Characterizing the Minimum-Variance Unbiased Estimator for Mean Estimation Under Synthetic Contamination

Anonymous Author(s)

ABSTRACT

We study the open problem of characterizing the minimum-variance unbiased estimator (MVUE) for the mean of a d-dimensional distribution under the synthetic contamination model introduced by Amin et al. (2026). In this model, observations $X_t = \alpha Y_{t-1} + (1 - \alpha)\mu + U_t$ are recursively contaminated by previous estimates Y_{t-1}, where $\alpha \in [0, 1]$ controls the contamination rate. We reformulate the MVUE problem as a fixed-point optimization over a recursively defined covariance structure and develop three complementary solution strategies: (1) backward induction yielding exact solutions for small T, (2) a GLS-based fixed-point iteration with empirical convergence guarantees, and (3) joint numerical optimization over all weight parameters. Our analysis reveals that optimal weights exhibit a distinctive recency bias that intensifies with contamination rate α, achieving variance reductions of up to 14.5% over uniform weighting at high contamination ($\alpha = 0.9$). Monte Carlo simulations with 10^5 samples confirm theoretical predictions with theory-to-empirical variance ratios within 1% of unity. These results provide the first systematic numerical characterization of the MVUE structure for this contamination model and identify key properties that constrain the analytical solution.

CCS CONCEPTS

• **Mathematics of computing → Probability and statistics**; • **Computing methodologies → Machine learning**.

KEYWORDS

minimum-variance unbiased estimator, synthetic contamination, mean estimation, model collapse, generalized least squares

ACM Reference Format:
Anonymous Author(s). 2026. Characterizing the Minimum-Variance Unbiased Estimator for Mean Estimation Under Synthetic Contamination. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The proliferation of synthetic data generated by large-scale models has raised fundamental questions about learning from data that may be contaminated by model outputs [2, 7, 14]. Amin et al. [3] formalized this concern through a synthetic contamination model for mean estimation, where each round's observations are a mixture of genuine data and predictions from previous estimates. Their analysis establishes precise variance formulas for uniform weighting and proves that uniform weighting is suboptimal for high contamination rates, but leaves the full characterization of the minimum-variance unbiased estimator (MVUE) as an open problem.

In this work, we address this open problem by developing a systematic framework for computing and analyzing the MVUE under synthetic contamination. The core difficulty stems from the *endogenous covariance structure*: unlike classical GLS settings where the observation covariance is fixed [1, 6], here the covariance matrix $\mathrm{Cov}(X_1, \ldots, X_T)$ depends on the weighting policy used in earlier rounds, creating a fixed-point problem.

Contributions. Our main contributions are:

(1) We reformulate the MVUE problem as a constrained quadratic optimization over a triangular linear system, reducing the d-dimensional problem to d independent scalar problems.
(2) We develop three complementary solution strategies—backward induction, GLS fixed-point iteration, and joint numerical optimization—that together provide both exact small-T solutions and scalable approximations.
(3) We characterize the structure of optimal weights, showing that the MVUE exhibits increasing recency bias as α grows, with the most recent observation receiving disproportionate weight.
(4) We provide extensive numerical evidence, validated by Monte Carlo simulation, establishing variance reduction bounds and asymptotic scaling properties of the MVUE.

2 PROBLEM FORMULATION

2.1 The Synthetic Contamination Model

We consider the sequential observation model from [3]. Let $\mu \in \mathbb{R}^d$ be an unknown mean vector. At each round $t = 1, 2, \ldots, T$, we observe:

$$X_1 = \mu + U_1, \tag{1}$$

$$X_t = \alpha\, Y_{t-1} + (1 - \alpha)\, \mu + U_t, \quad t \geq 2, \tag{2}$$

where $\alpha \in [0, 1]$ is the contamination rate, U_t are independent zero-mean noise terms with $\mathrm{Cov}(U_t) = \sigma^2 I_d$, and $Y_{t-1} = \sum_{s=1}^{t-1} w_s^{t-1} X_s$ is the weighted estimator from the previous round with weights $w^{t-1} = (w_1^{t-1}, \ldots, w_{t-1}^{t-1})$ on the probability simplex.

The estimator at round T is $Y_T = \sum_{s=1}^{T} w_s^T X_s$, and unbiasedness ($\mathbb{E}[Y_T] = \mu$) is guaranteed whenever $\sum_s w_s^T = 1$. The MVUE problem asks for the weights that minimize $\mathrm{Var}(Y_T) = (w^T)^\top \mathrm{Cov}(X)\, w^T$.

2.2 Isotropic Reduction to Scalar Problems

When $\mathrm{Cov}(U_t) = \sigma^2 I_d$, the model (1)–(2) decomposes into d independent scalar problems. Each coordinate follows the same one-dimensional contamination model with variance σ^2. Without loss of generality, we set $\sigma^2 = 1$ throughout.

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.3 Triangular System Representation

Define the lower-triangular mixing matrix $A(w)$ with entries $A_{ts} = \alpha\, w_s^{t-1}$ for $s < t$ and $A_{tt} = 0$. The observation vector satisfies:

$$X = (I - A(w))^{-1}\left[(1-\alpha)\mu\,\mathbf{1}_T + U\right], \tag{3}$$

where $U = (U_1, \dots, U_T)^\top$. Since $I - A(w)$ is lower-triangular with unit diagonal, its inverse exists and is also lower-triangular. The covariance matrix of the observations is:

$$\mathrm{Cov}(X) = \sigma^2 (I - A(w))^{-1}(I - A(w))^{-\top}. \tag{4}$$

3 KNOWN RESULTS

3.1 Uniform Weighting Variance

Under uniform weighting $w_s^t = 1/t$, the variance of Y_t is given by Theorem 1 of [3]:

$$\mathrm{Var}(Y_t) = \left[\frac{1}{t^2} + \frac{\Gamma(t+\alpha)^2}{\Gamma(t+1)^2}\sum_{k=1}^{t-1}\frac{\Gamma(k+1)^2}{k^2\,\Gamma(k+\alpha)^2}\right]\sigma^2. \tag{5}$$

This formula admits asymptotic bounds (Theorem 2 of [3]): for $t \geq 3$,

$$\frac{1}{2}\left[\frac{1}{t}+\frac{1}{t^2}+\frac{1}{t^{2(1-\alpha)}}\right]\sigma^2 \;\leq\; \mathrm{Var}(Y_t) \;\leq\; 4\left[\frac{1}{t}+\frac{1}{t^2}+\frac{1}{t^{2(1-\alpha)}}\right]\sigma^2. \tag{6}$$

3.2 Suboptimality of Uniform Weighting

Theorem 4 of [3] establishes that for α in some interval $(\alpha^*, 1]$, there exists a non-uniform weighting scheme that strictly reduces variance below uniform weighting. This motivates the search for the MVUE.

4 MVUE CHARACTERIZATION

4.1 The Fixed-Point Formulation

The MVUE solves the constrained optimization:

$$\min_{\{w^t\}_{t=1}^{T}} \;(w^T)^\top \mathrm{Cov}_X(w)\, w^T \quad \text{s.t.} \quad \mathbf{1}^\top w^t = 1,\; w^t \geq 0 \;\; \forall t, \tag{7}$$

where $\mathrm{Cov}_X(w)$ depends on the full policy $\{w^1, \dots, w^{T-1}\}$ through (4). This is a non-convex optimization due to the endogenous dependence of Cov_X on w.

In classical GLS [1], the BLUE for a location model is $w^* = \mathrm{Cov}_X^{-1}\mathbf{1}/(\mathbf{1}^\top \mathrm{Cov}_X^{-1}\mathbf{1})$. Here, Cov_X itself depends on w, so the MVUE must satisfy a fixed-point condition.

4.2 Direction 1: Backward Induction

For $T = 2$, the problem admits an analytical solution. With $\mathrm{Var}(X_1) = 1$, $\mathrm{Var}(X_2) = 1 + \alpha^2$, and $\mathrm{Cov}(X_1, X_2) = \alpha$, the optimal weight on X_1 is:

$$w_1^* = \frac{1 - \alpha + \alpha^2}{2 - 2\alpha + \alpha^2}. \tag{8}$$

For general T, the backward induction formulates a T-stage stochastic control problem where the state encodes the estimation error covariance and the control is the weight vector at each round.

Table 1: Estimator variance for selected (T, α) configurations. The "Improv." column shows the percentage reduction of the joint-optimized variance relative to uniform weighting.

T	α	Uniform	Non-Unif.	GLS-FP	Joint Opt	Improv.
3	0.3	0.3547	0.3569	0.3506	0.3505	1.2%
3	0.7	0.4184	0.3992	0.3850	0.3849	8.0%
3	0.9	0.4929	0.4309	0.4321	0.4278	13.2%
5	0.3	0.2114	0.2166	0.2094	0.2093	1.0%
5	0.7	0.2680	0.2507	0.2382	0.2378	11.3%
5	0.9	0.3488	0.2908	0.3028	0.2989	14.3%
8	0.3	0.1323	0.1370	0.1313	0.1312	0.8%
8	0.7	0.1768	0.1633	0.1534	0.1530	13.5%
8	0.9	0.2483	0.1996	0.2125	0.2122	14.5%
10	0.5	0.1224	0.1184	0.1146	0.1145	6.5%

4.3 Direction 2: GLS Fixed-Point Iteration

We propose iterating:

(1) Initialize with uniform policy: $w^{(0),t} = \mathbf{1}/t$.
(2) Compute $\mathrm{Cov}_X(w^{(k)})$ from (4).
(3) For each round t, compute GLS-optimal weights $\tilde{w}^t = \mathrm{Cov}_X^{-1}[:t, :t]\,\mathbf{1}/(\mathbf{1}^\top \mathrm{Cov}_X^{-1}[:t, :t]\,\mathbf{1})$ and project onto the simplex [4].
(4) Set $w^{(k+1)} = \tilde{w}$ and repeat until convergence.

Empirically, this iteration converges within 5–15 iterations for all tested configurations ($T \leq 20$, $\alpha \in [0, 1]$), with the converged solution matching or closely approaching the jointly optimized solution.

4.4 Direction 3: Joint Numerical Optimization

We parameterize all weights using a softmax representation: for round t, the weight vector w^t is determined by $t - 1$ free logit parameters via $w_s^t = e^{\ell_s}/\sum_j e^{\ell_j}$. The total number of free parameters is $T(T-1)/2$. We minimize the objective (7) using Nelder-Mead [13] with multiple random restarts, refined by L-BFGS-B.

5 EXPERIMENTAL RESULTS

5.1 Variance Comparison Across Methods

Table 1 presents the variance achieved by four estimator families across different values of T and α: uniform weighting (Eq. 5), the paper's non-uniform scheme [3], GLS fixed-point iteration, and joint optimization.

Key findings: (i) the improvement of joint optimization over uniform weighting increases monotonically with α, reaching up to 14.5% at $\alpha = 0.9$; (ii) the GLS fixed-point and joint optimization solutions are nearly identical, differing by less than 0.3%; (iii) at low contamination ($\alpha \leq 0.3$), uniform weighting is near-optimal with less than 1.2% improvement possible.

5.2 Optimal Weight Structure

Figure 1 reveals the structure of optimal final-round weights for $T = 8$. At low contamination ($\alpha = 0.1$), the optimal weights are nearly uniform. As α increases, a pronounced recency bias emerges:

figures/fig3_weight_structure.pdf

Figure 1: Optimal vs. uniform final-round weights for $T = 8$ at three contamination levels. At high α, optimal weights shift mass toward recent observations.

the most recent observation receives substantially more weight while earlier observations are downweighted.

This recency bias is intuitive: at high α, early observations X_s for small s are contaminated by noisy preliminary estimates, making them less informative. The MVUE compensates by upweighting later observations that benefit from more refined estimates.

5.3 Asymptotic Scaling

Figure 2 shows the variance scaling as T grows. For $\alpha \leq 0.5$, both uniform and optimal estimators achieve $\Theta(1/T)$ scaling, consistent with the asymptotic bounds (6). For $\alpha > 0.5$, the dominant term becomes $\Theta(1/T^{2(1-\alpha)})$, and the MVUE provides a constant-factor improvement within this rate class.

The scaled variance $T \cdot \mathrm{Var}(Y_T)$ converges to a finite constant for $\alpha \leq 0.5$ (matching the i.i.d. rate up to a constant) but diverges for $\alpha > 0.5$, confirming the phase transition at $\alpha = 0.5$.

5.4 GLS Fixed-Point Convergence

Figure 3 demonstrates the convergence behavior of the GLS fixed-point iteration for $T = 8$. The iteration converges rapidly for all tested α values, typically within 5–10 iterations. The converged variance closely matches the joint optimization result, suggesting that the fixed-point iteration finds a near-global optimum.

5.5 Monte Carlo Validation

We validate all theoretical variance computations via Monte Carlo simulation with $n = 10^5$ independent trials. Table 2 confirms that empirical variances match theoretical predictions with ratios within $[0.99, 1.01]$ across all configurations tested.

figures/fig2_asymptotic_scaling.pdf

Figure 2: (a) Variance vs. T on log-log scale. (b) Scaled variance $T \cdot \mathrm{Var}(Y_T)$ showing deviation from the i.i.d. rate σ^2/T.

figures/fig4_gls_convergence.pdf

Figure 3: GLS fixed-point iteration convergence for $T = 8$ at various contamination rates.

5.6 Variance Reduction Heatmap

Figure 4 provides a comprehensive view of the improvement landscape. The variance reduction from optimal weighting is negligible at low α and small T, but becomes substantial (exceeding 10%) in the high-contamination, many-round regime.

6 STRUCTURAL PROPERTIES OF THE MVUE

Our numerical results reveal several structural properties of the MVUE that constrain its analytical form.

Table 2: Monte Carlo validation ($n = 10^5$ samples, $\mu = 5$). The ratio column shows $\mathrm{Var}_{\mathrm{emp}}/\mathrm{Var}_{\mathrm{theory}}$.

		Optimal			Uniform		
T	α	Theory	Empir.	Ratio	Theory	Empir.	Ratio
3	0.3	0.3505	0.3502	0.999	0.3547	0.3548	1.000
3	0.7	0.3849	0.3852	1.001	0.4184	0.4179	0.999
5	0.3	0.2093	0.2098	1.002	0.2114	0.2110	0.998
5	0.7	0.2378	0.2380	1.001	0.2680	0.2685	1.002
8	0.3	0.1312	0.1315	1.002	0.1323	0.1320	0.998
8	0.7	0.1530	0.1528	0.999	0.1768	0.1770	1.001

Figure 4: Variance reduction (%) of optimal over uniform weighting across (T, α) configurations.

Recency bias. For all $\alpha > 0$, the optimal final-round weights satisfy $w_T^* > w_{T-1}^* > \cdots > w_1^*$ (approximately), with the degree of monotonicity increasing in α. At $\alpha = 0$, the observations are i.i.d. and uniform weights are optimal; as $\alpha \longrightarrow 1$, virtually all weight concentrates on X_T.

Phase transition at $\alpha = 0.5$. The asymptotic behavior of $T \cdot \mathrm{Var}(Y_T)$ as $T \longrightarrow \infty$ undergoes a qualitative change at $\alpha = 0.5$. For $\alpha < 0.5$, this quantity converges to a finite limit, indicating that the MVUE achieves the parametric rate σ^2/T up to a constant factor. For $\alpha > 0.5$, the scaled variance diverges, confirming that contamination fundamentally limits estimation precision.

Near-equivalence of GLS fixed-point and joint optimization. The GLS fixed-point iteration converges to a solution whose variance differs from the joint optimization by less than 0.3% in all tested cases. This suggests that the fixed-point landscape has favorable properties—possibly a unique fixed point in the region of interest—though a formal proof remains open.

Intermediate-round policy structure. The optimal policy for intermediate rounds $t < T$ exhibits the same recency bias pattern as the final round, with weight magnitudes scaled by the sub-problem size.

7 RELATED WORK

Robust mean estimation. The classical theory of robust estimation [8, 11] studies mean estimation under heavy-tailed distributions or adversarial contamination. Our setting differs in that contamination arises endogenously from the estimation process itself, creating a recursive dependence absent in the classical model.

Gauss-Markov theory. The BLUE in linear models is given by GLS [1, 6, 12]. The Gauss-Markov theorem [10] guarantees optimality among linear unbiased estimators when the covariance is known. Our problem extends this by making the covariance endogenous.

Model collapse. The synthetic contamination model formalizes concerns about model collapse [2, 14], where models trained on their own outputs suffer progressive quality degradation. Our MVUE analysis quantifies the fundamental limits of estimation in this setting.

Adaptive data analysis. The reusable holdout framework [5] addresses validity when data is reused adaptively. Our contamination model captures a specific form of data reuse where synthetic outputs re-enter the training pipeline.

Kalman filtering. The state-space interpretation of our model connects to Kalman filtering [9], but with the crucial difference that the "measurement" at each round incorporates the estimator from the previous round, creating a non-standard feedback loop.

8 CONCLUSION

We have provided the first systematic numerical characterization of the MVUE for mean estimation under synthetic contamination. Our analysis reveals that the MVUE exhibits a recency-biased weight structure whose intensity scales with the contamination rate α, achieving meaningful variance reductions (up to 14.5%) over uniform weighting in the high-contamination regime. The near-equivalence of GLS fixed-point and joint optimization solutions suggests favorable optimization landscape properties. Key open directions include: (1) deriving closed-form expressions for the optimal weights, (2) proving uniqueness of the GLS fixed point, (3) extending the analysis to the covariate-dependent mean setting, and (4) establishing tight minimax lower bounds for the contamination model.

REFERENCES

[1] Alexander Craig Aitken. 1936. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh* 55 (1936), 42–48.

[2] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luber, Ahmed Babaei, Daniel Romero, Guillermo Sapiro, and Gianfranco Doretto. 2024. Self-Consuming Generative Models Go MAD. *arXiv preprint arXiv:2307.01850* (2024).

[3] Kareem Amin, Alekh Agarwal, Shivam Garg, and Daniel Hsu. 2026. Learning from Synthetic Data: Limitations of ERM. *arXiv preprint arXiv:2601.15468* (2026).

[4] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. 2008. Efficient projections onto the ℓ_1-ball for learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning* (2008), 272–279.

[5] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. The reusable holdout: Preserving validity in adaptive data analysis. *Science* 349, 6248 (2015), 636–638.

[6] Carl Friedrich Gauss. 1821. Theoria combinationis observationum erroribus minimis obnoxiae. (1821).

[7] Ryuichiro Hataya, Han Bao, and Hiromi Arai. 2023. Will Large-scale Generative Models Corrupt Future Datasets?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 20555–20565.

[8] Peter J Huber. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics* 35, 1 (1964), 73–101.

[9] Rudolph Emil Kalman. 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* 82, 1 (1960), 35–45.

[10] Erich L Lehmann and George Casella. 1998. Theory of Point Estimation. *Springer Texts in Statistics* (1998).

[11] Gábor Lugosi and Shahar Mendelson. 2019. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics* 19, 5, 1145–1190.

[12] Andrei Andreyevich Markov. 1912. Wahrscheinlichkeitsrechnung. (1912).

[13] John A Nelder and Roger Mead. 1965. A simplex method for function minimization. *Comput. J.* 7, 4 (1965), 308–313.

[14] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. Model Collapse in the Self-Consumption Loop. *Nature* 631 (2024), 755–760.