An Optimal and Scalable Matrix Mechanism for Noisy Marginals under Convex Loss Functions

Yingtai Xiao, yxx5224@psu.edu

Guanlin He, gbh5146@psu.edu

Danfeng Zhang, dbz5017@psu.edu

Daniel Kifer, duk17@psu.edu

PennState

NEURAL INFORMATION PROCESSING SYSTEMS

Differentially Private Marginals

- Marginals are tables of counts on a set of attributes
  - e.g., how many people there are for each combination of race and gender.
  - The most common formats for
    - dissemination of statistical data
    - correlations between attributes
    - sufficient statistics for Bayesian networks and Markov random fields

- Matrix Mechanism answers linear queries under differential privacy
  - Enables valid confidence intervals and hypothesis tests
  - Noisy marginals can be used to generate differentially private synthetic data

Previous SOTA: HDMM

- Select
  - Select a Gaussian linear mechanism $M(x) = Bx + N(0, \Sigma)$
  - B is a linear combination of marginals

- Measure
  - Get the noisy output $\omega = M(x)$

- Reconstruct
  - Compute an unbiased estimate of Wx
  - Least Square Estimation is very slow for HDMM
  - Only works for domain size $d \leq 10^{15}$

NEURAL INFORMATION PROCESSING SYSTEMS

PennState

Our Method: Residual Planner

- Select
  - Select a set of base mechanisms $M_A(x) = R_A x + N(0, \sigma_A^2 R_A R_A^T)$
  - R_A are mutually orthogonal, they form a linearly independent basis
  - Solve an optimization problem to get optimal noise level σ_A^2
- Measure
  - Get the noisy outputs $\omega_{\mathrm{A}} = M_{\mathrm{A}}(x)$
- Reconstruct
  - R_A^+ has a closed form expression
  - Works with domain size $\mathrm{d} = 10^{100}$

NEURAL INFORMATION PROCESSING SYSTEMS

PennState

Residual Planner Advantages

- Optimize for a wide variety of convex objective functions
  - Guaranteed to be optimal under Gaussian noise.

- It is highly scalable
  - Run in seconds even when other scalable algorithms run out of memory.

- Return the variance and covariances of the desired marginals.

Scalability

Table 1: **Time for Selection Step in seconds** on Synth$-n^d$ dataset. $n = 10$ and the number of attributes d varies. The workload consists of all marginals on ≤ 3 attributes each. Times for HDMM are reported with ± 2 standard deviations.

d	HDMM RMSE Objective	ResidualPlanner RMSE Objective	ResidualPlanner Max Variance Objective
2	0.013 ± 0.003 sec	0.001 ± 0.0008 sec	0.007 ± 0.001 sec
6	0.065 ± 0.012 sec	0.002 ± 0.0008 sec	0.009 ± 0.001 sec
10	0.639 ± 0.059 sec	0.009 ± 0.001 sec	0.018 ± 0.001 sec
12	4.702 ± 0.315 sec	0.015 ± 0.001 sec	0.028 ± 0.001 sec
14	46.054 ± 12.735 sec	0.025 ± 0.002 sec	0.041 ± 0.001 sec
15	201.485 ± 13.697 sec	0.030 ± 0.017 sec	0.050 ± 0.001 sec
20	Out of memory	0.079 ± 0.017 sec	0.123 ± 0.023 sec
30	Out of memory	0.247 ± 0.019 sec	0.461 ± 0.024 sec
50	Out of memory	1.207 ± 0.047 sec	4.011 ± 0.112 sec
100	Out of memory	9.913 ± 0.246 sec	121.224 ± 3.008 sec

Optimizing Max Variance

Table 3: Max Variance Comparisons with ResidualPlanner and HDMM (showing that being restricted to optimizing only RMSE is not a good approximation of Max Variance optimization).

	Adult Dataset		CPS Dataset		Loans Dataset	
Workload	ResPlan	HDMM	ResPlan	HDMM	ResPlan	HDMM
1-way Marginals	12.047	41.772	4.346	13.672	10.640	33.256
2-way Marginals	67.802	599.843	7.897	47.741	52.217	437.478
3-way Marginals	236.843	5675.238	7.706	71.549	156.638	3095.997
$\leq$ 3-way Marginals	253.605	6677.253	13.216	415.073	180.817	4317.709

Thank you!

PennState

NEURAL INFORMATION PROCESSING SYSTEMS