1}U_j(C + U_j^T\hat{G}^{j-1}U_j)^{-1}U_j^T\hat{G}^{j-1}, \quad U_j := [\mathbf{e}_j, \Delta_j^+].$$

We can do the same for G^{j-1}, which leads to

$$G^{j-1} = \hat{G}^{j-1} - \hat{G}^{j-1}V_j(C + V_j^T\hat{G}^{j-1}V_j)^{-1}V_j^T\hat{G}^{j-1}, \quad V_j := [\mathbf{e}_j, \Delta_j^-].$$

The matrix $C + V_j^T\hat{G}^{j-1}V_j$ that appears in the middle need to be close to an invertible matrix. This requires some input. The presence of the quadratic form suggests that this will concentrate around:

$$C + V_j^T\hat{G}^{j-1}V_j \approx Q_j := \mathbb{E}(C + V_j^T\hat{G}^{j-1}V_j \mid \mathscr{F}_{j-1}) \quad \text{and} \quad \|Q_j^{-1}\|_{op} \leq P,$$

where $\mathscr{F}_j$ is the sigma-algebra generated by all $\{W^k : k \leq j\}$. The P is introduced here and needs to be bounded; this is related to stability.

For the moment suppose that we can work on the event

$$E_j := \{\|Q_j^{-1}\|_{op} \leq P\}.$$

Now by a moment in generalized entries, we mean

$$M^j = \prod_{\ell=1}^{m_1} \left(\mathrm{Tr}(G^j A^{(\ell)})\right) \prod_{\ell=1}^{m_2} \left(\mathrm{Tr}(\overline{G^j} B^{(\ell)})\right)$$

for some real symmetric matrices $A^{(\ell)}$ and $B^{(\ell)}$, and we say that $m = m_1 + m_2$ is the order of of the moment. Then we also define

$$\hat{M}^j = \prod_{\ell=1}^{m_1} \left(\mathrm{Tr}(\hat{G}^j A^{(\ell)})\right) \prod_{\ell=1}^{m_2} \left(\mathrm{Tr}(\overline{\hat{G}^j} B^{(\ell)})\right).$$

The moment M^j can be Taylor expanded around $\hat{M}^j$.

Now at first pass, we shall actually just suppose that we can replace $C + V_j^T \hat{G}^{j-1} V_j$ by Q_j in the definitions above. We can then expand, as a terminating Taylor series

$$M^j = \hat{M}^j + \langle \Delta_j^+, T^{j,1} \rangle + \frac{1}{2} \langle \Delta_j^+ \otimes \Delta_j^+, T^{j,2} \rangle + \frac{1}{6} \langle \Delta_j^+ \otimes \Delta_j^+ \otimes \Delta_j^+, T^{j,3} \rangle + \cdots.$$

The tensors $T^{j,k}$ are independent of Δ_j^+. We can do the same thing for M^{j-1}, producing the same coefficients

$$M^{j-1} = \hat{M}^j + \langle \Delta_j^-, T^{j,1} \rangle + \frac{1}{2} \langle \Delta_j^- \otimes \Delta_j^-, T^{j,2} \rangle + \frac{1}{6} \langle \Delta_j^- \otimes \Delta_j^- \otimes \Delta_j^-, T^{j,3} \rangle + \cdots.$$

Now we simply want to compare conditional expectations of these two terms. The key is that the matching first-second moment structure implies the first two terms vanish.

$$\mathbb{E}(M^j - M^{j-1} \mid \mathscr{F}_{j-2}) = \sum_{k=3}^{m} \langle \mathbb{E}\left((\Delta_j^+)^{\otimes k} - (\Delta_j^-)^{\otimes k}\right), T^{j,k} \rangle.$$

Now we can give a naïve bound by noting that any moment can be bounded by $|\,\mathrm{Im}\, z|^{-m}$ for m given by the order of the moment, which for M^j is $m = m_1 + m_2$. The factor $T^{j,k}$ is a sum of tensors which look like outer products of

$$\hat{G}^{j-1} A^{(\ell)} \hat{G}^{j-1}, \quad \hat{G}^{j-1} A^{(\ell)} \hat{G}^{j-1} \mathbf{e}_j,$$

and similar terms involving complex conjugates and $B^{(\ell)}$, which are then multiplied by a moment of order $m - k$ and entries of Q^{-1}. For the third moment, these terms must look like outer products of one of each of these types of tensors. Moreover, the expected difference $\mathbb{E}\left((\Delta_j^+)^{\otimes 3} - (\Delta_j^-)^{\otimes 3}\right)$ is a diagonal tensor, and hence we have a bound[19]

[19] There will also be B terms, which will be bounded the same way.

$$\begin{aligned}&\left|\langle \mathbb{E}\left((\Delta_j^+)^{\otimes 3}-(\Delta_j^-)^{\otimes 3}\right), T^{j,3}\rangle\right|\\ &\le \frac{C(m)P^2|\operatorname{Im} z|^{-m+3}}{n^{3/2}}\Bigg(\max_{\ell_1,\ell_2}\sum_a |(\hat{G}^{j-1}A^{(\ell_1)}\hat{G}^{j-1})_{aa}||(\hat{G}^{j-1}A^{(\ell_2)}\hat{G}^{j-1})_{aj}|\\ &+\max_{\ell_1,\ell_2,\ell_3}\sum_a |(\hat{G}^{j-1}A^{(\ell_1)}\hat{G}^{j-1})_{aj}||(\hat{G}^{j-1}A^{(\ell_2)}\hat{G}^{j-1})_{aj}||(\hat{G}^{j-1}A^{(\ell_3)}\hat{G}^{j-1})_{aj}|\Bigg)\end{aligned}$$

Now the maximum entry of $|(\hat{G}^{j-1}A^{(\ell_2)}\hat{G}^{j-1})_{aj}|$ can be bounded using operator norms. The sum over aa we can bound using the nuclear norm using[20]

$$\sum_a |(BAC)_{aa}| \le \|A\|_* \|B\|_{op}\|C\|_{op}.$$

We can do the same for the terms in the second line, bounding [21]

$$\sum_a |D_{aj}(BAC)_{aj}| \le \|D\|_{op}\|A\|_*\|B\|_{op}\|C\|_{op},$$

and we are led to (using the nuclear norm bound on A.)[22]

$$\begin{aligned}&\left|\langle \mathbb{E}\left((\Delta_j^+)^{\otimes 3}-(\Delta_j^-)^{\otimes 3}\right), T^{j,3}\rangle\right|\\ &\le \frac{C(m)P^2|\operatorname{Im} z|^{-m+3}}{n^{3/2}}\left(|\operatorname{Im} z|^{-4}\|A\|_{op}+|\operatorname{Im} z|^{-6}\|A\|_{op}^2\right).\end{aligned}$$

Now the error is good enough that we can now sum over all n. There are lots of details to fill in (higher terms in the series, bounding P, adding control over the terms we dropped from Q_j^{-1} – for which we can use concentration of quadratic forms), but these are all the same type of estimate. This brings us to an estimate of the form, for $\|A\|_* \le 1$, [23]

$$\mathbb{E}|\operatorname{Tr}(\xi A)|^m \le C(m, |\operatorname{Im} z|^{-1}, \max_{ij}\|X_{ij}\|_m)n^{-m/2}.$$

Stability for P+GOE

This section was written as an exercise, and can be safely ignored for those looking to learn random matrix theory, random matrix theory and optimization, or just about anything that does not directly concern the P+GOE.

In this case (26) becomes, with $\mathfrak{s}(t)=\frac{1}{n}\operatorname{Tr}(\mathcal{M}(t))$,

$$\frac{\mathrm{d}}{\mathrm{d}t}F(\mathcal{M}(t);z)=\dot{\mathcal{M}}(P-(\mathfrak{s}(t)+z)\operatorname{Id})+\mathcal{M}(-\dot{\mathfrak{s}}(t)).$$

We are looking to once more show that this flow is well-posed and provide some estimates on how it evolves. We are in particular looking for which estimates we should make on ξ to bound the difference of $\mathcal{M}(0)$ and $\mathcal{M}(\infty)$.

[20] The sum of absolute values $\sum |M_{aa}|$ can be represented as $\operatorname{Tr}(MS)$ for a diagonal matrix S of norm 1 and hence we can bound it by $\|M\|_*$.

[21] The left hand side again is $\operatorname{Tr}((\mathbf{e}_j^{\otimes}2)DSBAC)$ for a diagonal matrix of phases S. Now apply the prevous inequality.

[22] Here we have pessimistically bounded the contribution of the lower order moment by $|\operatorname{Im} z|^{-m+3}$. But we could do much better by working by induction on the order of the moment, and show that we have the same order of magnitude of the moment as the Gaussian case, as the error we get in comparison are no larger.

[23] At global scale ($\operatorname{Im} z$) bounded away from the axis, this is the best possible with this type of formulation, but there is a lot to gain by considering better estimates for good A (such as Id/n) and sharpening the dependence in $|\operatorname{Im} z|^{-1}$.

Let $P = \sum_{j=1}^{n} \lambda_j \Pi_j$ be a spectral decomposition of P, with Π_j rank-1 projection matrices. Set $\mathfrak{p}_j(t) = \frac{1}{n}\operatorname{Tr}(\mathcal{M}(t)\Pi_j)$. Then

$$\operatorname{Tr}\left(\frac{\mathrm{d}}{\mathrm{d}t}F(\mathcal{M}(t);z)\right) = \dot{\mathfrak{p}}_j(t)(\lambda_j - \mathfrak{s}(t) - z) - \mathfrak{p}_j(t)\dot{\mathfrak{s}}(t).$$

Thus the Newton flow becomes

$$\dot{\mathfrak{p}}_j(t)(\lambda_j - \mathfrak{s}(t) - z) - \mathfrak{p}_j(t)\dot{\mathfrak{s}}(t) = -e^{-t}\tfrac{1}{n}\operatorname{Tr}(\xi\Pi_j) =: -e^{-t}\psi_j.$$

We further have that $\mathfrak{s} = \sum_j \mathfrak{p}_j$, and hence we can write a matrix-vector equation

$$L(t)\dot{\mathfrak{p}}(t) = -e^{-t}\psi,$$

where the entries of the matrix L are given by

$$L_{jk} = \delta_{jk}(\lambda_j - \mathfrak{s}(t) - z) - \mathfrak{p}_j(t).$$

Using the Woodbury formula (Corollary 4), we have (provided we are away from all the poles)

$$\begin{aligned}(L^{-1})_{jk} &= \delta_{jk}\frac{1}{\lambda_j - \mathfrak{s}(t) - z} \\ &\quad - \frac{\mathfrak{p}_j(t)}{(\lambda_j - \mathfrak{s}(t) - z)(\lambda_k - \mathfrak{s}(t) - z)}\frac{1}{1 - \sum \frac{\mathfrak{p}_j(t)}{\lambda_j - \mathfrak{s}(t) - z}}.\end{aligned}$$

Hence we introduce the summed error

$$\begin{aligned}\Psi(z) &= \sum_j \frac{\psi_j}{(\lambda_j - z)} \\ &= \tfrac{1}{n}\operatorname{Tr}\left(\xi(P - z\operatorname{Id})^{-1}\right).\end{aligned}$$

We also let

$$\begin{aligned}\mathfrak{r}(t) &= \sum_j \frac{\mathfrak{p}_j(t)}{(\lambda_j - \mathfrak{s}(t) - z)} \\ &= \tfrac{1}{n}\operatorname{Tr}(\mathcal{M}(t)(P - (\mathfrak{s}(t) + z)\operatorname{Id})^{-1}).\end{aligned}$$

Then we conclude

$$\dot{\mathfrak{p}}_j(t) = -e^{-t}\left(\frac{\psi_j}{\lambda_j - \mathfrak{s}(t) - z} - \frac{\mathfrak{p}_j(t)}{\lambda_j - \mathfrak{s}(t) - z}\Psi(t;\mathfrak{s}(t)+z)\left(\frac{1}{1 - \mathfrak{r}(t)}\right)\right), \tag{27}$$

and hence

$$\dot{\mathfrak{s}}(t) = -e^{-t}\Psi(t;\mathfrak{s}(t)+z)\left(1 - \tfrac{\mathfrak{r}(t)}{1-\mathfrak{r}(t)}\right). \tag{28}$$

We also need to estimate $\dot{\mathfrak{r}}(t)$, as $\dot{\mathfrak{s}}$ is sensitive to $\mathfrak{r}(t)$. Taking time derivatives of $\mathfrak{r}$,

$$\begin{aligned}\dot{\mathfrak{r}}(t) &= \tfrac{1}{n}\operatorname{Tr}\left(\dot{\mathcal{M}}(t)(P - (\mathfrak{s}(t)+z)\operatorname{Id})^{-1} + \mathcal{M}(t)(P - (\mathfrak{s}(t)+z)\operatorname{Id})^{-2}\dot{\mathfrak{s}}(t)\right) \\ &= \tfrac{1}{n}\operatorname{Tr}\left(-e^{-t}\xi(P - (\mathfrak{s}(t)+z)\operatorname{Id})^{-2} + 2\mathcal{M}(t)(P - (\mathfrak{s}(t)+z)\operatorname{Id})^{-2}\dot{\mathfrak{s}}(t)\right).\end{aligned} \tag{29}$$

So we fix a $\delta > 0$ and define for $p \geq 2$

$$\Xi := \max\{|\Psi(s+z)| : s \in \mathbb{C}, |s - \mathfrak{s}(\infty)| < \delta\},$$
$$\Xi_p := \max\{|\tfrac{\mathrm{d}^p}{(\mathrm{d}z)^p}\Psi(s+z)| : s \in \mathbb{C}, |s - \mathfrak{s}(\infty)| < \delta/2\}.$$

From the Cauchy integral formula $\Xi_1 \leq \frac{2}{\delta}\Xi$. Define

$$\Delta(t) := \sup_{u \in [t,\infty]} \left(\max_j \left\{ \frac{1}{|\lambda_j - \mathfrak{s}(u) - z|} \right\}, \frac{1}{|1 - \mathfrak{r}(u)|}, \frac{|1 - 2\mathfrak{r}(u)|}{|1 - \mathfrak{r}(u)|}, 1 \right).$$

From (28),

$$|\mathfrak{s}(t) - \mathfrak{s}(\infty)| \leq \Delta(t)\Xi,$$

provided $\Delta(t)\Xi \leq \delta$. Set

$$\mathfrak{p}^*(t) = \sup_{u \in [t,\infty]} \max_j \|\mathfrak{p}_j(u)\| \quad \text{and} \quad \psi^* = \max_j \|\psi_j\|.$$

Then again if $\Delta(t)\Xi \leq \delta$,

$$|\dot{\mathfrak{r}}(t)| \leq e^{-t} \left(\Xi_1 + \Delta(t)^3 \mathfrak{p}^*(t)\Xi. \right),$$

and hence

$$|\mathfrak{r}(t) - \mathfrak{r}(\infty)| \leq \left(\Xi_1 + \Delta(t)^3 \mathfrak{p}^*(t)\Xi. \right).$$

Similarly,

$$\max_j |\mathfrak{p}_j(t) - \mathfrak{p}_j(\infty)| \leq \Delta^2(\psi^* + \mathfrak{p}^*(t)\Xi).$$

This allows us to show that good conditions at $t = \infty$ (measured by $\Delta(\infty)$ being small) will persist over time. So, provided that $\Delta(\infty)$, $\mathfrak{p}^*(\infty)$, then if ψ^* and Ξ and are small enough, we conclude

$$\Delta(0) \leq 2\Delta(\infty) \quad \text{and} \quad \mathfrak{p}^*(0) \leq 2\mathfrak{p}^*(\infty),$$

provided $2\Delta(\infty)\Xi \leq \delta$.

The constants that suffice for this are:

$$\begin{aligned} &4\Delta(\infty)^2(\psi^* + 2\mathfrak{p}^*(\infty)\Xi) \leq \mathfrak{p}^*(\infty), \\ &4\Delta(\infty)(\tfrac{2}{\delta}\Xi + 8\Delta(\infty)^3\mathfrak{p}^*(\infty)\Xi) \leq 1, \\ &8\Delta(\infty)^2\Xi \leq 1. \end{aligned} \tag{30}$$

The first bound ensures that $\mathfrak{p}_j$ does not grow too much. The second ensures that $1 - \mathfrak{r}(t)$ does not get too small, and the last ensures that $\mathfrak{s}(t)$ does not get too close to $\lambda_j - z$. Then under (30) and provided $2\Delta(\infty)\Xi \leq \delta$,

$$|\mathfrak{s}(0) - \mathfrak{s}(\infty)| \leq 2\Delta(\infty)\Xi.$$

We conclude the following proposition:

Lemma 16 (Stability for $P + GOE$): Suppose that $z \in \overline{\mathbb{H}}$ and suppose that $\varrho(z) = \varrho(z; P)$ is the distance of z to the spectral support of $P + GOE$. Let $s(z)$ be the Stieltjes transform of $P + GOE$. Then there are constants C, δ, depending only on $\varrho(z)$, so that if the solution of the perturbed Dyson equation satisfies

$$C(\varrho(z))(n\psi^* + \Xi) \leq 1$$

then the Newton flow is well-posed and satisfies

$$|\mathfrak{s}(0) - \mathfrak{s}(\infty)| \leq (4 + \varrho(z)^{-1})^2 \Xi.$$

If bulk convergence is all that is desired, then we can just stay at a fixed distance from the real line. For support convergence statements, it is more helpful to have the distance to the spectrum.

Proof. The solution of the Dyson equation is defined in terms of

$$s(z) = \frac{1}{n} \operatorname{Tr} \left(P - (s(z) + z) \operatorname{Id}\right)^{-1},$$

which leads to

$$M = (P - (s(z) + z) \operatorname{Id})^{-1}.$$

Hence, the spectrum of M we can define by Stieltjes inversion as

$$\mu_P(\mathrm{d}x) = \lim_{\epsilon \to 0} \frac{1}{\pi} \operatorname{Im} s(z).$$

This is a probability measure as $s(z)$ is asymptotic to $\frac{-1}{z}$ as $z = it$ and $t \to \infty$ by Theorem 6.[24] We can also define the distance of a point z to this spectrum by

$$\varrho(z) = d(z, \operatorname{Supp}(P))) \geq |\operatorname{Im} z|.$$

[24] Show that the mapping $s \mapsto \frac{1}{n} \operatorname{Tr}(P - (s(z) + z) \operatorname{Id})^{-1}$ maps a neighborhood of $\frac{-1}{z}$ to itself for $z = it$ with large t.

Differentiating the fixed point equation in z we arrive at

$$s'(z) = \frac{1}{n} \operatorname{Tr} \left(P - (s(z) + z) \operatorname{Id}\right)^{-2} \left(s'(z) + 1\right)$$

At time ∞ we have

$$\mathfrak{r}(\infty) = \tfrac{1}{n} \operatorname{Tr}(\mathcal{M}(\infty)(P - (\mathfrak{s}(\infty) + z) \operatorname{Id})^{-1}).$$

Then using the solution of Dyson equation

$$\mathfrak{r}(\infty) = \tfrac{1}{n} \operatorname{Tr}((P - (\mathfrak{s}(\infty) + z) \operatorname{Id})^{-2}).$$

Hence

$$\mathfrak{r}(\infty) = \frac{s'(z)}{s'(z) + 1},$$

and so

$$\frac{1}{1 - \mathfrak{r}(\infty)} = s'(z) + 1.$$

Now the Stieltjes transform $s(z)$ is

$$s(z) = \int_{\mathbb{R}} \frac{1}{x-z} \mu_P(\mathrm{d}x)$$

for the self-consistent spectral measure μ_P, and hence this has a z-derivative bounded in terms of the distance of z to the spectral support of μ_P. In particular,

$$|s'(z)| \leq \varrho(z)^{-2}.$$

Hence we arrive at a bound in terms of the imaginary part of z only

$$|\Delta(\infty)| \leq \max\left\{\varrho(z)^{-1}, 1+\varrho(z)^{-2}, 2+\varrho(z)^{-1}\right\}.$$

We can also bound $\mathfrak{p}^*(\infty) \leq \frac{1}{n}\varrho(z)^{-1}$. We conclude there is a constant depending only on $\varrho(z)^{-1}$, so that if

$$(2+\varrho(z)^{-1})^9(\Xi + n\psi^*) \leq 1,$$

then with $\delta = \varrho(z)/2$

$$|\mathfrak{s}(0) - \mathfrak{s}(\infty)| \leq (4+\varrho(z)^{-1})^2\Xi.$$

□

2.3 *Sample covariance matrices and linearizations*

The key random matrix for our applications will be the *sample covariance matrix*. This is a random matrix A built as the empirical covariance matrix of n random vectors in $\mathbb{R}^d$

> **Definition 29 (Sample covariance matrix):** Suppose that $X_1, X_2, \ldots, X_n$ are n random vectors in $\mathbb{R}^d$. Let $X \in \mathbb{M}(n,d)$ be the random matrix whose k-th row is given by X_k, the $(d \times d)$ sample covariance matrix
>
> $$A = \tfrac{1}{n} X^T X,$$
>
> or sometimes the *feature-feature covariance matrix*. We also consider the *sample-sample* covariance matrix
>
> $$\hat{A} = \tfrac{1}{n} X X^T.$$

We will usually consider the $\{X_k\}$ as being iid and centered, in which case A is the natural statistical estimator of the covariance matrix of X_1.

The random matrix theory of A follows a nearly parallel development to the affine GOE. The nonlinear dependence of A on the underlying random matrix X has a substantial effect upon the spectrum of A, chiefly that it makes the matrix positive semidefinite. Beyond this, we will put the good probabilistic assumptions on X and not on A (especially $\{X_k\}$ iid with good distributional assumptions, e.g. Gaussian), and so we need to properly handle the non-linear dependence of the matrix on the underlying randomness.

For the purpose of developing a Dyson equation, this is handled by an important idea of a *linearization*.

Definition 30 (Linearization): Let $p \in \mathbb{M}(n)$ be a rational expression in matrix variables $(A_1, A_2, ..., A_k)$, in vector spaces $V_1, V_2, ..., V_d$. A *linearization* of p is an (affine) linear function $L : V_1 \times V_2 \times \cdot \times V_d$ to $\mathbb{M}(n+d)$

$$\left(L(A_1, ..., A_k)^{-1}\right)_{11} = p(A_1, A_2, \dots, A_k),$$

where we have partitioned the rows/columns of $\mathbb{M}$ into two blocks.

There is a larger theory of linearizations and how to find them (see [MS17] for a discussion), as well as algorithms for finding them. We shall focus on a single linearization in this section:

$$L := \begin{bmatrix} -z\,\mathrm{Id}_d & \frac{1}{\sqrt{n}}X^T \\ \frac{1}{\sqrt{n}}X & -\mathrm{Id}_n \end{bmatrix}. \tag{31}$$

We can a direct computation of this inverse, using the Schur complement formula (Lemma 1), which leads to

$$L^{-1} = \begin{bmatrix} (\frac{1}{n}X^TX - z\,\mathrm{Id}_d)^{-1} & -\frac{1}{\sqrt{n}}X^T(\mathrm{Id}_n - \frac{1}{nz}XX^T)^{-1} \\ \frac{1}{\sqrt{n}}X(\frac{1}{n}X^TX - z\,\mathrm{Id}_d)^{-1} & -(\mathrm{Id}_n - \frac{1}{nz}XX^T)^{-1} \end{bmatrix},$$

and so is a linearization of the resolvent $(A - z\,\mathrm{Id}_d)^{-1}$.

Now in the Gaussian case, with $X \overset{\text{law}}{=} N(0, \mathrm{Id}_n \otimes \Sigma)$, [25] we may apply Lemma 14

$$\mathbb{E}(L^{-1}(L - \mathbb{E}L)) = -\mathbb{E}(L^{-1}(\tilde{L} - \mathbb{E}L)L^{-1}(\tilde{L} - \mathbb{E}L)),$$

where $\tilde{L} \overset{\text{law}}{=} L$ and is independent of L. Evaluating this expectation, we have (replacing L^{-1} by a matrix M for notational simplicity)

$$\mathbb{E}((\tilde{L} - \mathbb{E}L)M(\tilde{L} - \mathbb{E}L)) = \frac{1}{n}\begin{bmatrix} \mathbb{E}(X^TM_{22}X) & \mathbb{E}(X^TM_{21}X^T) \\ \mathbb{E}(XM_{12}X) & \mathbb{E}(XM_{11}X^T) \end{bmatrix}.$$

This we can further evaluate as

$$\mathbb{E}((\tilde{L} - \mathbb{E}L)M(\tilde{L} - \mathbb{E}L)) = \frac{1}{n}\begin{bmatrix} \Sigma\,\mathrm{Tr}(M_{22}) & \Sigma M_{21}^T \\ M_{12}^T\Sigma & \mathrm{Tr}(M_{11}\Sigma) \end{bmatrix}.$$

[25] This is to say X has independent rows, which are normal and have covariance Σ.

To define the Dyson equation, we again drop terms we expect to be lower order, here given by the off-diagonals. This leads us to the following:

Definition 31 (Dyson equation for SCM of independent samples): The Dyson equation for the matrix (31), in the case that X has independent centered rows of covariance Σ, is

$$M(\mathbb{E}L - \mathcal{S}(M)) = \mathrm{Id}_{n+d}, \tag{32}$$

where the mean and superoperator $\mathcal{S}$ are given by

$$\mathbb{E}L = \begin{bmatrix} -z\,\mathrm{Id}_d & 0 \\ 0 & -\mathrm{Id}_n \end{bmatrix},$$

and

$$\mathcal{S}(M) = \begin{bmatrix} \Sigma(\frac{1}{n}\operatorname{Tr}(M_{22})) & 0 \\ 0 & \frac{1}{n}\operatorname{Tr}(M_{11}\Sigma)\,\mathrm{Id}_n \end{bmatrix}.$$

The matrix $M \in \mathbb{M}(n+d)$ satisfies that $\operatorname{Im} M_{11} \succeq 0$.

The existence and uniqueness of the solution of this equation no longer immediately follows from the general fixed-point theorem (Theorem 14). However, we only need to do a few changes.

Theorem 17: Deformed Marchenko-Pastur Law

The solution of the Dyson equation for the independent sample SCM (31) is determined by

$$-m(z) = \frac{1}{1 + \frac{1}{n}\operatorname{Tr}\left(\frac{\Sigma}{-\Sigma m - z\,\mathrm{Id}_d}\right)},$$

which has a unique solution for $z \in \mathbb{H}$ with $\operatorname{Im} m(z) > 0$. In terms of $m(z)$, the solution of the Dyson equation is given by

$$M = \begin{bmatrix} M_{11} & 0 \\ 0 & M_{22} \end{bmatrix}, \quad \begin{cases} M_{11} = (-\Sigma m(z) - z\,\mathrm{Id})^{-1}, \\ M_{22} = m(z)\,\mathrm{Id}_n\,. \end{cases}$$

The Stieltjes transform is given in terms of m by

$$\frac{1}{d}\operatorname{Tr}(M_{11}) = \left(\frac{n}{d} - 1\right)\frac{1}{z} + \frac{n}{d}\frac{m(z)}{z}.$$

Proof. We set $m(z) = \frac{1}{n}\operatorname{Tr}(M_{22})$. We expand (32)

$$\begin{aligned} M_{11}(-\Sigma m(z) - z\,\mathrm{Id}_d) &= \mathrm{Id}_d \\ M_{22}(-\mathrm{Id}_n - \tfrac{1}{n}\operatorname{Tr}(M_{11}\Sigma)\,\mathrm{Id}_n) &= \mathrm{Id}_n \end{aligned}$$

Then we can solve for M_{11} and substitute it into the second equation, which produces

$$M_{22}\left(-\operatorname{Id}_n - \tfrac{1}{n}\operatorname{Tr}\left(\frac{\Sigma}{-\Sigma m(z) - z\operatorname{Id}_d}\right)\operatorname{Id}_n\right) = \operatorname{Id}_n .$$

Taking the normalized trace on both sides, we arrive at

$$-m(z)\left(1 + \tfrac{1}{n}\operatorname{Tr}\left(\frac{\Sigma}{-\Sigma m(z) - z\operatorname{Id}_d}\right)\right) = 1.$$

Now the mapping

$$m \mapsto \frac{-1}{1 + \frac{1}{n}\operatorname{Tr}\left(\frac{\Sigma}{-\Sigma m - z\operatorname{Id}_d}\right)}$$

is an analytic self map from $\mathbb{H} \to \mathbb{H}$. Hence it is a (possibly non-strict) contraction in the hyperbolic metric by the Schwarz-Lemma. Such a mapping is strict contraction if and only if the map is not a Möbius transformation fixing $\mathbb{H}$. This is if and only if

$$m \mapsto \operatorname{Tr}\left(\frac{\Sigma}{-\Sigma m - z\operatorname{Id}_d}\right)$$

is not a Möbius transformation fixing $\mathbb{H}$ (having applied a hyperbolic isometry). However, this maps the real line properly into the upper half plane by the positivity of $\operatorname{Im} z$, and so the mapping was a strict contraction in the hyperbolic metric.[26]

Hence there is a unique fixed point of

$$m \mapsto \frac{1}{1 + \frac{1}{n}\operatorname{Tr}\left(-\frac{\Sigma}{\Sigma m - z\operatorname{Id}_d}\right)}$$

in the upper half plane, and so $\operatorname{Im} m > 0$. This means M_{11} exists and is given by

$$M_{11} = (-\Sigma m(z) - z\operatorname{Id})^{-1},$$

and also M_{22} exists and is given by

$$M_{22} = (-\operatorname{Id}_n - \tfrac{1}{n}\operatorname{Tr}(M_{11}\Sigma)\operatorname{Id}_n)^{-1} = m(z)\operatorname{Id}_n .$$

Finally, we note that the Stieltjes transform can be extracted from m, since

$$\begin{aligned}
s(z) &= \tfrac{1}{d}\operatorname{Tr}(M_{11}) \\
&= \tfrac{1}{dz}\operatorname{Tr}(zM_{11}) \\
&= \tfrac{1}{dz}\operatorname{Tr}((z + \Sigma m(z) - \Sigma m(z))M_{11}) \\
&= (-\tfrac{1}{z} + \tfrac{1}{dz}\operatorname{Tr}(-\Sigma m(z) M_{11})) \\
&= (-\tfrac{1}{z} + \tfrac{n}{dz}(m(z)+1)) \\
&= (\tfrac{n}{d} - 1)\tfrac{1}{z} + \tfrac{n}{dz}m(z).
\end{aligned}$$

□

[26] This also shows that the fixed point iteration, started from $m = -1$ always converges. Note that for large z, $m \approx -1$ is a good approximation.

Example 6: Marchenko-Pastur law

Suppose that $\Sigma = \mathrm{Id}_d$. Then we have the equation for m

$$m = \frac{-1}{1 + \frac{d}{n}\frac{-1}{m+z}}.$$

We can solve this equation for m, first simplifying the equation to be

$$-(m+z) = m(m+z-\tfrac{d}{n}),$$

and hence

$$m^2 + m(z - \tfrac{d}{n} + 1) + z = 0,$$

so that

$$m = \frac{-(z - \frac{d}{n} + 1) \pm \sqrt{(z-\lambda_+)(z-\lambda_-)}}{2},$$

where $\lambda_\pm = (\sqrt{\frac{d}{n}} \pm 1)^2$.
Hence the Stieltjes transform is given by

$$\begin{aligned} s(z) &= (\tfrac{n}{d} - 1)\tfrac{1}{z} + \tfrac{n}{dz} m(z) \\ &= \frac{-(z + \frac{d}{n} - 1) \pm \sqrt{(z-\lambda_+)(z-\lambda_-)}}{2\frac{d}{n}z}. \end{aligned}$$

The branches can be determined by the sign requirements on the imaginary part and the condition that $s(z)$ vanishes as $z \to \infty$ (note the sign depends on z if taking the principal branch of $\sqrt{\cdot}$). Performing Stieltjes inversion, this leads to

$$s(x + \tfrac{i}{t})\frac{\mathrm{d}x}{\pi} \xrightarrow[t\to\infty]{\text{law}} (1 - \tfrac{n}{d})_+ \delta_0 + \frac{\sqrt{[(\lambda_+ - x)(x-\lambda_-)]_+}}{2\frac{d}{n}x},$$

which is the Marchenko-Pastur law.

2.4 *Stability of the Dyson equation for sample covariance matrices*

We again use the Newton flow to show that the solution of the perturbed Dyson equation is close to the unperturbed equation. So we define the operator

$$F(M; z) := M(\mathbb{E}L - \mathcal{S}(M)) - \mathrm{Id}_{n+d},$$

and we introduce the Newton flow

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mathcal{M}(t); z) = -F(\mathcal{M}(t); z) \quad \text{where} \quad F(\mathcal{M}(0); z) = \xi.$$

Once more, we need to show that this flow is well-posed (so the equation is uniquely solvable for all time), and we would like to bound the difference between its value at $t = 0$ and $t = \infty$ in terms of ξ. This will allow us to show the solution of the Dyson equation (which is given by Theorem 17) is close to the linearization of the sample covariance matrix. We note that provided it is well-posed, the value of the objective function $F(\mathcal{M}(t); z)$ can be easily integrated along the flow to give $e^{-t}\xi$. (And so on a maximal interval of existence of the flow, either started from $t = \infty$ or from $t = 0$, we can simply use this formula, on way to justifying this maximal interval is $[0, \infty]$)

We introduce the scalar variables

$$\begin{aligned} \mathfrak{s} &:= \mathfrak{s}(t) := \tfrac{1}{n}\operatorname{Tr}(\mathcal{M}_{11}(t)), \\ \mathfrak{q} &:= \mathfrak{q}(t) := \tfrac{1}{n}\operatorname{Tr}(\mathcal{M}_{11}(t)\Sigma), \\ \mathfrak{m} &:= \mathfrak{m}(t) := \tfrac{1}{n}\operatorname{Tr}(\mathcal{M}_{22}(t)). \end{aligned}$$

This we express (using $A \oplus B$ to denote the block matrix $\left(\begin{smallmatrix} A & 0 \\ 0 & B \end{smallmatrix}\right)$),

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F(\mathcal{M}(t); z) = \dot{\mathcal{M}}\left((-\Sigma\mathfrak{m} - z\operatorname{Id}) \oplus (-\operatorname{Id})(1+\mathfrak{q})\right) \\ + \mathcal{M}(-\Sigma\dot{\mathfrak{m}} \oplus -\dot{\mathfrak{q}}). \end{aligned}$$

We let $\Sigma = \sum \lambda_j \Pi_j$ and introduce

$$\mathfrak{p}_j := \mathfrak{p}_j(t) := \tfrac{1}{n}\operatorname{Tr}(\mathcal{M}_{11}\Pi_j).$$

Then using that $F(\mathcal{M}(t); z) = -e^{-t}\xi$ along the flow,

$$\begin{aligned} -e^{-t}\tfrac{1}{n}\operatorname{Tr}(\xi_{11}\Pi_j) &= \dot{\mathfrak{p}}_j(-\lambda_j\mathfrak{m} - z) + \mathfrak{p}_j(-\lambda_j\dot{\mathfrak{m}}) \\ -e^{-t}\tfrac{1}{n}\operatorname{Tr}(\xi_{22}) &= \dot{\mathfrak{m}}(-(1+\mathfrak{q})) + \mathfrak{m}(-\dot{\mathfrak{q}}). \end{aligned} \tag{33}$$

Note that we can write

$$\mathfrak{s} = \sum_j \mathfrak{p}_j \quad \text{and} \quad \mathfrak{q} = \sum_j \lambda_j \mathfrak{p}_j.$$

Now we detour briefly to cover the simpler case of the Marchenko-Pastur law, which is to say that $\Sigma = \operatorname{Id}_d$. In this case, $\mathfrak{q} = \mathfrak{s}$ we can consider the two parameter dynamical system $(\mathfrak{s}, \mathfrak{m})$. Here we can simplfy (33) to become

$$\begin{aligned} -e^{-t}\tfrac{1}{n}\operatorname{Tr}(\xi_{11}) &= \dot{\mathfrak{s}}(-\mathfrak{m} - z) + \mathfrak{s}(-\dot{\mathfrak{m}}) \\ -e^{-t}\tfrac{1}{n}\operatorname{Tr}(\xi_{22}) &= \dot{\mathfrak{m}}(-(1+\mathfrak{s})) + \mathfrak{m}(-\dot{\mathfrak{s}}). \end{aligned} \tag{34}$$

Hence we can write this as matrix equation

$$\begin{bmatrix} \mathfrak{m} + z & \mathfrak{s} \\ \mathfrak{m} & 1 + \mathfrak{s} \end{bmatrix} \begin{bmatrix} \dot{\mathfrak{s}} \\ \dot{\mathfrak{m}} \end{bmatrix} = e^{-t} \begin{bmatrix} \frac{1}{n}\operatorname{Tr}(\xi_{11}) \\ \frac{1}{n}\operatorname{Tr}(\xi_{22}) \end{bmatrix},$$

which we formally invert to give

$$\begin{bmatrix} \dot{\mathfrak{s}} \\ \dot{\mathfrak{m}} \end{bmatrix} = \frac{e^{-t}}{\mathfrak{a}(t)} \begin{bmatrix} 1 + \mathfrak{s} & -\mathfrak{s} \\ -\mathfrak{m} & \mathfrak{m} + z \end{bmatrix} \begin{bmatrix} \frac{1}{n}\operatorname{Tr}(\xi_{11}) \\ \frac{1}{n}\operatorname{Tr}(\xi_{22}) \end{bmatrix},$$

where $\mathfrak{a} = (\mathfrak{m}+z)(1+\mathfrak{s}) - \mathfrak{s}\mathfrak{m}$. Thus, we need to ensure that $\mathfrak{a}$ stays away from 0, and if it does, then we are are going to be happy.

This $\mathfrak{a}$ will, as in the case of the semicircle law, measure how close the flow comes to the spectral edge. The following works in the same way as the Lemma 15.

Lemma 17 (MP Stability): Let $D(z) := \max\{|m(z)|, |s(z)|, 1, |z|\}$ for $z \in \mathbb{H}$. There is an absolute constant $c > 0$ so that if

$$D(z)^2(|\tfrac{1}{n}\operatorname{Tr}(\xi_{11})| + |\tfrac{1}{n}\operatorname{Tr}(\xi_{22})|) \leq c|m(z) - \tfrac{z}{m(z)}|^2,$$

then the ODE for $\mathfrak{s}$ and $\mathfrak{q}$ is well-posed for all $t \in [0,\infty]$ and

$$\max\{|\dot{\mathfrak{s}}|, |\dot{\mathfrak{m}}|\} \leq \frac{e^{-t}6D}{|m(z) - \frac{z}{m(z)}|^2}(|\tfrac{1}{n}\operatorname{Tr}(\xi_{11})| + |\tfrac{1}{n}\operatorname{Tr}(\xi_{22})|).$$

Hence, to prove the Marchenko-Pastur law, it suffices to bound both $|\frac{1}{n}\operatorname{Tr}(\xi_{11})|$ and $|\frac{1}{n}\operatorname{Tr}(\xi_{22})|$.

The parameter $\mathfrak{a}(\infty) = m(z) - \frac{z}{m(z)}$ vanishes precisely at the edges of the Marchenko-Pastur law, and it can be checked to be given by

$$m(z) - \tfrac{z}{m(z)} = \pm\sqrt{(z-\lambda_+)(z-\lambda_-)},$$

with the sign conventions the same as in the m that appears in the Marchenko-Pastur law.

General case

We return to (33), in which we can substitute the second equation into the first equation to produce

$$e^{-t}\tfrac{1}{n}\left(-\operatorname{Tr}(\xi_{11}\Pi_j) + \operatorname{Tr}(\xi_{22})\lambda_j\tfrac{\mathfrak{p}_j}{1+\mathfrak{q}}\right) = \dot{\mathfrak{p}}_j(-\lambda_j\mathfrak{m} - z) + \dot{\mathfrak{q}}\tfrac{\lambda_j\mathfrak{p}_j\mathfrak{m}}{1+\mathfrak{q}}.$$

The right hand side of the equation we can express as a linear combination of the vector $\dot{\mathfrak{p}}$. In particular there is a matrix $\mathfrak{L}$ depending on $\mathfrak{m}, \mathfrak{q}$ and $\mathfrak{p}$ so that for all j

$$(\mathfrak{L}\dot{\mathfrak{p}})_j = e^{-t}\tfrac{1}{n}\left(-\operatorname{Tr}(\xi_{11}\Pi_j) + \operatorname{Tr}(\xi_{22})\lambda_j\tfrac{\mathfrak{p}_j}{1+\mathfrak{q}}\right).$$

Moreover we can formally invert this matrix, as it is a rank-1 perturbation of the diagonal matrix $(-\lambda_j\mathfrak{m} - z)$ using the Woodbury identity. In particular, we have

$$\mathfrak{L} = \operatorname{Diag}(-\lambda_j\mathfrak{m} - z) + \left(\tfrac{\lambda_j\mathfrak{p}_j\mathfrak{m}}{1+\mathfrak{q}} : j\right) \otimes (\lambda_k : k).$$

This produces

$$\begin{aligned}
&\mathfrak{L}^{-1} \\
&= \mathrm{Diag}\left(\tfrac{-1}{\lambda_j \mathfrak{m}+z}\right) \\
&- \left(\tfrac{\lambda_j \mathfrak{p}_j \mathfrak{m}}{(1+\mathfrak{q})(\lambda_j\mathfrak{m}+z)} : j\right) \otimes \left(\tfrac{\lambda_k}{\lambda_k \mathfrak{m}+z} : k\right) \left(\frac{1}{1-\sum_j \frac{\lambda_j^2 \mathfrak{p}_j \mathfrak{m}}{(1+\mathfrak{q})(\lambda_j \mathfrak{m}+z)}} \right).
\end{aligned}$$

As it arises naturally, we define

$$\mathfrak{r} := \sum_j \tfrac{\lambda_j^2 \mathfrak{p}_j}{(1+\mathfrak{q})(\lambda_j \mathfrak{m}+z)} \quad \text{and} \quad \psi_j := \tfrac{1}{n}\operatorname{Tr}(\xi_{11}\Pi_j).$$

Thus we arrive at the equation for $\dot{\mathfrak{p}}$

$$\dot{\mathfrak{p}}_j = e^{-t}\left(\frac{\psi_j - \frac{1}{n}\operatorname{Tr}(\xi_{22})\frac{\lambda_j\mathfrak{p}_j}{1+\mathfrak{q}}}{\lambda_j \mathfrak{m}+z} + \mathfrak{p}_j \frac{\lambda_j \mathfrak{m}\left(\frac{1}{n}\operatorname{Tr}(\xi_{11}\Sigma(\Sigma\mathfrak{m}+z)^{-1}) - \frac{1}{n}\operatorname{Tr}(\xi_{22})\mathfrak{r}\right)}{(1+\mathfrak{q})(1-\mathfrak{r}\mathfrak{m})(\lambda_j\mathfrak{m}+z)} \right). \tag{35}$$

Summing against λ_j in j produces

$$\dot{\mathfrak{q}} = e^{-t}\left(\tfrac{1}{n}\operatorname{Tr}(\xi_{11}\Sigma(\Sigma\mathfrak{m}+z)^{-1}) \left(1 + \frac{\mathfrak{r}\mathfrak{m}}{(1+\mathfrak{q})(1-\mathfrak{r}\mathfrak{m})}\right) - \tfrac{1}{n}\operatorname{Tr}(\xi_{22})\mathfrak{r}\left(\frac{(1-\mathfrak{r}\mathfrak{m})+\mathfrak{r}\mathfrak{m}}{(1+\mathfrak{q})(1-\mathfrak{r}\mathfrak{m})}\right)\right).$$

So we define

$$\Delta(t) := \sup_{u\in[t,\infty]} \max\left\{ \|(\Sigma\mathfrak{m}(u)+z)^{-1}\|_{op}, \tfrac{1}{|1+\mathfrak{q}(u)|}, \tfrac{1+|\mathfrak{r}(u)|+|\mathfrak{m}(u)|}{\min\{|1-\mathfrak{r}(u)\mathfrak{m}(u)|,1\}} \right\}.$$

We also define for any $\delta > 0$

$$\Xi = \max\left\{ |\tfrac{1}{n}\operatorname{Tr}\left(\xi_{11}(\Sigma(\Sigma m+z)^{-1})^p\right)| + |\tfrac{1}{n}\operatorname{Tr}(\xi_{22})| : p\in\{1,2\}, |m-\mathfrak{m}(\infty)| \le \delta \right\},$$

and finally, we define

$$\mathfrak{p}^*(t) = \sup_{u\in[t,\infty]} \max_j |\mathfrak{p}_j(u)| \quad \text{and} \quad \psi^* = \max_j |\psi_j|.$$

Now we suppose that $\Delta(\infty) < \infty$, and we look to give conditions (bounds on Ξ and ψ^*) which will ensure the flow exists for all time $[0,\infty]$.

Note that provided that $|\mathfrak{m}(u) - \mathfrak{m}(\infty)| \le \delta$ for all $u \ge t$ we have

$$|\dot{\mathfrak{q}}(t)| \le e^{-t}\Xi(|1+\Delta(t)|)^2,$$

and hence also

$$|\dot{\mathfrak{m}}(t)| \le e^{-t}\Xi(|1+\Delta(t)|)^3.$$

To ensure that Δ remains bounded along the flow, we also need to bound the evolution of $\mathfrak{p}_j$ and $\mathfrak{r}$. This gives

$$|\dot{\mathfrak{p}}_j(t)| \le e^{-t}\left(\Delta(t)\psi^* + (\Delta(t)^2\Xi\|\Sigma\|_{op} + \Delta(t)^3\Xi)|\mathfrak{p}_j(t)|\right),$$

and so

$$|\dot{\mathfrak{r}}(t)| \leq e^{-t}\frac{1}{1+\mathfrak{q}(t)}\frac{1}{n}\left(\operatorname{Tr}\left(\xi_{11}\Sigma^2(\Sigma\mathfrak{m}(t)+z)^{-2}\right)\right)$$
$$+e^{-t}n\Delta(t)^2\|\Sigma^2\|_{op}\left(\Delta(t)^2\Xi\|\Sigma\|_{op}\mathfrak{p}^*+\mathfrak{p}^*\Delta(t)^3\Xi\right)$$
$$+e^{-t}n\Delta(t)^3\|\Sigma^2\|_{op}\mathfrak{p}^*\left(\Xi(1+\Delta(t))^3\right).$$

Now putting all these bounds together, we can conclude that there is an absolute constant $c>0$ so that if[27]

$$\begin{aligned}&\Xi((1+\|\Sigma\|_{op})\Delta(\infty))^4 \leq c,\\ &\Xi((1+\|\Sigma\|_{op})\Delta(\infty))^6(n\psi^*+n\mathfrak{p}^*(\infty)) \leq c,\end{aligned} \tag{36}$$

and $\Xi(1+2\Delta(\infty))^3 \leq \delta$, then the whole Newton flow is well-posed and[28]

$$\Delta(0) \leq 2\Delta(\infty).$$

Furthermore, we have

$$|\mathfrak{m}(0)-\mathfrak{m}(\infty)| \leq \Xi(1+2\Delta(\infty))^2.$$

We conclude with the following general stability statement.

We conclude the following proposition:

Lemma 18 (Stability for deformed MP): Suppose that $z \in \overline{\mathbb{H}}$ and suppose that $\varrho(z) = \varrho(z;\Sigma)$ is the distance of z to the spectral support of the deformed MP law with parameters (n,Σ). Then there are constants C,δ, depending only on $(\frac{d}{n},\varrho(z)^{-1},\|\Sigma\|_{op},|z|)$, so that if the solution of the perturbed Dyson equation satisfies

$$C(\tfrac{d}{n},\varrho(z)^{-1},\|\Sigma\|_{op},|z|)(n\psi^*+1)\Xi \leq 1$$

then the Newton flow is well-posed and satisfies for all $t \in [0,\infty]$,

$$\max\{|\dot{\mathfrak{s}}|,|\dot{\mathfrak{m}}|,|\dot{\mathfrak{q}}|\} \leq e^{-t}C(\tfrac{d}{n},\varrho(z)^{-1},\|\Sigma\|_{op},|z|)\Xi.$$

Furthermore, for every $\epsilon \in (0,\epsilon_0)$ there is a finite collection of matrices $\{A_j\}$ of cardinality at most C/ϵ^2, commuting with Σ and having norm at most $\frac{\|\Sigma\|_{op}}{n\varrho(z)^2}$ so that

$$\Xi \leq C(\tfrac{d}{n},\varrho(z)^{-1},\|\Sigma\|_{op},|z|)n^2\epsilon\psi^*+\max_j|\operatorname{Tr}(\xi_{11}A_j)|+|\tfrac{1}{n}(\operatorname{Tr}\xi_{22})|.$$

[27] The first inequality ensures the differences of $\mathfrak{q},\mathfrak{m}$ over time is small enough that Δ cannot grow. The first line of $\dot{\mathfrak{r}}$ is controlled as well by this. Assuming $\Delta(t) \leq 2\Delta(\infty)$ one then bounds by Gronwall's inequality $\mathfrak{p}^*$. Finally the last inequality ensures the terms in second/third terms of $\dot{\mathfrak{r}}$ do not cause Δ to grow too much.

[28] To see these bounds, start from trying to show that $\Delta(t) \leq 2\Delta(\infty)$. Now so long as this is true, we can bound the derivatives of $\mathfrak{q},\mathfrak{m},\mathfrak{p}_j,\mathfrak{r}$, and hence we can bound (in terms of $2\Delta(\infty)$) how much Δ can grow, and hence verify over the lifetime of the flow that $\Delta(t) \leq 2\Delta(\infty)$. This also implies the flow exists over all $t \in [0,\infty]$.

Informally, this theorem says: if generalized entries of ξ_{11} are small then the deformed Marchenko-Pastur law is true. More precisely, we need ψ^* to be not-too-big (in fact we should expect it to shrink), and we need some collection of other test matrices to be small. Note that we can take ϵ much smaller than $1/n$, provided we have good enough bounds in probability on the error term.

Proof. We need to estimate some properties of the solution of the Dyson equation. We have from Theorem 17,

$$t(z) := \mathfrak{q}(\infty) = \frac{1}{n}\operatorname{Tr}(M_{11}\Sigma) = -\frac{1}{n}\operatorname{Tr}(\Sigma(\Sigma\mathfrak{m}(\infty) + z\,\mathrm{Id}_d)^{-1}),$$
$$m(z) = \mathfrak{m}(\infty) = \frac{-1}{1+t(z)}.$$

We also have from Theorem 17

$$s(z) := \frac{1}{d}\operatorname{Tr}(M_{11}) = \left(\frac{n}{d} - 1\right)\frac{1}{z} + \frac{n}{d}\frac{m(z)}{z}.$$

The spectrum of M_{11} we can define by Stieltjes inversion as

$$\mu_\Sigma(\mathrm{d}x) = \lim_{\epsilon\to 0}\frac{1}{\pi}\frac{1}{d}\operatorname{Tr}(\mathcal{M}_{11}).$$

We can also define the distance of a point z to this spectrum by

$$\varrho(z) = d(z, \operatorname{Supp}(\mu_\Sigma) \cup \{0\}) \geq |\operatorname{Im} z|.$$

Then we have $s(z)$ can be trivially bounded in terms of $\varrho(z)$, since

$$s(z) = \int_{\mathbb{R}} \frac{\mu_\Sigma(\mathrm{d}x)}{x - z},$$

and so that

$$|s(z)| \leq \frac{1}{\varrho(z)}.$$

Now we can also use Stieltjes representation on $p_j = \frac{1}{n}\operatorname{Tr}(\mathcal{M}_{11}\Pi_j)$, which gives another set of probability measures $\{\mu_\Sigma^j\}$

$$\frac{1}{n}\frac{1}{-\lambda_j m(z) - z} = p_j(z) = \int_{\mathbb{R}} \frac{1}{n}\frac{\mu_\Sigma^j(\mathrm{d}x)}{x - z}.$$

Then we have

$$\mu_\Sigma = \frac{1}{n}\sum_{j=1}^{d}\mu_\Sigma^j,$$

and hence we can also bound p_j in terms of the distance ϱ. In particular

$$\frac{1}{|\lambda m(z) + z|} = n|p_j(z)| \leq \varrho^{-1}(z).$$

We shall therefore take $\delta = \frac{\varrho(z)}{2\|\Sigma_{op}\|}$ (used in the definition of Ξ) which ensures that $(\Sigma m + z)^{-1}$ will remain invertible.

As a corollary, we can represent

$$t(z) = \frac{1}{n}\sum_{j=1}^{n}\lambda_j p_j(z),$$

and so bound

$$|t(z)| \leq \frac{\mathrm{Tr}(\Sigma)}{n} \frac{1}{\varrho(z)}.$$

Solving for m in terms of s, we have

$$|m(z)| \leq \frac{d}{n} \left(\frac{|z|}{\varrho(z)} \right) + \left| \frac{d}{n} - 1 \right|.$$

Finally, we let

$$r(z) = \mathfrak{r}(\infty) = -\frac{1}{n} \operatorname{Tr} \left(\Sigma^2 m(z) (\Sigma m(z) + z)^{-2} \right).$$

We begin by observing that

$$\begin{aligned} r(z) &= \frac{1}{n} \operatorname{Tr} \left(\Sigma (\Sigma m(z) + z)^{-1} \right) - \frac{z}{n} \operatorname{Tr} \left(\Sigma m(z) (\Sigma m(z) + z)^{-2} \right) \\ &= -t(z) - z \left(t'(z) - \frac{m'(z) r(z)}{m(z)} \right), \end{aligned}$$

where we have differentiated $t(z)$ in z to produce

$$\begin{aligned} t'(z) &= \frac{1}{n} \operatorname{Tr}(\Sigma(\Sigma m'(z) + 1)(\Sigma m(z) + z)^{-2}) \\ &= \frac{m'(z) r(z)}{m}(z) + \frac{1}{n} \operatorname{Tr}(\Sigma(\Sigma m(z) + z)^{-2}). \end{aligned}$$

We also have that $t'(z) = m'(z)/m(z)^2$, and so we conclude that

$$r(z) m(z) = 1 + m(z) - \frac{z}{m(z)} \left(1 - r(z) m(z)\right) m'(z),$$

and hence solving for $1 - r(z)m(z)$,

$$1 - r(z) m(z) = \frac{m(z)}{1 - \frac{z}{m(z)} m'(z)},$$

or that (using $\frac{1}{m(z)} = -(1 + t(z))$)

$$\frac{1}{1 - r(z) m(z)} = -(1 + z(1 + t(z)) m'(z))(1 + t(z)).$$

Thus this can only explode where $m'(z)$, but it easily seen this is bounded in terms of $\frac{d}{n}, \|\Sigma\|_{op}, \frac{1}{\varrho(z)}, |z|$. Similar bounds hold for $r(z)$. All of these together bound $\Delta(\infty)$, and so from (36) the first two conclusions of the lemma follow.

As for the second part, we choose the matrices $A_{j,p}$ to be

$$A_{j,p} \coloneqq \frac{1}{n} \Sigma (\Sigma m_j + z)^{-p},$$

where m_j runs over an ϵ-net of the complex disk $\mathbb{D} \coloneqq |m - \mathfrak{m}(\infty)| \leq \delta$ and $p \in \{1, 2\}$. Now we note for any other m in $\mathbb{D}$, if we let m_0 be the closest point in the net,

$$\begin{aligned}
&\left|\frac{1}{n}\operatorname{Tr}(\xi_{11}\Sigma(\Sigma m_0+z)^{-p})-\frac{1}{n}\operatorname{Tr}(\xi_{11}\Sigma(\Sigma m+z)^{-p})\right| \\
&\leq \psi^*\times\sum_j\left|\frac{\lambda_j}{(\lambda_j m_0+z)^p}-\frac{\lambda_j}{(\lambda_j m+z)^p}\right| \\
&\leq C_p\psi^*\epsilon\varrho^{-p-1}(z)\left(\sum_j\lambda_j^2\right),
\end{aligned}$$

from how δ was chosen.

□

2.5 *Proof of the deformed MP law*

Now we turn to bounding the errors that appear in the Dyson equation for the deformed MP law.

This is to say, that we wish to show the linearization L^{-1} approximately satisfies the Dyson equation. The error term ξ can be expressed by

$$\begin{aligned}
\xi_{11}&=(\tfrac{1}{n}X^TX-z\operatorname{Id}_d)^{-1}\left(\Sigma\tfrac{1}{n}\operatorname{Tr}(-(\operatorname{Id}_n-\tfrac{1}{nz}XX^T)^{-1})-z\operatorname{Id}_d\right)-\operatorname{Id}_d,\\
\xi_{22}&=(\operatorname{Id}_n-\tfrac{1}{nz}XX^T)^{-1}(1+\tfrac{1}{n}\operatorname{Tr}((\tfrac{1}{n}X^TX-z\operatorname{Id}_d)^{-1}\Sigma))-\operatorname{Id}_n\,.
\end{aligned}$$

Set $m_n(z)=\frac{1}{n}\operatorname{Tr}(-(\operatorname{Id}_n-\frac{1}{nz}XX^T)^{-1})$ and $t_n(z)=\frac{1}{n}\operatorname{Tr}((\frac{1}{n}X^TX-z\operatorname{Id}_d)^{-1}\Sigma)$.

Note that by considering the eigenvalues of XX^T, we can also express $m_n(z)$ by

$$\begin{aligned}
\tfrac{1}{z}m_n(z)&=\tfrac{1}{n}\operatorname{Tr}((\tfrac{1}{n}XX^T-z\operatorname{Id}_n)^{-1})\\
&=-\frac{(n-d)}{n}\frac{1}{z}+\frac{1}{n}\operatorname{Tr}((\tfrac{1}{n}X^TX-z\operatorname{Id}_d)^{-1}).
\end{aligned}\tag{37}$$

Now we will only need to consider traces of ξ_{22} to get the simplest case of the deformed Marchenko-Pastur law. So that using this notation, we have

$$\begin{aligned}
\xi_{11}&=(\tfrac{1}{n}X^TX-z\operatorname{Id}_d)^{-1}(\Sigma m_n(z)-z\operatorname{Id}_d)-\operatorname{Id}_d\\
\tfrac{1}{n}\operatorname{Tr}(\xi_{22})&=-m_n(z)(1+t_n(z))-1.
\end{aligned}\tag{38}$$

We define $Q=\frac{1}{n}X^TX$, which we can represent alternatively, as a tensor product, as

$$Q=\frac{1}{n}\sum_{j=1}^n X_j\otimes X_j.$$

We also define $Q^{[i]}$ as the same without the i-th sample, so

$$Q^{[i]}:=\frac{1}{n}\sum_{j\neq i}^n X_j\otimes X_j.$$

Then using the Woodbury identity,

$$(Q - z\,\mathrm{Id})^{-1} = (Q^{[i]} - z\,\mathrm{Id})^{-1} - \frac{1}{n}\frac{((Q^{[i]} - z\,\mathrm{Id})^{-1}X_i) \otimes ((Q^{[i]} - z\,\mathrm{Id})^{-1}X_i)}{1 + \frac{1}{n}\langle X_i, (Q^{[i]} - z\,\mathrm{Id})^{-1}X_i\rangle}.$$

Starting from the tautology

$$\begin{aligned}
\mathrm{Id} &= R(z;Q)(Q - z\,\mathrm{Id}) \\
&= -zR(z;Q) + \frac{1}{n}\sum_{i=1}^{n} R(z;Q)(X_i \otimes X_i) \\
&= -zR(z;Q) + \frac{1}{n}\sum_{i=1}^{n} R(z;Q^{[i]})(X_i \otimes X_i) - \frac{1}{n}\sum_{i=1}^{n}\frac{a_i}{1+a_i}((Q^{[i]} - z\,\mathrm{Id})^{-1}X_i) \otimes X_i, \\
&= -zR(z;Q) + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{R(z;Q^{[i]})}{1+a_i}\right)(X_i \otimes X_i),
\end{aligned}$$

where we have set $a_i = \frac{1}{n}\langle X_i, (Q^{[i]} - z\,\mathrm{Id})^{-1}X_i\rangle$.

In the second term, we expect that we can trade $X_i \otimes X_i$ for Σ, so we introduce an error matrix

$$\alpha_1 := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{R(z;Q^{[i]})}{1+a_i}\right)(X_i \otimes X_i - \Sigma). \tag{39}$$

Then resumming, we produce a second error term

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(\frac{R(z;Q^{[i]})}{1+a_i}\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{R(z;Q)}{1+a_i}\right) \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n}\frac{((Q^{[i]} - z\,\mathrm{Id})^{-1}X_i) \otimes ((Q^{[i]} - z\,\mathrm{Id})^{-1}X_i)}{(1+a_i)^2} \\
&=: \frac{R(z;Q)}{1+t_n(z)} + \alpha_2.
\end{aligned} \tag{40}$$

In summary, we have

$$\begin{aligned}
\mathrm{Id} &= -zR(z;Q) + \frac{R(z;Q)}{1+t_n(z)}\Sigma + \alpha_2\Sigma + \alpha_1 \\
&= -zR(z;Q) - m_n(z)R(z;Q)\Sigma\left(\frac{1}{1+\frac{1}{n}\operatorname{Tr}\xi_{22}}\right) + \alpha_2\Sigma + \alpha_1
\end{aligned} \tag{41}$$

Hence using (38) we conclude

$$\xi_{11} = m_n(z)R(z;Q)\Sigma\left(\frac{\frac{1}{n}\operatorname{Tr}\xi_{22}}{1+\frac{1}{n}\operatorname{Tr}\xi_{22}}\right) + \alpha_2\Sigma + \alpha_1. \tag{42}$$

If we take $\frac{1}{n}\operatorname{Tr}(\cdot)$ of both sides of (41) and apply (37), we also arrive at

$$\frac{d}{n} = -m_n(z) - \left(1 - \frac{d}{n}\right) + \frac{t_n(z)}{1+t_n(z)} + \frac{1}{n}\operatorname{Tr}(\alpha_2\Sigma + \alpha_1),$$

so that putting everything together,

$$\frac{1}{n}\operatorname{Tr}(\xi_{22}) = -m_n(z)(1+t_n(z)) - 1 = -(1+t_n(z))\frac{1}{n}\operatorname{Tr}(\alpha_2\Sigma + \alpha_1). \tag{43}$$

In light of everything, the deformed Marchenko-Pastur law follows simply by estimating the errors α_1 and α_2.

Theorem 18: Deformed MP law

Suppose that $\{X_j\}$ are iid centered random vectors with covariance matrix Σ that satisfy the concentration inequality that for some $q \geq 1$,

$$\|\langle X_1, AX_1\rangle - \operatorname{Tr}(A\Sigma)\|_q \leq C_q' \|A\|_{HS}.$$

Then for all $q \geq 8$, there is a $C_q(|\operatorname{Im} z|^{-1}, |z|, \|\Sigma\|_{op}, C_q')$ so that

$$\|\operatorname{Tr}(R(z;Q)A) - \operatorname{Tr}(M_{11}A)\|_q \leq C_q\left(n^{1/q}\|A\|_{HS} + n^{4/q-1/2}\right).$$

The Hilbert–Schmidt norm control on A is the easiest. We would ideally trade this for $n^{-1/2}\|A\|_*$. Another direction of improvement is to dig into the dependence on $\operatorname{Im} z$, which should *almost* be replaced by $\varrho(z)$, as in Lemma 18. As the support of the eigenvalues of the random matrix can be on all of $\mathbb{R}$, this can only be true in a high-probability sense or alternatively allowing for error terms of the form of $(n|\operatorname{Im} z|)$.

We need a lower bound on $\{1+a_i\}$ and $1+t_n$ to be get started, as these appear in the denominator.[29].

[29] Note that for the deformed MP law, these are well-behaved, but to get started in comparing the random resolvent to this deterministic equivalent, we first need some a priori control

Lemma 19 (Safe denominator): Suppose $\Sigma, Q \succeq 0$. Then with $\tau = \operatorname{Tr}(R(z;Q)\Sigma)$

$$|1+\tau| = |1+\operatorname{Tr}(R(z;Q)\Sigma)| \geq \frac{|\operatorname{Im} z|}{|z|},$$

and if $\operatorname{Re} z \leq 0$, $|1+\tau| \geq 1$.

Proof. Taking the real and imaginary parts of this quadratic form, we get

$$x := \operatorname{Re}\operatorname{Tr}(R(z;Q)\Sigma) = \frac{1}{2}\operatorname{Tr}((R(z;Q) + R(\bar{z};Q))\Sigma).$$

Diagonalizing Q with eigenvalues λ_j and eigenvectors u_j

$$\frac{1}{2}\operatorname{Tr}((R(z;Q) + R(\bar{z};Q))\Sigma) = \sum_j \frac{(\lambda_j - \operatorname{Re} z)\langle u_j^{\otimes 2}, \Sigma\rangle}{|\lambda_j - z|^2}.$$

We can do the same with the imaginary part, which produces

$$y := \operatorname{Im}\operatorname{Tr}(R(z;Q)\Sigma) = \sum_j \frac{(\operatorname{Im} z)\langle u_j^{\otimes 2}, \Sigma\rangle}{|\lambda_j - z|^2}.$$

So we note that the real part cannot be too negative:

$$(\operatorname{Re}\operatorname{Tr}(R(z;Q)\Sigma))_- \leq \frac{\operatorname{Re} z}{\operatorname{Im} z}\operatorname{Im}\operatorname{Tr}(R(z;Q)\Sigma).$$

Now if $\operatorname{Re} z < 0$, we have $x \geq 0$ and $|1+\tau| \geq 1$. Similarly if $x < -2$, then we get the same bound. Otherwise we have $x \in (-2,0)$ and we have $-x < \frac{\operatorname{Re} z}{\operatorname{Im} z} y$, and so

$$(1+x)^2 + y^2 \geq (1+x)^2 + \left(\frac{\operatorname{Im} z}{\operatorname{Re} z}\right)^2 x^2.$$

This parabola we can minimize in x; its vertex is always in $(-1,0)$, and so by explicit computation, we get for $x < 0$

$$(1+x)^2 + y^2 \geq \frac{(\operatorname{Im} z)^2}{(\operatorname{Re} z)^2 + (\operatorname{Im} z)^2}.$$

□

Now we return to the proof of the deformed MP law.

Proof. We will control both $\frac{1}{n}\operatorname{Tr}(\xi_{22})$ and generalized entries of ξ_{11} with high probability. In light of (42) and (43), this problem reduces to bounding generalized entries of α_2 and α_1. Note that we can apply Lemma 19 can be applied to $1+t_n$ as well as $1+a_i$. We can also bound m_n deterministically by $|z|/|\operatorname{Im} z|$ and $|1+t_n|$ above by $\frac{\operatorname{Tr}(\Sigma)}{n}\frac{1}{|\operatorname{Im} z|}$.

For a generalized entry of α_1, even if we make no attempt at cancellation in the sum, we have

$$|\operatorname{Tr}(\alpha_1 A)| \leq \frac{|z|}{|\operatorname{Im} z|} \max_i \left| \langle X_i \otimes X_i - \Sigma, R(z; Q^{[i]})A \rangle \right|.$$

Hence for any $t \geq 1$ and any $q \geq 1$, conditioning on $Q^{[i]}$,

$$\begin{aligned} \Pr^{[i]}\left(\tfrac{|\operatorname{Im} z|}{|z|}|\operatorname{Tr}(\alpha_1 A)| \geq t\right) &\leq \frac{n \max_i \mathbb{E}\left| \langle X_i \otimes X_i - \Sigma, R(z; Q^{[i]})A \rangle \right|^q}{t^q} \\ &\leq \frac{n C_q^q \|AR(z; Q^{[i]})\|_{HS}^q}{t^q}. \end{aligned}$$

Integrating, against t^{q-2} over $[1,\infty]$ we conclude

$$\mathbb{E}^{[i]}\left(\frac{|\operatorname{Im} z|^2}{|z| C_q \|AR(z; Q^{[i]})\|_{HS}} |\operatorname{Tr}(\alpha_1 A)|\right)^{q-1} \leq (q-1)(1+n).$$

Adusting constants, we have

$$\|\operatorname{Tr}(\alpha_1 A)\|_q \leq C_q n^{1/q} \frac{|z|}{|\operatorname{Im} z|^2} \|A\|_{HS}.$$

Turning to α_2 we divide it into three parts (cf (40)) which are given by

$$\begin{aligned} \alpha_2' &= R(z;Q) \sum_{i=1}^n \frac{1}{n} \left(\frac{1}{1+a_i} - \frac{1}{1+\operatorname{Tr}(R(z;Q^{[i]})\Sigma)} \right) \\ \alpha_2'' &= R(z;Q) \sum_{i=1}^n \frac{1}{n} \left(\frac{1}{1+\operatorname{Tr}(R(z;Q^{[i]})\Sigma)} - \frac{1}{1+t_n(z)} \right) \\ \alpha_2''' &= \frac{1}{n^2} \sum_{i=1}^n \frac{((Q^{[i]} - z\operatorname{Id})^{-1}X_i) \otimes ((Q^{[i]} - z\operatorname{Id})^{-1}X_i)}{(1+a_i)^2} \end{aligned}$$

For the last part,

$$|\operatorname{Tr}(\alpha_2''' A)| \leq \frac{|z|^2}{|\operatorname{Im} z|^2} \frac{1}{n} \max_i |\langle (R(z;Q^{[i]})X_i)^{\otimes 2}, A\rangle|.$$

Each of these quadratic forms can be bounded in moments. Moreover taking the Hölder norm on both sides, for any $q \geq 1$

$$\|\operatorname{Tr}(\alpha_2''' A)\|_q \leq \frac{|z|^2}{|\operatorname{Im} z|^4} \frac{1}{n} \left(\|A\|_* \|\Sigma\|_{op} + C_q \|A\|_{HS}\right).$$

The α_2' terms, the a_i are quadratic forms that nearly concentrate, and so it can be bounded the same way as α_1, albeit with an extra factor of $|z|/|\operatorname{Im} z|^2$ accounting for the resolvent of Q and the additional $(1+a_i)$ terms in the denominator. For the α_2'' terms, we note

$$\operatorname{Tr}((R(z;Q^{[i]}) - R(z;Q))\Sigma)$$

can be expressed using the Woodbury identity, which produces a term that is bounded the same way as α_2'''.

We conclude that for any q and for a sufficiently large C_q depending on $\|\Sigma\|_{op}$, $\frac{1}{|\operatorname{Im} z|}$ and $|z|$

$$\|\tfrac{1}{n}\operatorname{Tr}(\xi_{22})\|_q \leq C_q n^{1/q-1/2} \quad \text{and} \quad \|\operatorname{Tr}(\xi_{11}A)\|_q \leq C_q n^{1/q}(n^{-1/2} + \|A\|_{HS}).$$

Finally we use the stability Lemma 18 to bound the difference

$$\Delta := |\operatorname{Tr}(R(z;Q)A) - \operatorname{Tr}(M_{11}A)|$$

for A with small $\|A\|_{HS}$. We note that when

$$C(n\psi^* + 1)\Xi \geq 1,$$

the stability gives nothing. However this event will have probability smaller than any power of n, and moreover we can simply bound Δ by $\|A\|_* \leq \sqrt{n}\|A\|_{HS}$. As for $n\psi^*$, these are generalized entries with test matrices of rank 1 and operator norm 1. Hence for some constant C_q (using the same tail bounds developed above)

$$\|n\psi^*\|_q \leq C_q n^{2/q}.$$

Hence for some constant (with the same dependencies as in Lemma 18)

$$\|\Xi\|_q \leq C_q \left(n^{2/q+1}\epsilon + \epsilon^{-2/q} n^{1/q-1/2}\right).$$

Taking $\epsilon = n^{3/2}$ we conclude

$$\|\Xi\|_q \leq C_q n^{4/q-1/2},$$

and hence

$$\|\operatorname{Tr}(R(z;Q)A) - \operatorname{Tr}(M_{11}A)\|_q \leq C_q \left(n^{1/q}\|A\|_{HS} + n^{4/q-1/2}\right).$$

□

2.6 Example: Spike models

2.7 Example: Power-law spectrum

2.8 Example: The conjugate kernel

3 SGD and optimization theory

SGD (stochastic gradient descent) [30] 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. [31] 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.

[30] 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].

[31] 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].

To introduce SGD, we will consider the *finite-sum framework*. This is an example of a *structure*[32] we impose on the objective function f to be optimized.

[32] 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.

Definition 32 (Finite sum): An optimization problem is *finite sum* if its objective function f can be given by

$$\min_{x\in\mathbb{R}^d}\{f(x)\} \quad \text{where} \quad f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad x\in\mathbb{R}^d. \tag{44}$$

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

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

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 $x, y \in \mathbb{R}^d$

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

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 34 (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 $x \neq x^*$.

3.1 SGD on the finite-sum

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

Definition 35 (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* (33), 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.

[34] 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 35. 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].

3.2 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 algorithmic 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 36 (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 37 (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 38 (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 7: 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

$$\mathscr{R}(x) = \tfrac{1}{2}\mathbb{E}(\epsilon + \langle a, \beta \rangle - \langle a, x \rangle)^2 = \tfrac{1}{2}\mathbb{E}\epsilon^2 + \tfrac{1}{2}\langle \beta - x, \Sigma(\beta - x) \rangle. \tag{45}$$

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 $1 \leq i \leq n$. 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

$$\mathscr{L}(x) = \tfrac{1}{2n} \sum_{i=1}^{n} (\langle a_i, x \rangle - b_i)^2 = \tfrac{1}{2n} \|Ax - b\|^2. \tag{46}$$

Example 8: 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

$$\mathscr{L}(x) = \tfrac{1}{2n}\|Ax - b\|^2 + \tfrac{\lambda}{2}\|x\|^2. \tag{47}$$

It should be noted that in context, one may wish to consider either the regularized population risk, or alternatively the unregularized population risk (45).

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 9: 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

$$\phi(x) = |x|. \tag{48}$$

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

Another is *binary logistic regression* in which case

$$\phi(x) = \frac{e^x}{1 + e^x}. \tag{49}$$

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

$$\mathscr{P}(x) := \tfrac{1}{2}\mathbb{E}(|\langle a, x \rangle| - |\langle a, \beta \rangle|)^2 \tag{50}$$

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. ([35])

[36] 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 10: Generalized linear models II

Generalized linear models also naturally can take $p > 1$. This allows natural generalizations of the Example 9 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:

$$M(x,a) := \frac{e^{\langle x,a\rangle_m}}{\langle e^{\langle x,a\rangle_m}, \mathbb{1}\rangle}, \tag{51}$$

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 ([37])class vector and a is an element of $\mathbb{R}^m$. For the loss, one may take once more the KL–divergence

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

[38] 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 11: The two layer neural network

Neural networks generalize Example 10 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 10, if we set (see [39])

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

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 (51),

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

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.

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

3.3 Streaming/Online stochastic gradient descent

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

Definition 39 (Streaming SGD): Streaming (aka online) SGD is the algorithm with updates given by (52).

$$X_{k+1} = X_k - \gamma_k \nabla_{X_k} \ell(X_k, M(X_k, a_{k+1}), b_{k+1}). \tag{52}$$

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.

3.4 *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

$$X_{k+1} := X_k - \gamma_k (\nabla F(X_k) + \xi_{k+1}) \tag{53}$$

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 19: 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 [41]

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

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. □

[41] 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.

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

$$F(X_{k+1}) = F(X_k) - \gamma_k \langle \nabla F(X_k), \nabla F(X_k) + \xi_{k+1} \rangle + R_{k+1}. \tag{54}$$

Theorem 20: 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 (53) 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 $\mathbb{E}(\|\xi_{k+1}\|^2 \mid \mathscr{F}_k) \leq K$. 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 19.

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 (54), 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

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

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 $F_\infty := \lim_{k\to\infty} F(X_k)$ exists almosts surely.

So we introduce the event $\mathcal{E}_P := \{\sup_k F(X_k) < P\}$ 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 $K = \{x : F(x) \le P\}$ 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 $\{x \in K : \|\nabla F(x)\|^2 \ge \epsilon\}$. 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

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

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 $\max_k \|(Z_k - Z_T)\| \le \delta/8$. If $\tau > T$ is a time at which X_k is in $\mathcal{U}_\epsilon$, then

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

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

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

Now on this time window, we have

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

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 [KY06].

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 $\mathcal{S} \cap K$.

3.5 *The pessimism of almost sure convergence*

The good part about the Robbins-Monro type condition on $\{\gamma_k\}$ in Theorem 20 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.

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.

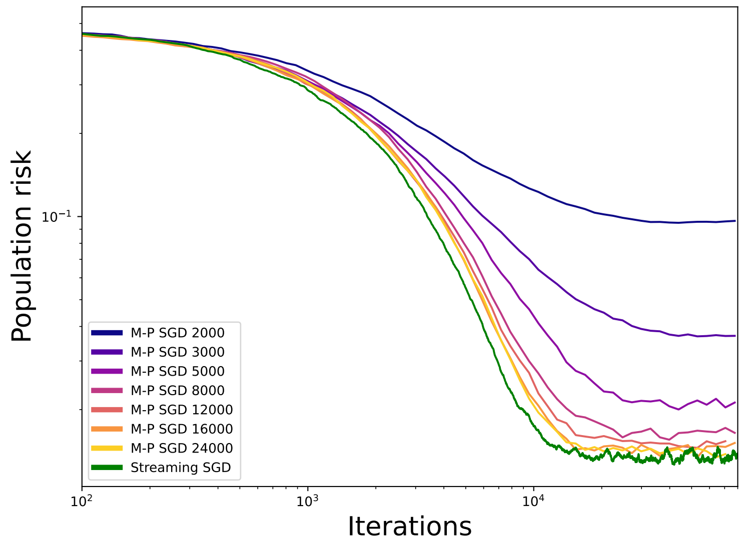


Figure 1: Linear regression (see Example 7). Constant step-size SGD with step-size within a factor of 3 of the largest stable step-size. Fixed dimension $d = 2000$, identity data covariance. Increasing numbers of samples, with multipass SGD. Streaming is the "infinite data" version.

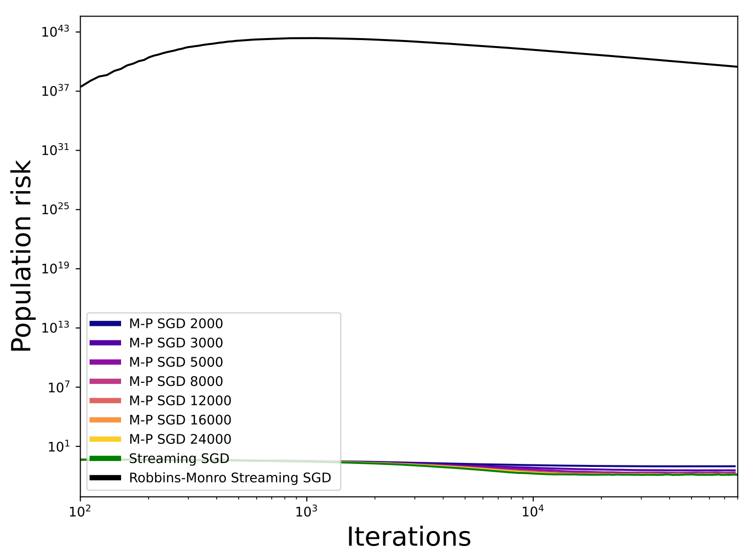


Figure 2: Linear regression (see Example 7). Same setup as Figure 1 with one additional curve, the Robbins-Monro stepsize $\gamma_k = 1/k$. By Theorem 20, the black curve converges. Rescaling the step-size (for example dividing by d) gives a curve which is effectively constant over the same time-scale.

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

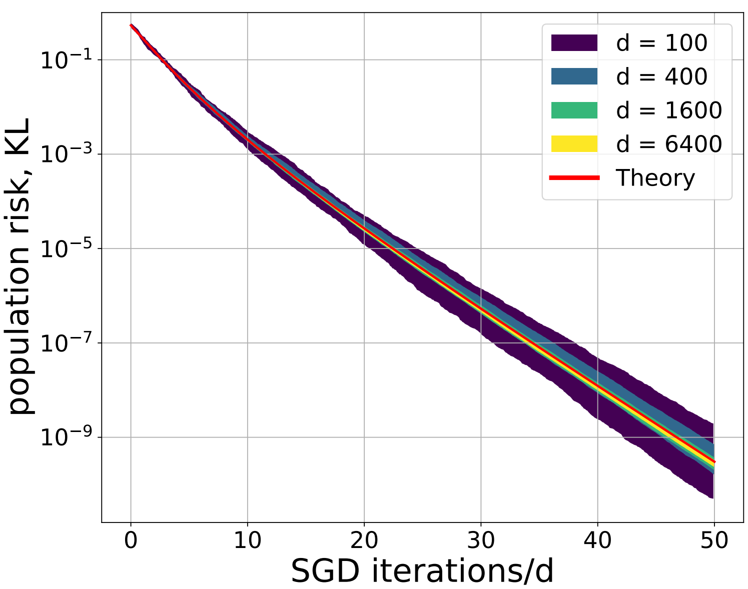


Figure 3: Population risk of logistic regression (see Example 9). In each dimension, 10 runs of streaming SGD for logistic regression are performed. We then display 80% confidence intervals over time (i.e. we discard the largest and smallest at error at each point in time). The curves concentrate around 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.)

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 12: Isotropic Gaussian Linear Regression

We follow Example 7, 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 13: Dynamical analysis of Isotropic Gaussian Linear Regression

Let $\mathscr{F}_k$ be the σ-algebra generated by $((a_j, b_j) : 0 \le j \le k)$. 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

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

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.

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

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

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

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

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

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

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

4.1 *Hidden finite dimensional risk manifold*

The key to the previous example (Example 12) 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 13, in which the risk itself describes an autonomous evolution.

Definition 40 (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 locally Lipschtiz 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 5 we we show one way to handle this.

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

Example 14: 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 13 show that these ERMs satisfy Part 2 and 3 of Definition 40 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 21: 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 [42]

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

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.

[42] 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 40 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 40, 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 40, 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 40

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

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 40. □

Example 15: GLMs with isotropic features

We suppose that we have GLM (Example 9) 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 16: 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 $(1/\sqrt{1 + A_{ii}} : 1 \leq i \leq \dim(\mathcal{V}))$,

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

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.

5 *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 17: 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

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

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 1.6).

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

$$\Pr\left(\left|a^T B a - \mathbb{E} a^T B a\right| \geq t\right) \leq 2 \exp\left(-\min\left\{\frac{t^2 d^{-4\varepsilon}}{\|B\|^2}, \frac{t d^{-2\varepsilon}}{\|B\|_\sigma}\right\}\right).$$

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

$$x_k - \beta = (\mathrm{Id}_d - \tfrac{\gamma}{d} a_k a_k^T)(x_{k-1} - \beta) + \tfrac{\gamma}{d} \eta w_k a_k, \tag{55}$$

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

$$\mathrm{d}\mathbf{X}_t = -\gamma \nabla \mathscr{P}(\mathbf{X}_t)\,\mathrm{d}t + \gamma \sqrt{\tfrac{2}{d} \mathscr{P}(\mathbf{X}_t) K}\,\mathrm{d}B_t \tag{56}$$

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

$$\mathscr{P}(x) := \tfrac{1}{2} \mathbb{E}_{(a,b)}(\langle a, x \rangle - b)^2, \quad (a, b) \sim \mathcal{D}. \tag{57}$$

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

Definition 41 (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 42 (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 22: 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 $n \leq d \log d / C(\|K\|_\sigma)$

$$\sup_{0 \leq k \leq n} \left| q(x_k) - q(\mathbf{X}_{k/d}) \right| < \|q\|_{C^2} \cdot e^{C(\|K\|_\sigma)\frac{n}{d}} \cdot d^{-\frac{1}{2}+9\varepsilon} \tag{58}$$

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 $\mathbb{E}q(\mathbf{X}_{k/d})$.

5.1 *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)

$$Q_t(z) := \tfrac{1}{2}\langle R(z;K), (\mathbf{X}_t - \beta) \otimes (\mathbf{X}_t - \beta)\rangle \quad z \in \mathbb{C}. \tag{59}$$

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

$$\mathscr{P}(\mathbf{X}_t) := \tfrac{1}{2}\langle K, (\mathbf{X}_t - \beta) \otimes (\mathbf{X}_t - \beta)\rangle = \frac{-1}{2\pi i} \oint_\Gamma z Q_t(z)\, \mathrm{d}z. \tag{60}$$

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

$$\begin{aligned} \mathrm{d}Q_t(z) = &-\gamma\langle KR(z;K), (\mathbf{X}_t - \beta) \otimes (\mathbf{X}_t - \beta)\rangle\, \mathrm{d}t \\ &+ \gamma\langle R(z;K), (\mathbf{X}_t - \beta) \otimes \sqrt{\tfrac{2K\mathscr{P}(\mathbf{X}_t)}{d}} dB_t\rangle \\ &+ \tfrac{\gamma^2 \mathscr{P}(\mathbf{X}_t)}{d}\langle R(z;K), \sqrt{K} dB_t \otimes \sqrt{K} dB_t\rangle. \\ &= -\gamma\langle zR(z;K) + \mathrm{Id}_d, (\mathbf{X}_t - \beta) \otimes (\mathbf{X}_t - \beta)\rangle\, \mathrm{d}t + \mathrm{d}M_t(z) \\ &+ \tfrac{\gamma^2 \mathscr{P}(\mathbf{X}_t)}{d}\langle R(z;K), K\,\mathrm{Id}_d\rangle\, \mathrm{d}t. \end{aligned} \tag{61}$$

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 (60)

$$\mathscr{P}(\mathbf{X}_t) = \frac{-1}{2\pi i} \oint_\Gamma z \left(Q_0(z) e^{-2\gamma z t} + \int_0^t e^{-2\gamma z (t-s)} \tfrac{\gamma^2 \mathscr{P}(\mathbf{X}_s)}{d}\, \mathrm{Tr}(KR(z;K))\, \mathrm{d}s + \mathrm{E}_t(z) \right) \mathrm{d}z. \tag{62}$$

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 43 (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 20 (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

$$\begin{aligned}\frac{-1}{2\pi i}\oint_\Gamma z Q_0(z) e^{-2\gamma z t}\,\mathrm{d}z &= \frac{1}{2}\left\langle \frac{-1}{2\pi i}\oint_\Gamma z e^{-2\gamma z t} R(z;K)\,\mathrm{d}z, (\mathbf{X}_0 - \beta)^{\otimes 2}\right\rangle \\ &= \frac{1}{2}\langle K e^{-2\gamma K t}, (\mathbf{X}_0 - \beta)^{\otimes 2}\rangle \\ &= \frac{1}{2}\langle K, (\mathscr{X}_{\gamma t} - \beta)^{\otimes 2}\rangle.\end{aligned}$$

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 (62) can be expressed as

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

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

Definition 44 ((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

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

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 23: 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.

5.2 *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 21 (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 20, we have the explicit integral curve $\mathscr{X}_t - \beta = e^{-tK}(X - \beta)$, with X the initialization of gradient flow. It follows that

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

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 45 (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 24: 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.

5.3 *Infinite dimensional Volterra equation*

The Volterra model in Definition 44 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 46 (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 47 (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 24 remains true, with $2\,\mathrm{Tr}\,K/d = 2\int_0^\infty x\mu(dx)$. 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 24.

Theorem 25: 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).

5.4 *Proof sketch of the homogenized SGD comparison*

We give a reduced version of the proof of Theorem 22. 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

$$v_k := x_k - \beta \quad \text{and} \quad V_t := \mathbf{X}_t - \beta. \tag{63}$$

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 (55), for a quadratic q

$$\begin{aligned} q(v_k) - q(v_{k-1}) &= -\gamma(\nabla q(v_{k-1}))^T(\Delta_k) + \tfrac{\gamma^2}{2}(\Delta_k)^T(\nabla^2 q)(\Delta_k), \\ \text{where} \quad m_k &= a_k/\sqrt{d} \\ \Delta_k &= m_k(m_k^T v_{k-1} - \eta w_k) \end{aligned} \tag{64}$$

The equation above can each be decomposed as a predictable part

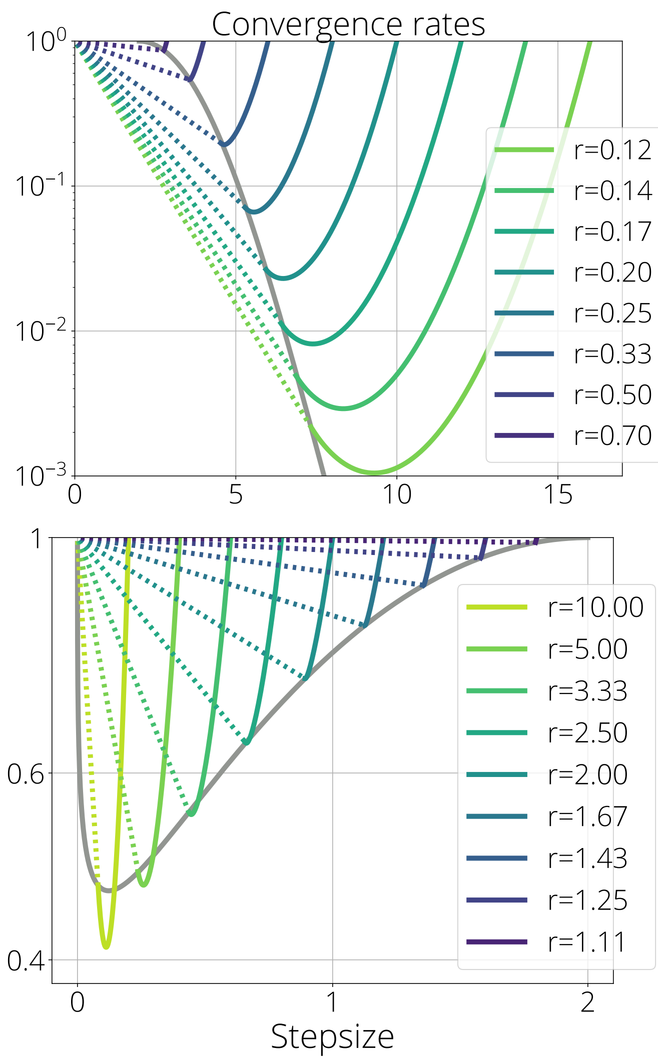


Figure 4: **Phase transition of the convergence rate** (y-axis) as a function of the stepsize (x-axis, γ) for the isotropic features model at infinite dimensions. Thus μ is Marchenko-Pastur (depending on aspect ratio r) and gradient flow is given by an isotropic starting vector. Smaller stepsizes (dotted) 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.

and two martingale increments

$$\begin{aligned} q(v_k) - q(v_{k-1}) = &- \gamma(\nabla q(v_{k-1}))^T \left(\tfrac{1}{d} K v_{k-1}\right) + \Delta \mathcal{M}_k^{\text{lin}} \\ &+ \tfrac{\gamma^2}{2} \operatorname{Tr}(\tfrac{1}{d} K (\nabla^2 q)) \left(\tfrac{1}{d} v_{k-1}^T K v_{k-1} + \mathbb{E}[\eta_k^2]\right) \\ &+ \Delta \mathcal{E}_k^{\text{quad}} + \Delta \mathcal{M}_k^{\text{quad}}, \\ \text{where} \quad \Delta \mathcal{M}_k^{\text{quad}} := &\Delta_k^T \nabla^2 q \Delta_k - \mathbb{E}[\Delta_k^T \nabla^2 q \Delta_k \mid \mathcal{F}_{k-1}]. \end{aligned} \tag{65}$$

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

$$\begin{aligned} &\mathcal{Q}_n(q) := \mathcal{Q}_n(q, K) = \\ &\Big\{ q(x), \quad (\nabla q(x))^T R(z; K) x, \quad x^T R(y; K)(\nabla^2 q) R(z; K) x, \\ &\quad (\nabla q(x))^T R(z; K) \beta, \quad x^T R(y; K)(\nabla^2 q) R(z; K) \beta, \quad \forall z, y \in \Gamma \Big\}. \end{aligned} \tag{66}$$

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

$$\tau := \inf \{k : ||v_k|| > d^{\varepsilon}\} \cup \{td : ||V_t|| > d^{\varepsilon}\}, \tag{67}$$

and we introduce the corresponding stopped processes

$$v_k^{\tau} = v_{k \wedge \tau}, \quad V_t^{\tau} = V_{t \wedge (\tau/d)}. \tag{68}$$

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

$$\begin{aligned} &\max_{0 \leq t \leq \frac{n}{d}} |q(v_{td}^{\tau}) - q(V_t^{\tau})| \leq \\ &\sup_{0 \leq t \leq n/d} \left(|\mathcal{M}_{\lfloor td \rfloor}^{\text{lin},\tau}| + |\mathcal{M}_{\lfloor td \rfloor}^{\text{quad},\tau}| + |\mathcal{E}_{\lfloor td \rfloor}^{\text{quad},\tau}| + |\mathcal{M}_t^{HSGD,\tau}| \right) \\ &+ C(\|K\|_{\sigma}) \cdot \sup_{g \in \mathcal{Q}} \int_0^{n/d} |g(v_{sd}^{\tau}) - g(V_s^{\tau})| ds. \end{aligned} \tag{69}$$

> 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 (65), 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 (66)), 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

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

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

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 $\mathcal{M}_k^{\text{lin},\tau}, \mathcal{M}_k^{\text{quad},\tau}, \mathcal{E}_k^{\text{quad},\tau}$, which are defined analogously to (68).

Lemma 24 (Martingale bounds): For any quadratic q with $\|q\|_{C^2} \leq 1$, the terms $\mathcal{M}_k^{\text{lin},\tau}, \mathcal{M}_k^{\text{quad},\tau}, \mathcal{E}_k^{\text{quad},\tau}$ satisfy the following bounds with overwhelming probability (with a bound which is uniform in q) for $n \leq d \log d$

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

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

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

Combining Lemmas 22 and 23, along with the above, we conclude that, for any $\bar{q} \in \bar{Q}$ with $\|q\|_{C^2} = 1$,

$$|\bar{q}(v_{td}^\tau) - \bar{q}(V_t^\tau)| \leq 4d^{-\frac{1}{2}+9\varepsilon} + C(\|K\|_\sigma) \max_{g\in Q} \int_0^t |g(v_{sd}^\tau) - g(V_s^\tau)| ds. \tag{70}$$

Hence by Lemma 23 and by bounding $\|g\|_{C^2}$ over all Q,

$$\max_{g\in Q} |q(v_{td}^\tau) - q(V_t^\tau)| \leq C(\|K\|_\sigma) \left(d^{-2} + d^{-\frac{1}{2}+9\varepsilon} + \int_0^t \max_{g\in Q} |g(v_{sd}^\tau) - g(V_s^\tau)| ds \right). \tag{71}$$

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

$$\max_{g\in Q} \max_{0\leq t\leq n/d} |g(v_{td}^\tau) - g(V_t^\tau)| \leq C(\|K\|_\sigma)(d^{-2} + 4d^{-\frac{1}{2}+9\varepsilon}) e^{C(\|K\|_\sigma)n/d}. \tag{72}$$

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 $\mathcal{E} = \{\max_{0\leq s\leq n/d} \|V_s\| \leq d^{\varepsilon/2}\}$, then we have

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

This is because on the event $\{\tau \leq n/d\} \cap \mathcal{E}$ 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 25 (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

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

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

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

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

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

□

5.5 *Controlling the errors*

The main goal of this section is to control the martingale terms and error terms; in particular we prove Lemma 24. We will also record for future use an estimate on ∇q that follows from $\|\cdot\|_{C^2}$ control.

$$||\nabla q(x)|| \le \|\nabla^2 q\|_\sigma \cdot ||x|| + ||\nabla q(0)|| \le \|q\|_{C^2} \cdot (||x|| + 1). \tag{73}$$

Martingale for gradient part of recurrence.

Proof. Comparing (64) and (65), we see that for $k \le \tau$

$$\begin{aligned} \Delta\mathcal{M}_k^{\text{lin},\tau} &= \left[\left(w_{k-1}^T m_k\right)\left(m_k^T v_{k-1}^\tau - \eta_k\right) - \tfrac{1}{d} w_{k-1}^T K v_{k-1}^\tau\right] \\ &=: [\Delta\mathcal{M}_k^{\text{lin}\,1,\tau} - \Delta\mathcal{M}_k^{\text{lin}\,2,\tau}], \\ &\text{where} \quad w_{k-1} := -\gamma\nabla q(v_{k-1}^\tau) + \tfrac{\gamma^2 \mathrm{d}}{d}(v_{k-1}^\tau + \beta). \end{aligned} \tag{74}$$

Note for $k > \tau$, the stopped martingale increment is 0. Using (73), $\|w_{k-1}\| \le C(\gamma, \mathrm{d}) d^{\varepsilon}$. 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

$$||\Delta\mathcal{M}_k^{\text{lin}\,1,\tau}||_{\psi_1} \le \left\|w_{k-1}^T m_k\right\|_{\psi_2} \left\|m_k^T v_{k-1}^\tau - \eta_k\right\|_{\psi_2} \le Cd^{-\frac{1}{2}+2\varepsilon} \cdot d^{-\frac{1}{2}+2\varepsilon} \tag{75}$$

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

$$|\Delta\mathcal{M}_k^{\text{lin}\,2,\tau}| = |\tfrac{1}{d} w_{k-1}^T K v_{k-1}^\tau| \le Cd^{-1+2\varepsilon}. \tag{76}$$

Combining these, we see that, for every k,

$$\sigma_{k,1} := \inf\{t > 0 : \mathbb{E}[\exp(|\Delta\mathcal{M}_k^{\text{lin}\,1,\tau} - \Delta\mathcal{M}_k^{\text{lin}\,2,\tau}|/t)|\mathcal{F}_{k-1}] \le 2\} \le Cd^{-1+4\varepsilon} \tag{77}$$

and, by the martingale Bernstein inequality,

$$\begin{aligned} &\Pr\left(\sup_{1\le k\le n} |\mathcal{M}_k^{\text{lin},\tau} - \mathbb{E}\mathcal{M}_0^{\text{lin}}| \ge t\right) \\ &\le 2\exp\left(-\min\left\{\frac{t}{c\max\sigma_{k,1}}, \frac{t^2}{c\sum_{k=1}^n \sigma_{k,1}}\right\}\right) \\ &\le 2\exp\left(-\min\left\{Ctd^{1-4\varepsilon}, Ct^2 d^{2-8\varepsilon} n^{-1}\right\}\right). \end{aligned} \tag{78}$$

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

$$\sup_{1\le k\le n} |\mathcal{M}_k^{\text{lin},\tau}| \le d^{-\frac{1}{2}+5\varepsilon} \tag{79}$$

with overwhelming probability. □

Martingale for Hessian part of recurrence.

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

$$\begin{aligned} &\tfrac{\gamma^2}{2}(m_k m_k^T v_{k-1}^\tau - m_k\eta_k)^T(\nabla^2 q)(m_k m_k^T v_{k-1}^\tau - m_k\eta_k) \\ &= \mathbb{E}\left[\tfrac{\gamma^2}{2}(m_k m_k^T v_{k-1}^\tau - m_k\eta_k)^T(\nabla^2 q)(m_k m_k^T v_{k-1}^\tau - m_k\eta_k)|\mathcal{F}_{k-1}\right] + \Delta\mathcal{M}_k^{\text{quad}}. \end{aligned} \tag{80}$$

Rearranging the terms, we get

$$\Delta\mathcal{M}_k^{\text{quad}} = A_k B_k - \mathbb{E}[A_k B_k|\mathcal{F}_{k-1}] \tag{81}$$

where

$$A_k := m_k^T(\nabla^2 q)m_k, \quad B_k := (m_k^T v_{k-1}^\tau - \eta_k)^2. \tag{82}$$

This can be expanded as

$$\begin{aligned} \Delta\mathcal{M}_k^{\text{quad}} =& (A_k - \mathbb{E}[A_k])(B_k - \mathbb{E}[B_k]) + \mathbb{E}[A_k]\mathbb{E}[B_k] - \mathbb{E}[A_k B_k] \\ &+ (A_k - \mathbb{E}[A_k])\mathbb{E}[B_k] + (B_k - \mathbb{E}[B_k])\mathbb{E}[A_k], \end{aligned} \tag{83}$$

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

$$\begin{aligned} &\Pr(|A_k - \mathbb{E}A_k| \geq t) \\ &\leq 2\exp\left[-c\min\left(\frac{t^2}{d^{-2+4\varepsilon}||\nabla^2 q||_{HS}^2}, \frac{t}{d^{-1+2\varepsilon}||\nabla^2 q||}\right)\right] \\ &\leq 2\exp[-c'\min(t^2 d^{1-4\varepsilon}, td^{1-2\varepsilon})] \leq 2\exp[-c'' t d^{\frac{1}{2}-2\varepsilon}] \end{aligned} \tag{84}$$

and thus we have the subexponential bound

$$||A_k - \mathbb{E}[A_k]||_{\psi_1} < Cd^{-\frac{1}{2}+2\varepsilon}. \tag{85}$$

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

$$\begin{aligned} &\Pr\left(\left|m_k^T v_{k-1}^\tau (v_{k-1}^\tau)^T m_k - \mathbb{E} m_k^T v_{k-1}^\tau (v_{k-1}^\tau)^T m_k\right| \geq t\right) \\ &\leq 2\exp\left[-c\min\left(\frac{t^2}{d^{-2+4\varepsilon}||v_{k-1}^\tau (v_{k-1}^\tau)^T||_{HS}^2}, \frac{t}{d^{-1+2\varepsilon}||v_{k-1}^\tau (v_{k-1}^\tau)^T||}\right)\right] \\ &\leq 2\exp[-c\min(t^2 d^{2-8\varepsilon}, td^{1-4\varepsilon})]. \end{aligned} \tag{86}$$

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

$$\begin{aligned} ||m_k^T v_{k-1}^\tau \eta_k||_{\psi_1} &\leq ||m_k^T v_{k-1}^\tau||_{\psi_2} \cdot ||\eta_k||_{\psi_2} = d^{-\frac{1}{2}+2\varepsilon} d^{-\frac{1}{2}+\varepsilon} \\ &= d^{-1+3\varepsilon}. \end{aligned} \tag{87}$$

Since also $||\eta_k^2||_{\psi_1} = d^{-1+2\varepsilon}$ combining the bounds (86) and (87), we have

$$||B_k - \mathbb{E}[B_k]||_{\psi_1} < Cd^{-1+4\varepsilon}. \tag{88}$$

Furthermore, we have

$$\mathbb{E}[A_k] = O(1), \qquad \mathbb{E}[B_k] = O(d^{-1}), \tag{89}$$

uniformly for all k based on the assumptions on η_k and m_k. We now use (85), (88), (89) to bound each term of (83) 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,

$$\sum_{k=1}^{n} \left|(A_k - \mathbb{E}[A_k])(B_k - \mathbb{E}[B_k])\right| < nd^{-\frac{3}{2}+8\varepsilon} < d^{-\frac{1}{2}+9\varepsilon}. \tag{90}$$

For the second term of (83) we have

$$\Big|\mathbb{E}[A_k]\mathbb{E}[B_k] - \mathbb{E}[A_kB_k]\Big| = \Big|\mathbb{E}\big[(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\big]\Big| \leq \mathbb{E}\Big|(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\Big|. \tag{91}$$

We can bound this quantity using

$$\begin{aligned} &\Pr\left(\big|(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\big| \geq t\right) \\ &\leq \Pr(|A_k - \mathbb{E}A_k| \geq \sqrt{t}) + \Pr(|B_k - \mathbb{E}B_k| \geq \sqrt{t}) \\ &\leq 4\exp\left[-c\min(td^{1-4\varepsilon}, \sqrt{t}d^{1-4\varepsilon})\right] \end{aligned} \tag{92}$$

where the bound in the last line comes from combining (84) and (88).
Using this bound, we obtain

$$\begin{aligned} \Big|\mathbb{E}[A_k]\mathbb{E}[B_k] - \mathbb{E}[A_kB_k]\Big| &\leq \int_0^\infty x\Pr\left(\big|(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\big| \geq x\right)dx \\ &\leq \int_0^1 4x\exp(-cxd^{1-4\varepsilon})dx + \int_1^\infty 4x\exp(-c\sqrt{x}d^{1-4\varepsilon})dx \end{aligned} \tag{93}$$

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

$$4d^{-2+8\varepsilon}\int_0^{d^{1-4\varepsilon}} y\exp(-cy)dy + 4d^{-4+8\varepsilon}\int_{d^{1-4\varepsilon}}^\infty z^2\exp(-cz)dz = O(d^{-2+8\varepsilon}). \tag{94}$$

Thus,

$$\sum_{k=1}^n \Big|\mathbb{E}[A_k]\mathbb{E}[B_k] - \mathbb{E}[A_kB_k]\Big| = O(nd^{-2+8\varepsilon}). \tag{95}$$

Finally, we note that the remaining terms of (83), 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

$$||(A_k - \mathbb{E}[A_k])\mathbb{E}[B_k]||_{\psi_1} \leq Cd^{-\frac{3}{2}+2\varepsilon}, \qquad ||(B_k - \mathbb{E}[B_k])\mathbb{E}[A_k]||_{\psi_1} \leq Cd^{-1+4\varepsilon}. \tag{96}$$

Applying the Martingale Bernstein inequality, we conclude

$$\begin{aligned} &\Pr\left(\sup_{1\leq k\leq n}\left|\sum_{j=1}^k (A_j - \mathbb{E}[A_j])\mathbb{E}[B_j] + (B_j - \mathbb{E}[B_j])\mathbb{E}[A_j]\right| \geq t\right) \\ &\leq 2\exp\left(-\min\left\{\frac{t}{c\max\sigma_{k,1}}, \frac{t^2}{c\sum_{k=1}^n \sigma_{k,1}}\right\}\right) \\ &\leq 2\exp\left(-\min\left\{Ctd^{1-4\varepsilon}, Ct^2d^{2-8\varepsilon}n^{-1}\right\}\right). \end{aligned} \tag{97}$$

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

$$\sup_{1\leq k\leq n}\left|\sum_{j=1}^k (A_j - \mathbb{E}[A_j])\mathbb{E}[B_j] + (B_j - \mathbb{E}[B_j])\mathbb{E}[A_j]\right| \leq d^{-\frac{1}{2}+5\varepsilon} \tag{98}$$

with overwhelming probability. Finally, combining the bounds from (90), (95), (98), we conclude that, for $n \leq d \log d$,

$$\sup_{1 \leq k \leq n} |\mathcal{M}_k^{\text{quad},\tau}| \leq d^{-\frac{1}{2}+8\varepsilon} \tag{99}$$

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

For part (iii), we observe that $\Delta\mathcal{E}_k^{\text{quad},\tau} = \mathbb{E}[A_k B_k] - \mathbb{E}[A_k]\mathbb{E}[B_k] + O(d^{-2+4\epsilon})$, the error terms arising from u_k cross terms, so that the bound of $\mathcal{E}_k^{\text{quad},\tau}$ follows immediately from (95). $\square$

6 *Homogenization of Multipass SGD on the least squares*

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

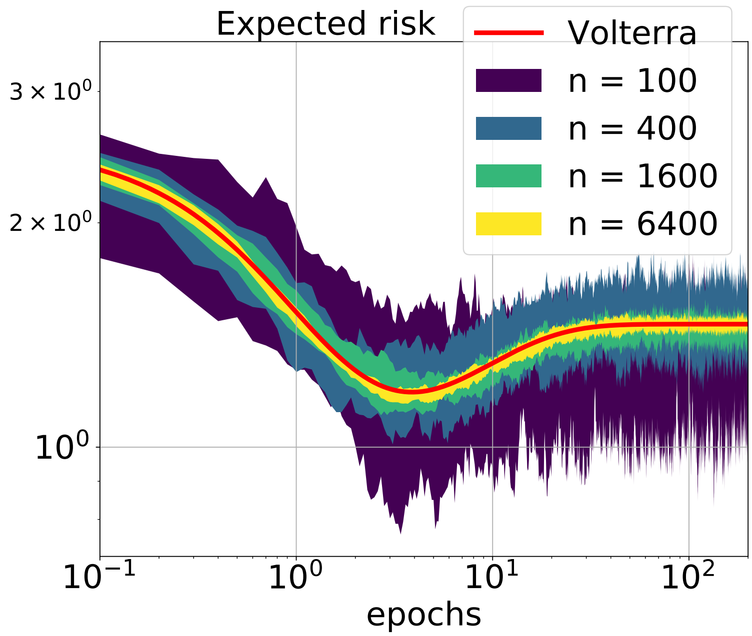


Figure 5: Risk curves of SGD across different dimensions. In each dimension, 10 runs of multi-pass constant step-size SGD are performed on a least squares problem, and the test error is computed over time. We then display 80% confidence intervals over time (i.e. we discard the largest and smallest at 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

$$x_{k+1} = x_k - \gamma_k(\langle a_{i_{k+1}}, x\rangle - b_{i_{k+1}})a_{i_{k+1}}, \quad \{i_k\} \text{ iid } \mathrm{Unif}(\{1, 2, \cdots, n\}). \tag{100}$$

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

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

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

$$\hat{K} := \tfrac{1}{n} A^T A \quad \text{and} \quad \check{K} := \tfrac{1}{n} A A^T, \tag{101}$$

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 (56) as a guide, then with $\gamma_k = \gamma(k/d)/d$

$$\mathrm{d}\mathbf{X}_t = -\gamma(t)\big(\nabla \mathscr{L}(\mathbf{X}_t) + \sqrt{\tfrac{2}{d}\mathscr{L}(\mathbf{X}_t)\hat{K}} dB_t\big). \tag{102}$$

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

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

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

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

$$\mathscr{R}(x) = \tfrac{1}{2}x^T T x + u^T x + c. \tag{103}$$

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

$$\begin{aligned} &\max_{z,y \in \Gamma} \max_{1 \leq i \leq n} \tfrac{1}{n} |\boldsymbol{e}_i^T A \widehat{T} A^T \boldsymbol{e}_i - \operatorname{Tr}(\hat{K}\widehat{T})| \leq \|T\| n^{-\theta}, \quad \text{where} \\ &\widehat{T} = R(z) T R(y) + R(y) R(z), \quad R(z) = R(z; \hat{K}). \end{aligned} \tag{104}$$

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

Theorem 26: Homogenization of multi-pass SGD

Suppose $n \geq d^{\tilde{\varepsilon}}$ and $n \leq d^C$ 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 $T > 0$

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

with overwhelming probability.

Example 18: 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 $((a_i, b_i) : 1 \leq i \leq n)$ 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 $\mathscr{L}$,

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

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 $N(0, \text{Id}_m \otimes \text{Id}_p)$, 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 26; see [Paq+22a] for details.

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

Definition 48 ((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

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

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 23. 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 26 (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 7 (Generalization error model): The generalization error $\mathscr{P}(\mathbf{X}_t)$ evolves according to the risk curve

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

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

6.1 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

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

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

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

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.

6.2 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.

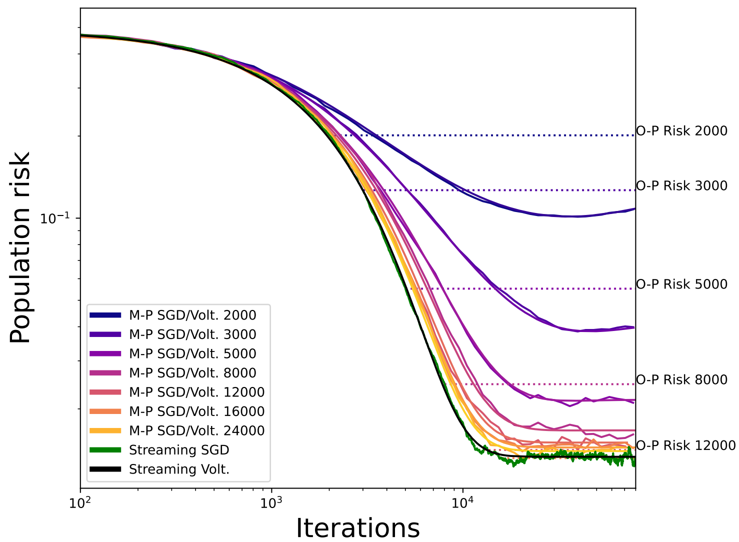


Figure 6: Risk curves for a simple linear regression problem. Multi-pass SGD, its high dimensional equivalent (i.e. "Volterra"), Streaming SGD (i.e. one-pass with varying dataset size), and the expected risk of homogenized SGD ("Streaming Volterra") are all plotted. Risk levels for streaming SGD at various 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 (100)

$$\begin{aligned} q(x_{k+1}) - q(x_k) &= -\gamma_k \langle \nabla q(x_k), a_{i_{k+1}} \rangle (\langle a_{i_{k+1}}, x_k \rangle - b_{i_{k+1}}) \\ &\quad + \tfrac{\gamma_k^2}{2} \langle \nabla^2 q(x_k), a_{i_{k+1}}^{\otimes 2} \rangle (\langle a_{i_{k+1}}, x_k \rangle - b_{i_{k+1}})^2. \end{aligned} \tag{105}$$

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

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

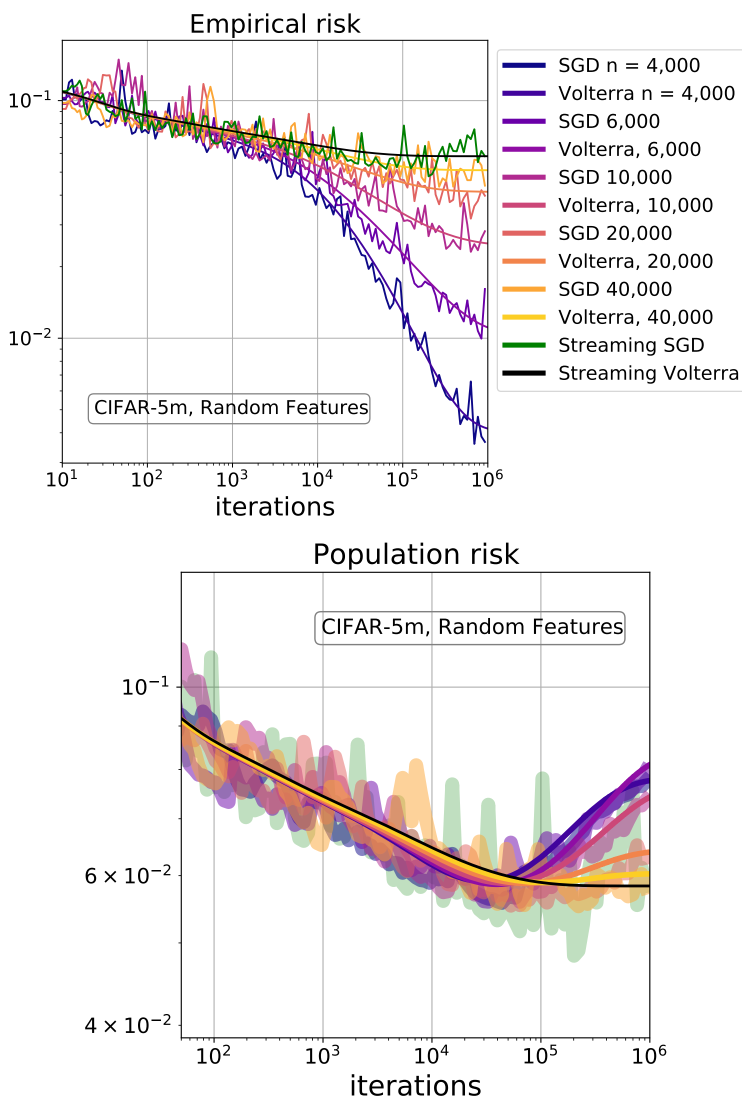


Figure 7: Risk curves for fully-connected random features model (with $d = 6000$) built on CIFAR-5m, empirical (top) and test-loss (bottom). The CIFAR-5m dataset [NNS21] is a synthetically generated 5 million data-point set of images, with the same class structure and image geometry as CIFAR-10. We compare 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

$$\begin{aligned} q(x_{k+1}^{\tau}) - q(x_k^{\tau}) &= -\gamma_k \langle \nabla q(x_k^{\tau}), \nabla \mathscr{L}(x_k^{\tau})\rangle + \gamma_k \Delta M_k^{\text{lin}} \\ &\quad + \tfrac{\gamma_k^2}{2}\langle \nabla^2 q, \hat{K}\rangle \mathscr{L}(x_k^{\tau}) + \gamma_k \mathrm{KL}_k + \gamma_k^2 \Delta M_k^{\text{quad}}, \end{aligned} \tag{106}$$

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 (66), 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:

$$\begin{aligned} &\mathcal{Q}_n(q) := \mathcal{Q}_n(q, \hat{K}) = \\ &\Big\{ q(x), \quad (\nabla q(x))^T R(z; \hat{K}) x, \quad x^T R(y; \hat{K})(\nabla^2 q) R(z; K) x, \\ &\quad (\nabla q(x))^T R(z; \hat{K}) A^T b, \quad x^T R(y; \hat{K})(\nabla^2 q) R(z; K) A^T b, \quad \forall z, y \in \Gamma \Big\}, \end{aligned} \tag{107}$$

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

[AEK19] Oskari H Ajanki, László Erdős, and Torben Krüger. "Stability of the matrix Dyson equation and random matrices with correlations". In: *Probability Theory and Related Fields* 173 (2019), pp. 293–373.

[AEK20] Johannes Alt, László Erdős, and Torben Krüger. "The Dyson equation with linear self-energy: spectral bands, edges and cusps". In: *Documenta Mathematica* 25 (2020), pp. 1421–1539.

[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).

[Arn71] L. Arnold. "On Wigner's semicircle law for the eigenvalues of random matrices". In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 19 (1971), pp. 191–198. DOI: 10.1007/BF00534107. URL: https://doi.org/10.1007/BF00534107.

[AT07] Robert J. Adler and Jonathan E. Taylor. *Random fields and geometry*. Springer Monographs in Mathematics. Springer, New York, 2007, pp. xviii+448.

[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].

[CY12] Loïc Chaumont and Marc Yor. *Exercises in probability*. Second. Vol. 35. Cambridge Series in Statistical and Probabilistic Mathematics. A guided tour from measure theory to random processes, via conditioning. Cambridge University Press, Cambridge, 2012, pp. xx+279. DOI: 10.1017/CBO9781139135351. URL: https://doi.org/10.1017/CBO9781139135351.

[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).

[DK05] Michael Demuth and M. Krishna. *Determining spectra in quantum theory*. Vol. 44. Progress in Mathematical Physics. Birkhäuser Boston, 2005, pp. xii+219.

[Erd19] Laszlo Erdos. *The matrix Dyson equation and its applications for random matrices*. 2019. arXiv: 1903.10060 [math.PR].

[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.

[HFS07] J William Helton, Reza Rashidi Far, and Roland Speicher. "Operator-valued semicircular elements: solving a quadratic matrix equation with positivity constraints". In: *International Mathematics Research Notices* 2007.9 (2007), rnm086–rnm086.

[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.

[KY06] H. Kushner and G.G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer New York, 2006.

[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.

[MP67] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. "Distribution of eigenvalues for some sets of random matrices". In: *Matematicheskii Sbornik* 114.4 (1967), pp. 507–536.

[MS17] James A. Mingo and Roland Speicher. *Free probability and random matrices*. Vol. 35. Fields Institute Monographs. Springer, 2017, pp. xiv+336. DOI: 10.1007/978-1-4939-6942-5. URL: https://doi.org/10.1007/978-1-4939-6942-5.

[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, arXiv:2102.04396, 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, arXiv:2106.03696, 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.

[Whi60] Peter Whittle. "Bounds for the moments of linear and quadratic forms in independent variables". In: *Theory of Probability & Its Applications* 5.3 (1960), pp. 302–305.

[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).