Global Convergence of Multiplicative Updates for the Matrix Mechanism: A Collaborative Proof with Gemini 3.

Keith Rush
Google DeepMind

March 23, 2026

Abstract

We analyze a fixed-point iteration $v \leftarrow \phi(v)$ arising in the optimization of a regularized nuclear norm objective involving the Hadamard product structure, posed in [DMR+22] in the context of an optimization problem over the space of algorithms in private machine learning. We prove that the iteration $v^{(k+1)} = \operatorname{diag}((D_{v^{(k)}}^{1/2} M D_{v^{(k)}}^{1/2})^{1/2})$ converges monotonically to the unique global optimizer of the potential function $J(v) = 2\operatorname{Tr}((D_v^{1/2} M D_v^{1/2})^{1/2}) - \sum v_i$, closing a problem left open there.

The bulk of this proof was provided by Gemini 3, subject to some corrections and interventions. Gemini 3 also sketched the initial version of this note. Thus, it represents as much a commentary on the practical use of AI in mathematics as it represents the closure of a small gap in the literature. As such, we include a small narrative description of the prompting process, and some resulting principles for working with AI to prove mathematics.

1 The Problem and its Brief History

We consider the problem of finding the fixed point $v^* \in \mathbb{R}^N_{++}$ satisfying the equation:

$$v^* = \operatorname{diag}\left(\left(D_{v^*}^{1/2} M D_{v^*}^{1/2}\right)^{1/2}\right) \tag{1}$$

where $M \in \mathbb{R}^{N\times N}$ is a symmetric positive definite matrix ($M \succ 0$), and $D_v = \operatorname{diag}(v)$.

This fixed-point equation arises naturally as the stationarity condition ($\nabla J(v) = 0$) for the potential function $J : \mathbb{R}^N_{++} \to \mathbb{R}$:

$$J(v) = 2\operatorname{Tr}\left(\left(D_v^{1/2} M D_v^{1/2}\right)^{1/2}\right) - \sum_{i=1}^N v_i \tag{2}$$

To our knowledge, the first occurrence of this potential can be found in [DMR+22], where it arose in the context of designing matrix factorizations that are in some sense *optimal* for the matrix mechanism [LMH+15] under a specific notion of error. [DMR+22] was able to show that iterating the mapping Eq. (1) converged locally, and provided empirical evidence that this fixed-point iteration converged extremely rapidly, outpacing pure gradient-based approaches, but global convergence of these iterates was left open.

The followup papers in the series of [DMR+22] (see the literature review in [PUCC+25]) generally moved away from the unconstrained optimization problem that yields the potential Eq. (2),

in order to impose structural constraints on matrices which address practical concerns like runtime efficiency and ability to prove privacy amplification results. The specification of the BLT class [DMP+24] providing the ability to compute near-optimal factorizations with complexity independent of matrix size rendered the convergence of the iterates of Eq. (1) even more academic. Still, I always viewed it as unsatisfactory that we were unable to prove global convergence despite reasonably serious attempts by reasonably serious mathematicians. This global convergence always had the feel to me of something which was just an issue of putting together the right tools, which probably all existed.

I therefore kept it in my back pocket, and tested it repeatedly on new versions of various AI models as they were released. Given my affiliation, most of these attempts used Gemini. With a few iterations, Gemini 3 was able to identify the crucial piece that we had missed: a variational characterization of the nuclear norm that allows the problem to be diagonalized.

This diagonalization of our crucial object essentially required the ability to call certain linear algebra to hand that neither myself nor my collaborators demonstrated, but it was not exactly revolutionary. This is not unreasonable – mathematics is quite large, and not every mathematician can recall every inequality or characterization of an object under study, particularly if this object is far from their main field. We hope that this note can serve as a small example of utilization of AI in mathematics, echoing several that have appeared elsewhere in the literature, and also represent the closure of a problem that had always nagged me.

We study the convergence of the parameter-free multiplicative update rule:

$$v^{(k+1)} = \phi(v^{(k)}) = \operatorname{diag}\left(\left(D_{v^{(k)}}^{1/2} M D_{v^{(k)}}^{1/2}\right)^{1/2}\right) \tag{3}$$

Our main contribution is a proof that this iteration satisfies the monotonic ascent property $J(v^{(k+1)}) \geq J(v^{(k)})$, ensuring global convergence to the unique maximum. We also include discussion of the experience of working interactively with natural-language AI models to produce nontrivial mathematics.

2 Theoretical Analysis

2.1 Basic Properties

The result here depends critically on a representation of the first term in our potential as the *nuclear norm* of a matrix, and a particular variational characterization for this norm.

Definition 1 (Nuclear Norm)**.** *For any real matrix A, define the nuclear norm $\|A\|_*$ as:*

$$\|A\|_* = \operatorname{Tr}((A^\top A)^{1/2})$$

We will collect the properties of the nuclear norm that we need before we begin.

Lemma 1. *For any real A, the nuclear norm can be characterized in the following manner:*

$$\|A\|_* = \sup_{U \in \mathcal{U}(N)} \operatorname{Tr}(U^\top A)$$

where $\mathcal{U}(N)$ is the set of unitary matrices. The supremum is attained when U is the unitary factor from the polar decomposition of A (see Section 7.4 of [HJ90]).

Proof. We recall the polar factorization of A, the representation $A = UP$ where $P = (A^\top A)^{1/2}$ and U is unitary. (Since A is real, U is in fact orthogonal). Therefore $U^\top A = P$, and:

$$\|A\|_* = \operatorname{Tr}\left((A^\top A)^{1/2}\right) = \operatorname{Tr}(P) = \operatorname{Tr}(U^\top A).$$

This shows that $\|A\|_* \geq \sup_{U \in \mathcal{U}(N)} \operatorname{Tr}(U^\top A)$. The reverse inequality follows from the Von Neumann trace inequality; for O unitary, its singular values are of absolute value 1, and therefore

$$|\operatorname{Tr}(O^\top A)| \leq \sum |\omega_i \alpha_i| = \sum |\alpha_i| = \operatorname{Tr}\left((A^\top A)^{1/2}\right)$$

□

Lemma 2 (Strict Concavity)**.** *For a symmetric positive definite matrix $M \succ 0$, the potential function*

$$J(v) = 2 \operatorname{Tr}\left(\left(D_v^{1/2} M D_v^{1/2}\right)^{1/2}\right) - \sum_{i=1}^{N} v_i$$

is strictly concave on the positive orthant $\mathbb{R}^N_{++}$.

Proof. For any matrices A and B, AB and BA share the same non-zero eigenvalues (this is [HJ90], Theorem 1.3.22). Let $A = D_v^{1/2} M^{1/2}$ and $B = M^{1/2} D_v^{1/2}$.

As AB and BA share eigenvalues, the trace of their square roots is identical:

$$\operatorname{Tr}\left((D_v^{1/2} M D_v^{1/2})^{1/2}\right) = \operatorname{Tr}\left((M^{1/2} D_v M^{1/2})^{1/2}\right)$$

Since the mapping $v \mapsto M^{1/2} D_v M^{1/2}$ is linear in v and injective, Lieb's concavity theorem ([Lie73]) combined with concavity for a composition of an injective linear map with a concave map finish the proof. □

Lemma 3 (Coercivity)**.** *The potential function $J(v)$ is coercive on $\mathbb{R}^N_+$ in the sense that $J(v) \to -\infty$ as $\|v\|_1 \to \infty$.*

Proof. Let $A = D_v^{1/2} M D_v^{1/2}$. Note that $\operatorname{Tr}(A^{1/2})$ is the nuclear norm of $D_v^{1/2} M^{1/2}$. By the relationship between the nuclear norm and the Frobenius norm, $\operatorname{Tr}(A^{1/2}) \leq \sqrt{N \operatorname{Tr}(A)}$. We have:

$$\operatorname{Tr}(A) = \operatorname{Tr}(D_v^{1/2} M D_v^{1/2}) = \operatorname{Tr}(M D_v) = \sum_{i=1}^{N} M_{ii} v_i \leq \max_i (M_{ii}) \|v\|_1$$

Thus, $J(v) \leq 2\sqrt{N \max_i(M_{ii}) \|v\|_1} - \|v\|_1$. As $\|v\|_1 \to \infty$, the linear term dominates the square-root term, hence $J(v) \to -\infty$. This, combined with strict concavity and the fact that $J(v)$ is bounded at the boundary $\partial \mathbb{R}^N_+$, ensures the existence of a unique maximizer $v^* \in \mathbb{R}^N_{++}$. □

Lemma 4 (Positive Invariance)**.** *If $v^{(k)} \in \mathbb{R}^N_{++}$ and $M \succ 0$, then $v^{(k+1)} = \phi(v^{(k)}) \in \mathbb{R}^N_{++}$.*

Proof. Let $X = D_{v^{(k)}}^{1/2} M D_{v^{(k)}}^{1/2}$. Since $v^{(k)} > 0$ element-wise, $D_{v^{(k)}}^{1/2}$ is a non-singular diagonal matrix. Because $M \succ 0$, the congruence transformation X is also symmetric positive definite ($X \succ 0$). The unique positive definite square root $X^{1/2}$ is also strictly positive definite. Since the diagonal entries of any strictly positive definite matrix are strictly positive, $v_i^{(k+1)} = [X^{1/2}]_{ii} > 0$ for all i. □

Counterintuitive as it may seem given its triviality, we will begin by analyzing the scalar case. Surprisingly, using the nuclear norm representation above, we will be able to, in effect, reduce to it.

2.2 The Scalar Case ($N = 1$)

To build intuition, we first analyze the scalar case where $M = 1$ and $D_v = x$. The potential simplifies to $J(x) = 2\sqrt{x} - x$. The update rule is $x_{new} = \sqrt{x}$.

Let $\Delta J = J(x_{new}) - J(x)$. Substituting $x_{new} = x^{1/2}$:

$$\begin{aligned}\Delta J &= (2x^{1/4} - x^{1/2}) - (2x^{1/2} - x) \\ &= 2x^{1/4} - 3x^{1/2} + x\end{aligned}$$

To analyze the sign of ΔJ, we perform a change of variables $x = z^4$ (with $z > 0$). This transforms the expression into a polynomial in z:

$$\Delta J(z) = z^4 - 3z^2 + 2z \tag{4}$$

By inspecting the root at $z = 1$ (the fixed point), we can factorize the polynomial:

$$\Delta J(z) = z(z-1)^2(z+2) \tag{5}$$

Since $z > 0$, we have $z > 0$, $(z-1)^2 \geq 0$, and $(z+2) > 0$. Thus, $\Delta J \geq 0$, with equality if and only if $z = 1$ ($x = 1$). This factorization confirms unconditional monotonicity for the scalar case.

2.3 The Matrix Case

Theorem 1. *For positive-definite M, the algorithm defined by iterating the map ϕ,*

$$\phi(v) = \operatorname{diag}\left(\left(D_v^{1/2} M D_v^{1/2}\right)^{1/2}\right), \tag{6}$$

converges to the unique fixed point of ϕ for any initialization in $\mathbb{R}^n_{++}$.

Proof. In the general matrix case, D_v and M do not commute. Consequently, we cannot apply the scalar polynomial factorization directly to the eigenvalues. Note that the fixed point of ϕ is exactly the maximizer of the potential Eq. (2). To prove convergence of the iterates of ϕ, we employ the **Majorization-Minimization (MM)** principle by constructing a variational surrogate function: we will show that this surrogate function lower bounds the potential J, and iterating ϕ makes strict progress on this surrogate function every step, and thus that the iterates make monotonic progress on the potential. Elementary analysis allows us to conclude that the iterates indeed converge to v^*: given coercivity, since we stay away from the boundary (IE, no $v_i \to 0$ under this iteration, as can be shown simply), we can extract a subsequence which converges to a limit point, but this could be nothing but a fixed point. And monotonic ascent tells us that we cannot leave a neighborhood once we have entered it. Alternatively, we may invoke Zangwill's theorem [Zan69] to conclude.

Thus we proceed with the MM framework.

2.3.1 Step 1: Construction of the Minorizer (Surrogate Function)

Let $A(v) = M^{1/2} D_v^{1/2}$. By Lemma 1, the first term of our potential is:

$$2\operatorname{Tr}\left((D_v^{1/2} M D_v^{1/2})^{1/2}\right) = 2\|M^{1/2} D_v^{1/2}\|_* = \sup_{U \in \mathcal{U}(N)} 2\operatorname{Tr}\left(U^\top M^{1/2} D_v^{1/2}\right) \tag{7}$$

Let $v^{(k)}$ be the current iterate. Define the matrix $A_k = M^{1/2} D_{v^{(k)}}^{1/2}$. We compute the polar decomposition $A_k = U_k P_k$, where $P_k = (A_k^\top A_k)^{1/2}$ is positive semidefinite and U_k is unitary.

Note that strictly U_k aligns the singular vectors; implicitly this solves the variational problem $\sup_U \operatorname{Tr}(U^\top A_k)$.

We define the surrogate function $G(v; v^{(k)})$ by fixing the unitary U_k:

$$G(v; v^{(k)}) = 2 \operatorname{Tr}\left(U_k^\top M^{1/2} D_v^{1/2}\right) - \sum_{i=1}^{N} v_i \tag{8}$$

Properties of the Surrogate:

1. **Lower Bound:** By the variational definition of the nuclear norm, $J(v) \geq G(v; v^{(k)})$ for all v.
2. **Tightness (Value Match):** At $v = v^{(k)}$, the optimal unitary for the variational form is precisely U_k. Thus, $J(v^{(k)}) = G(v^{(k)}; v^{(k)})$.
3. **Tangency (Gradient Match):** We verify that the surrogate shares the same first-order behavior as the objective at $v^{(k)}$.

   The gradient of the original potential $J(v)$ is:

$$\nabla J(v) = \operatorname{diag}\left((D_v^{1/2} M D_v^{1/2})^{1/2}\right) \oslash v - \mathbf{1} \tag{9}$$

   At $v^{(k)}$, note that $(D_{v^{(k)}}^{1/2} M D_{v^{(k)}}^{1/2})^{1/2} = (A_k^\top A_k)^{1/2} = P_k$. Using the polar decomposition $P_k = U_k^\top A_k$:

$$\nabla J(v^{(k)}) = \operatorname{diag}(U_k^\top A_k) \oslash v^{(k)} - \mathbf{1} \tag{10}$$

   Now, consider the gradient of the surrogate $G(v; v^{(k)})$. The trace term is linear in $D_v^{1/2}$:

$$2 \operatorname{Tr}(U_k^\top M^{1/2} D_v^{1/2}) = 2 \sum_i [U_k^\top M^{1/2}]_{ii} \sqrt{v_i} \tag{11}$$

   The gradient with respect to v_i is:

$$\frac{\partial G}{\partial v_i} = \frac{[U_k^\top M^{1/2}]_{ii}}{\sqrt{v_i}} - 1 \tag{12}$$

   Evaluating at $v^{(k)}$, we substitute $[U_k^\top M^{1/2}]_{ii}$ back into matrix form. Note that, as $A_k = M^{1/2} D_{v^{(k)}}^{1/2}$, we have $\operatorname{diag}(U_k^\top A_k)_i = [U_k^\top M^{1/2}]_{ii} \sqrt{v_i^{(k)}}$. Thus:

$$\nabla G(v^{(k)}; v^{(k)}) = \frac{\operatorname{diag}(U_k^\top A_k)}{\sqrt{v^{(k)}} \odot \sqrt{v^{(k)}}} - \mathbf{1} = \nabla J(v^{(k)}) \tag{13}$$

   This confirms $v^{(k)}$ is a valid touch-point for Majorization-Minimization.

2.3.2 Step 2: Monotonic Ascent via the Surrogate

To prove monotonic ascent, we analyze the behavior of the surrogate function constructed in the previous step. Recall that the surrogate, constructed by freezing the unitary U_k from the polar decomposition of $M^{1/2} D_{v^{(k)}}^{1/2}$, is given by:

$$G(v; v^{(k)}) = \sum_{i=1}^{N} (2 c_i \sqrt{v_i} - v_i) \tag{14}$$

where the coefficients are $c_i = \left[U_k^\top M^{1/2}\right]_{ii}$.

The Surrogate Maximizer vs. The Algorithm Update. The global maximizer of this separable surrogate, denoted v^{opt}, is found by setting the derivative to zero:

$$\frac{\partial G}{\partial v_i} = \frac{c_i}{\sqrt{v_i}} - 1 = 0 \implies v_i^{opt} = c_i^2 \tag{15}$$

However, we must verify how this relates to our actual fixed-point update $\phi(v^{(k)})$. The gradient of the true potential $J(v)$ at $v^{(k)}$ is given by $\nabla J(v) = \phi(v) \oslash v - \mathbf{1}$. The gradient of the surrogate $G(v)$ at $v^{(k)}$ is $\nabla G(v) = c \oslash \sqrt{v} - \mathbf{1}$. By the tangency condition above, $\nabla J(v^{(k)}) = \nabla G(v^{(k)}; v^{(k)})$. Equating the terms reveals the explicit form of the update in terms of the surrogate coefficients:

$$\frac{\phi(v^{(k)})_i}{v_i^{(k)}} = \frac{c_i}{\sqrt{v_i^{(k)}}} \implies v_i^{(k+1)} = c_i \sqrt{v_i^{(k)}} \tag{16}$$

Substituting $c_i = \sqrt{v_i^{opt}}$, we observe that the algorithm update is the **element-wise geometric mean** of the current iterate and the surrogate's optimum:

$$v_i^{(k+1)} = \sqrt{v_i^{(k)} \cdot v_i^{opt}} \tag{17}$$

Proof of Ascent. Since the algorithm does not jump strictly to the maximizer v^{opt} but instead takes a damped step, we must show that this step still increases the surrogate value. Let $\Delta G = G(v^{(k+1)}) - G(v^{(k)})$. Since the terms separate, we analyze the scalar contribution for a single index i (omitting the subscript for brevity). Let v be the current value and c be the coefficient. The update is $v_{new} = c\sqrt{v}$.

$$\begin{aligned}
\Delta G_{scalar} &= (2c\sqrt{v_{new}} - v_{new}) - \left(2c\sqrt{v} - v\right) \\
&= \left(2c\sqrt{c\sqrt{v}} - c\sqrt{v}\right) - 2c\sqrt{v} + v \\
&= 2c^{3/2}v^{1/4} - 3cv^{1/2} + v
\end{aligned}$$

And we see the relation to the scalar case above. Again we substitute variables. Let $v = c^2 z^4$ (for $z > 0$). This implies $v^{1/2} = cz^2$ and $v^{1/4} = c^{1/2}z$.

$$\begin{aligned}
\Delta G_{scalar} &= 2c^{3/2}(c^{1/2}z) - 3c(cz^2) + c^2 z^4 \\
&= c^2\left(2z - 3z^2 + z^4\right)
\end{aligned}$$

We must factorize the polynomial $P(z) = z^4 - 3z^2 + 2z$. But note, we already did this, in the scalar case. Thus, referring to Eq. (4), we have

$$\Delta G_{scalar} = c^2 \cdot z \cdot (z-1)^2 \cdot (z+2) \tag{18}$$

Since $c, z > 0$ (where $c > 0$ relies on positive-definiteness of M) and $(z-1)^2 \geq 0$, we have $\Delta G \geq 0$ unconditionally.

Combining these results yields the ascent inequality chain:

$$J(v^{(k+1)}) \geq G(v^{(k+1)}; v^{(k)}) \geq G(v^{(k)}; v^{(k)}) = J(v^{(k)}) \tag{19}$$

This confirms that the update rule guarantees monotonic ascent of the objective function, with strict increase whenever $v^{(k)}$ is not a fixed point (as, by Eq. (16), $z = 1$ again corresponds to the fixed point). □

Remark: This same proof can show linear (exponential) convergence to the objective. But this was already shown asymptotically (IE, close to the fixed point) by [DMR+22], and convergence rates are typically formulated asymptotically.

3 Conclusion

I don't consider the result shown here to be particularly noteworthy; I view it as the closure of a small problem that appeared in one of my past papers. It was something of a shame to me that we did not prove global convergence initially, despite spending a nontrivial amount of time searching for a surrogate function to fit in the MM framework (and some reasonably heavy analysis by Sergey Denisov which showed ϕ was a local contraction around its fixed point). The key fact we were missing was the representation Lemma 1, which, as we saw above, allowed us to effectively diagonalize the problem. The effectiveness of a potential method of "diagonalizing the problem" was quite clear to us; but like most ex-mathematicians working in industry, I would be hard-pressed to recall every matrix equation I have ever seen.

I have a deep an abiding love for mathematics; I entered computing because I believed that it would be a significant source of problems and approaches for the mathematics of the 21st century, in roughly the way physics was for so many years. I remember seeing Jordan Ellenberg at the 2025 joint mathematics meetings, and discussing with him my fear that advances in AI would be used primarily for companionship chatbots rather than progress in fundamental mathematics and science. He rightly pointed out to me the slightly strange priorities my comment implied: that I would rather prove more theorems than bring some comfort to lonely people around the world!

I initially did this work in December of 2025; since then, several other examples of work of this flavor have appeared ([Knu26, Ope26, GLS+26, BEM+26]). I am happy to see that greater minds than my own have found utility in the current models in a roughly similar way to me.

I decided to write up this note, rather than simply let sit the fact that I was able to close a not-entirely-trivial problem primarily through prompting of an AI model, in the hope that I might be able to play some small part in bringing the communities of pure mathematics and AI closer in these, the early days of our journey together.

4 Acknowledgments

Gemini 3 made me aware of the representation Lemma 1, and sketched the major steps of this proof. I filled in some details and fixed some argumentation (including some subtle gaps), but I consider this proof to have been effectively fully generated by Gemini 3. Moreover, Gemini 3 wrote a significant amount of this note; after some rounds of edits, it is difficult to say exactly how much, but its contribution was substantial.

I owe a debt to Abhradeep Thakurta and Brendan McMahan, with whom I worked on formulating the problem that eventually led here. Certainly also our other co-authors on that first paper, Adam Smith and my advisor, Sergey Denisov.

Finally, thank you to my wife Breanne Lynch, who has provided me with unfailing support, even if I can't quite articulate *why* I need to be typing away at odd hours.

References

[BEM+26] Jim Bryan, Balázs Elek, Freddie Manners, George Salafatinos, and Ravi Vakil. The motivic class of the space of genus 0 maps to the flag variety. *arXiv preprint arXiv:2601.07222*, January 2026. In collaboration with the Google DeepMind Blueshift team, including Adam Brown and Vinay Ramasesh.

[DMP+24] Krishnamurthy Dj Dvijotham, H. Brendan McMahan, Krishna Pillutla, Thomas Steinke, and Abhradeep Thakurta. Efficient and Near-Optimal Noise Generation

for Streaming Differential Privacy . In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2306–2317, Los Alamitos, CA, USA, October 2024. IEEE Computer Society.

[DMR+22] Sergey Denisov, H. Brendan McMahan, John Rush, Adam Smith, and Abhradeep Guha Thakurta. Improved differential privacy for sgd via optimal private linear operators on adaptive streams. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5910–5924. Curran Associates, Inc., 2022.

[GLS+26] Alfredo Guevara, Alexandru Lupsasca, David Skinner, Andrew Strominger, and Kevin Weil. Single-minus gluon tree amplitudes are nonzero. *arXiv preprint arXiv:2602.12176*, February 2026. On behalf of OpenAI.

[GP12] F. E. Guerra-Pujol. Gödel's loophole. *SSRN Electronic Journal*, 2012.

[HJ90] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[Knu26] Donald E. Knuth. Claude's cycles. Unpublication available at `https://www-cs-faculty.stanford.edu/~knuth/papers/claude-cycles.pdf`, February 2026. Revised March 16, 2026.

[Lie73] Elliott H Lieb. Convex trace functions and the wigner-yanase-dyson conjecture. *Advances in Mathematics*, 11(3):267–288, 1973.

[LMH+15] Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *VLDB J.*, 24(6):757–781, 2015.

[Ope26] OpenAI. Our first proof submissions. OpenAI Blog, February 2026. Published February 20, 2026.

[PUCC+25] Krishna Pillutla, Jalaj Upadhyay, Christopher A. Choquette-Choo, Krishnamurthy Dvijotham, Arun Ganesh, Monika Henzinger, Jonathan Katz, Ryan McKenna, H. Brendan McMahan, Keith Rush, Thomas Steinke, and Abhradeep Thakurta. Correlated noise mechanisms for differentially private learning, 2025.

[Zan69] W.I. Zangwill. *Nonlinear Programming: A Unified Approach*. Prentice-Hall international series in management. Prentice-Hall, 1969.

A A User's Journey

Given my background, I am relatively uniquely positioned to effectively interact with current AI systems and produce effective mathematics. At present, it is a nontrivial procedure, and it is not entirely surprising that, with a different mental model of contemporary AI systems, interactions may go nowhere useful. In this section, I will attempt to extract the relevant aspects of my experience resulting in the proof above, and highlight some of the ways the process could have gone wrong.

First, it must be stated that the models that currently exist are generally *incredibly credulous*. Mathematics is a process which requires surfacing all relevant reasoning steps. As much as formal methods indicate that we may be somewhat deceiving ourselves about the tightness of

our reasoning, mathematics is, in my experience, the domain with the highest bar of reasoning in human activity.

Thus, keeping in mind the view of LLMs as a "steerable probability distribution over all human text" (which is probably correct to first order), it is no surprise that models will tend, from the mathematician's perspective, to leap to conclusions. These leaps, like our leaps, are often incorrect. Currently, models don't demonstrate significant ability to spot their own gaps and critique their own leaps, though I expect this to change in the near future. This leads to the first pattern of effective use that I have found:

Principle 1: Push hard on gaps. Models will often leap over the requirements of strict reasoning. Mathematicians know this is common practice in human interaction, and we must work quite hard to achieve the state of mind that allows for precise reasoning. Mathematicians, then, are often excellent (though far from perfect!) at "sniffing out" logical gaps. (In an infamous story, the great logician Kurt Gödel was said to have found an "internal contradiction" in the US Constitution which permits the US to legally convert itself from a republic to a dictatorship [GP12]. Luckily the judge administering the citizenship examination to Gödel declined to pursue the matter.)

Interacting with an AI with the goal of proving mathematics often stretches this logical-gap-sniffing capacity to the limit – there are so many! I will give one example that I found in the log of my interaction with Gemini that resulted in this note. After I pointed to [DMR+22] and asked for a proof of the convergence of iterates of ϕ, Gemini responded with:

Gemini: ... Scalar Analogy: Consider the 1D case $J(x) = 2\sqrt{ax} - x$. The iteration is $x_{new} = \sqrt{ax}$. Let $u = (x/a)^{1/4}$. The gain $\Delta J = J(x_{new}) - J(x)$ factors as:

$$\Delta J \propto u(u-1)^2(u+2)$$

This is strictly positive for all $u \neq 1$. The matrix case satisfies a similar inequality.

Clearly, this is insufficient. (Though, as we have seen above, indeed correct). Thus I asked:

Keith: It seems like the crux of your argument is step 3, particularly, "the matrix case satisfies a similar inequality". please cut no corners and carefully show step 3. then: is step 3 a well-known fact? if so, please give a reference.

Unfortunately, even to this I was left with a "matrix case satisfies a similar inequality" response. Gemini even admitted it could not prove this to me, and suggested I try numerics. At this point I opened a new chat, pasting in the argument Gemini gave me as context, asking:

Keith: It seems like the crux of the argument to follow is step 3, particularly, "the matrix case satisfies a similar inequality". please cut no corners and carefully show step 3. then: is step 3 a well-known fact? if so, please give a reference.

Here we finally, in hindsight, started to get closer. But much reasoning remained suppressed:

Gemini: ... Summary: The scalar proof $(z(z-1)^2(z+2) \geq 0)$ provides the intuition. The matrix case holds because the Trace acts on the singular values, and the non-commutativity terms are handled by the unitary invariance of the Nuclear Norm, effectively "diagonalizing" the problem into the scalar case at the optimal alignment.

By now we see implicit references to the proof above, but it was quite difficult in the moment to disambiguate a correct hand-waving around unitary invariance from a pure hallucination.

Keith: so, you haven't given me a proof, right? Please try harder.

This sequence of interactions yielded the representation of the nuclear norm we have in this note, and the basic outline of an MM argument. With a few more iterations and fixing a few

more gaps (initially, Gemini claimed that the iterations of ϕ jumped directly to the maximum of the surrogate; upon my attempt to rewrite the logic, I discovered this was incorrect), we had the result you see above. But the interaction above contained a crucial easy-to-miss step, which can serve as our second principle:

Principle 2: Beware of your context. Transformer-based LLMs typically attend to their context (in this case, the chat history) on every turn. This means that ideas present in the context can in some sense "pollute" new outputs, even if you are hoping to steer the model away from these ideas. Remember, an LLM is in "just" a steerable distribution over text, with the context serving as the conditioning variables for the generation process. Thus, the model is generating for you text that is likely to follow the context, presumably modulated by some posttraining to steer it towards free-flowing interactions. (I do not work in posttraining, so I do not have intimate familiarity with such details). But for this reason, things in the context can be very hard to make the model ignore. If it has responded to you with an answer you don't like multiple times, you may be better off manually curating the context by starting a new chat, and feeding in your summary of the things you want the model to know, filtering out the things you wish for it to avoid.

One aspect of this interaction which I suppressed above was the generation of colab snippets to check that we could not generate counterexamples to the claimed monotonic ascent. The validity of these snippets (plus my recollection that this was empirically true during the writing of [DMR+22]) served as a critical grounding in continuing to push the model towards a proof. Here Gemini was reasonably inventive, numerically attempting to *minimize the gain* in the vector v. Gemini generated the colab to follow, which I then altered slightly while running, to ensure the claimed ascent was not simply a low-dimensional phenomenon.

```
import numpy as np
from scipy.linalg import eigh
from scipy.optimize import minimize
import matplotlib.pyplot as plt

def get_sym_sqrt(A):
    vals, vecs = eigh(A)
    # Numerical stability: Clip small eigenvalues
    vals = np.maximum(vals, 1e-14)
    return vecs @ np.diag(np.sqrt(vals)) @ vecs.T

def potential_J(v, M):
    v_sqrt = np.sqrt(v)
    D_sqrt = np.diag(v_sqrt)
    Inner = D_sqrt @ M @ D_sqrt

    # Tr(A^{1/2}) is sum(sqrt(eigenvalues(A)))
    vals = np.linalg.eigvalsh(Inner)
    vals = np.maximum(vals, 1e-14)

    return 2 * np.sum(np.sqrt(vals)) - np.sum(v)

def phi_map(v, M):
    v_sqrt = np.sqrt(v)
    D_sqrt = np.diag(v_sqrt)
    Inner = D_sqrt @ M @ D_sqrt
    S = get_sym_sqrt(Inner)
    return np.diag(S)

def run_adversarial_search(N=3):

    # Parameterize M = L L^T (to ensure PD) and v = exp(u) (to ensure v > 0)
```

```
    # Params: [u (size N), L_flat (size N*N)]

    def pack(v, L):
        return np.concatenate([np.log(v), L.flatten()])

    def unpack(params):
        u = params[:N]
        L_flat = params[N:]
        v = np.exp(u)
        L = L_flat.reshape(N, N)
        M = L @ L.T + 1e-5 * np.eye(N)
        return v, M

    def objective(params):
        v, M = unpack(params)
        v_next = phi_map(v, M)

        # We want to minimize gain
        gain = potential_J(v_next, M) - potential_J(v, M)
        return gain

    # Initial random guess
    A_init = np.random.randn(N, N)
    v_init = np.random.rand(N) + 0.1
    x0 = pack(v_init, A_init)

    # Run Optimization (L-BFGS-B handles smooth unconstrained problems well)
    res = minimize(objective, x0, method='L-BFGS-B',tol=1e-9)

    min_gain = res.fun

    if min_gain < -1e-8:
        raise ValueError(min_gain)
    else:
        pass

# Run the search

# Test on a few dimensions
run_adversarial_search(N=2)
run_adversarial_search(N=3)
run_adversarial_search(N=5)
```

This colab verification pattern I have found useful in several other mathematics-heavy contexts when working with AI models; e.g., if the model claims to prove convexity of a certain function, the quickest route to falsifying this claim is often to simply ask the model to generate a code snippet which tests this claim numerically. This leads to our third and final principle:

Principle 3: Ask for empirical verification of theory. If this test fails, it is often enough to feed this evidence back to the model to encourage it to revise its approach. Given the tendency of models to either hallucinate or suppress reasoning steps, quick empirical verification of claimed theoretical statements has saved me much time in the past.

I have hope that other mathematicians and scientists can save themselves time and frustration, and achieve a fair sense of what the current models are capable of, by following and expanding on these three principles.

B On the subtleties of checking correctness

In an earlier version of this note, before posting publicly, there was an incorrect "proof" of Lemma 2 which slipped past me.

It is reproduced below.

"Proof". Let $\Psi(v) = D_v^{1/2} M D_v^{1/2}$ and note that $\Psi(v) = (\sqrt{v}\sqrt{v}^\top) \odot M$. Since the square root function is strictly concave on $\mathbb{R}_{++}$, the map $v \mapsto \sqrt{v}\sqrt{v}^\top$ is strictly operator concave. Because $M \succ 0$, the Schur Product Theorem implies that $\Psi(v)$ is also strictly operator concave.

Finally, since the trace-square-root functional $X \mapsto \mathrm{Tr}(X^{1/2})$ is operator monotone and strictly concave on the positive definite cone, the composition $J(v)$ is strictly concave. The linear term $-\sum v_i$ has a zero Hessian, thus the strict concavity of the first term yields $\nabla^2 J(v) \prec 0$ for all $v \in \mathbb{R}^N_{++}$. □

The fundamental flow with this argument is in the claim that $v \mapsto \sqrt{v}\sqrt{v}^\top$ is operator concave. A simple example choosing $v = (1, 4)$ and $w = (4, 1)$ is enough to contradict this claim. However, I did not spot it. Almost certainly I had asked for some kind of numerical verification of the claim of Lemma 2, similar to the snippet above, and this led me to be a bit too trusting in the details of the argument.

Later on, I caught a subtle mismatch between the definition of A used in Section 2.3.1 and that which corresponds to the actual literal potential function. (The model used $D_v^{1/2} M^{1/2}$ instead of $M^{1/2} D_v^{1/2}$). Technically this was only a missed logical step, and not incorrect, because of the equivalence of the eigenvalues of AB and BA as noted in the (new) proof of Lemma 2, but it resulted in some strange gymnastics (e.g. the use of the left polar factorization instead of the more traditional right polar factorization). I rewrote the argument in the current form you find it, which I find much more straightforward.

In some ways, each of these is an effective illustration of our principle 1 above. But it also highlights a real difficulty in using AI to accelerate mathematics in this way, at least, without coupling with formal methods. On the other hand, it is a belief of mine (shared by at least one other mathematician I know) that every math paper has gaps; the only question that counts is whether or not those gaps matter. Neither of these gaps matter. And indeed, once Gemini presented me with the variational characterization of the nuclear norm Lemma 1, I knew the problem was solved, and the rest was details.

Time will tell, I think, whether the *ease* of producing a gap with AI-assisted mathematics will proliferate their relative occurrence, and whether the human checking will be overwhelmed by the volume of argumentation and its apparent simplicity, or whether this is a phantom concern that will either be solved by better models or improved ways of working.