

Heteroskedastic Errors in Learning

- **Motivation:** Data or **observations** collected from multiple users (or sensing devices), each has different **reliability** (noise level)
- Commonly referred to as *heteroskedastic* errors in statistics; several heuristics studied for problems like regression
- Standard learning algorithm yields guarantee that depends on *largest* error
- **Broad Research Question:** Can we design algorithms that do better when some observations have lower error than others?

Model: Regression with Heteroskedastic Errors

There are n sources. Source i provides t observations

$$x_i^{(j)} \sim D_i, \quad y_i^{(j)} = \langle x_i^{(j)}, \beta \rangle + \eta_i^{(j)}, \quad \eta_i^{(j)} \sim \mathcal{N}(0, \sigma_i^2).$$

- Regression vector $\beta \in \mathbb{R}^d$ is shared
- Variances σ_i^2 are arbitrary, different, and **unknown**.

Our goal: design and analyze algorithms that improve on vanilla regression when *many of the variances* are *small*.

Subset of Signals Model. For m of the signals, we know that $\sigma_i \leq 1$; but crucially, we do not know *which signals* have low variance!

Prior Work and the Regression Baseline

- Long line of statistical work on regression with heteroskedastic errors: weighted least squares – shown to be optimal [1]
- Weighted least squares needs estimates/priors on σ_i for each i
- Ordinary (unweighted) least squares —a trivial baseline— incurs error that depends on $\sigma_{\max}$; even a single large variance can throw it off!
- Easier problem of *mean estimation* with heteroskedastic noise well studied [2]
- Compton and Valiant: tight bounds in 1D for subset-of-signals model [3]

*Can we show similar guarantees for **regression** in the subset-of-signals model?*

Assumptions

Recall: source i provides t observations for each $i \in [n]$

- 1 **Scaling:** $\|x_i^{(j)}\|_2 \leq 1$. [Necessary for $\sigma_i \leq 1$ to make sense.]
- 2 **Well-spread covariates:** sample a $2d$ -tuple of indices $(i_1, i_2, \dots, i_{2d})$; the corresponding covariate vectors x_{i_ℓ} are well-conditioned w.h.p.
- 3 **Independence of errors:** Errors are independent across sources i and observations within source.

Assumptions hold with high probability for random Gaussian $x_i^{(j)}$

Results

Theorem 1. ($t = 1$ observation per user.) If $m \geq n^{1-1/(4d)}$, then the estimate $\widehat{\beta}$ produced by algorithm satisfies, w.h.p.,

$$\|\widehat{\beta} - \beta\|_2 \leq \tilde{O}\left(\frac{d^{3/2}n^{2d-1/2}}{m^{2d}}\right).$$

- Bound is $o(1)$, for $m \geq n^{1-\frac{1}{4d}}$
- Algorithm runs in time roughly $O(n^{2d})$ — effective only for $d = O(1)$:(

Theorem 2. ($t = 2$ observations per user.) If $m \geq n^{5/6}$, then with probability at least 0.9, output $\widehat{\beta}$ of the algorithm satisfies

$$\|\widehat{\beta} - \beta\|_2 \leq \tilde{O}\left(d^3 \left(\frac{n^5}{m^6}\right)^{1/3}\right).$$

Once again, bound is $o(1)$ for chosen regime of m ; run time is roughly $O(n^{2d})$.

Note: If set of reliable sources were known, we can achieve error $\Theta(\sqrt{d/(mt)})$.

Key Idea: Regression-to-Mean Reduction.

Take any $2d$ -tuple of indices $\tau = (i_1, i_2, \dots, i_{2d})$, for which the corresponding vectors $\{x_{i_\ell}\}_{\ell=1}^{2d}$ are “well-conditioned”. For any standard basis vector e_r , we can write

$$e_r = \sum_{j=1}^{2d} \alpha_{j,r}(\tau) x_{i_j}.$$

$$\text{Thus,} \quad u_{\tau,r} := \sum_{j=1}^{2d} \alpha_{j,r}(\tau) y_{i_j} = \beta_r + \sum_{j=1}^{2d} \alpha_{j,r}(\tau) \eta_{i_j}.$$

Thus every $2d$ -tuple produces an estimate of coordinate β_r . *What if we feed these estimates into a mean estimation (with heteroskedastic errors) algorithm!?*

Algorithm Outline ($t = 1$)

- 1 **Enumerate stable tuples.** Collect all well-conditioned $2d$ -tuples τ .
- 2 **Construct coordinate estimates.** Form $u_{\tau,r}$ for all such τ and $r \in [d]$.
- 3 **Iteratively truncate.** Repeatedly clip estimates around the current center at decreasing scales. (Known algorithm for Mean Estimation [4].)
- 4 **Aggregate coordinates.** Return $\widehat{\beta} = (\widehat{\beta}_1, \dots, \widehat{\beta}_d)$.

Key Technical Obstacle: Tuple estimates are **dependent** because overlapping tuples re-use the same noise. Needs new argument using Janson’s inequality on concentration of sums of dependent r.v.s

Algorithm Outline ($t = 2$)

Main Idea: With **two** observations per source, can perform “sample splitting” and end up with simpler pruning-based algorithm!

- 1 **Screen with the first replicate.** Guess a reliable $2d$ -tuple and “test” the remaining sources against it.
- 2 **Retain a pruned set T .** At least $0.9m$ reliable sources survive, while very high-variance sources become fewer (but **do not** go away fully).
- 3 **Build fresh tuple estimates.** Using only second-replicate observations from T , construct $u_{\tau,r}$.
- 4 **Take the empirical mode.** For window width 1,

$$\widehat{\beta}_r = \arg \max_{u \in \mathbb{R}} \sum_{\tau \in \binom{T}{2d} \cap T_{\text{good}}} \mathbf{1}\{u_{\tau,r} \in [u - 1, u + 1]\}.$$

Technical Contribution: Analysis of the classic modal estimator for mean estimation in the setting with *correlated* noise. The initial *pruning* step is crucial: can only analyze modal estimation when number of high-noise sources is small enough.

Takeaways and open questions

- These are the first sub-constant recovery guarantees for regression in this subset-of-signals model.
- **Question:** Is the distinction between $t = 1$ and $t = 2$ only a technical one?
- The “threshold” at which β is recoverable to $o(1)$ error is $m \sim n^{1/4}$ for the mean estimation problem. Even with $t = 2$, we can only do it for $m \sim n^{5/6}$. Can we improve the threshold?
- **Open Question:** What if we restrict to $\text{poly}(n, d)$ time algorithms?
- Lower bounds for recovery in terms of m, d ?

References

- [1] Halbert White. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity”. In: *Econometrica* (1980).
- [2] Flavio Chierichetti et al. “Learning entangled single-sample Gaussians”. In: *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. (SODA). 2014.
- [3] Spencer Compton and Gregory Valiant. “Near-Optimal Mean Estimation with Unknown, Heteroskedastic Variances”. In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. (STOC). 2024.
- [4] Yingyu Liang and Hui Yuan. “Learning Entangled Single-Sample Gaussians in the Subset-of-Signals Model”. In: *Proceedings of Thirty Third Conference on Learning Theory*. (COLT). 2020.