Empirical Investigation of an Open Conjecture: Can Single-Shuffle SGD be Better than Reshuffling SGD and GD?

Agentic NL→Lean 4 Pipeline
Job #18

April 26, 2026

Abstract

This report documents the empirical investigation of an open mathematical conjecture that could not be formally proved or disproved in Lean 4 with Mathlib. Numerical experiments were conducted to gather evidence for or against the conjecture. The empirical verdict is: **Empirically Supported**. The conjecture remains formally open.

1 Conjecture Statement

Conjecture 1.

```
Can Single-Shuffle SGD be Better than Reshuffling SGD and GD?

Let n >= 2, K >= 1, and d >= 1. Let S_n denote the set of all permutations
    of {1,...,n}. For real symmetric d x d matrices X, Y, write X   Y for the
    Loewner order (i.e., Y - X is positive semidefinite), and let ||·||_2 be
    the spectral (operator) norm.

Given symmetric matrices A_1, ..., A_n, define for each    S_n the without-
    replacement product P_ := _{i=1}^{n} A_ {(i)} = A_ {(n)}   A_ {(1)}.

Define:
  W_SS := (1/n!) Σ_ {   S_n} ( P_ )^K        (single-shuffle)
  W_RS := ((1/n!) Σ_ {   S_n} P_ )^K        (random reshuffling)
  W_GD := ((1/n) Σ_{i=1}^{n} A_i)^{nK}       (gradient descent)

CONJECTURE: For every n >= 2 and K >= 1 there exists a constant _{n,K}
    (0,1] (depending only on n and K, not on d or on the specific matrices)
    such that, whenever (1 - _{n,K})I   A_i   I for all i   {1,...,n}, one has
    :

  ||W_SS||_2   ||W_RS||_2   ||W_GD||_2

Equivalently: under uniform near-identity well-conditioning, single-shuffle
    never yields a larger spectral norm than random reshuffling, and random
    reshuffling is never worse than the full-gradient proxy W_GD, with
    constants uniform over dimension and instances.
```

2 Status

Formal Status: OPEN — no Lean 4 proof or disproof was found.

Empirical Verdict: Empirically Supported

The pipeline attempted formal verification in Lean 4 with Mathlib but was unable to produce a compiling proof or disproof. Empirical testing was then conducted to gather numerical evidence.

3 Basic Empirical Testing

The following output was produced by the basic numerical experiment:

```
=== EXPERIMENT PLAN ===

Conjecture (COLT 2021 open problem): For every n >= 2, K >= 1, there exists
eta_{n,K} in (0,1] (independent of d and of the specific symmetric matrices)
such that whenever (1-eta)I <= A_i <= I (Loewner) for all i in {1,...,n},
the spectral-norm inequality
        ||W_SS||_2  <=  ||W_RS||_2  <=  ||W_GD||_2
holds, where
  W_SS = (1/n!) sum_sigma (A_{sigma(n)} ... A_{sigma(1)})^K
  W_RS = ((1/n!) sum_sigma  A_{sigma(n)} ... A_{sigma(1)})^K
  W_GD = ((1/n) sum_i A_i)^{nK}.

Experimental strategy (multiple angles):

(1) Near-identity random sampling (the regime where the conjecture claims to
    hold):
    Draw A_i = I - eta * H_i where H_i is a random symmetric matrix with
    spectrum in [0,1]. Vary (n,K,d,eta) and estimate the violation rate.

(2) Sweep of eta: for fixed (n,K,d), vary eta from tiny (1e-3) to 1 and
    record the proportion of violations of each of the two inequalities.
    If the conjecture holds, we should see a regime (small eta) where the
    violation rate is 0 uniformly in d.

(3) Dimension scaling: for a chosen small eta, scale d from 2 to ~50 to
    confirm that "no violations" persists as d grows (a necessary condition
    since eta must not depend on d).

(4) Edge / structured cases: commuting matrices (simultaneously diagonal),
    rank-one perturbations of I, 2x2 closed-form families.

(5) Adversarial counterexample search: a random-restart coordinate search
    which tries to maximize the violation (||W_SS|| - ||W_RS||) and
    (||W_RS|| - ||W_GD||) inside the feasible box (1-eta)I <= A_i <= I.

(6) Asymptotic / scaling behaviour: study the gap ||W_GD|| - ||W_RS||
    and ||W_RS|| - ||W_SS|| as n, K grow (for n up to 5 to keep n! tractable
    ).

We report per-(n,K,eta) violation counts, maximum observed violations,
and confidence-via-counterexample: if any reproducible counterexample
```

```
survives verification to high precision in the small-eta regime, we
refute; otherwise we report support.


--- (1/2) Random sampling sweep over eta ---
 n  K     eta  v(RS>GD)  v(SS>RS) Δ   maxRS-GD Δ   maxSS-RS
 2  1   0.001         0         0    0.000e+00    0.000e+00
 2  1    0.01         0         0    0.000e+00    0.000e+00
 2  1    0.05         0         0    0.000e+00    0.000e+00
 2  1     0.1         0         0    0.000e+00    0.000e+00
 2  1    0.25         0         0    0.000e+00    0.000e+00
 2  1     0.5         0         0    0.000e+00    0.000e+00
 2  1    0.75         0         0    0.000e+00    0.000e+00
 2  1       1         0         0    0.000e+00    0.000e+00
 2  2   0.001         0         0    0.000e+00    0.000e+00
 2  2    0.01         0         0    0.000e+00    0.000e+00
 2  2    0.05         0         0    0.000e+00    0.000e+00
 2  2     0.1         0         0    0.000e+00    0.000e+00
 2  2    0.25         0         0    0.000e+00    0.000e+00
 2  2     0.5         0         0    0.000e+00    0.000e+00
 2  2    0.75         0         0    0.000e+00    0.000e+00
 2  2       1         0         0    0.000e+00    0.000e+00
 2  3   0.001         0         0    0.000e+00    0.000e+00
 2  3    0.01         0         0    0.000e+00    0.000e+00
 2  3    0.05         0         0    0.000e+00    0.000e+00
 2  3     0.1         0         0    0.000e+00    0.000e+00
 2  3    0.25         0         0    0.000e+00    0.000e+00
 2  3     0.5         0         0    0.000e+00    0.000e+00
 2  3    0.75         0         0    0.000e+00    0.000e+00
 2  3       1         0         0    0.000e+00    0.000e+00
 3  1   0.001         0         0    0.000e+00    0.000e+00
 3  1    0.01         0         0    0.000e+00    0.000e+00
 3  1    0.05         0         0    0.000e+00    0.000e+00
 3  1     0.1         0         0    0.000e+00    0.000e+00
 3  1    0.25         0         0    0.000e+00    0.000e+00
 3  1     0.5         0         0    0.000e+00    0.000e+00
 3  1    0.75         0         0    0.000e+00    0.000e+00
 3  1       1         0         0    0.000e+00    0.000e+00
 3  2
... [truncated]
```

4 Experiment Code (Basic)

```python
import itertools
import math
import random
import time
import numpy as np
import matplotlib
matplotlib.use("Agg")
import matplotlib.pyplot as plt
```

```
np.random.seed(20260421)
random.seed(20260421)

print("===␣EXPERIMENT␣PLAN␣===")
print("""
Conjecture (COLT 2021 open problem): For every n >= 2, K >= 1, there exists
eta_{n,K} in (0,1] (independent of d and of the specific symmetric matrices)
such that whenever (1-eta)I <= A_i <= I (Loewner) for all i in {1,...,n},
the spectral-norm inequality
        ||W_SS||_2  <=  ||W_RS||_2  <=  ||W_GD||_2
holds, where
  W_SS = (1/n!) sum_sigma (A_{sigma(n)} ... A_{sigma(1)})^K
  W_RS = ((1/n!) sum_sigma  A_{sigma(n)} ... A_{sigma(1)})^K
  W_GD = ((1/n) sum_i A_i)^{nK}.

Experimental strategy (multiple angles):

(1) Near-identity random sampling (the regime where the conjecture claims to
    hold):
    Draw A_i = I - eta * H_i where H_i is a random symmetric matrix with
    spectrum in [0,1]. Vary (n,K,d,eta) and estimate the violation rate.

(2) Sweep of eta: for fixed (n,K,d), vary eta from tiny (1e-3) to 1 and
    record the proportion of violations of each of the two inequalities.
    If the conjecture holds, we should see a regime (small eta) where the
    violation rate is 0 uniformly in d.

(3) Dimension scaling: for a chosen small eta, scale d from 2 to ~50 to
    confirm that "no violations" persists as d grows (a necessary condition
    since eta must not depend on d).

(4) Edge / structured cases: commuting matrices (simultaneously diagonal),
    rank-one perturbations of I, 2x2 closed-form families.

(5) Adversarial counterexample search: a random-restart coordinate search
    which tries to maximize the violation (||W_SS|| - ||W_RS||) and
    (||W_RS|| - ||W_GD||) inside the feasible box (1-eta)I <= A_i <= I.

(6) Asymptotic / scaling behaviour: study the gap ||W_GD|| - ||W_RS||
    and ||W_RS|| - ||W_SS|| as n, K grow (for n up to 5 to keep n! tractable
        ).

We report per-(n,K,eta) violation counts, maximum observed violations,
and confidence-via-counterexample: if any reproducible counterexample
survives verification to high precision in the small-eta regime, we
refute; otherwise we report support.
""")

t_start = time.time()

# ---------- Helpers ----------

def random_sym_in_box(d, eta, rng):
```

```python
    """Random symmetric matrix with spectrum in [1-eta, 1]."""
    Q, _ = np.linalg.qr(rng.standard_normal((d, d)))
    lam = 1.0 - eta * rng.random(d)          # uniform in [1-eta, 1]
    return (Q * lam) @ Q.T

def random_sym_in_box_extreme(d, eta, rng):
    """Eigenvalues at the box extremes 1-eta or 1 (harder cases)."""
    Q, _ = np.linalg.qr(rng.standard_normal((d, d)))
    lam = np.where(rng.random(d) < 0.5, 1 - eta, 1.0)
    return (Q * lam) @ Q.T

def compute_W(A_list, K):
    """Return (W_SS, W_RS, W_GD) for a list of n symmetric matrices."""
    n = len(A_list)
    d = A_list[0].shape[0]
    perms = list(itertools.permutations(range(n)))
    # Pre-compute all P_sigma
    P_sigmas = []
    for sigma in perms:
        P = np.eye(d)
        for idx in sigma:            # multiply A_{sigma(1)} first, on the right
            P = A_list[idx] @ P      # final P = A_{sigma(n)} ... A_{sigma(1)}
        P_sigmas.append(P)
    mean_P = np.mean(P_sigmas, axis=0)
    # W_SS: mean of P_sigma^K
    W_SS = np.mean([np.linalg.matrix_power(P, K) for P in P_sigmas], axis=0)
    # W_RS:
    W_RS = np.linalg.matrix_power(mean_P, K)
    # W_GD:
    mean_A = np.mean(A_list, axis=0)
    W_GD = np.linalg.matrix_power(mean_A, n * K)
    return W_SS, W_RS, W_GD

def spectral(M):
    # For symmetric-ish matrices use eigvalsh on symmetric part would be unsafe
    # because W_SS/W_RS/W_GD are symmetric (products of symmetric commuting? no)
    # Actually W_GD is symmetric (power of symmetric), W_SS is symmetric
    # (average of P_sigma^K plus its transpose since sum over sigma pairs
    # sigma and reverse gives conjugate? -- not necessarily symmetric).
    # Safer to use general spectral norm (largest singular value).
    return np.linalg.svd(M, compute_uv=False)[0]

def trial(n, K, d, eta, rng, extreme=False):
    if extreme:
        A_list = [random_sym_in_box_extreme(d, eta, rng) for _ in range(n)]
    else:
        A_list = [random_sym_in_box(d, eta, rng) for _ in range(n)]
    W_SS, W_RS, W_GD = compute_W(A_list, K)
    s_ss, s_rs, s_gd = spectral(W_SS), spectral(W_RS), spectral(W_GD)
    return s_ss, s_rs, s_gd, A_list
```

```
# ---------- (1)(2) Random sampling sweep over eta ----------

print("\n--- (1/2) Random sampling sweep over eta ---")

configs = [(2,1), (2,2), (2,3), (3,1), (3,2), (4,1), (5,1), (3,3)]
etas = [1e-3, 1e-2, 5e-2, 0.1, 0.25, 0.5, 0.75, 1.0]
d_default = 4
trials_per_cell = 300  # per config & eta

summary_rows = []
grid_violation_rs_le_gd = np.zeros((len(configs), len(etas)))
grid_violation_ss_le_rs = np.zeros((len(configs), len(etas)))
worst_violation_overall = {"rs_gd": (0.0, None), "ss_rs": (0.0, None)}

rng = np.random.default_rng(1234)

for i, (n, K) in enumerate(confi
# ... [truncated]
```

5 Conclusion

The conjecture remains formally open. Numerical experiments **support** the conjecture — no counterexamples were found across all tested parameter ranges. Further investigation (both formal and empirical) is warranted.