Robust Estimation Under Heterogeneous Corruption Rates

Syomantak Chaudhuri[1], Jerry Li[2], Thomas A. Courtade[1]

[1]University of California Berkeley, [2]University of Washington

NeurIPS 2025, San Diego

Background

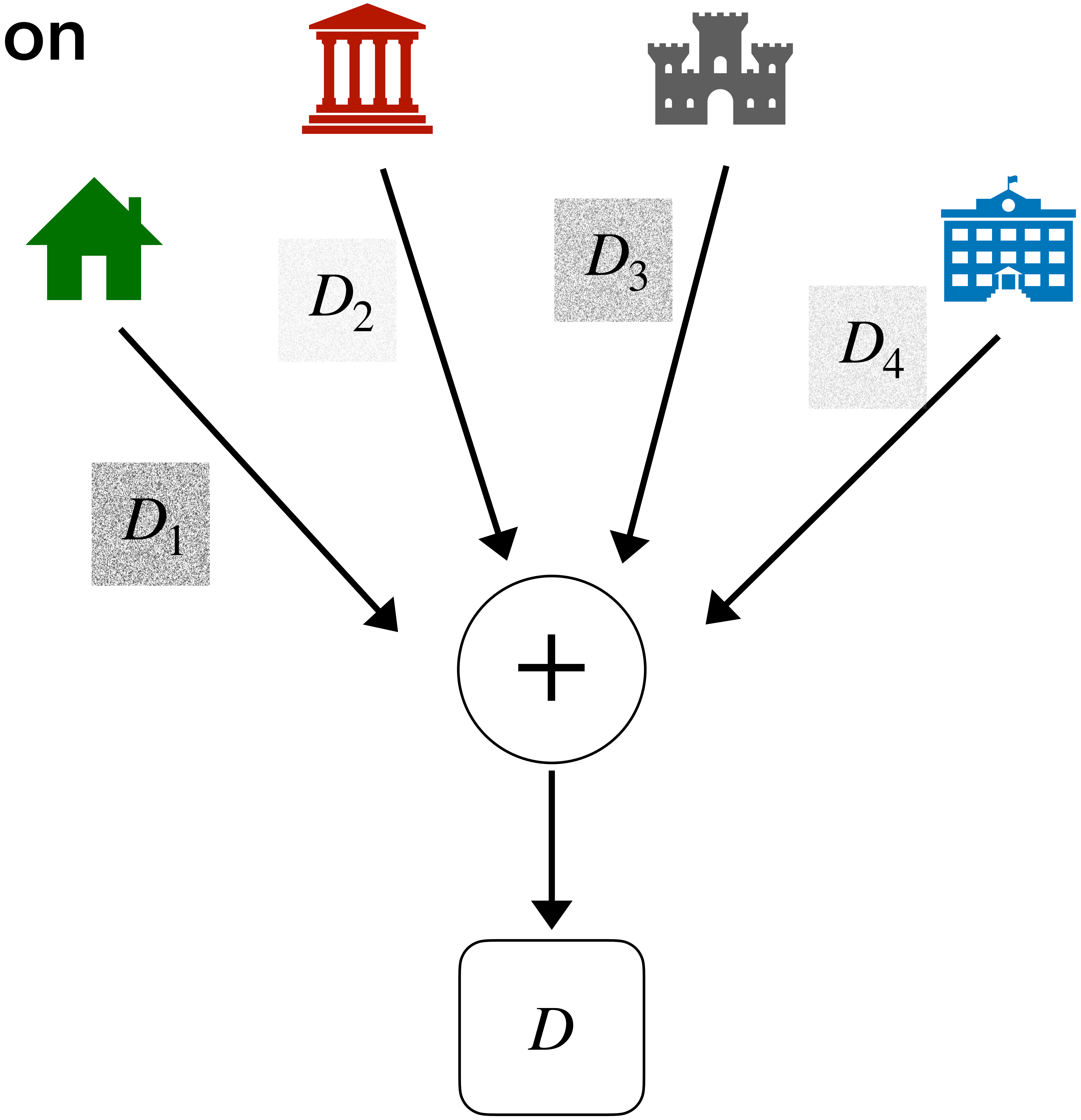
Robust Statistics Setup

- Conventional statistics → degrade rapidly when distributional assumptions are violated
- Robust statistics → work under distributional deviations or ‘contamination’
- *Huber* contamination model:
  - True data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$
  - Corruption rate $\lambda \in [0,1]$
  - Observations $Z_1, \ldots, Z_n$ with $Z_i = \begin{cases} X_i & \text{wp } 1 - \lambda, \\ \tilde{X}_i & \text{else.} \end{cases}$
  - $\tilde{X}_1, \ldots, \tilde{X}_n$ are modeled as worst-case (adversarial) for the statistical task

Heterogeneity

Motivation

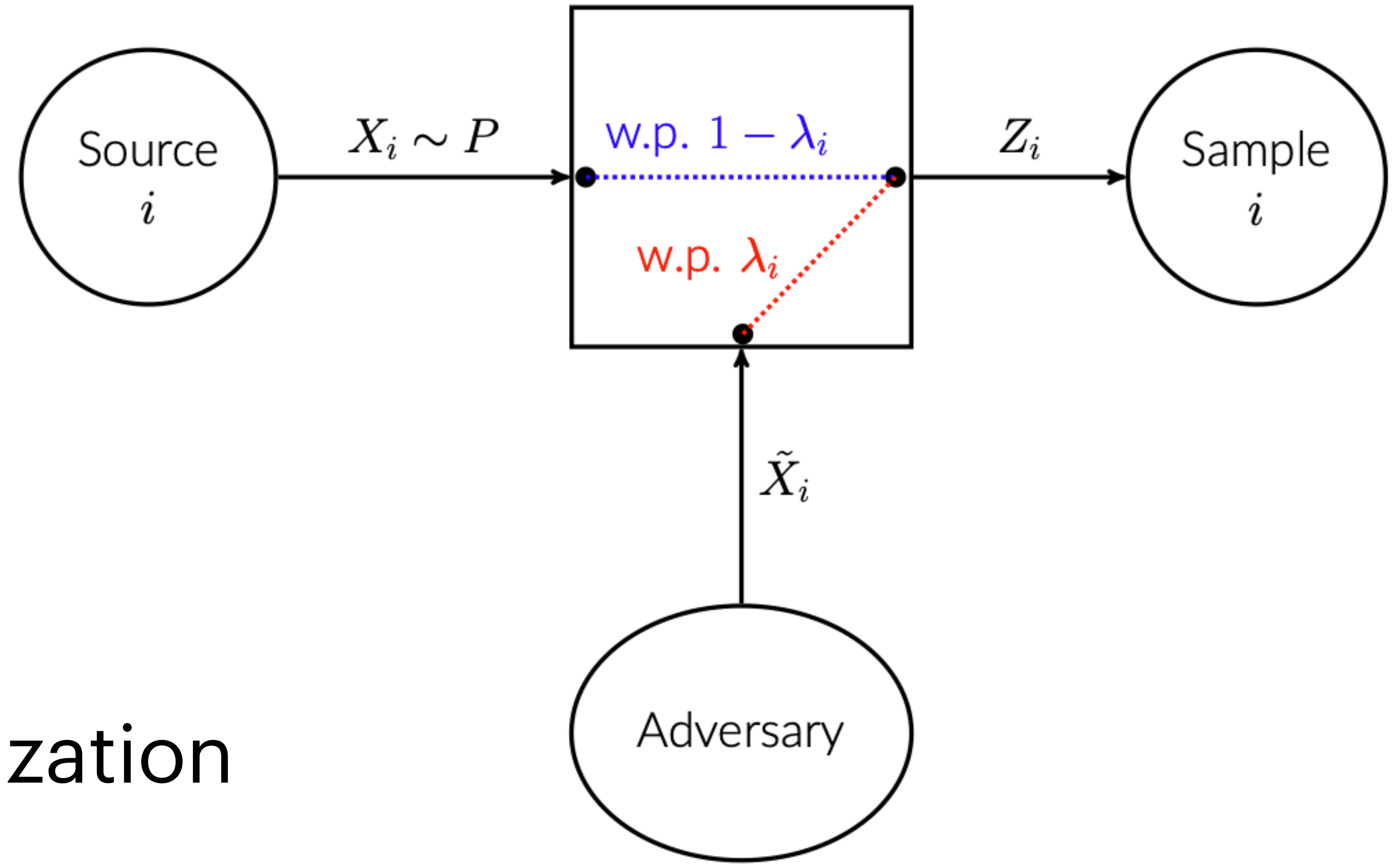
- Modern machine learning: federated setup
  - Dataset obtained from multiple sources
- Different sources → different reliability
- Example: temperature measured from different IoT sensors over the city
  - Cheaper, less reliable, IoT sensors in residential sources



Formal Setup

Notations

- Most general setup → every datapoint has different corruption rate
- True data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$
- Corruption rates $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$, $B_i \sim \mathrm{Bern}(\lambda_i)$ independently
- Observations $Z_1, \ldots, Z_n$ with $Z_i = (1 - B_i)X_i + B_i\tilde{X}_i$
- Choose $\tilde{X}_1, \ldots, \tilde{X}_n$ worst-case conditioned on the realization $\{(X_i, B_i)\}_{i\in[n]}$
- Notation: $\boldsymbol{Z} \sim_{\boldsymbol{\lambda}} P$



Bounded Mean Estimation

Setup and Results

- Let $\mathscr{D}_r$ be the set of probability distributions in $\mathbb{R}^d$ on the l_2-ball of radius r
- Define $\boldsymbol{\lambda}$-instance specific minimax MSE

$$L(\boldsymbol{\lambda}) = \inf_{M} \sup_{P \in \mathscr{D}_r} \mathbb{E}_{\mathbf{Z} \sim_{\lambda} P} \left[\left\| M(\mathbf{Z}) - \mathbb{E}_{X \sim P}[X] \right\|_{l_2}^2 \right]$$

- Let $f(\boldsymbol{\lambda}, k) = \min_{t \in [0,1]} \left(\frac{k}{|\{i : \lambda_i \leq t\}|} + t^2 \right)$
- We show $L(\boldsymbol{\lambda}) \simeq r^2 f(\boldsymbol{\lambda}, 1)$
- Corollary: extra data above a certain level of corruption does not help reduce MSE

Gaussian Mean Estimation

Setup and Results

- Let $\mathscr{D}_G$ be the set of all Gaussian distributions on $\mathbb{R}^d$ with identity covariance
- Define λ-instance specific minimax PAC error

$$L_{\text{PAC}}(\lambda) = \inf_{M} \sup_{P \in \mathscr{D}_r} Q\left(\| M(\mathbf{Z}) - \mathbb{E}_{X \sim P}[X] \|_{l_2}^2, \frac{1}{5} \right), \text{ where}$$

$$Q(Y, \delta) = \inf \{ t \in [0, \infty) : \mathrm{P}[Y \geq t] \leq \delta \}.$$

- We show $\frac{1}{\sqrt{d}} f(\lambda, d) \lesssim L_{\text{PAC}}(\lambda) \lesssim f(\lambda, d)$
- The gap of $O(\sqrt{d})$ is non-trivial

Gaussian Mean Estimation

Continued

- Challenges:
  - Standard robust estimation lower bound techniques do not work for heterogeneous λ -instance specific minimax rate
  - Use an interpolation of Assouad's method and Le Cam's method
- More in the paper:
  - Upper bound techniques using weighted Tukey median
  - Gaussian linear regression under heterogeneous corruption