Preliminary experiements for Streaming SVD

Ahmer Nadeem Khan

Florida State University

March 4, 2026

Outline

Outline

Cold-Start rSVD Algorithm

Algorithm Randomized SVD (Halko et al.)

Require: $A \in \mathbb{R}^{m \times n}$, rank k, oversampling p, iterations q
Ensure: $U \in \mathbb{R}^{m \times k}$, $\mathbf{s} \in \mathbb{R}^{k}$, $V^T \in \mathbb{R}^{k \times n}$

Draw $\Omega \in \mathbb{R}^{n \times (k+p)} \sim \mathcal{N}(0, 1)$
$Y \leftarrow A\Omega$
for $i = 1$ to q **do**
    $Y \leftarrow A(A^T Y)$
end for
$Q, \_ \leftarrow \mathrm{QR}(Y)$
$B \leftarrow Q^T A$
$U_{\mathrm{hat}}, \mathbf{s}, V^T \leftarrow \mathrm{SVD}(B)$
$U \leftarrow QU_{\mathrm{hat}}$
return $U_{:,1:k}, \mathbf{s}_{1:k}, V^T_{1:k,:}$

Outline

Warm-Started rSVD Algorithm

Algorithm Warm-Start Randomized SVD

Require: $A \in \mathbb{R}^{m \times n}$, previous basis U_{prev}, rank k, oversampling p, iterations q

Ensure: $U \in \mathbb{R}^{m \times k}$, $\mathbf{s} \in \mathbb{R}^{k}$, $V^T \in \mathbb{R}^{k \times n}$

$G \leftarrow A^T U_{\text{prev}}$ ▷ Warm-start projection

$Y_1 \leftarrow AG$ ▷ Prior subspace component

Draw $\Omega \in \mathbb{R}^{n \times p} \sim \mathcal{N}(0, 1)$

$Y_2 \leftarrow A\Omega$ ▷ Exploration component

$Y \leftarrow [Y_1 \mid Y_2]$ ▷ Concatenate

for $i = 1$ to q **do**

  $Y \leftarrow A(A^T Y)$

end for

$Q, \_ \leftarrow \text{QR}(Y)$

$B \leftarrow Q^T A$

$U_{\text{hat}}, \mathbf{s}, V^T \leftarrow \text{SVD}(B)$

$U \leftarrow QU_{\text{hat}}$

return $U_{:,1:k}, \mathbf{s}_{1:k}, V^T_{1:k,:}$

Outline

Initial Matrix Generation

- Generate low-rank matrix: $S = U\text{diag}(\mathbf{s})V^T$
  - $U \in \mathbb{R}^{m \times r}$: random orthonormal basis (QR)
  - $V \in \mathbb{R}^{n \times r}$: random orthonormal basis (QR)
  - Singular values with exponential decay: $s_i = \exp(-\text{decay} \cdot i)$

- Parameters:
  - Dimensions: m, n (rows, columns)
  - True rank: r
  - Decay rate: decay (controls spectral concentration)

Matrix Perturbation Process

- At each time step: $S^{(t+1)} = S^{(t)} + E^{(t)}$
- Perturbation magnitude controlled by: $\|E\|_F = \eta \cdot \|S^{(t)}\|_F$
- Noise structure options:
  - Full-rank noise: $E \in \mathbb{R}^{m \times n}$ (general perturbation)
  - Low-rank noise: $E = U_E V_E^T$ (structured perturbation)
- Allows systematic evaluation of algorithm robustness under different perturbation regimes

Outline

Experiment Parameters

Parameter	Value
Matrix size (m $\times$ n)	1000 $\times$ 1000
Target rank (k)	20
Time steps (T)	10
Perturbation magnitude (η)	0.05
Cold-start oversampling (p_{cold})	10
Warm-start oversampling (p_{warm})	5
Power iterations (q)	0
Random seed	42

Metrics

- **Relative Frobenius error**

$$\frac{\|A - U\text{diag}(\mathbf{s})V^T\|_F}{\|A\|_F}$$

- **Relative spectral error estimate**

$$\frac{\|(I - UU^T)A\|_2}{\|A\|_2}$$

estimated in code by power iteration (`rel_spec_error_est`).

Metrics

- **Subspace distance** (principal-angle based)

$$\sin\theta(U_1, U_2) = \sqrt{1 - \sigma_{\min}(U_1^T U_2)^2}$$

and a Frobenius-angle variant for aggregate mismatch.

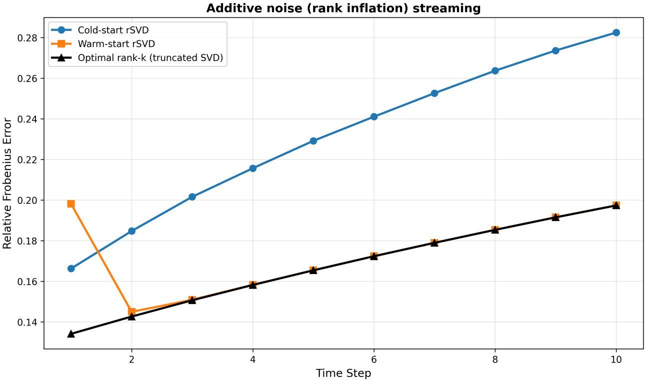
- **Optimal rank-k baseline**

$$\frac{\|A - A_k\|_F}{\|A\|_F} = \sqrt{\frac{\sum_{i>k} \sigma_i^2}{\sum_i \sigma_i^2}}$$

computed via singular values (when `compute_optimal=True`).

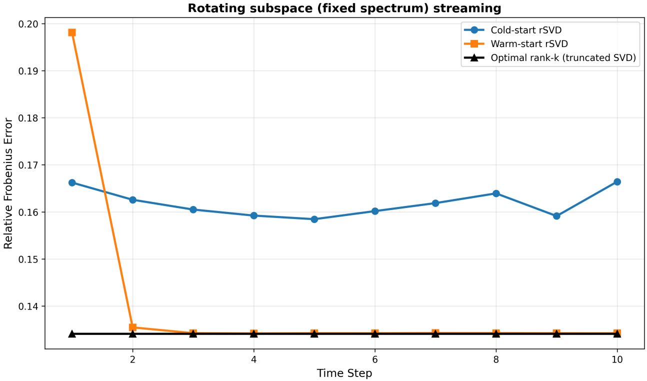
Experiment 1: Additive Perturbations

- $S_t = S_{t-1} + E_t$, with $\|E_t\|_F = \eta \|S_{t-1}\|_F$.
- Consequence: energy beyond rank k accumulates, so even the **optimal rank-**k error increases over time.
- Cold- and warm-start are compared against the best possible truncated-SVD baseline.

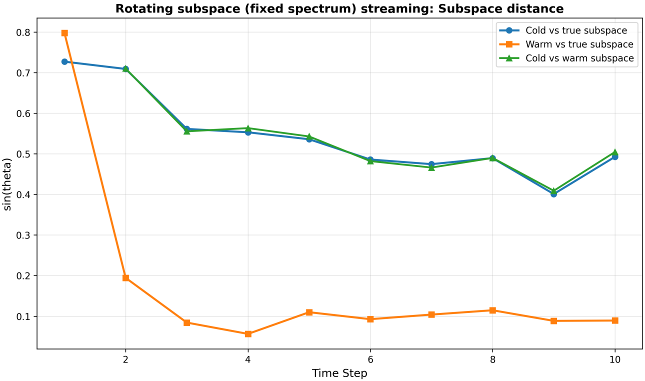


Experiment 2: Rotating Subspace

- $S_t = U_t \,\mathrm{diag}(\mathbf{s})\, V_t^T$ with fixed $\mathbf{s}$; U_t and V_t rotate slowly via small orthogonal updates.
- Consequence: **optimal rank-**k error remains approximately constant; method performance reflects tracking ability rather than rank inflation.



Rotating Regime: Subspace Tracking



Key Findings

- **Warm-start improves approximation quality** across the stream
  - Relative Frobenius error remains consistently below cold-start
  - Relative spectral error is also lower over timesteps
- **Lower approximation error:** Warm-start achieves 24.6% better error on average
  - Leverages prior subspace information
  - Better exploration of perturbation direction
- **Consistent behavior:** Error ratio remains stable across 10 time steps (0.754)
  - Demonstrates robustness to streaming perturbations
  - Prior basis remains informative over time

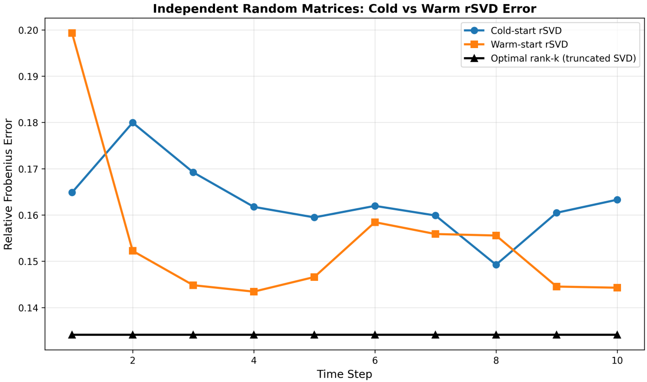
Summary Results

Reported quantities: all error values are averages over timesteps of relative Frobenius error; improvement is the warm-vs-cold relative reduction with warm/cold ratio in parentheses.

Regime	Cold	Warm	Optimal
Additive noise streaming	0.2310	0.1743	0.1676
Rotating subspace streaming	0.1619	0.1408	0.1341
Additive warm improvement	**24.6% (ratio 0.771)**		
Rotating warm improvement	**13.0% (ratio 0.869)**		

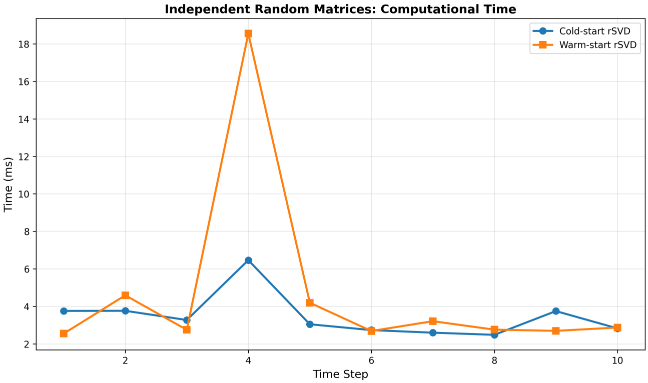
Outline

Independent Random Matrices: Error Comparison



- Matrices at each timestep are generated independently.
- This removes temporal correlation between snapshots.
- Warm-start rSVD should therefore provide little or no advantage.

Independent Random Matrices: Runtime



- Cold and warm algorithms perform similar work.
- Any improvement seen in correlated experiments should disappear here.

Interpretation

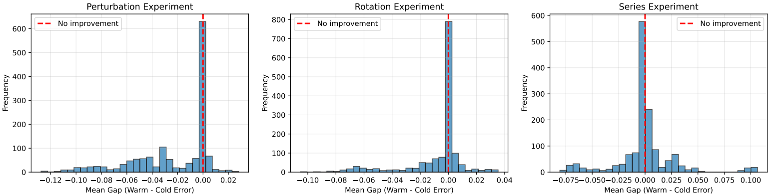
- In the independent matrix setting there is no shared structure between timesteps.
- The previous singular subspace contains no useful information for the next snapshot.
- As expected, warm-start rSVD performs similarly to cold-start rSVD.
- This experiment serves as a control validating the warm-start hypothesis.

Outline

Sweep Setup (Completed Run)

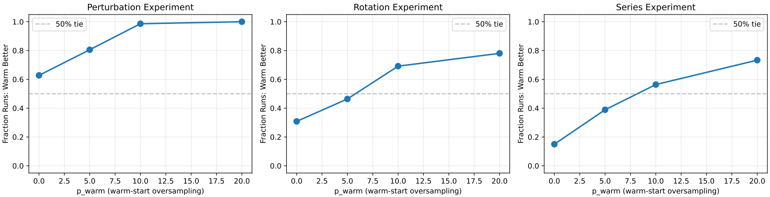
- Full grid executed across **series**, **perturbation**, and **rotation** regimes.
- Total runs: **4320** (all combinations of $m, n, k, p_{\text{cold}}, p_{\text{warm}}, q$, with 5 seeds).
- Metrics shown are error-only: mean gap (warm − cold), mean ratio (warm/cold), and fraction warm better.
- CSV outputs: `results/sweep_raw.csv` and `results/sweep_summary.csv`.

Sweep Results: Error-Gap Distribution



- Negative mean gap indicates warm-start has lower error than cold-start.
- Perturbation and rotation show broader mass on the negative side than the independent-series control.

Sweep Results: Warm Advantage vs p_{warm}



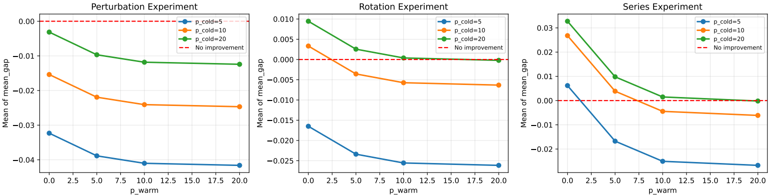
- Increasing warm exploration generally improves consistency of warm-start wins.
- Effect is strongest in perturbation and more moderate in rotation; control series is weakest overall.

Sweep Summary Across All Configurations

Regime	**Avg gap**	**Avg ratio**	**Warm-win frac**
Perturbation	-0.0231	0.9046	0.8549
Rotation	-0.0076	0.9370	0.5611
Series (control)	0.0001	0.9870	0.4590

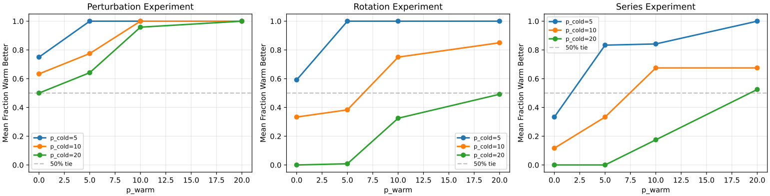
- Warm-start gives the clearest average error improvement in **perturbation**.
- **Series** is near parity, consistent with the no-temporal-correlation control.

Sampling Dimensions: Effect on Mean Gap



- Lines show mean gap versus p_{warm}, grouped by p_{cold}.
- More negative values are better for warm-start.

Sampling Dimensions: Effect on Warm-Win Frequency



- Warm-win frequency rises with larger p_{warm} in all three regimes.
- Lower p_{cold} generally favors warm-start consistency.

Largest Parameter Effects (from Sweep CSV)

Parameter	Range of global avg mean-gap
p_{cold}	0.0272
q	0.0225
p_{warm}	0.0173
k	0.0109
m	0.0004
n	0.0003

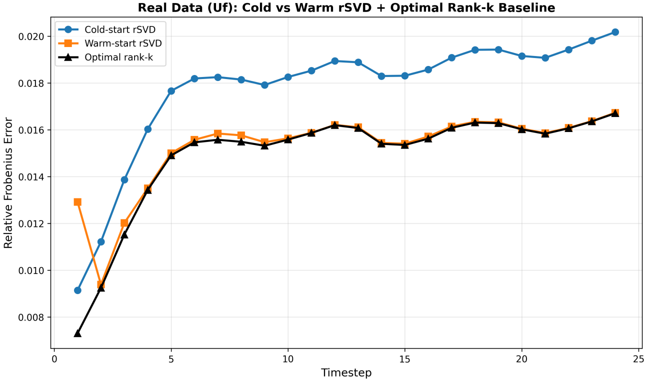
- Biggest overall effect is from **sampling dimensions**, led by p_{cold}, then p_{warm}.
- Global trend for p_{warm}: $0 \rightarrow 20$ moves avg gap from $+0.0012$ to -0.0160.
- Global trend for p_{cold}: $5 \rightarrow 20$ moves avg gap from -0.0256 to $+0.0016$.

Outline

Real Data Setup

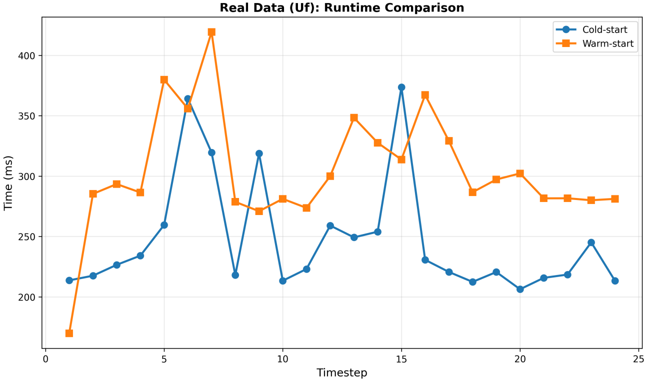
- **Data source:** 24 timesteps of 3D volumetric data (Uf01–Uf24.bin, full 24-hour cycle)
  - Each file: $100 \times 500 \times 500$ float32 (z, y, x order)
  - Missing/outlier values: pre-processed to 0
- **Matrix unfolding:** reshape volume to unstructured matrix $A_t \in \mathbb{R}^{250000 \times 100}$
  - Rows: spatial (x,y) points, $500 \times 500 = 250{,}000$
  - Columns: z-levels, 100 levels
- **Optimal baseline:** computed exactly from Gram matrix $G = A_t^T A_t$ (100×100 only)
  - Avoids costly full SVD of $250{,}000 \times 100$ matrix
  - Exact truncated-SVD Frobenius error tail

Cold vs Warm rSVD on Real Data



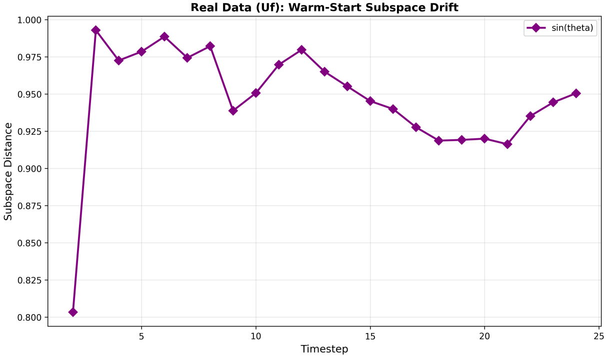
- Cold-start: learns from scratch; error gap indicates evolving rank structure
- Warm-start: reuses prior basis; adapts to subspace drift
- Optimal: fundamental limit for rank-k approximation

Weather Results: Runtime



- Cold-start: ~ 0.247 s/step
- Warm-start: ~ 0.304 s/step
- Trade-off: 23.0% slower per-step, but lower error on streaming data

Weather Results: Subspace Drift



- Subspace distance (sin of principal angles) between consecutive warm-start bases
- High drift (~ 0.95) indicates large changes in dominant subspace structure
- Warm initialization remains useful despite this drift

Weather Results Summary

Real Data (Uf01–Uf24): Average performance over 24 timesteps (full day cycle)

Method	Avg. Rel. Fro. Error	Avg. Time (s)
Cold rSVD	0.01774	0.2469
Warm rSVD	0.01524	0.3037
Optimal rank-k	0.01492	—

- **Warm-start improvement:** 14.1% error reduction (ratio 0.859 vs cold)
- **Gap to optimal:** Warm 2.2% above optimal, Cold 18.9% above optimal
- **Runtime trade-off:** Warm-start increases per-step time by 23.0%
- **Key insight:** Over full diurnal cycle, warm-start benefits accumulate; maintains robustness despite persistent subspace drift (≈ 0.95) from evolving meteorological patterns