A. Appendix

A.1. Proofs

Proposition A.1. Let $\tilde{D} \in \mathcal{X}^n$ and $\widehat{D} \in \mathcal{X}^m$. Let $\mathcal{Q}$ be a finite set of statistical queries $q : \mathcal{X} \to [0,1]$. Let $\tilde{\varepsilon} > 0$ and $T \in \mathbb{N}$. Let A be the output of Algorithm 1 with parameters $\tilde{\varepsilon}$ and T, query class $\mathcal{Q}$, and inputs $\tilde{D}$ as the private dataset and $\widehat{D}$ as the public dataset. Then A is a distribution on $\widehat{\mathcal{X}} = \mathrm{supp}(\widehat{D}) \subset \mathcal{X}$. For all $\beta \in (0,1)$, if $T \geq 7\log(3/\beta)$, then

$$\Pr\left[\begin{array}{r} \max_{q\in\mathcal{Q}} |q(\widetilde{D}) - q(A)| \leq 2f_{\widetilde{D},\mathcal{Q}}(\widehat{\mathcal{X}}) + \sqrt{\frac{4\log m}{T} + \frac{4T}{\tilde{\varepsilon}^2 n^2} + \frac{4\sqrt{\log(3/\beta)}}{\tilde{\varepsilon} n}} \\ + \frac{2\sqrt{2T}}{\tilde{\varepsilon} n}\log|\mathcal{Q}| + \sqrt{\frac{1}{2T}\log\left(\frac{3}{\beta}\right)} \end{array}\right] \geq 1-\beta.$$

If we set $T = \Theta\left(\frac{\tilde{\varepsilon} n\sqrt{\log m}}{\log|\mathcal{Q}|} + \log(1/\beta)\right)$, then the bound above becomes

$$\Pr\left[\max_{q\in\mathcal{Q}} |q(\widetilde{D}) - q(A)| \leq O\left(f_{\widetilde{D},\mathcal{Q}}(\widehat{\mathcal{X}}) + \sqrt{\frac{\log|\mathcal{Q}|}{\tilde{\varepsilon} n}\cdot\left(\sqrt{\log m} + \log(1/\beta)\right)}\right)\right] \geq 1-\beta,$$

thus proving Theorem 4.4.

Proof. We follow the analysis of Hardt et al. (2012). Let A_t, q_t, a_t be as in Algorithm 1. Let $\alpha_0 = f_{\widetilde{D},\mathcal{Q}}(\widehat{\mathcal{X}})$ be the error of the optimal reweighting of the public data. Let

$$D^* = \arg\min_{D\in\Delta(\widehat{\mathcal{X}})} \max_{q\in\mathcal{Q}} |q(D) - q(\widetilde{D})|$$

be the optimal reweighting so that $\max_{q\in\mathcal{Q}} |q(D^*) - q(\widetilde{D})| = \alpha_0$. We define a potential function $\Psi : \Delta(\widehat{\mathcal{X}}) \to \mathbb{R}$ by

$$\Psi(A) = \mathrm{D}_1(D^*\|A) = \sum_{x\in\widehat{\mathcal{X}}} D^*(x)\log\left(\frac{D^*(x)}{A(x)}\right).$$

Since Ψ is a KL divergence, it follows that, for all $A \in \Delta(\widehat{\mathcal{X}})$,

$$0 \leq \Psi(A) \leq \log\left(\frac{1}{\min_{x\in\widehat{\mathcal{X}}} A(x)}\right).$$

In particular, $\Psi(A_T) \geq 0$ and $\Psi(A_0) \leq \log m$, since any $x \in \widehat{\mathcal{X}}$ must be one of the m elements of $\widehat{D}$ and hence has $A_0(x) \geq 1/m$.

Fix an arbitrary $t \in [T]$. For all $x \in \widehat{\mathcal{X}}$, we have $A_t(x) = \frac{A_{t-1}(x)\exp(q_t(x)(a_t - q_t(A_{t-1}))/2)}{\sum_{y\in\widehat{\mathcal{X}}} A_{t-1}(y)\exp(q_t(y)(a_t - q_t(A_{t-1}))/2)}$. Thus

$$
\begin{aligned}
&\Psi(A_{t-1}) - \Psi(A_t) \\
&= \sum_{x\in\widehat{\mathcal{X}}} D^*(x)\log\left(\frac{A_t(x)}{A_{t-1}(x)}\right) \\
&= \sum_{x\in\widehat{\mathcal{X}}} D^*(x)\log\left(\frac{\exp(q_t(x)(a_t - q_t(A_{t-1}))/2)}{\sum_{y\in\widehat{\mathcal{X}}} A_{t-1}(y)\exp(q_t(y)(a_t - q_t(A_{t-1}))/2)}\right) \\
&= \sum_{x\in\widehat{\mathcal{X}}} D^*(x)q_t(x)\frac{a_t - q_t(A_{t-1})}{2} - \log\left(\sum_{y\in\widehat{\mathcal{X}}} A_{t-1}(y)\exp\left(q_t(y)\frac{a_t - q_t(A_{t-1})}{2}\right)\right) \\
&\geq q_t(D^*)\frac{a_t - q_t(A_{t-1})}{2} + 1 - \sum_{y\in\widehat{\mathcal{X}}} A_{t-1}(y)\exp\left(q_t(y)\frac{a_t - q_t(A_{t-1})}{2}\right) \qquad (\forall x > 0 \;\; \log x \leq x - 1) \\
&\geq q_t(D^*)\frac{a_t - q_t(A_{t-1})}{2} + 1 - \sum_{y\in\widehat{\mathcal{X}}} A_{t-1}(y)\left(1 + q_t(y)\frac{a_t - q_t(A_{t-1})}{2} + q_t(y)^2\frac{(a_t - q_t(A_{t-1}))^2}{4}\right) \\
&\qquad\qquad (\forall x \leq 1 \;\; \exp(x) \leq 1 + x + x^2) \\
&= q_t(D^*)\frac{a_t - q_t(A_{t-1})}{2} + 1 - 1 - q_t(A_{t-1})\frac{a_t - q_t(A_{t-1})}{2} - \mathop{\mathbb{E}}_{X\leftarrow A_{t-1}}\left[q_t(X)^2\right]\frac{(a_t - q_t(A_{t-1}))^2}{4} \\
&= (q_t(D^*) - q_t(A_{t-1}))\frac{a_t - q_t(A_{t-1})}{2} - \mathop{\mathbb{E}}_{X\leftarrow A_{t-1}}\left[q_t(X)^2\right]\frac{(a_t - q_t(A_{t-1}))^2}{4} \\
&\geq (q_t(D^*) - q_t(A_{t-1}))\frac{a_t - q_t(A_{t-1})}{2} - \frac{(a_t - q_t(A_{t-1}))^2}{4} \\
&= \frac{1}{4}(2q_t(D^*) - a_t - q_t(A_{t-1}))(a_t - q_t(A_{t-1})) \\
&= \frac{1}{4}(q_t(\widetilde{D}) - q_t(A_{t-1}))^2 + \frac{1}{2}(q_t(D^*) - q_t(\widetilde{D}))(a_t - q_t(A_{t-1})) - \frac{1}{4}(a_t - q_t(\widetilde{D}))^2 \\
&= \frac{1}{4}(q_t(\widetilde{D}) - q_t(A_{t-1}))^2 + \frac{1}{2}(q_t(D^*) - q_t(\widetilde{D}))(q_t(\widetilde{D}) - q_t(A_{t-1})) \\
&\qquad + \frac{1}{2}(q_t(D^*) - q_t(\widetilde{D}))(a_t - q_t(\widetilde{D})) - \frac{1}{4}(a_t - q_t(\widetilde{D}))^2 \\
&\geq \frac{1}{4}(q_t(\widetilde{D}) - q_t(A_{t-1}))^2 - \frac{1}{2}\alpha_0|q_t(\widetilde{D}) - q_t(A_{t-1})| \\
&\qquad + \frac{1}{2}(q_t(D^*) - q_t(\widetilde{D}))(a_t - q_t(\widetilde{D})) - \frac{1}{4}(a_t - q_t(\widetilde{D}))^2,
\end{aligned}
$$

where the final inequality follows from the fact that $|q_t(D^*) - q_t(\widetilde{D})| \leq \alpha_0$ by the definition of D^*.

Putting together what we have so far gives

$$\begin{aligned}
\frac{2}{T}\log m &\geq \frac{2}{T}\left(\Psi(A_0) - \Psi(A_T)\right) \\
&= \frac{2}{T}\sum_{t\in[T]} \Psi(A_{t-1}) - \Psi(A_t) \\
&\geq \frac{2}{T}\sum_{t\in[T]} \frac{1}{4}(q_t(\widetilde{D}) - q_t(A_{t-1}))^2 - \frac{2}{T}\sum_{t\in[T]} \frac{1}{2}\alpha_0 |q_t(\widetilde{D}) - q_t(A_{t-1})| \\
&\qquad + \frac{2}{T}\sum_{t\in[T]} \frac{1}{2}(q_t(D^*) - q_t(\widetilde{D}))(a_t - q_t(\widetilde{D})) - \frac{2}{T}\sum_{t\in[T]} \frac{1}{4}(a_t - q_t(\widetilde{D}))^2 \\
&\geq \frac{1}{2}\left(\frac{1}{T}\sum_{t\in[T]} |q_t(\widetilde{D}) - q_t(A_{t-1})|\right)^2 - \frac{\alpha_0}{T}\sum_{t\in[T]} |q_t(\widetilde{D}) - q_t(A_{t-1})| \\
&\qquad + \frac{1}{T}\sum_{t\in[T]} (q_t(D^*) - q_t(\widetilde{D}))(a_t - q_t(\widetilde{D})) - \frac{1}{2T}\sum_{t\in[T]} (a_t - q_t(\widetilde{D}))^2,
\end{aligned}$$

where the final inequality uses the relationship between the 1-norm and 2-norm.

Now, for each $t \in [T]$ independently, $a_t - q_t(\widetilde{D})$ is distributed according to $\mathcal{N}(0, 1/\varepsilon_0^2 n^2)$. Thus the sum $\sum_{t\in[T]}(a_t - q_t(\widetilde{D}))^2$ follows a chi-square distribution with T degrees of freedom and mean $\frac{T}{\varepsilon_0^2 n^2}$. This yields the tail bound

$$\forall \kappa \geq 1 \quad \Pr\left[\sum_{t\in[T]} (a_t - q_t(\widetilde{D}))^2 \geq \kappa \cdot \frac{T}{\varepsilon_0^2 n^2}\right] \leq \left(\kappa \cdot e^{1-\kappa}\right)^{T/2}.$$

In addition, the noise $a_t - q_t(\widetilde{D})$ is independent from $q_t(D^*) - q_t(\widetilde{D})$. Hence, the sum $\sum_{t\in[T]}(q_t(D^*) - q_t(\widetilde{D}))(a_t - q_t(\widetilde{D}))$ follows a σ^2-subgaussian distribution with $\sigma^2 = \frac{1}{\varepsilon_0^2 n^2}\sum_{t\in[T]}(q_t(D^*) - q_t(\widetilde{D}))^2 \leq \frac{T\alpha_0^2}{\varepsilon_0^2 n^2}$. In particular,

$$\forall \lambda \geq 0 \quad \Pr\left[\sum_{t\in[T]} (q_t(D^*) - q_t(\widetilde{D}))(a_t - q_t(\widetilde{D})) \geq \lambda \frac{\alpha_0\sqrt{T}}{\varepsilon_0 n}\right] \leq e^{-\lambda^2/2}.$$

Set $V := \frac{1}{T}\sum_{t\in[T]} |q_t(\widetilde{D}) - q_t(A_{t-1})|$.

Thus, for all $\kappa, \lambda \geq 0$,

$$\Pr\left[\frac{1}{2}V^2 - \alpha_0 V \leq \frac{2\log m}{T} + \frac{\kappa}{2\varepsilon_0^2 n^2} + \frac{\lambda\alpha_0}{\varepsilon_0 n\sqrt{T}}\right] \geq 1 - \left(\kappa \cdot e^{1-\kappa}\right)^{T/2} - e^{-\lambda^2/2}.$$

The above expression contains the quadratic inequality $\frac{1}{2}V^2 - \alpha_0 V \leq \frac{2\log m}{T} + \frac{\kappa}{2\varepsilon_0^2 n^2} + \frac{\lambda\alpha_0}{\varepsilon_0 n\sqrt{T}}$. This equation implies

$$V \leq \alpha_0 + \sqrt{\alpha_0^2 + \frac{4\log m}{T} + \frac{\kappa}{\varepsilon_0^2 n^2} + \frac{2\lambda\alpha_0}{\varepsilon_0 n\sqrt{T}}} \leq 2\alpha_0 + \sqrt{\frac{4\log m}{T} + \frac{\kappa}{\varepsilon_0^2 n^2} + \frac{2\lambda\alpha_0}{\varepsilon_0 n\sqrt{T}}}.$$

Now we invoke the properties of the permute-and-flip or exponential mechanism that selects q_t. For each $t \in [T]$, we have (**?**)Lemma 7.1]BassilyNSSSU16

$$\mathop{\mathbb{E}}_{q_t}\left[|q_t(\widetilde{D}) - q_t(A_{t-1})|\right] \geq \max_{q\in\mathcal{Q}} |q(\widetilde{D}) - q(A_{t-1})| - \frac{2}{\varepsilon_0 n}\log|\mathcal{Q}|.$$

Since $0 \leq |q_t(\widetilde{D}) - q_t(A_{t-1})| \leq 1$, we can apply Azuma's inequality to obtain

$$\Pr\left[\frac{1}{T}\sum_{t\in[T]} |q_t(\widetilde{D}) - q_t(A_{t-1})| \geq \frac{1}{T}\sum_{t\in[T]} \max_{q\in\mathcal{Q}} |q(\widetilde{D}) - q(A_{t-1})| - \frac{2}{\varepsilon_0 n}\log|\mathcal{Q}| - \nu\right] \geq 1 - e^{-2\nu^2 T}$$

for all $\nu \geq 0$. Finally, for $A = \frac{1}{T}\sum_{t\in[T]} A_{t-1}$, we have

$$\begin{aligned}
\max_{q\in\mathcal{Q}} |q(\widetilde{D}) - q(A)| &\leq \frac{1}{T}\sum_{t\in[T]} \max_{q\in\mathcal{Q}} |q(\widetilde{D}) - q(A_{t-1})| \\
&\leq \frac{1}{T}\sum_{t\in[T]} |q_t(\widetilde{D}) - q_t(A_{t-1})| + \frac{2}{\varepsilon_0 n}\log|\mathcal{Q}| + \nu \qquad \text{(with probability } \geq 1 - e^{-2\nu^2 T}\text{)} \\
&\leq 2\alpha_0 + \sqrt{\frac{4\log m}{T} + \frac{\kappa}{\varepsilon_0^2 n^2} + \frac{2\lambda\alpha_0}{\varepsilon_0 n\sqrt{T}}} + \frac{2}{\varepsilon_0 n}\log|\mathcal{Q}| + \nu. \\
&\qquad\qquad \text{(with probability } \geq 1 - \left(\kappa\cdot e^{1-\kappa}\right)^{T/2} - e^{-\lambda^2/2}\text{)}
\end{aligned}$$

Now we set $\nu = \sqrt{\frac{1}{2T}\log\left(\frac{3}{\beta}\right)}$, $\kappa = 2$, and $\lambda = \sqrt{2\log(3/\beta)}$ and apply a union bound. If $T \geq 7\log(3/\beta)$, then

$$\Pr\left[\begin{array}{r} \max_{q\in\mathcal{Q}} |q(\widetilde{D}) - q(A)| \leq 2\alpha_0 + \sqrt{\frac{4\log m}{T} + \frac{2}{\varepsilon_0^2 n^2} + \frac{2\sqrt{2\log(3/\beta)}\alpha_0}{\varepsilon_0 n\sqrt{T}}} \\ + \frac{2}{\varepsilon_0 n}\log|\mathcal{Q}| + \sqrt{\frac{1}{2T}\log\left(\frac{3}{\beta}\right)} \end{array}\right] \geq 1-\beta.$$

Substituting in $\alpha_0 = f_{\widetilde{D},\mathcal{Q}}(\widehat{\mathcal{X}}) \leq 1$ and $\varepsilon_0 = \frac{\tilde{\varepsilon}}{\sqrt{2T}}$ yields the result. □

We remark that the proof above uses the bound $\Psi(A_0) = \mathrm{D}_1\left(D^* \middle\| \widehat{D}\right) \leq \log m$. This is tight in the worst case, but is likely to be loose in practice, as the private and public datasets are likely to be relatively similar. We could also alter Algorithm 1 to initialize A_0 to be uniform on $\widehat{\mathcal{X}}$, in which case we can replace $\log m$ with $\log|\widehat{\mathcal{X}}|$ in the final bound.

Lemma A.2. For any support $S \in 2^{\mathcal{X}}$ and set of linear queries Q, the best mixture error function $f_{D,Q}$ is $\frac{1}{n}$ sensitive. That is for any pair of neighboring datasets D, D' of size n, $|f_{D,Q}(S) - f_{D',Q}(S)| \leq \frac{1}{n}$.

Proof. First, we show that the maximum of s-sensitive functions is an s-sensitive function and by symmetry the minimum of s-sensitive functions is s-sensitive. For any $s \leq 1$, let $G = \{g : \mathcal{X} \to [0,1]\}$ be a class of s-sensitive functions and define a function $f : \mathcal{X} \to [0,1]$ as $f(X) = \max_{g\in G} g(X)$, for $X \in \mathcal{X}$.

Fix any support $S \in 2^{\mathcal{X}}$ and neighboring dataset D, D' with size n. Also fix the set Q and note each query $q \in Q$ is bounded in [0,1] and it's $\frac{1}{n}$-sensitive. Let $g' = \arg\max_{g\in G} g(D')$ and $g = \arg\max_{g\in G} g(D)$, then for neighboring D, D' we have

$$\begin{aligned}
f(D) - f(D') &\leq f(D) - g(D') && \text{Since } f(D') \geq g(D') \\
&\leq f(D) - g(D) + s && \text{Since } |g(D) - g(D')| \leq s \\
&= s && \text{Since } f(D) = g(D)
\end{aligned}$$

Similarly, we can show that $f(D') - f(D) \leq s$, therefore f is s-sensitive.

Since a marginal query $q \in \mathcal{Q}$, is $\frac{1}{n}$-sensitive, after fixing any μ the expression

$$\max_{q\in Q}\left|q(D) - \sum_{x\in S}\mu_x q(x)\right|$$

is a max of $\frac{1}{n}$ sensitive functions, then by the argument above it is a $\frac{1}{n}$-sensitive function. It follows that $f_{D,Q}(S)$ is a minimum of $\frac{1}{n}$-sensitive functions therefore $f_{D,Q}(S)$ is $\frac{1}{n}$-sensitive.

□

A.2. Data

Attributes for our experiments on ACS, ACS (reduced), and ADULT:

- **ACS:** ACREHOUS, AGE, AVAILBLE, CITIZEN, CLASSWKR, DIFFCARE, DIFFEYE, DIFFHEAR, DIFFMOB, DIFFPHYS, DIFFREM, DIFFSENS, DIVINYR, EDUC, EMPSTAT, FERTYR, FOODSTMP, GRADEATT, HCOVANY, HCOVPRIV, HINSCAID, HINSCARE, HINSVA, HISPAN, LABFORCE, LOOKING, MARRINYR, MARRNO, MARST, METRO, MIGRATE1, MIGTYPE1, MORTGAGE, MULTGEN, NCHILD, NCHLT5, NCOUPLES, NFATHERS, NMOTHERS, NSIBS, OWNERSHP, RACAMIND, RACASIAN, RACBLK, RACE, RACOTHER, RACPACIS, RACWHT, RELATE, SCHLTYPE, SCHOOL, SEX, SPEAKENG, VACANCY, VEHICLES, VET01LTR, VET47X50, VET55X64, VET75X90, VET90X01, VETDISAB, VETKOREA, VETSTAT, VETVIETN, VETWWII, WIDINYR, WORKEDYR
- **ACS (reduced):** DIFFEYE, DIFFHEAR, EMPSTAT, FOODSTMP, HCOVPRIV, HINSCAID, HINSCARE, OWNERSHP, RACAMIND, RACASIAN, RACBLK, RACOTHER, RACPACIS, RACWHT, SEX
- **ADULT:** sex, income>50K, race, relationship, marital-status, workclass, occupation, education-num, native-country, capital-gain, capital-loss, hours-per-week, age

In addition, we discretize the following continuous attributes (with the number of bins after preprocessing) into categorical attributes:

- **ACS:** AGE (10)
- **ACS (reduced):** AGE (10)
- **ADULT:** capital-gain (16), capital-loss (6), hours-per-week (10), age (10)

A.3. Hyperparameters

We report hyperparameters used across all experiments in Table 2.

Table 2: Hyperparameter selection for experiments on all datasets.

Method	Parameter	Values
PMW$^{\text{Pub}}$	T	300, 250, 200, 150, 125, 100, 75, 50, 25, 10, 5
MWEM	T	300, 250, 200, 150, 125, 100, 75, 50, 25, 10, 5
DualQuery	samples η	500 250 100 50 5 4 3 2

A.4. Experimental results

A.4.1. Evaluations using mean error and RMSE

In Figures 4 and 5, we evaluate on ACS PA-18 and ACS GA-18 with respect to mean error and mean squared error respectively. Like in Figure 1, in which we present results with respect to max error, PMW$^{\text{Pub}}$ performs well when using public datasets with low best mixture error. In the case where PMW$^{\text{Pub}}$ uses a poor dataset (i.e., CA-18), we observe the performance of PMW$^{\text{Pub}}$ suffers, and the algorithm does not improve even when the privacy budget is increased. However, we note that PMW$^{\text{Pub}}$ still performs better than DualQuery in this setting (where $\varepsilon \leq 1.0$).

A.4.2. Additional results

In Figure 6, we plot results for ACS PA-18 and ACS GA-18 comparing PMW$^{\text{Pub}}$ using the 2010 ACS data (PA-10 and GA-10) with the remaining public datasets (TX-18, FL-18, IL-18) not presented in Figure 1. In addition, we present results

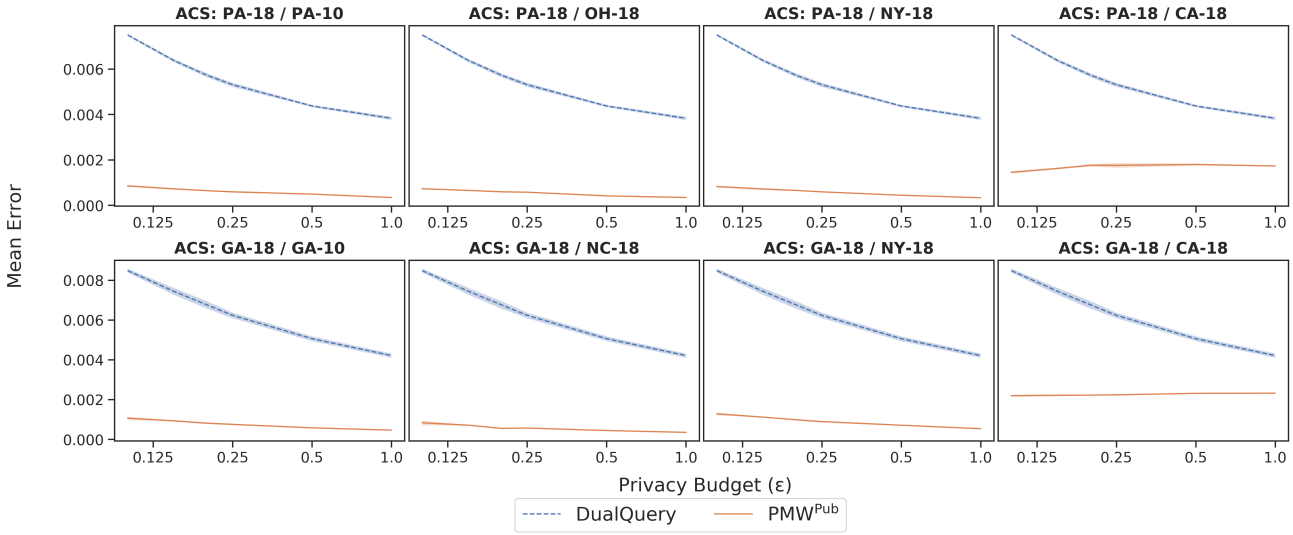

Figure 4: Mean error for $\varepsilon \in \{0.1, 0.15, 0.2, 0.25, 0.5, 1\}$ and $\delta = \frac{1}{n^2}$, evaluated on 3-way marginals with a workload size of 4096. Results are averaged over 5 runs, and error bars represent one standard error. The *x-axis* uses a logarithmic scale. **Top row:** 2018 ACS for Pennsylvania. **Bottom row:** 2018 ACS for Georgia.

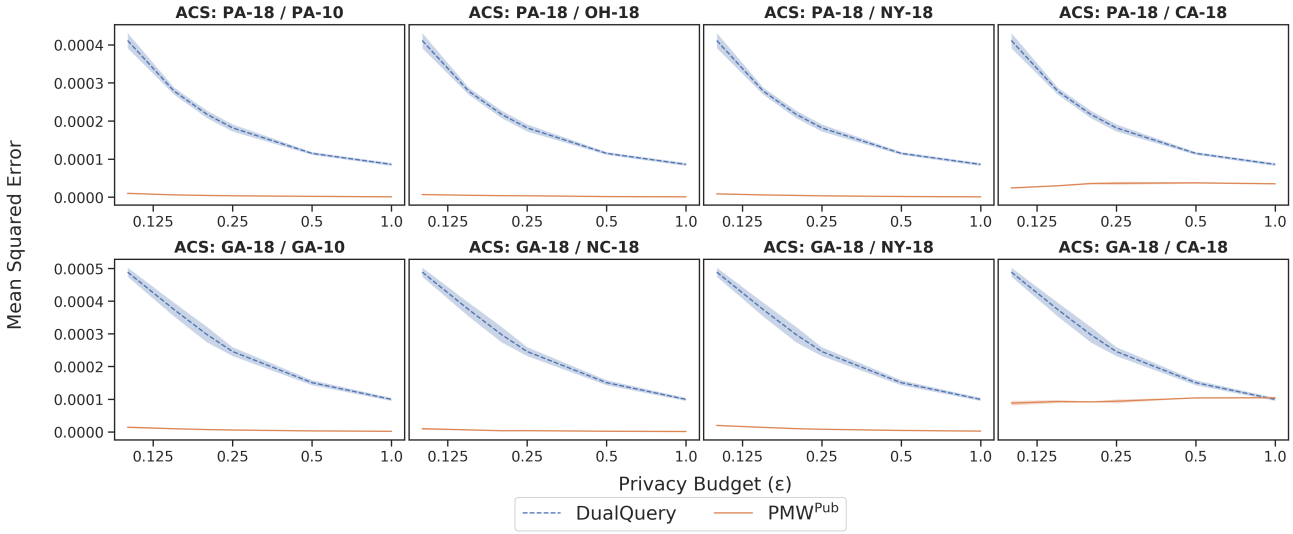

Figure 5: Mean squared error for $\varepsilon \in \{0.1, 0.15, 0.2, 0.25, 0.5, 1\}$ and $\delta = \frac{1}{n^2}$, evaluated on 3-way marginals with a workload size of 4096. Results are averaged over 5 runs, and error bars represent one standard error. The *x-axis* uses a logarithmic scale. **Top row:** 2018 ACS for Pennsylvania. **Bottom row:** 2018 ACS for Georgia.

on 2018 ACS data for the states of New York (NY-18) and California (CA-18). To run PMW$^{\text{Pub}}$, we choose Texas (TX-18), Florida (FL-18), and Illinois (IL-18) for New York and choose Texas (TX-18), Nevada (NV-18), and New York (NY-18) for California.

A.4.3. Using the last iterate

In this work, we present theoretical guarantees of PMW$^{\text{Pub}}$ in which we output the average distribution $A = \text{avg}_{t \le T} A_t$ (see Algorithm 1), mimicking the output in the original formulation of MWEM. However, Hardt et al. (2012) note that while they prove guarantees for this variant of MWEM, in practical settings, one can often achieve better results by outputting the distribution from the last iterate, A_T. In Figure 8, we compare PMW$^{\text{Pub}}$ to the variant of PMW$^{\text{Pub}}$ that outputs A_T and observe that indeed, outputting the last iterate achieves better performance across all experiments (excluding those in which

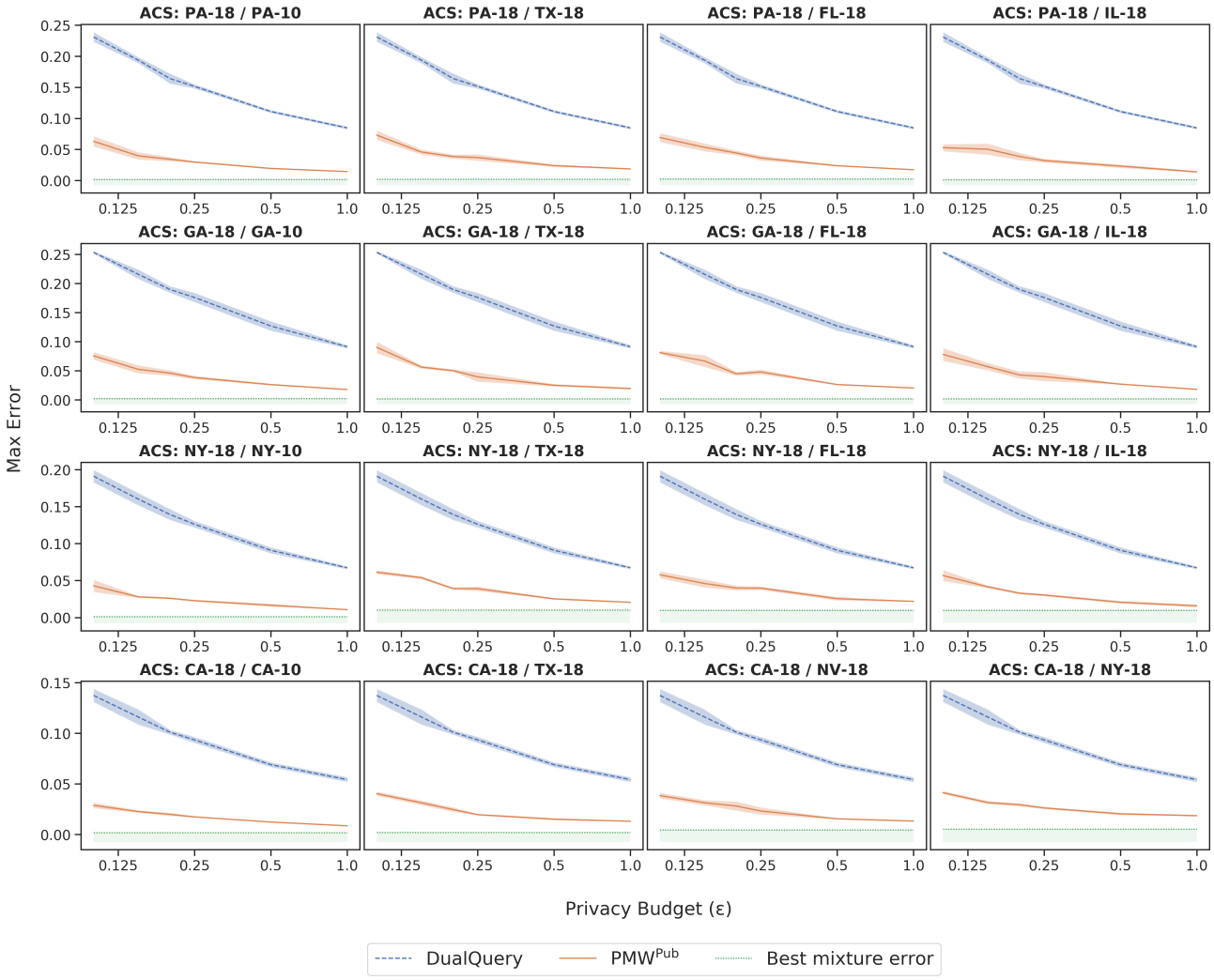

Figure 6: Additional plots of the max error (3-way marginals and workload size of 4096) for $\varepsilon \in \{0.1, 0.15, 0.2, 0.25, 0.5, 1\}$ and $\delta = \frac{1}{n^2}$ on PA-18 (Row 1), GA-18 (Row 2), NY-18 (Row 3), and CA-18 (Row 4). Results are averaged over 5 runs, and error bars represent one standard error. The *x-axis* uses a logarithmic scale. Given the support of each public dataset, we shade the area below the *best mixture error* to represent max error values that are unachievable by PMW$^{\text{Pub}}$.

Table 3: Max error (averaged over 5 runs, best results in **bold**) comparison on the 2018 ACS (reduced)-PA, 2018 ACS-PA, and 2018 ACS-GA datasets. At each privacy budget parametrized by $\varepsilon \in \{0.1, 0.15, 0.2, 0.25, 0.5, 1\}$ and $\delta = \frac{1}{n^2}$, PMW$^{\text{Pub}}$ uses the public dataset (and corresponding hyperparameter T) that achieves the lowest max error on the validation set.

Dataset	Algo.	$\varepsilon = 0.1$	$\varepsilon = 0.15$	$\varepsilon = 0.2$	$\varepsilon = 0.25$	$\varepsilon = 0.5$	$\varepsilon = 1$
ACS (red.)-PA	PMW$^{\text{Pub}}$	**0.0301**	**0.0197**	**0.0196**	**0.0172**	**0.0097**	**0.0067**
	DualQuery	0.1115	0.0871	0.0816	0.0625	0.0473	0.0330
ACS-PA	PMW$^{\text{Pub}}$	**0.0499**	**0.0458**	**0.0332**	**0.0298**	**0.0195**	**0.0141**
	DualQuery	0.2289	0.1908	0.1639	0.1526	0.1086	0.0816
ACS-GA	PMW$^{\text{Pub}}$	**0.0753**	**0.0523**	**0.0470**	**0.0380**	**0.0244**	**0.0175**
	DualQuery	0.2615	0.2117	0.1904	0.1709	0.1212	0.0910

the best mixture error of the public dataset's support is high, i.e. CA-18).

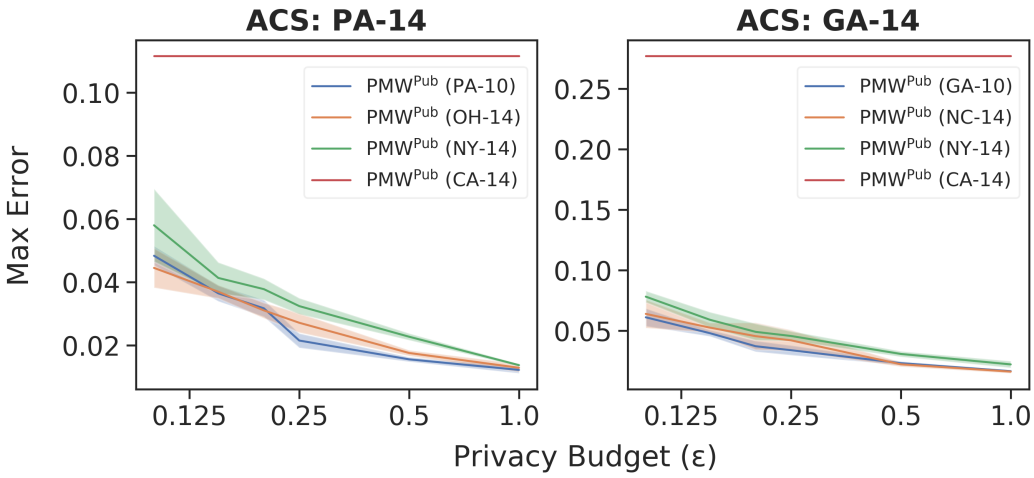

Figure 7: Max error on the ACS validation sets for 3-way marginals with a workload size of 4096 with privacy $\varepsilon \in \{0.1, 0.15, 0.2, 0.25, 0.5, 1\}$ and $\delta = \frac{1}{n^2}$. Results are averaged over 5 runs, and error bars represent one standard error. The *x-axis* uses a logarithmic scale. **Left:** 2014 ACS for Pennsylvania. **Right:** 2014 ACS for Georgia.

A.4.4. Identifying public datasets with poor support

In Section 5.5.1, we describe how using Laplace noise, one can get determine the quality of a support by getting a noisy estimate of the best mixture error for any public dataset. While we emphasize that this strategy is the most principled approach to ensuring the public data is viable for PMW$^{\text{Pub}}$, we note that in settings like ours in which we have a validation set, one can apply additional sanity checks. For instance, in Figure 7, we observe that PMW$^{\text{Pub}}$ performs poorly on the validation set when using CA-14, both in absolute terms and relative to the other public datasets. For demonstration purposes, we show in Table 3 that if we select the public dataset (at each privacy budget ε) based solely on which public dataset performed best on the validation set, we achieve very strong results. Thus in practical settings, one can use validation sets in conjunction with the best mixture error function to find a suitable public dataset (for example, one can first filter out poor public datasets using a validation set and then find the best mixture error of any remaining candidates).

A.5. Discussion of other baselines

Ji & Elkan (2013). While Ji & Elkan (2013)'s method reweights the support of a public dataset, their goal is not tailored towards query release. Ji & Elkan (2013) instead measure the success their algorithm by evaluating the parameters learned from training a support vector machine on the synthetic dataset. Furthermore, they specifically contrast their method's goal with that of MWEM, whose objective is to optimize over a set of predefined queries. Thus, one would expect the method to perform worse than PMW$^{\text{Pub}}$ for query release. To verify this hypothesis, we implement their algorithm with hyperparameter $\lambda \in \{0.005, 0.01, 0.025, 0.05, 0.1, 0.5\}$. PMW$^{\text{Pub}}$ outperforms across all metrics (max, mean, and mean squared error) and privacy budgets on ACS PA-18 and GA-18 using each public dataset used to evaluate PMW$^{\text{Pub}}$ in Section 5.5.1. For example, w.r.t. max error on PA-18, PMW$^{\text{Pub}}$ outperforms by between $1.38\times$ and $5.07\times$ (depending on ε) when using PA-10 as the public dataset.

HDMM. Unlike MWEM and DualQuery, which solve the query release problem by generating synthetic data, the High-Dimensional Matrix Mechanism (McKenna et al., 2018) is designed to directly answer a workload of queries. By representing query workloads compactly, HDMM selects a new set of "strategy" queries that minimize the estimated error with respect to the input workload. The algorithm then answers the "strategy" queries using the *Laplace mechanism* and reconstructs the answers to the input workload queries using these noisy measurements.

We had originally evaluated against HDMM. However, having consulted McKenna et al. (2018), we learned that currently, running HDMM is infeasible for the ACS and ADULT datasets. There does not exist a way to solve the least square problem described in the paper for domain sizes larger than approximately 10^9. While a variant of HDMM using local least squares could potentially circumvent such computational issues, there does not exist support for this version of HDMM for more

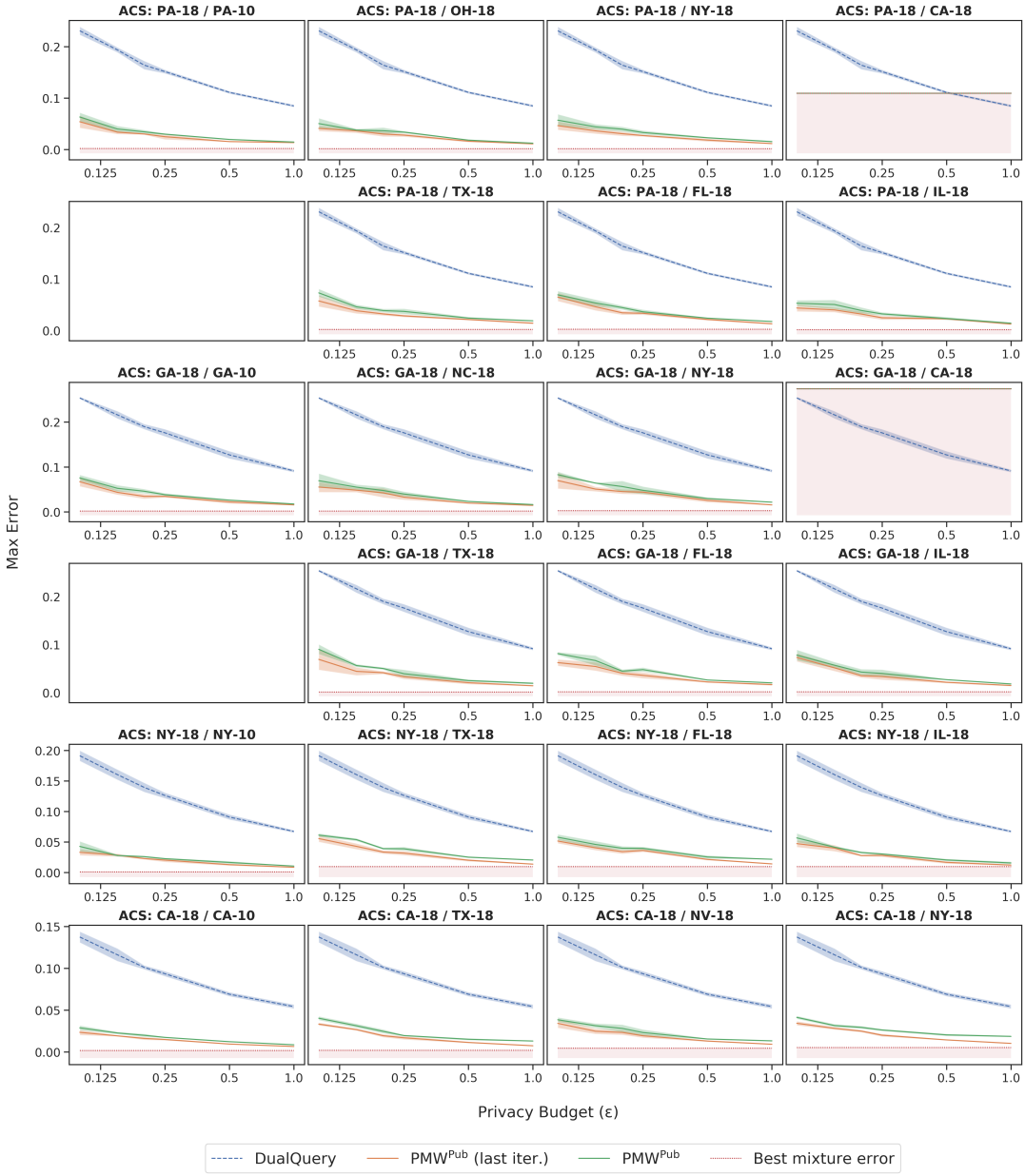

Figure 8: We compare PMW$^{\text{Pub}}$ with the variant of PMW$^{\text{Pub}}$ that outputs the last iterate A_T for all experiments (3-way marginals and workload size of 4096) on the (full-sized) 2018 ACS dataset, plotting max error for $\varepsilon \in \{0.1, 0.15, 0.2, 0.25, 0.5, 1\}$ and $\delta = \frac{1}{n^2}$. Results are averaged over 5 runs, and error bars represent one standard error. The *x-axis* uses a logarithmic scale. Given the support of each public dataset, we shade the area below the *best mixture error* to represent max error values that are unachievable by PMW$^{\text{Pub}}$. Using the last iterate in PMW$^{\text{Pub}}$ improves performance across all experiments.

general query workloads outside of those predefined in McKenna et al. (2018)'s codebase. Currently, there is no timeline for when the authors will begin developing this modification that would allow us to use HDMM as a baseline for the ADULT and ACS datasets.