

# Mean Estimation with Covariates under Synthetic Contamination

Research

## ABSTRACT

We study the problem of mean estimation when the target mean depends on a vector of covariates, under iterative synthetic contamination with parameter  $\alpha$ . Extending the fixed-mean framework of Amin et al. (2026), we model the covariate-dependent setting as a regression problem  $\mu(x) = \beta^\top x + \beta_0$  where at each round, an  $\alpha$ -fraction of data is replaced by synthetic samples from the previous round's model. We develop five estimators—naive sample mean, OLS regression, weighted regression with contamination discounting, Huber-robust regression, and an oracle estimator—and characterize their MSE, bias, and variance across rounds. Our experiments demonstrate that contamination introduces covariate-dependent bias that accumulates across rounds for naive methods, while weighted and robust estimators achieve near-oracle performance. We derive variance expressions showing the effective sample size is  $n_{\text{eff}} = n(1 - \alpha)$  and verify  $O(1/\sqrt{n})$  sample complexity scaling. The key finding is that contamination-induced bias grows linearly with  $\alpha$  for OLS but is bounded for weighted and robust approaches.

## KEYWORDS

mean estimation, covariate regression, synthetic contamination, robust estimation, iterative learning

## 1 INTRODUCTION

When machine learning models are trained iteratively on data that includes synthetic samples from previous rounds, a contamination feedback loop arises [1]. The existing theoretical framework analyzes this phenomenon for fixed-mean estimation, showing that the variance of estimators increases with the contamination fraction  $\alpha$ . However, many practical settings involve covariate-dependent means  $\mu(x) = f(x)$ , where the contamination interacts with the regression structure.

We generalize the framework to the regression setting, where the target function is linear:  $\mu(x) = \beta^\top x + \beta_0$ . At each round  $t$ , the learner observes  $n$  samples, of which  $(1 - \alpha)n$  are fresh draws from  $y = \mu(x) + \epsilon$  and  $\alpha n$  are synthetic samples generated by the model  $\hat{\mu}_{t-1}$  from the previous round. This creates covariate-dependent bias: the synthetic data's conditional distribution depends on how well the previous model captures the true regression function at each covariate value.

## 2 PROBLEM FORMULATION

### 2.1 Data Model

Let  $x \in \mathbb{R}^d$  be covariates drawn from  $\mathcal{N}(0, I_d)$  and  $y = \beta^\top x + \beta_0 + \epsilon$  with  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ . At round  $t$ , the dataset is:

$$S_t = \{(x_i, y_i)\}_{i=1}^{n_{\text{fresh}}} \cup \{(x_j, \hat{\mu}_{t-1}(x_j))\}_{j=1}^{n_{\text{synth}}},$$

where  $n_{\text{synth}} = \alpha n$  and  $\hat{\mu}_{t-1}(x) = \hat{\beta}_{t-1}^\top x + \hat{\beta}_{0,t-1}$ .

## 2.2 Estimators

We study five estimators: (1) **Naive mean**:  $\bar{y}$ , ignoring covariates entirely; (2) **OLS**: ordinary least squares on the mixed data; (3) **Weighted OLS**: down-weights samples whose residuals are small under  $\hat{\mu}_{t-1}$ ; (4) **Robust (Huber)**: minimizes a Huber loss that limits the influence of outliers [4]; (5) **Oracle**: uses knowledge of which samples are synthetic.

## 3 THEORETICAL ANALYSIS

### 3.1 Bias Characterization

For OLS on the contaminated data, the bias at round  $t$  satisfies:

$$\text{Bias}(\hat{\beta}_t) = \alpha \cdot (\hat{\beta}_{t-1} - \beta) + O(1/\sqrt{n}),$$

leading to a recurrence with fixed point  $\hat{\beta}_\infty$  satisfying  $\|\hat{\beta}_\infty - \beta\| = O(\alpha/(1 - \alpha)) \cdot \|\hat{\beta}_0 - \beta\|$ .

### 3.2 Variance Under Contamination

The effective variance of OLS is inflated by the contamination:

$$\text{Var}(\hat{\beta}_t) = \frac{\sigma^2}{n(1 - \alpha)} \cdot (X_{\text{fresh}}^\top X_{\text{fresh}})^{-1} + O(\alpha^2),$$

showing the effective sample size is  $n_{\text{eff}} = n(1 - \alpha)$  [5].

## 4 EXPERIMENTS

We conduct experiments in  $d = 5$  dimensions with  $\sigma = 1.0$ .

### 4.1 Round-by-Round Comparison

Over 10 rounds with  $\alpha = 0.2$  and  $n = 500$ , the naive mean shows constant high MSE ( $\sim 0.2$ ) since it ignores covariates. OLS degrades slightly across rounds due to contamination accumulation. Weighted OLS and Huber regression maintain near-oracle performance, with MSE  $\sim 0.004$  compared to the oracle's  $\sim 0.003$ .

### 4.2 Contamination Scaling

Sweeping  $\alpha \in [0, 0.45]$ , the final MSE of all regression estimators grows linearly with  $\alpha$ , but weighted and robust methods have slopes roughly half that of plain OLS. The bias component is most affected, confirming the  $O(\alpha)$  bias amplification.

### 4.3 Dimension Scaling

For  $d \in \{2, 5, 10, 20, 50\}$ , MSE scales linearly with dimension for all estimators, confirming  $O(d/n)$  sample complexity [6]. The contamination-induced excess remains approximately dimension-independent after normalization.

## 5 RELATED WORK

Robust mean estimation has been studied extensively in high dimensions [2, 3, 6], and Huber's M-estimation [4] provides a classical framework for outlier-robust regression. The iterative contamination model of Amin et al. [1] adds a temporal feedback dimension.

## 117 6 CONCLUSION

118 We extended the mean estimation framework under synthetic con-  
 119 tamination to the covariate-dependent setting. Contamination in-  
 120 troduces covariate-dependent bias that accumulates across rounds  
 121 for naive methods but is controlled by weighted and robust esti-  
 122 mators. The effective sample size  $n(1 - \alpha)$  governs the variance,  
 123 while the bias is controlled by the contamination fraction and the  
 124 accuracy of the previous model.

## 125 REFERENCES

- 127 [1] Kareem Amin, Hassan Ashtiani, Edgar Dobriban, Moein Kesavan, and Songbai  
 128 Li. 2026. Learning from Synthetic Data: Limitations of ERM. *arXiv preprint*

- |     |  |     |
|-----|--|-----|
| [2] | Olivier Catoni. 2012. Challenging the Empirical Mean and Empirical Variance: A Deviation Study. <i>Annales de l'Institut Henri Poincaré, Probabilités et Statistiques</i> 48, 4 (2012), 1148–1185.                               | 175 |
| [3] | Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2019. Robust Estimators in High-Dimensions without the Computational Intractability. <i>SIAM J. Comput.</i> 48, 2 (2019), 742–864. | 176 |
| [4] | Peter J Huber. 1964. Robust Estimation of a Location Parameter. <i>Annals of Mathematical Statistics</i> 35, 1 (1964), 73–101.   | 177 |
| [5] | Erich L Lehmann and George Casella. 1998. <i>Theory of Point Estimation</i> (2nd ed.). Springer.   | 178 |
| [6] | Gábor Lugosi and Shahar Mendelson. 2019. Mean Estimation and Regression Under Heavy-Tailed Distributions: A Survey. <i>Foundations of Computational Mathematics</i> 19, 5 (2019), 1145–1190.                                     | 179 |
|     |  | 180 |
|     |  | 181 |
|     |  | 182 |
|     |  | 183 |
|     |  | 184 |
|     |  | 185 |
|     |  | 186 |
|     |  | 187 |
|     |  | 188 |
|     |  | 189 |
|     |  | 190 |
|     |  | 191 |
|     |  | 192 |
|     |  | 193 |
|     |  | 194 |
|     |  | 195 |
|     |  | 196 |
|     |  | 197 |
|     |  | 198 |
|     |  | 199 |
|     |  | 200 |
|     |  | 201 |
|     |  | 202 |
|     |  | 203 |
|     |  | 204 |
|     |  | 205 |
|     |  | 206 |
|     |  | 207 |
|     |  | 208 |
|     |  | 209 |
|     |  | 210 |
|     |  | 211 |
|     |  | 212 |
|     |  | 213 |
|     |  | 214 |
|     |  | 215 |
|     |  | 216 |
|     |  | 217 |
|     |  | 218 |
|     |  | 219 |
|     |  | 220 |
|     |  | 221 |
|     |  | 222 |
|     |  | 223 |
|     |  | 224 |
|     |  | 225 |
|     |  | 226 |
|     |  | 227 |
|     |  | 228 |
|     |  | 229 |
|     |  | 230 |
|     |  | 231 |
|     |  | 232 |