Collaborative Learning of Discrete Distributions under Heterogeneity and Communication Constraints

Xinmeng Huang[1*], Donghwan Lee[1*], Edgar Dobriban[2], Hamed Hassani[3]

[1]AMCS@UPenn [2]STATS@UPenn [3]ESE@UPenn

NeurIPS 2022

Federated Analytics

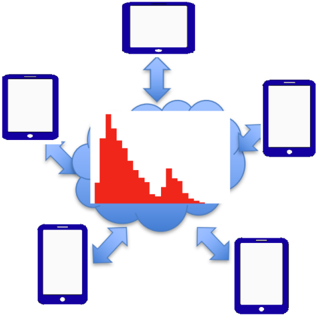


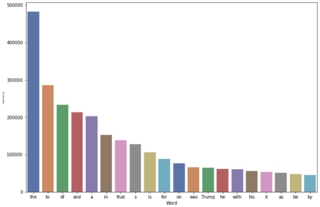
Figure: Data analysis on users' devices, locally[1]

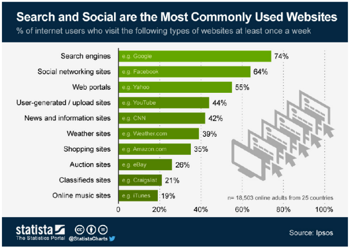
Two main challenges:

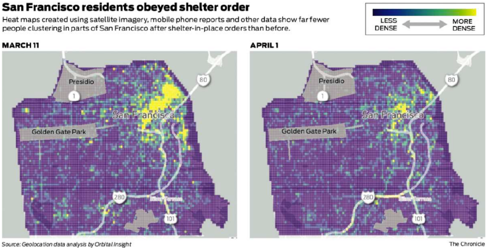
- Small communication bandwidth
- Heterogeneity among users

[1] Image credits to Prof. Ayfer Özgür

Distributed Estimation







We study learning discrete distributions under communication constraints and sparse heterogeneity

Setup

- T clusters $\{(X^{t,j})_{j\in[n]}\}_{t=1}^{T}$, each of which contains n datapoints
- Each client has one local one-hot datapoint $X^{t,j} \in [d]$ following Cat(p^t)
- Sparse heterogeneity: there is a global distribution $p^\star$ such that

$$\|p^t - p^\star\|_0 \le s, \quad \forall\, 1 \le t \le T$$

- Communication constraint: b-bits $(b \ll \log_2(d))$ budget for each client to communicate with a central server
- Goal: to design estimators $\widehat{p}^t : \{(Y^{t,j})_{j\in[n]}\}_{t=1}^{T} \to \mathbb{R}^d$ to minimize

$$\mathbb{E}[\|\widehat{p}^t - p^t\|_2^2], \quad \forall\, 1 \le t \le T$$

where $Y^{t,j}$ is the observed message transmitted from $X^{t,j}$

Setup

- T clusters $\{(X^{t,j})_{j\in[n]}\}_{t=1}^{T}$, each of which contains n datapoints
- Each client has one local one-hot datapoint $X^{t,j} \in [d]$ following $\mathrm{Cat}(p^t)$
- Sparse heterogeneity: there is a global distribution $p^\star$ such that

$$\|p^t - p^\star\|_0 \le s, \quad \forall\, 1 \le t \le T$$

- Communication constraint: b-bits $(b \ll \log_2(d))$ budget for each client to communicate with a central server
- Goal: to design estimators $\widehat{p}^t : \{(Y^{t,j})_{j\in[n]}\}_{t=1}^{T} \to \mathbb{R}^d$ to minimize

$$\mathbb{E}[\|\widehat{p}^t - p^t\|_2^2], \quad \forall\, 1 \le t \le T$$

where $Y^{t,j}$ is the observed message transmitted from $X^{t,j}$

Setup

- T clusters $\{(X^{t,j})_{j\in[n]}\}_{t=1}^{T}$, each of which contains n datapoints
- Each client has one local one-hot datapoint $X^{t,j} \in [d]$ following $\mathsf{Cat}(p^t)$
- Sparse heterogeneity: there is a global distribution $p^\star$ such that

$$\|p^t - p^\star\|_0 \le s, \quad \forall\, 1 \le t \le T$$

- Communication constraint: b-bits ($b \ll \log_2(d)$) budget for each client to communicate with a central server
- Goal: to design estimators $\widehat{p}^t : \{(Y^{t,j})_{j\in[n]}\}_{t=1}^{T} \to \mathbb{R}^d$ to minimize

$$\mathbb{E}[\|\widehat{p}^t - p^t\|_2^2], \quad \forall\, 1 \le t \le T$$

where $Y^{t,j}$ is the observed message transmitted from $X^{t,j}$

Setup

- T clusters $\{(X^{t,j})_{j\in[n]}\}_{t=1}^{T}$, each of which contains n datapoints
- Each client has one local one-hot datapoint $X^{t,j} \in [d]$ following $\mathrm{Cat}(p^t)$
- Sparse heterogeneity: there is a global distribution $p^\star$ such that

$$\|p^t - p^\star\|_0 \leq s, \quad \forall\, 1 \leq t \leq T$$

- Communication constraint: b-bits ($b \ll \log_2(d)$) budget for each client to communicate with a central server
- Goal: to design estimators $\widehat{p}^t : \{(Y^{t,j})_{j\in[n]}\}_{t=1}^{T} \to \mathbb{R}^d$ to minimize

$$\mathbb{E}[\|\widehat{p}^t - p^t\|_2^2], \quad \forall\, 1 \leq t \leq T$$

where $Y^{t,j}$ is the observed message transmitted from $X^{t,j}$

Big Picture

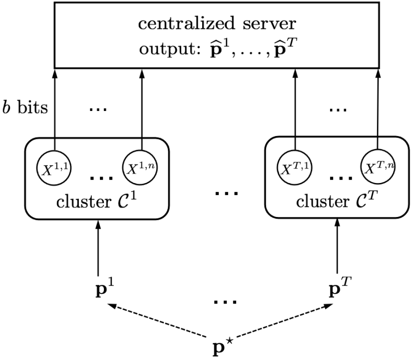


Figure: Learning distributions with heterogeneity and communication constraints

Algorithm: Uniform Hashing

Our algorithm is built upon uniform hashing:

(`Encoding`):

Send the message $Y^{t,j} = h^{t,j}(X^{t,j})$ encoded by a hash function $h^{t,j} : [d] \to [2^b]$

(`Decoding`):

Count $N_k^t(Y^{t,[n]}) = |\{j \in [n] : h^{t,j}(k) = Y^{t,j}\}|$ and return $\check{b}_k^t = N_k^t / n$

Algorithm: SHIFT

Algorithm 1 SHIFT: Sparse Heterogeneity Inspired Collaboration and Fine-Tuning

input: individual hashed estimates $\breve{b}^1, \ldots, \breve{b}^T$, threshold parameter α

▷ Stage I: Collaborative Learning

Estimate $b^\star$ via robust statistical methods: $\breve{b}^\star \leftarrow \mathrm{robust\_estimate}(\{\breve{b}^t : t \in [T]\})$

▷ Stage II: Fine-Tuning

for $k = 1, \ldots, d$ **do**

  for $t = 1, \ldots, T$ **do**

    $[\widehat{b}^t]_k \leftarrow [\breve{b}^\star]_k$ **if** $|[\breve{b}^\star]_k - [\breve{b}^t]_k| \leq \sqrt{\alpha[\breve{b}^t]_k/n}$, **else** $[\breve{b}^t]_k$

    $[\widehat{p}^t]_k \leftarrow \mathrm{Proj}_{[0,1]}(\frac{2^b[\widehat{b}^t]_k - 1}{2^b - 1})$

  end for

end for

output: estimates $\widehat{p}^1, \ldots, \widehat{p}^T$

- robust_estimate can be any robust estimator, *e.g.*, median, trimmed mean
- SHIFT requires the info. of cluster membership but is s-agnostic

Algorithm: SHIFT

Algorithm 1 SHIFT: Sparse Heterogeneity Inspired Collaboration and Fine-Tuning

input: individual hashed estimates $\breve{b}^1, \ldots, \breve{b}^T$, threshold parameter α

▷ Stage I: Collaborative Learning

Estimate $b^\star$ via robust statistical methods: $\breve{b}^\star \leftarrow \text{robust\_estimate}(\{\breve{b}^t : t \in [T]\})$

▷ Stage II: Fine-Tuning

for $k = 1, \ldots, d$ **do**

  for $t = 1, \ldots, T$ **do**

    $[\widehat{b}^t]_k \leftarrow [\breve{b}^\star]_k$ **if** $|[\breve{b}^\star]_k - [\breve{b}^t]_k| \le \sqrt{\alpha[\breve{b}^t]_k/n}$, **else** $[\breve{b}^t]_k$

    $[\widehat{p}^t]_k \leftarrow \text{Proj}_{[0,1]}\left(\frac{2^b[\widehat{b}^t]_k - 1}{2^b - 1}\right)$

  end for

end for

output: estimates $\widehat{p}^1, \ldots, \widehat{p}^T$

- robust_estimate can be any robust estimator, *e.g.*, median, trimmed mean
- SHIFT requires the info. of cluster membership but is s-agnostic

Algorithm: SHIFT

Algorithm 1 SHIFT: Sparse Heterogeneity Inspired Collaboration and Fine-Tuning

input: individual hashed estimates $\breve{b}^1, \ldots, \breve{b}^T$, threshold parameter α

▷ Stage I: Collaborative Learning

Estimate $b^\star$ via robust statistical methods: $\breve{b}^\star \leftarrow \text{robust\_estimate}(\{\breve{b}^t : t \in [T]\})$

▷ Stage II: Fine-Tuning

for $k = 1, \ldots, d$ **do**

  for $t = 1, \ldots, T$ **do**

    $[\widehat{b}^t]_k \leftarrow [\breve{b}^\star]_k$ **if** $|[\breve{b}^\star]_k - [\breve{b}^t]_k| \leq \sqrt{\alpha[\breve{b}^t]_k/n}$, **else** $[\breve{b}^t]_k$

    $[\widehat{p}^t]_k \leftarrow \text{Proj}_{[0,1]}(\frac{2^b[\widehat{b}^t]_k - 1}{2^b - 1})$

  end for

end for

output: estimates $\widehat{p}^1, \ldots, \widehat{p}^T$

- robust_estimate can be any robust estimator, *e.g.*, median, trimmed mean
- SHIFT requires the info. of cluster membership but is s-agnostic

Upper Bound: Median-based SHIFT

For the median-based SHIFT, $\mathrm{robust\_estimate}(\{\breve{b}^t : t \in [T]\})$ is taken as

$$\breve{b}_k^\star = \mathrm{median}\Big(\{\breve{b}_k^t : t \in [T]\}\Big), \quad \forall\, k \in [d].$$

Theorem

Suppose $n \geq 2^{b+6} \ln(n)$ and $\alpha = \Theta(\ln(n))$. Then, for the median-based SHIFT method, for any $1 \leq t \leq T$,

$$\mathbb{E}\left[\|\widehat{p}^t - p^t\|_2^2\right] = \tilde{O}\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n} + \frac{d}{n^2}\right).$$

Upper Bound: Median-based SHIFT

For the median-based SHIFT, $\mathrm{robust\_estimate}(\{\check{b}^t : t \in [T]\})$ is taken as

$$\check{b}_k^\star = \mathrm{median}\Big(\{\check{b}_k^t : t \in [T]\}\Big), \quad \forall\, k \in [d].$$

Theorem

Suppose $n \geq 2^{b+6} \ln(n)$ and $\alpha = \Theta(\ln(n))$. Then, for the median-based SHIFT method, for any $1 \leq t \leq T$,

$$\mathbb{E}\left[\|\widehat{p}^t - p^t\|_2^2\right] = \tilde{O}\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n} + \frac{d}{n^2}\right).$$

Upper Bound: Median-based SHIFT

- $n = 2^b \Omega\left(\ln(n), \min\left\{T, \frac{d}{\max\{2^b, s\}}\right\}\right)$ results in $\tilde{O}\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n}\right)$

- Term $\tilde{O}\left(\frac{\max\{2^b, s\}}{2^b n}\right)$ is independent of d, benefiting from sparse heterogeneity, *i.e.*, when $s \ll d$

- Term $\tilde{O}\left(\frac{d}{2^b T n}\right)$, while relating to d, is T times smaller because of smart data collaboration

- In the paper, we also show that $p^\star$ can be recovered when the heterogeneity is evenly distributed

Upper Bound: Median-based SHIFT

- $n = 2^b \Omega \left(\ln(n), \min \left\{ T, \frac{d}{\max\{2^b, s\}} \right\} \right)$ results in $\tilde{O} \left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n} \right)$
- Term $\tilde{O} \left(\frac{\max\{2^b, s\}}{2^b n} \right)$ is independent of d, benefiting from sparse heterogeneity, *i.e.*, when $s \ll d$
- Term $\tilde{O} \left(\frac{d}{2^b T n} \right)$, while relating to d, is T times smaller because of smart data collaboration
- In the paper, we also show that $p^\star$ can be recovered when the heterogeneity is evenly distributed

Upper Bound: Median-based SHIFT

- $n = 2^b \Omega \left(\ln(n), \min \left\{ T, \frac{d}{\max\{2^b, s\}} \right\} \right)$ results in $\tilde{O}\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n} \right)$
- Term $\tilde{O}\left(\frac{\max\{2^b, s\}}{2^b n} \right)$ is independent of d, benefiting from sparse heterogeneity, *i.e.*, when $s \ll d$
- Term $\tilde{O}\left(\frac{d}{2^b T n} \right)$, while relating to d, is T times smaller because of smart data collaboration
- In the paper, we also show that $p^\star$ can be recovered when the heterogeneity is evenly distributed

Upper Bound: Median-based SHIFT

- $n = 2^b \Omega\left(\ln(n), \min\left\{T, \frac{d}{\max\{2^b, s\}}\right\}\right)$ results in $\tilde{O}\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n}\right)$

- Term $\tilde{O}\left(\frac{\max\{2^b, s\}}{2^b n}\right)$ is independent of d, benefiting from sparse heterogeneity, *i.e.*, when $s \ll d$

- Term $\tilde{O}\left(\frac{d}{2^b T n}\right)$, while relating to d, is T times smaller because of smart data collaboration

- In the paper, we also show that $p^\star$ can be recovered when the heterogeneity is evenly distributed

Minimax Lower Bound

Theorem

For any—possibly interactive—estimation method and $1 \leq t \leq T$, we have

$$\inf_{(W^{r,[n]},\hat{p}^r)_{r\in[T]}} \sup_{\|p^r-p^\star\|_0\leq s} \mathbb{E}[\|\widehat{p}^t - p^t\|_2^2] = \Omega\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n}\right).$$

- The supremum is over all possible $p^\star$ and $\{p^r\}_{r=1}^T$ with $\|p^r - p^\star\|_0 \leq s$
- The infimum is over all possible communication mechanisms and estimates
- Our median-based SHIFT is minimax optimal

Minimax Lower Bound

Theorem

For any—possibly interactive—estimation method and $1 \leq t \leq T$, we have

$$\inf_{(W^{r,[n]},\hat{p}^r)_{r\in[T]}} \sup_{\|p^r - p^\star\|_0 \leq s} \mathbb{E}[\|\widehat{p}^t - p^t\|_2^2] = \Omega\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n}\right).$$

- The supremum is over all possible $p^\star$ and $\{p^r\}_{r=1}^T$ with $\|p^r - p^\star\|_0 \leq s$
- The infimum is over all possible communication mechanisms and estimates
- Our median-based SHIFT is minimax optimal

Minimax Lower Bound

Theorem

For any—possibly interactive—estimation method and $1 \le t \le T$*, we have*

$$\inf_{(W^{r,[n]},\hat{p}^r)_{r\in[T]}} \sup_{\|p^r - p^\star\|_0 \le s} \mathbb{E}[\|\widehat{p}^t - p^t\|_2^2] = \Omega\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n}\right).$$

- The supremum is over all possible $p^\star$ and $\{p^r\}_{r=1}^T$ with $\|p^r - p^\star\|_0 \le s$
- The infimum is over all possible communication mechanisms and estimates
- Our median-based SHIFT is minimax optimal

Minimax Lower Bound

Theorem

For any—possibly interactive—estimation method and $1 \leq t \leq T$, we have

$$\inf_{(W^{r,[n]},\hat{p}^r)_{r\in[T]}} \sup_{\|p^r - p^\star\|_0 \leq s} \mathbb{E}[\|\widehat{p}^t - p^t\|_2^2] = \Omega\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n}\right).$$

- The supremum is over all possible $p^\star$ and $\{p^r\}_{r=1}^T$ with $\|p^r - p^\star\|_0 \leq s$
- The infimum is over all possible communication mechanisms and estimates
- Our median-based SHIFT is minimax optimal

Experiments: Synthetic Data

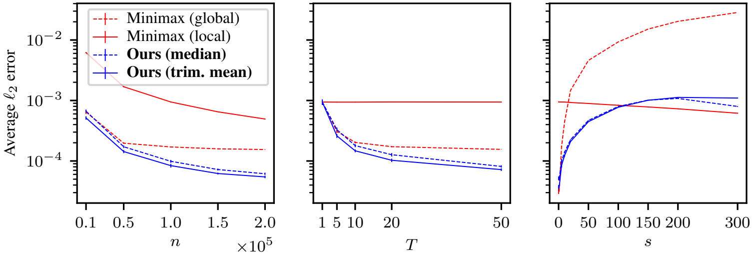


Figure: Average ℓ_2 estimation error in synthetic experiment. (Left): Fixing $s = 5$, $T = 30$ and varying n. (Middle): Fixing $s = 5$, $n = 100,000$ and varying T. (Right): Fixing $T = 30$, $n = 100,000$ and varying s.

- SHIFT outperforms the baseline methods for most choices of n, T, s
- The ℓ_2 error of SHIFT decreases as T and s increases

Experiments: Synthetic Data

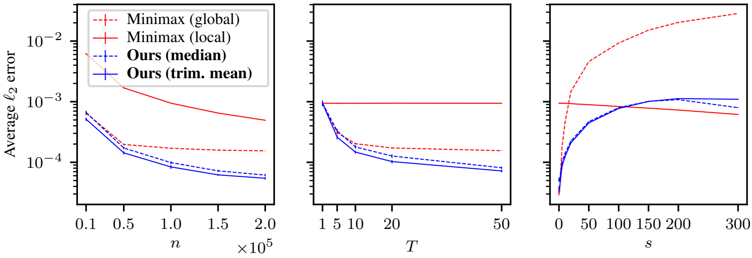


Figure: Average ℓ_2 estimation error in synthetic experiment. (Left): Fixing $s = 5$, $T = 30$ and varying n. (Middle): Fixing $s = 5$, $n = 100,000$ and varying T. (Right): Fixing $T = 30$, $n = 100,000$ and varying s.

- SHIFT outperforms the baseline methods for most choices of n, T, s
- The ℓ_2 error of SHIFT decreases as T and s increases

Experiments: Synthetic Data

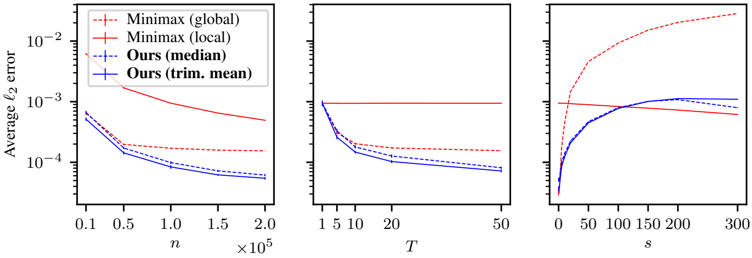


Figure: Average ℓ_2 estimation error in synthetic experiment. (Left): Fixing $s = 5$, $T = 30$ and varying n. (Middle): Fixing $s = 5$, $n = 100,000$ and varying T. (Right): Fixing $T = 30$, $n = 100,000$ and varying s.

- SHIFT outperforms the baseline methods for most choices of n, T, s
- The ℓ_2 error of SHIFT decreases as T and s increases

Experiments: Shakespeare Data

$k=2$	$b=2$	$b=4$	$b=6$	$b=8$
Minimax (local)	640 ± 6.0	142 ± 1.2	40 ± 0.40	14 ± 0.13
Minimax (global)	33 ± 1.8	17 ± 0.37	14 ± 0.081	13 ± 0.037
SHIFT (median)	47 ± 2.4	21 ± 0.66	14 ± 0.17	11 ± 0.10
SHIFT (trimean)	36 ± 2.2	19 ± 0.51	13 ± 0.24	10 ± 0.062
$k=3$	$b=2$	$b=4$	$b=6$	$b=8$
Minimax (local)	15000 ± 21	3000 ± 5.9	720 ± 2.1	180 ± 0.39
Minimax (global)	4400 ± 5.7	100 ± 1.4	38 ± 0.35	23 ± 0.090
SHIFT (median)	7300 ± 9.6	180 ± 2.1	53 ± 1.0	20 ± 0.18
SHIFT (trimean)	5100 ± 6.3	140 ± 2.3	43 ± 0.66	18 ± 0.18

Table: Average ℓ_2 error for estimating distributions of k-grams in the Shakespeare dataset. Numbers are scaled by 10^{-5}.

- SHIFT shows competitive performance on the empirical dataset, even though we do not rigorously know if the sparse heterogeneity model applies

Experiments: Shakespeare Data

$k=2$	$b=2$	$b=4$	$b=6$	$b=8$
Minimax (local)	640 ± 6.0	142 ± 1.2	40 ± 0.40	14 ± 0.13
Minimax (global)	33 ± 1.8	17 ± 0.37	14 ± 0.081	13 ± 0.037
SHIFT (median)	47 ± 2.4	21 ± 0.66	14 ± 0.17	11 ± 0.10
SHIFT (trimean)	36 ± 2.2	19 ± 0.51	13 ± 0.24	10 ± 0.062
$k=3$	$b=2$	$b=4$	$b=6$	$b=8$
Minimax (local)	15000 ± 21	3000 ± 5.9	720 ± 2.1	180 ± 0.39
Minimax (global)	4400 ± 5.7	100 ± 1.4	38 ± 0.35	23 ± 0.090
SHIFT (median)	7300 ± 9.6	180 ± 2.1	53 ± 1.0	20 ± 0.18
SHIFT (trimean)	5100 ± 6.3	140 ± 2.3	43 ± 0.66	18 ± 0.18

Table: Average ℓ_2 error for estimating distributions of k-grams in the Shakespeare dataset. Numbers are scaled by 10^{-5}.

- SHIFT shows competitive performance on the empirical dataset, even though we do not rigorously know if the sparse heterogeneity model applies

Conclusion

- Heterogeneity and communication matter in learning distributions from distributed data
- We propose the SHIFT method that leverages sparse heterogeneity smartly under communication constraints
- We provide the minimax lower bound to justify the optimality of SHIFT
- Experiments corroborate the excellent performance of SHIFT

Thanks you!