Journal of Machine Learning Research 26 (2025) 1-65 Submitted 10/23; Revised 12/24; Published 4/25

# On Consistent Bayesian Inference from Synthetic Data

**Ossi Räisä** OSSI.RAISA@HELSINKI.FI
**Joonas Jälkö** JOONAS.JALKO@HELSINKI.FI
**Antti Honkela** ANTTI.HONKELA@HELSINKI.FI
*Department of Computer Science*
*University of Helsinki*
*P.O. Box 68 (Pietari Kalmin katu 5)*
*00014 University of Helsinki, Finland*

**Editor:** Mohammad Emtiyaz Khan

## Abstract

Generating synthetic data, with or without differential privacy, has attracted significant attention as a potential solution to the dilemma between making data easily available, and the privacy of data subjects. Several works have shown that consistency of downstream analyses from synthetic data, including accurate uncertainty estimation, requires accounting for the synthetic data generation. There are very few methods of doing so, most of them for frequentist analysis. In this paper, we study how to perform consistent Bayesian inference from synthetic data. We prove that mixing posterior samples obtained separately from multiple large synthetic data sets, that are sampled from a posterior predictive, converges to the posterior of the downstream analysis under standard regularity conditions when the analyst's model is compatible with the data provider's model. We also present several examples showing how the theory works in practice, and showing how Bayesian inference can fail when the compatibility assumption is not met, or the synthetic data set is not significantly larger than the original.

**Keywords:** synthetic data, Bayesian inference, Bernstein-von Mises theorem, differential privacy

## 1. Introduction

Synthetic data has the potential of opening privacy-sensitive data sets for widespread analysis. The idea is to train a generative model with real data, and release synthetic data that has been generated from the model. The synthetic data does not contain records from real people, and ideally it preserves the population-level properties of the real data, making it useful for analysis. Synthetic data can remain vulnerable to modern privacy attacks (van Breugel et al., 2023b), but they can be provably mitigated with *differential privacy* (DP; Dwork et al., 2006b).

The most convenient and straightforward way for downstream analysts to analyse synthetic data is using the same method that would be used with real data. However, ignoring the additional stochasticity arising from the synthetic data generation will yield biased results and overconfident uncertainty estimates (Raghunathan et al., 2003; Wilde et al., 2021; Räisä et al., 2023). This is especially problematic under DP, which requires adding extra noise,

which will be ignored if the synthetic data is treated like real data. This problem creates the need for *noise-aware* analyses that account for the synthetic data generation.

For frequentist downstream analyses, it is possible to account for the synthetic data generation by generating and analysing multiple synthetic data sets (Raghunathan et al., 2003). Recent work has extended this to DP synthetic data (Räisä et al., 2023), which allows generating multiple synthetic data sets without compromising on privacy. These methods reuse the analysis method for the real data, and only require using simple combining rules to combine the results from the analyses on each synthetic data set, making them simple to apply.

For Bayesian downstream analyses, Wilde et al. (2021) have shown that the analyst can use additional samples of public real data to correct their analysis. However, their method requires targeting a generalised notion of the posterior (Bissiri et al., 2016) and needs additional public data for calibration. Ghalebikesabi et al. (2022) propose a correction using importance sampling to avoid the need of public data, but only prove convergence to a generalised posterior and do not clearly address the noise-awareness of the method.

The simple frequentist methods using multiple synthetic data sets were derived from methods in missing data imputation (Rubin, 1987), so our starting point is the method of Gelman et al. (2004, 2014) for Bayesian inference with missing data. They proposed inferring the downstream posterior by imputing multiple completed data sets, inferring the analysis posterior for each completed data set separately, and mixing the posteriors together. We investigate whether this method is also applicable to synthetic data, generated with or without DP by sampling a posterior predictive distribution. With DP synthetic data, we require that the posterior predictive accounts for the DP noise, i.e. it must be *noise-aware* (Bernstein and Sheldon, 2018; Räisä et al., 2023).

### 1.1 Contributions

1. We study inferring the downstream analysis posterior by generating multiple synthetic data sets from a posterior predictive distribution, inferring the analysis posterior for each synthetic data set as if it were the real data set, and mixing the posteriors together. We find two important conditions for consistent Bayesian inference with this method: synthetic data sets that are larger than the original one, and a notion of compatibility between the data provider's and analyst's models called *congeniality* (Meng, 1994).

2. We prove that when congeniality is met and the Bernstein–von Mises theorem applies, this method converges to the true posterior as the number of synthetic data sets and the size of the synthetic data sets grow. Under stronger assumptions, we prove a convergence rate for this method in the synthetic data set size, which we expect to match the rate that usually applies in the Bernstein–von Mises theorem (Hipp and Michel, 1976). These are presented in Section 3.

3. We evaluate this method with two examples in Sections 4 and 5: non-private univariate Gaussian mean or variance estimation, and DP Bayesian logistic regression. In the first example, we use the tractability of the model to derive further theoretical properties of the method, and in both examples, we verify that the method works in practice when

the assumptions are met, and examine what can happen when they are not met. Our code is available under an open-source license.[1]

### 1.2 Related Work

Generating synthetic data to preserve privacy was, as far as we know, originally proposed by Liew et al. (1985). Rubin (1993) proposed accounting for the synthetic data generation in frequentist downstream analyses by adapting *multiple imputation* (Rubin, 1987), which involves generating multiple synthetic data sets, analysing each of them, and combining the results with so called Rubin's rules (Raghunathan et al., 2003; Reiter, 2002). Recently, Räisä et al. (2023) have shown that multiple imputation also works for synthetic data generated under DP when the data generation algorithm is noise-aware.

Recently, van Breugel et al. (2023a) have studied uncertainty quantification for prediction tasks when training on synthetic data. They propose generating multiple synthetic data sets, like the multiple imputation line of work, and experimentally show that aggregating predictions from the multiple synthetic data sets improves generalisation performance and uncertainty quantification. They use a similar Bayesian framework as we do to justify using multiple synthetic data sets, but their theoretical study of the framework is very light. In particular, they do not consider the effect of the synthetic data set size theoretically.

Wilde et al. (2021) study downstream Bayesian inference from DP synthetic data by considering the analyst's model to be misspecified, and targeting a generalised notion of the posterior (Bissiri et al., 2016) to deal with the misspecification, which makes their method more difficult to apply than standard Bayesian inference. They also assume that the analyst has additional public data available to calibrate their method.

Ghalebikesabi et al. (2022) use importance sampling to correct for bias with DP synthetic data, and have Bayesian inference as an example application. However, they also target a generalised variant (Bissiri et al., 2016) of the posterior instead of the noise-aware posterior we target, and they do not evaluate uncertainty estimation, so the noise-awareness of their method is not clear.

We are not aware of any existing work adapting multiple imputation for Bayesian downstream analysis in the synthetic data setting. In the missing data setting without DP, where multiple imputation was originally developed (Rubin, 1987), Gelman et al. (2004, 2014) have proposed sampling the downstream posterior by mixing samples of the downstream posteriors from each of the multiple imputed data sets. We find that this is not sufficient in the synthetic data setting, and add one extra component: our synthetic data sets are larger than the original data set. We compare the two cases in more detail in Section 3.6, and in particular explain why large synthetic data sets are not needed in the missing data setting.

A line of work considers mixing posteriors from multiple bootstrap samples of the real dataset (Bühlmann, 2014; Huggins and Miller, 2020, 2023) in an approach called BayesBag. BayesBag is almost a special case of our approach, as bootstrapping is a form of generating synthetic data, but the standard bootstrap does not sample from a posterior predictive distribution.[2]

1. `https://github.com/DPBayes/NAPSU-MQ-bayesian-downstream-experiments`
2. The Bayesian bootstrap (Rubin, 1981) does sample from a posterior predictive, and is very close to the standard bootstrap.

Noise-aware DP Bayesian inference is critical for taking into account the DP noise in synthetic data, but only a few works address this even without synthetic data. Bernstein and Sheldon (2018) present an inference method for simple exponential family models. Their approach was extended to linear models (Bernstein and Sheldon, 2019) and generalised linear models (Kulkarni et al., 2021). Recently, Ju et al. (2022) developed an MCMC sampler that can sample the noise-aware posterior using a noisy summary statistic.

## 2. Background

In this section, we introduce some background needed for the rest of our work. We start by introducing Bayesian inference and the Bernstein–von Mises theorem in Section 2.1, and then introduce differential privacy in Section 2.2 and noise-aware synthetic data generation in Section 2.3.

### 2.1 Bayesian Inference

Bayesian inference is a paradigm of statistical inference where the data analyst's uncertainty in a quantity $Q$ after observing data $X$ is represented using the posterior distribution $p(Q|X)$ (Gelman et al., 2014). The posterior is given by Bayes' rule:

$$p(Q|X) = \frac{p(X|Q)p(Q)}{\int p(X|Q')p(Q')\,\mathrm{d}Q'},$$

where $p(X|Q)$ is the likelihood of observing the data $X$ for a given value of $Q$, and $p(Q)$ is the analyst's prior of $Q$. Computing the denominator is typically intractable, so analysts often use numerical methods to sample $p(Q|X)$ (Gelman et al., 2014).

It turns out that in many typical settings, the prior's influence on the posterior vanishes when the data set $X$ is large. A basic example of this is the Bernstein–von Mises theorem (van der Vaart, 1998), which informally states that under regularity conditions, the posterior approaches a Gaussian that does not depend on the prior as the size of the data set increases.

A crucial component of the theorem, and also our theory, is the notion of *total variation distance* which is used to measure the difference between two random variables or probability distributions.

**Definition 1.** *The total variation distance between random variables (or distributions) $P_1$ and $P_2$ is*

$$\mathrm{TV}(P_1, P_2) = \sup_A |\Pr(P_1 \in A) - \Pr(P_2 \in A)|,$$

*where $A$ is any measurable set.*

As a slight abuse of notation, we allow the arguments of $\mathrm{TV}(\cdot, \cdot)$ to be random variables, probability distributions, or probability density functions interchangeably. We list some properties of total variation distance that we use in Lemma 18 in Appendix A.2.

Now we can state the theorem.

**Theorem 2** (Bernstein–von Mises, van der Vaart, 1998)**.** *Let $n$ denote the size of the data set $X_n$. Under regularity conditions stated in Condition 21 in Appendix A.3, for true parameter*

*value $Q_0$, the posterior $\bar{Q}(X_n) \sim p(Q|X_n)$ satisfies*

$$\mathrm{TV}\left(\sqrt{n}(\bar{Q}(X_n) - Q_0), \mathcal{N}(\mu(X_n), \Sigma)\right) \xrightarrow{P} 0$$

*as $n \to \infty$ for some $\mu(X_n)$ and $\Sigma$, that do not depend on the prior, where the convergence in probability is over sampling $X_n \sim p(X_n|Q_0)$.*

### 2.2 Differential Privacy and Noise-Aware Synthetic Data

*Differential privacy* (DP) (Dwork et al., 2006b) quantifies the privacy loss from releasing the results of analysing data. The quantification is done by looking at the output distributions of the analysis algorithm for two data sets that differ in a single data subject (Dwork and Roth, 2014):

**Definition 3.** *An algorithm $\mathcal{M}$ is $(\epsilon, \delta)$-DP if*

$$\Pr(\mathcal{M}(X) \in S) \leq e^{\epsilon} \Pr(\mathcal{M}(X') \in S) + \delta$$

*for all measurable sets $S$ and all data sets $X, X'$ that differ in one data subject.*

The choice of $\epsilon$ and $\delta$ is a matter of policy (Dwork, 2008). One should set $\delta \ll 1/n$ for $n$ datapoints, as $\delta \approx 1/n$ permits mechanisms that clearly violate privacy (Dwork and Roth, 2014).

A common primitive for making an algorithm DP is the *Gaussian mechanism* (Dwork et al., 2006a), which simply adds Gaussian noise to the output of a function:

**Definition 4.** *The Gaussian mechanism with noise variance $\sigma^2_{DP}$ and function $f$ outputs $f(X) + \mathcal{N}(0, \sigma^2_{DP} I)$ for input $X$.*

For a given $(\epsilon, \delta)$-bound and function $f$, the required value for $\sigma^2_{DP}$ can be computed tightly using the analytical Gaussian mechanism (Balle and Wang, 2018).

### 2.3 Noise-Aware Private Synthetic Data

To solve the uncertainty estimation problem for frequentist analyses from DP synthetic data, Räisä et al. (2023) developed a noise-aware algorithm for generating synthetic data called NAPSU-MQ. NAPSU-MQ takes discrete data and summarises it with marginal queries, which count how many data points have each possible combination of values for given variables. See Appendix A.1 for a precise definition. Then NAPSU-MQ releases the query values under DP with the Gaussian mechanism, and finally generates multiple synthetic data sets. The downstream analysis is done on each synthetic data set, and the results are combined using Rubin's rules for synthetic data (Raghunathan et al., 2003; Rubin, 1993), which use the multiple analysis results to account for the extra uncertainty coming from the synthetic data generation.

The synthetic data is generated by sampling the posterior predictive distribution

$$p(X^*|\tilde{s}) = \int p(X^*|\theta) p(\theta|\tilde{s}) \, \mathrm{d}\theta, \tag{1}$$

- $\theta$: data generating model parameters
- $X$: real data
- $X^*$: hypothetical data
- $Z$: observed summary of $X$ ($Z = X$ without DP)
- $X^{Syn}$: synthetic data, $X^{Syn} \sim p(X^*|Z, I_S)$
- $Q$: estimated quantity in downstream analysis
- $I_S$: synthetic data provider's background information
- $I_A$: analyst's background information

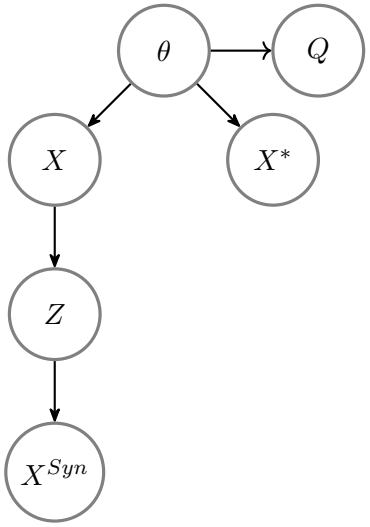



Figure 1: Left: random variables in noise-aware uncertainty estimation from synthetic data. Right: a Bayesian network describing the dependencies of the random variables. This network is conditional on either $I_S$ or $I_A$, depending on whether viewed by the data provider or the analyst.

where $\theta$ is the parameters of the synthetic data generator and $\tilde{s}$ is the noisy marginal query values. The conditioning on $\tilde{s}$ and including the Gaussian mechanism in the model is what makes NAPSU-MQ noise-aware, and allows Rubin's rules to accurately account for the synthetic data generation and DP noise in the downstream analysis.

As an alternative to NAPSU-MQ, one can use prior noise-aware Bayesian inference methods (Bernstein and Sheldon, 2018; Kulkarni et al., 2021) to generate synthetic data from a noise-aware posterior predictive distribution. However, these methods heavily restrict the possible models, so they are unlikely to be flexible enough to accurately model realistic data.

## 3. Bayesian Inference from Synthetic Data

When the downstream analysis is Bayesian, the analyst would ultimately want to obtain the posterior $p(Q|X, I_A)$ of some quantity $Q$ given real data $X$, where $I_A$ denotes the background knowledge such as priors of the analyst. We assume the analyst has a method to sample $p(Q|X, I_A)$ if they had access to real data, and study what they can do when they only have access to synthetic data.

In order to introduce the synthetic data into the posterior of interest, we can decompose the posterior using the law of total probability as

$$p(Q|X, I_A) = \int p(Q|X, X^*, I_A)p(X^*|X, I_A)\,\mathrm{d}X^*. \tag{2}$$

We are using a standard abuse of notation here, where inside the integral $X^*$ is the variable to integrate over, and outside the integral, $X^*$ is a random variable. The decomposition in (2) means that we can sample from $p(Q|X, I_A)$ by first sampling synthetic data $X^{Syn}$ from the posterior predictive $p(X^*|X, I_A)$, and then sampling $Q$ from $p(Q|X, X^* = X^{Syn}, I_A)$.

To make using the decomposition in (2) possible, the synthetic data provider must generate the synthetic data by sampling the posterior predictive $X^{Syn} \sim p(X^*|X, I_A)$. The most common way to accomplish this is to model $X$ with a parametric model with parameters $\theta$, taking a sample $\theta_s \sim p(\theta|X, I_A)$ from the posterior of this model, and then sampling $X^{Syn} \sim p(X^*|\theta = \theta_s, I_A)$.

Note that $X^*$ and $X^{Syn}$ are not the same random variable. In fact, their relationship is analogous to $\theta$ and $\theta_s$ above. The random variable $X^*$ represents a hypothetical real data set that could be obtained if more data was collected, as seen in Figure 1. The synthetic data set $X^{Syn}$ is a sample from the conditional distribution of $X^*$ given $X$, analogous to how $\theta_s$ is a sample from a conditional distribution of $\theta$. For this reason, $p(Q|X, X^*, I_A) \neq p(Q|X, I_A)$. To make our notation less cluttered, we use a standard abuse of notation, and write $\theta \sim p(\theta|\cdot)$, suppressing the separate $\theta_s$ symbol when the meaning is clear from now on. We will also use this convention with other random variables. However, due to the importance of $X^{Syn}$ and $X^*$ in this work, we will keep them separate, when needed for clarity.

The parameterisations of $Q$ and $\theta$ may be very different. For example, in the example of Section 5.3, $Q$ consists of the coefficients of logistic regression, while $\theta$ consists of the parameters of a Markov network (Räisä et al., 2023).

Under DP, the exact posterior $p(Q|X, I_A)$ is unobtainable, so we assume that $X$ is only available through a noisy summary $\tilde{s}$ (Ju et al., 2022; Räisä et al., 2023), with the posterior $p(Q|\tilde{s}, I_A)$. It would be ideal for $p(Q|\tilde{s}, I_A)$ to be as close to $p(Q|X, I_A)$ as possible, but they cannot be equal due to the noise that must be added to $\tilde{s}$. For this reason, our goal will be consistent inference of $p(Q|\tilde{s}, I_A)$, which must account for the noise that is added to $\tilde{s}$.

We can write the same decomposition as in (2) for $p(Q|\tilde{s}, I_A)$:

$$p(Q|\tilde{s}, I_A) = \int p(Q|\tilde{s}, X^*, I_A)p(X^*|\tilde{s}, I_A)\,\mathrm{d}X^*, \tag{3}$$

and use synthetic data in the same way to sample $p(Q|\tilde{s}, I_A)$. Note that the synthetic data will need to be sampled from the posterior predictive $p(X^*|\tilde{s}, I_A)$, which will need to account for the noise that is added to $\tilde{s}$, i.e. it needs to be noise-aware. This can be accomplished with the methods introduced in Section 2.3: NAPSU-MQ (Räisä et al., 2023) or more restricted methods (Bernstein and Sheldon, 2018; Kulkarni et al., 2021).

Since the only difference between the non-DP and DP cases is the symbol for the observed values, we unify the notations by using $Z$ to denote the observed values, so $Z = X$ in the non-DP case, $Z = \tilde{s}$ in the DP case, and the posterior of interest is $p(Q|Z, I_A)$. The decomposition of interest is then

$$p(Q|Z, I_A) = \int p(Q|Z, X^*, I_A)p(X^*|Z, I_A)\,\mathrm{d}X^*. \tag{4}$$

We summarise these random variables and their dependencies in Figure 1.

There are still two major issues with the decomposition in (4):

1. Sampling $p(Q|Z, X^*, I_A)$ requires access to $Z$, which defeats the purpose of using synthetic data.

2. The synthetic data needs to be sampled conditionally on the analyst's background information $I_A$, while the synthetic data provider could have different background information $I_S$.

To solve the first issue, in Section 3.2 we show that if we replace $p(Q|Z, X^*, I_A)$ inside the integral of (4) with $p(Q|X^*, I_A)$, the resulting distribution converges to the desired posterior,

$$\int p(Q|X^*, I_A)p(X^*|Z, I_A)\,\mathrm{d}X^* \to p(Q|Z, I_A) \tag{5}$$

in total variation distance as the size of $X^*$ grows. It should be noted that many synthetic data sets $X^{Syn} \sim p(X^*|Z, I_A)$ will be needed to account for the integral over $X^*$.

The second issue is known as *congeniality* in the multiple imputation literature (Meng, 1994; Xie and Meng, 2016). We look at congeniality in the context of Bayesian inference from synthetic data in Section 3.1, and find that we can obtain $p(Q|Z, I_A)$ under appropriate assumptions on the relationship between $I_A$ and $I_S$.

Exactly sampling the LHS of (5) requires generating a synthetic data set for each sample of $p(Q|Z, I_A)$, which is not practical. However, we can perform a Monte-Carlo approximation for $p(Q|Z, I_A)$ by independently generating $m$ synthetic data sets $X_1^{Syn}, \ldots, X_m^{Syn} \sim p(X^*|Z, I_A)$, drawing multiple samples from each of the $p(Q|X^* = X_i^{Syn}, I_A)$, and mixing these samples, which allows us to obtain more than one sample of $p(Q|Z, I_A)$ per synthetic data set. We look at some properties of this in Section 3.4, but we use the integral form in (5) in the rest of our theory.

### 3.1 Congeniality

In the decomposition (4) of the analyst's posterior, we need to sample the conditional distribution of $X^*$ while conditioning on the analyst's background information $I_A$, while in reality the synthetic data provider could have different background information $I_S$.

A similar distinction has been studied in the context of missing data (Meng, 1994; Xie and Meng, 2016), where the imputer of missing data has a similar role as the synthetic data provider. Meng (1994) found that Rubin's rules implicitly assume that the probability models of both parties are compatible in a certain sense, which Meng (1994) defined as *congeniality*.

As our examples with Gaussian distributions in Section 4 show, some notion of congeniality is also required in our setting. However, because we study synthetic data instead of imputation, and Bayesian instead of frequentist downstream analysis, we need a different formal definition. As the analyst only makes inferences on $Q$, it suffices that both the analyst and synthetic data provider make the same inferences of $Q$:

**Definition 5.** *The background information sets $I_S$ and $I_A$ are congenial for observation $Z$ if*

$$p(Q|X^*, I_S) = p(Q|X^*, I_A) \tag{6}$$

*for all $X^*$ and*

$$p(Q|Z, I_S) = p(Q|Z, I_A). \tag{7}$$

In the non-DP case, (7) is redundant, as it is implied by (6), but in the DP case, both are needed, as the parties may draw different conclusions on $X$ given $Z = \tilde{s}$.

Combining congeniality and (5),

$$\begin{aligned}\int p(Q|X^*, I_A)p(X^*|Z, I_S)\,\mathrm{d}X^* &= \int p(Q|X^*, I_S)p(X^*|Z, I_S)\,\mathrm{d}X^* \\ &\to p(Q|Z, I_S) = p(Q|Z, I_A),\end{aligned} \tag{8}$$

where the convergence is in total variation distance as the size of $X^*$ grows.

Congeniality is a strict condition, since it requires both parties to make exactly the same inference on $Q$. We will also study the following relaxation, where the parties only need to make the same inferences asymptotically. In the definition, we consider a countably infinite population $X^*_\infty$. We use $X^*_{n_{X^*}}$ to denote the first $n_{X^*}$ members of the population. We also use this notation with $Z_\infty$ and $Z_{n_Z}$, which denote the observed, potentially noisy, value computed from either the infinite population, or the first $n_Z$ members.

**Definition 6.** *The background information sets $I_S$ and $I_A$ are asymptotically congenial for population $Z_\infty$ if, for all populations $X^*_\infty$,*

$$\mathrm{TV}(p(Q|X^*_{n_{X^*}}, I_S), p(Q|X^*_{n_{X^*}}, I_A)) \to 0 \tag{9}$$

*as $n_{X^*} \to \infty$, and*

$$\mathrm{TV}(p(Q|Z_{n_Z}, I_S), p(Q|Z_{n_Z}, I_A)) \to 0 \tag{10}$$

*as $n_Z \to \infty$.*

If both parties make reasonable assumptions, they should arrive at the same posterior of $Q$ with an infinite population, since the infinite population determines the value of $Q$. They should also arrive at the same posterior if given $Z$ from an infinite population, even with added noise, since the process of adding the noise is known to both parties, and the signal-to-noise ratio grows to infinity when increasing $n$ in the DP applications we are interested in. As a result, we can expect asymptotic congeniality to hold in most practical cases with reasonable parties.

### 3.2 Consistency Proof

To recap, we want to prove that the posterior from synthetic data,

$$\bar{p}_n(Q) = \int p(Q|X^*_n)p(X^*_n|Z)\,\mathrm{d}X^*_n, \tag{11}$$

converges in total variation distance to $p(Q|Z)$ as the size $n$ of $X^*_n$ grows. We prove this in Theorem 9, which requires that both $p(Q|Z, X^*_n)$ and $p(Q|X^*_n)$ approach the same distribution as $n$ grows. We formally state this in Condition 7. In Lemma 8, we show that Condition 7 is a consequence of the Bernstein–von Mises theorem (Theorem 2) under some additional assumptions, so we expect it to hold in typical settings.

To make the notation more compact, let $\bar{Q}^+_n \sim p(Q|Z, X^*_n)$, and let $\bar{Q}_n \sim p(Q|X^*_n)$.

**Condition 7.** *For the observed $Z$ and all $Q$, there exist random variables $D_n$ such that*

$$\mathrm{TV}\left(\bar{Q}_n^+, D_n\right) \xrightarrow{P} 0 \quad \text{and} \quad \mathrm{TV}\left(\bar{Q}_n, D_n\right) \xrightarrow{P} 0$$

*as $n \to \infty$, where the convergence in probability is over sampling $X_n^* \sim p(X_n^*|Z, Q)$.*

Theorem 2 implies Condition 7 with some additional assumptions:

**Lemma 8.** *Let the assumptions of Theorem 2 (stated in Condition 21) hold for the downstream analysis for all $Q_0$, and the following assumptions:*

*(1) $Z$ and $X^*$ are conditionally independent given $Q$; and*

*(2) $p(Z|Q) > 0$ for all $Q$,*

*hold. Then Condition 7 holds.*

**Proof** The full proof is in Appendix B.1. Proof idea: when $Z$ and $X^*$ are conditionally independent given $Q$,

$$p(Q|Z, X^*) \propto p(X^*|Q)p(Z|Q)p(Q)$$

so $p(Q|Z, X^*)$ can be equivalently seen as the result of Bayesian inference with observed data $X^*$ and prior $p(Q|Z)$. As the only difference to $p(Q|X^*)$ is the prior, the Bernstein–von Mises theorem implies that both $p(Q|Z, X^*)$ and $p(Q|X^*)$ converge in total variation distance to the same distribution. ■

Assumption (1) of Lemma 8 will hold if the downstream analysis treats its input data as an i.i.d. sample from some distribution. Note that this i.i.d. assumption is on the real data, so using a DP mechanism with correlated noise does not affect it. Assumption (2) holds when the likelihood is always positive, and in the DP case when every possible output has positive probability regardless of the private data, which is the case for common DP mechanisms like the Gaussian and Laplace mechanisms (Dwork and Roth, 2014).

Next is the main theorem of this work: (5) holds under Condition 7.

**Theorem 9.** *Under congeniality and Condition 7,* $\mathrm{TV}\left(p(Q|Z), \bar{p}_n(Q)\right) \to 0$ *as $n \to \infty$.*

**Proof** The full proof is in Appendix B.1. Proof idea: the proof consists of three steps. The first two are in Lemma 22 and the third is in Lemma 23 in the Appendix. The first step is showing that $\mathrm{TV}(\bar{Q}_n, \bar{Q}_n^+) \xrightarrow{P} 0$ when $X_n^* \sim p(X_n^*|Z, Q)$ for fixed $Z$ and $Q$. This is a simple consequence of the triangle inequality and Condition 7, as total variation distance is a metric. In the second step, we show that $\mathrm{TV}(\bar{Q}_n, \bar{Q}_n^+) \xrightarrow{P} 0$ also holds when $X_n^* \sim p(X_n^*|Z)$. In the final step, we show that this implies the claim. ■

### 3.3 Convergence Rate

Under a stronger regularity condition, we can get a convergence rate for Theorem 9. The regularity condition depends on uniform integrability:

**Definition 10.** *A sequence of random variables $X_n$ is uniformly integrable if*

$$\lim_{M\to\infty} \sup_n \mathbb{E}(|X_n|\mathbb{I}_{|X_n|>M}) = 0.$$

Now we can state the regularity condition for a convergence rate $O(R_n)$:

**Condition 11.** *For the observed $Z$, there exist random variables $D_n$ such that for a sequence $R_1, R_2, \cdots > 0$, $R_n \to 0$ as $n \to \infty$,*

$$\frac{1}{R_n}\,\mathrm{TV}\left(\bar{Q}_n^+, D_n\right) \quad \text{and} \quad \frac{1}{R_n}\,\mathrm{TV}\left(\bar{Q}_n, D_n\right)$$

*are uniformly integrable when $X_n^* \sim p(X_n^*|Z)$.*

Note that $X_n^* \sim p(X_n^*|Z)$ conditions on $Z$, not $Q$ and $Z$ like in Condition 7.

Condition 11 is not a standard regularity condition, so it is not clear what settings it applies in. To ensure that it at least applies in some setting, we show that it is met in multivariate Gaussian mean estimation with $R_n = \frac{1}{\sqrt{n}}$. This is the rate that commonly appears in the Bernstein–von Mises theorem (Hipp and Michel, 1976).

**Theorem 12.** *When the up- and downstream models are $d$-dimensional Gaussian mean estimations with known covariance $\Sigma_k$, and $D_n \sim \mathcal{N}(\bar{X}_n^*, n^{-1}\Sigma_k)$,*

$$\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n^+, D_n\right) \quad \text{and} \quad \sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n, D_n\right)$$

*are uniformly integrable when $X_n^* \sim p(X_n^*|X)$.*

**Proof** The idea of the proof is to use Pinsker's inequality (Lemma 18) to upper bound the total variation distance with KL divergence, and prove the required uniform integrability for the KL divergence upper bound. This is a fairly lengthy exercise in upper bounding and computing the limit in the definition of uniform integrability for the various terms that appear in the KL-divergence between the two Gaussians in question. We defer the full proof to Appendix B.2. ■

Condition 11 implies an $O(R_n)$ convergence rate:

**Theorem 13.** *Under congeniality and Condition 11,* $\mathrm{TV}\left(p(Q|Z), \bar{p}_n(Q)\right) = O(R_n)$.

**Proof** The full proof is in Appendix B.2. Proof idea: first, we prove the uniform integrability of $\frac{1}{R_n}\mathrm{TV}(\bar{Q}_n, \bar{Q}_n^+)$ when $X_n^* \sim p(X_n^*|Z)$ by using the triangle inequality and properties of uniform integrability. Second, we prove that this implies the claimed convergence rate. ■

### 3.4 Finite Number of Synthetic Data Sets

We have now shown that the mixture of posteriors

$$\bar{p}_n(Q) = \int p(Q|X_n^*)p(X_n^*|Z)\,\mathrm{d}X_n^*,$$

converges to the target posterior $p(Q|Z)$ as $n$ grows. However, sampling $\bar{p}_n(Q)$ exactly requires one synthetic data set per sample, which is not practical in realistic settings. We can further approximate by generating a fixed number $m$ of synthetic datasets and using a Monte-Carlo approximation of the integral:

$$p(Q|Z) \approx \int p(Q|X_n^*)p(X_n^*|Z)\,\mathrm{d}X_n^* \approx \frac{1}{m}\sum_{i=1}^{m} p(Q|X_n^* = X_{i,n}^{Syn}), \tag{12}$$

with $X_{i,n}^{Syn} \sim p(X_n^*|Z)$. In the following, we shorten $p(Q|X_n^* = X_{i,n}^{Syn}) = p(Q|X_{i,n}^{Syn})$.

Total variation distance is a metric, so

$$\mathrm{TV}\left(\frac{1}{m}\sum_{i=1}^{m} p(Q|X_{i,n}^{Syn}), p(Q|Z)\right) \leq \mathrm{TV}\left(\frac{1}{m}\sum_{i=1}^{m} p(Q|X_{i,n}^{Syn}), \bar{p}_n(Q)\right) + \mathrm{TV}\left(\bar{p}_n(Q), p(Q|Z)\right).$$

Theorem 9 gives

$$\lim_{n\to\infty} \mathrm{TV}\left(\bar{p}_n(Q), p(Q|Z)\right) = 0.$$

If, for all $n$, $p(Q|X_n^*)$ is continuous for all $X_n^*$, $p(Q|X_n^*) \leq h_n(X_n^*)$ for an integrable function $h_n(X_n^*)$, and $\mathcal{Q} \subset \mathbb{R}^d$ is compact, the uniform law of large numbers (Jennrich, 1969, Theorem 2) gives

$$\sup_{Q\in\mathcal{Q}} \left|\frac{1}{m}\sum_{i=1}^{m} p(Q|X_{i,n}^{Syn}) - \bar{p}_n(Q)\right| \to 0 \tag{13}$$

almost surely as $m \to \infty$. If $\mathcal{Q} = \mathbb{R}^d$, we can represent $\mathcal{Q}$ as a countable union of compact sets $\mathcal{Q}_k$, apply the uniform law of large numbers on each $\mathcal{Q}_k$, and use the union bound to obtain (13) without the supremum for all $Q \in \mathcal{Q}$. A similar decomposition of $\mathcal{Q}$ can be done for many other constrained parameter sets encountered in practice.

This pointwise convergence implies (van der Vaart, 1998, Corollary 2.30)

$$\lim_{m\to\infty} \mathrm{TV}\left(\frac{1}{m}\sum_{i=1}^{m} p(Q|X_{i,n}^{Syn}), \bar{p}_n(Q)\right) = 0$$

for almost all $X_{i,n}^{Syn}$, so

$$\lim_{n\to\infty}\lim_{m\to\infty} \mathrm{TV}\left(\frac{1}{m}\sum_{i=1}^{m} p(Q|X_{i,n}^{Syn}), p(Q|Z)\right) = 0 \tag{14}$$

almost surely when $X_{i,n}^{Syn} \sim p(X_n^*|Z)$.

In Figure 3, we see that the mixture of posteriors becomes very spiky with small $m$ and large $n$, and does not fully converge to the target posterior when $n$ grows but $m$ is kept small. This suggests that

$$\lim_{m\to\infty}\lim_{n\to\infty} \mathrm{TV}\left(\frac{1}{m}\sum_{i=1}^{m} p(Q|X_{i,n}^{Syn}), p(Q|Z)\right) \neq 0 \tag{15}$$

because the distributions $p(Q|X_{i,n}^{Syn})$ become narrower as $n$ increases, so a fixed number of them is not enough to cover $p(Q|Z)$. This means that in practice, the number of synthetic data sets should be increased along with the size of the synthetic data sets.

### 3.5 Convergence with Asymptotic Congeniality

Recall that with asymptotic congeniality, we consider an infinite population $X_\infty$, and we observe a value $Z_{n_Z}$ computed from the first $n_Z$ elements of $X_\infty$. We also consider hypothetical infinite populations $X^*_\infty$, and their finite samples $X^*_{n_{X^*}}$ of size $n_{X^*}$. In this setting, we obtain an analogue of Theorem 9, with the difference that the size of the real data $n_Z$ also needs to approach infinity.

**Theorem 14.** *If asymptotic congeniality and Condition 7 for all $Z_{n_Z}$, with the probabilities of Condition 7 taken conditional to $I_S$, hold,*

$$\lim_{n_Z \to \infty} \lim_{n_{X^*} \to \infty} \mathrm{TV}\left(p(Q|Z_{n_Z}, I_A), \bar{p}_{n_{X^*}}(Q)\right) = 0. \tag{16}$$

**Proof** The full proof is in Appendix B.3. The idea is similar to the proof of Theorem 9, but we need to use the asymptotic congeniality assumption to show that the difference between posteriors with $I_A$ and $I_S$ approaches zero in the limit. ■

### 3.6 Relation to Missing Data Imputation

Combining inferences by mixing posteriors from multiple data sets in the style of (4) was originally proposed by Gelman et al. (2004, 2014) for Bayesian inference with missing data, with completed data sets corresponding to synthetic data sets of our setting. Large completed data sets are not required in the missing data setting. Next, we explain where this difference between the two settings arises.

In the missing data setting, only a part $X_{obs}$ of the complete data set $X$ is observed, while a part $X_{mis}$ is missing (Rubin, 1987). To facilitate downstream analysis, the missing data are imputed by sampling $X_{mis} \sim p(X_{mis}|X_{obs}, I_I)$. Analogously with synthetic data, $I_I$ represents the imputer's background knowledge.

Like with synthetic data, we have the decomposition (Gelman et al., 2014)

$$p(Q|X_{obs}, I_A) = \int p(Q|X_{obs}, X_{mis}, I_A) p(X_{mis}|X_{obs}, I_A) \,\mathrm{d}X_{mis}. \tag{17}$$

If the analyst's and imputer's models are congenial in the sense that

$$p(Q|X_{obs}, I_A) = p(Q|X_{obs}, I_I)$$

and

$$p(Q|X, I_A) = p(Q|X, I_I)$$

for any complete data set $X$, then

$$\begin{aligned} p(Q|X_{obs}, I_A) = p(Q|X_{obs}, I_I) &= \int p(Q|X_{obs}, X_{mis}, I_I) p(X_{mis}|X_{obs}, I_I) \,\mathrm{d}X_{mis} \\ &= \int p(Q|X_{obs}, X_{mis}, I_A) p(X_{mis}|X_{obs}, I_I) \,\mathrm{d}X_{mis}, \end{aligned} \tag{18}$$

so sampling $p(Q|X_{obs}, I_A)$ can be done by sampling $X_{mis} \sim p(X_{mis}|X_{obs}, I_I)$ multiple times, sampling $p(Q|X_{obs}, X_{mis}, I_A)$ for each $X_{mis}$, and combining the samples. Unlike with

| | Data Provider | Analyst |
|---|---|---|
| Known Variance | $\bar{\sigma}_k^2$ | $\hat{\sigma}_k^2$ |
| Prior Mean | $\bar{\mu}_0$ | $\hat{\mu}_0$ |
| Prior Variance | $\bar{\sigma}_0^2$ | $\hat{\sigma}_0^2$ |
| Data | $X$ | $X^*$ |
| Single Datapoint | $x$ | $x^*$ |
| Data Size | $n_X$ | $n_{X^*}$ |
| Data Average | $\bar{X}$ | $\bar{X}^*$ |
| Posterior Mean | $\bar{\mu}_{n_X}$ | $\hat{\mu}_{n_{X^*}}$ |
| Posterior Variance | $\bar{\sigma}_{n_X}^2$ | $\hat{\sigma}_{n_{X^*}}^2$ |
| Posterior Sample | $\bar{\mu}$ | $\hat{\mu}$ |
| Mixture of Posteriors Sample | | $\mu^*$ |

Table 1: Notation for known variance model in Section 4.1.

synthetic data, where sampling $p(Q|X, X^*, I_A)$ would require the original data and defeat the purpose of using synthetic data, sampling $p(Q|X_{obs}, X_{mis}, I_A)$ is simply the analysis for a complete data set, so generating large imputed data sets is not required.

## 4. Non-private Gaussian Examples

In this section, we look at the Bayesian inference of a univariate Gaussian mean or variance from mixing the posteriors from multiple synthetic data sets, which are also generated from the same model. This allows us to analytically examine various aspects of Bayesian inference from multiple synthetic data sets which are not visible in our high-level theory in Section 3, as all of the posteriors are analytically tractable and relatively simple. In particular, we find that the synthetic data set needs to be larger than the original data set (Section 4.3) and find a variance correction formula for the Gaussian setting that gets around this requirement (Section 4.4). We also look at two forms of uncongeniality, and find that they cause very different effects: estimating the mean with incorrect known variance converges to the data provider's posterior, but estimating the variance with incorrect known mean does not converge to either party's posterior. For reference, we list the posteriors for these Gaussian settings in Appendix A.4.

### 4.1 Gaussian Mean Estimation with Known Variance

Our first example is very simple: $x_i \sim \mathcal{N}(\mu, \sigma^2)$, $X = (x_1, \ldots, x_{n_X})$, and analyst infers the mean $\mu$ of a univariate Gaussian distribution with known variance from synthetic data that has been generated from the same model. To differentiate the variables for the analyst and data provider, we use bars for the data provider (like $\bar{\sigma}_0^2$) and hats for the analyst (like $\hat{\sigma}_0^2$). Table 1 summarises the notation used in this section.

When the synthetic data is generated from the model with known variance $\bar{\sigma}_k^2$, we sample from the posterior predictive $p(X^*|X)$ as

$$\bar{\mu}|X \sim \mathcal{N}(\bar{\mu}_{n_X}, \bar{\sigma}^2_{n_X}), \quad X^*|\bar{\mu} \sim \mathcal{N}^{n_{X^*}}(\bar{\mu}, \bar{\sigma}_k^2)$$
$$\bar{\mu}_{n_X} = \frac{\frac{1}{\bar{\sigma_0}^2}\bar{\mu}_0 + \frac{n_X}{\bar{\sigma}_k^2}\bar{X}}{\frac{1}{\bar{\sigma}_0^2} + \frac{n_X}{\bar{\sigma}_k^2}}, \quad \frac{1}{\bar{\sigma}^2_{n_X}} = \frac{1}{\bar{\sigma}_0^2} + \frac{n_X}{\bar{\sigma}_k^2}.$$

Here $\mathcal{N}^{n_{X^*}}$ denotes a Gaussian distribution over $n_{X^*}$ i.i.d. samples and $\bar{X}$ is the mean of $X$.

When downstream analysis is the model with known variance $\hat{\sigma}_k^2$, we have

$$\hat{\mu}|X^* \sim \mathcal{N}(\hat{\mu}_{n_{X^*}}, \hat{\sigma}^2_{n_{X^*}}), \quad \hat{\mu}_{n_{X^*}} = \frac{\frac{1}{\hat{\sigma_0}^2}\hat{\mu}_0 + \frac{n_{X^*}}{\hat{\sigma}_k^2}\bar{X}^*}{\frac{1}{\hat{\sigma}_0^2} + \frac{n_{X^*}}{\hat{\sigma}_k^2}}, \quad \frac{1}{\hat{\sigma}^2_{n_{X^*}}} = \frac{1}{\hat{\sigma}_0^2} + \frac{n_{X^*}}{\hat{\sigma}_k^2}.$$

The following proposition gives the behavior of the mixture of posteriors from synthetic data $\bar{p}_n(\mu)$ from (11) in the $n_{X^*} \to \infty$ limit.

**Proposition 15.** *If $\mu^*$ is a sample from $\bar{p}_n(\mu)$ in Gaussian mean estimation with known variance, $\mu^*$ has a Gaussian distribution, and as $n_{X^*} \to \infty$,*

$$\mathbb{E}_{\mu^*}(\mu^*) \to \bar{\mu}_{n_X}, \quad \mathrm{Var}_{\mu^*}(\mu^*) \to \bar{\sigma}^2_{n_X}. \tag{19}$$

**Proof** See Appendix B.4. ■

This means that $p(\mu^*) \xrightarrow{d} p(\mu|X, I_S)$. This is regardless of congeniality, which corresponds to both parties having equal known variances and priors ($\bar{\sigma}_k^2 = \hat{\sigma}_k^2, \bar{\mu}_0 = \hat{\mu}_0, \bar{\sigma}_0^2 = \hat{\sigma}_0^2$) in this setting. If the parties had equal known variances, but different priors, we would have asymptotic congeniality instead, and if they had different known variances, we would not have even asymptotic congeniality. These follow from upper and lower bounds on the total variation distance between one-dimensional Gaussian distributions (Devroye et al., 2022, Theorem 1.3).

We test the theory with a numerical simulation in Figure 2. We generated the real data $X$ of size $n_X = 100$ by i.i.d. sampling from $\mathcal{N}(1, 4)$. Both the analyst and data provider use $\mathcal{N}(0, 10^2)$ as the prior. The data provider uses the correct known variance ($\bar{\sigma}_k^2 = 4$), and the analyst either uses the correct known variance ($\hat{\sigma}_k^2 = 4$), or a too small known variance ($\hat{\sigma}_k^2 = 1$), which is an example of uncongeniality.

In the congenial case in the left panel of Figure 2, both parties have the same posterior given the real data $X$, and the mixture of posteriors from synthetic data is very close to that. In the uncongenial case in the right panel, where the analyst underestimates the variance, the parties have different posteriors given $X$, but the mixture of synthetic data posteriors is still close to the data provider's posterior.

In Figure 3, we examine the convergence of the mixture of posteriors from synthetic data under congeniality. We see that setting $n_{X^*} = n_X$ is not enough, as the mixture of posteriors is significantly wider than the analyst's posterior for all values of $m$. The synthetic data set needs to be larger than the original, with $n_{X^*} = 5n_X$ already giving a decent approximation and $n_{X^*} = 20n_X$ a rather good one with the larger values of $m$. We also see that $m$ must be sufficiently large, otherwise the method produces very jagged posteriors, for example the top right corner.

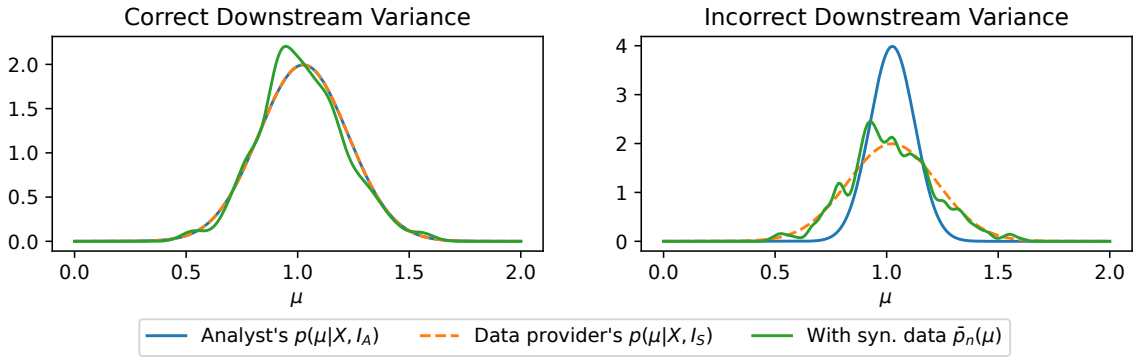



Figure 2: Simulation results for the Gaussian mean estimation example, showing that the mixture of posteriors from synthetic data in green converges. In the left panel, both the analyst and data provider have the correct known variance. The blue and orange lines overlap, as both parties have the same $p(\mu|X)$. On the right, the analyst's known variance is too small ($\hat{\sigma}_k^2 = \frac{1}{4}\bar{\sigma}_k^2$), so congeniality is not met, but the mixture of posteriors from synthetic data, $\bar{p}_n(\mu)$, still converges to the data provider's posterior. In both panels, $m = 400$ and $\frac{n_{X^*}}{n_X} = 20$.

| | Data Provider | Analyst |
|---|---|---|
| Variance Prior | Inv-$\chi^2(\bar{\nu}_0, \bar{\sigma}_0)$ | $\hat{\sigma}_k^2$ |
| Mean Prior | $\mathcal{N}(\bar{\mu}_0, \bar{\sigma}^2 \vert \bar{\kappa}_0)$ | $\mathcal{N}(\hat{\mu}_0, \hat{\sigma}_0^2)$ |
| Data | $X$ | $X^*$ |
| Single Datapoint | $x$ | $x^*$ |
| Data Size | $n_X$ | $n_{X^*}$ |
| Data Average | $\bar{X}$ | $\bar{X}^*$ |
| Data Sample Variance | $s^2$ | |
| Posterior Parameters | $\bar{\mu}_{n_X}, \bar{\kappa}_{n_X}, \bar{\nu}_{n_X}, \bar{\sigma}^2_{n_X}$ | $\hat{\mu}_{n_{X^*}}, \hat{\sigma}^2_{x_{X^*}}$ |
| Posterior Sample | $\bar{\mu}, \bar{\sigma}^2$ | $\hat{\mu}^2$ |
| Mixture of Posteriors Sample | | $\mu^*$ |

Table 2: Notation for unknown variance upstream, known variance downstream model in Section 4.2.

### 4.2 Gaussian with Unknown Variance Upstream, Known Variance Downstream

In this section, we look at a slightly more complex version of Gaussian mean estimation, where the data provider does not assume a known variance, but the analyst still assumes a known variance. We summarise the notation used in this section in Table 2.

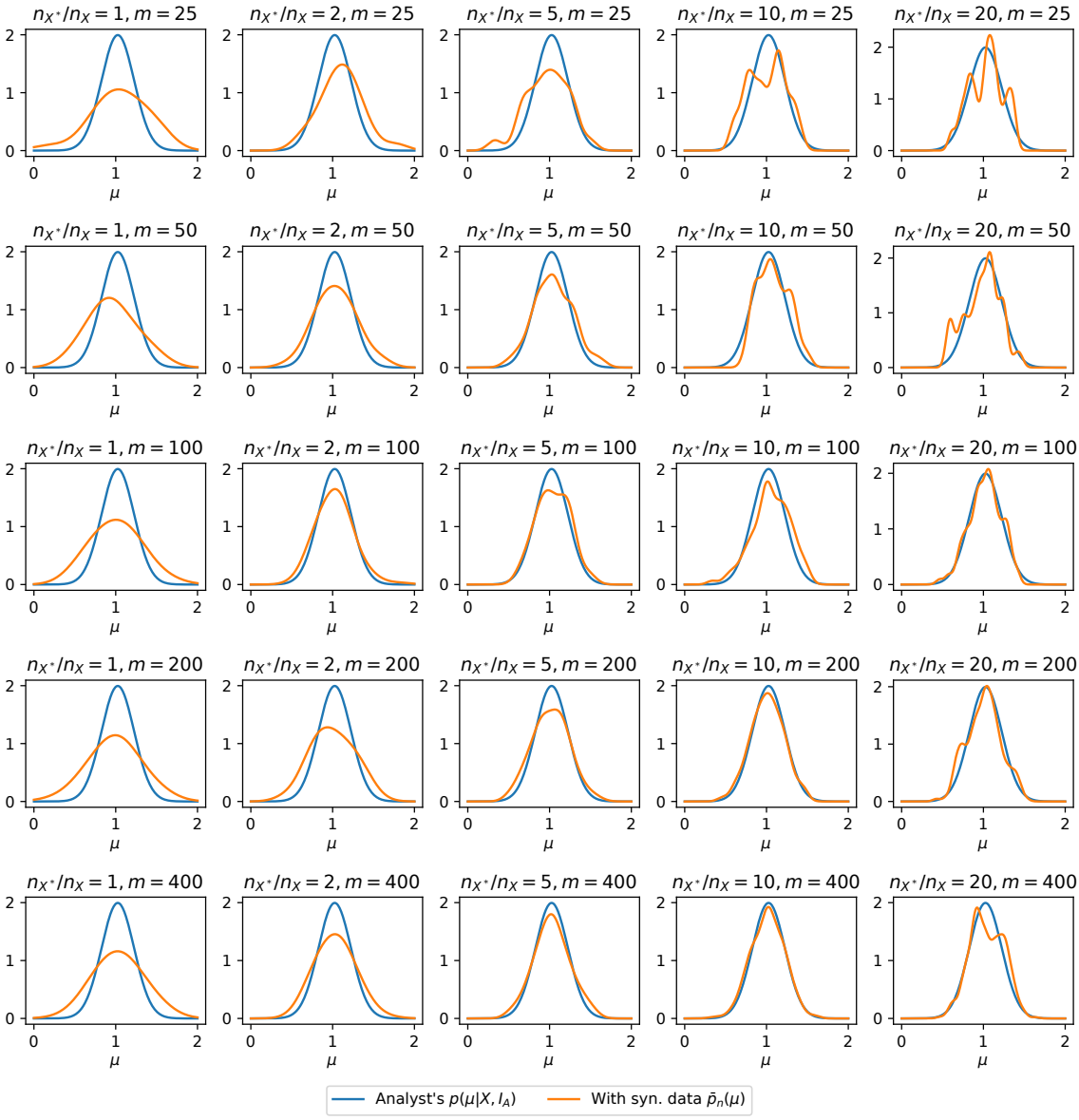



Figure 3: Convergence of the mixture of synthetic data posteriors (in orange) with different values of $m$ (increasing top-to-bottom) and $n_{X^*}$ (increasing left-to-right) in Gaussian mean estimation with known variance.

In this case, the synthetic data is generated from the Gaussian mean estimation with unknown variance model, so $p(X^*|X)$ is

$$\begin{aligned}\bar{\sigma}^2|X &\sim \text{Inv-}\chi^2(\bar{\nu}_{n_X}, \bar{\sigma}^2_{n_X})\\ \bar{\mu}|\bar{\sigma}^2, X &\sim \mathcal{N}\left(\bar{\mu}_{n_X}, \frac{\bar{\sigma}^2}{\bar{\kappa}_{n_X}}\right)\\ X^*|\bar{\mu}, \bar{\sigma}^2 &\sim \mathcal{N}^{n_{X^*}}(\bar{\mu}, \bar{\sigma}^2).\end{aligned}$$

with

$$\begin{aligned}s^2 &= \frac{1}{n_X - 1}\sum_{i=1}^{n}(x_i - \bar{X})^2\\ \bar{\mu}_n &= \frac{\bar{\kappa}_0}{\bar{\kappa}_0 + n_X}\bar{\mu}_0 + \frac{n_X}{\bar{\kappa}_0 + n_X}\bar{X}\\ \bar{\kappa}_n &= \bar{\kappa}_0 + n_X\\ \bar{\nu}_n &= \bar{\nu}_0 + n_X\\ \bar{\nu}_n\bar{\sigma}^2_n &= \bar{\nu}_0\bar{\sigma}^2_0 + (n_X - 1)s^2 + \frac{\bar{\kappa}_0 n_X}{\bar{\kappa}_0 + n_X}(\bar{X} - \bar{\mu}_0)^2.\end{aligned}$$

When downstream analysis is the model with known variance $\hat{\sigma}^2_k$, $p(\mu^*|X^*)$ is

$$\begin{aligned}\mu^*|X^* &\sim \mathcal{N}(\hat{\mu}_{n_{X^*}}, \hat{\sigma}^2_{n_{X^*}})\\ \hat{\mu}_{n_{X^*}} &= \frac{\frac{1}{\hat{\sigma_0}^2}\hat{\mu}_0 + \frac{n_{X^*}}{\hat{\sigma}^2_k}\bar{X}^*}{\frac{1}{\hat{\sigma}^2_0} + \frac{n_{X^*}}{\hat{\sigma}^2_k}}\\ \frac{1}{\hat{\sigma}^2_{n_{X^*}}} &= \frac{1}{\hat{\sigma}^2_0} + \frac{n_{X^*}}{\hat{\sigma}^2_k}.\end{aligned}$$

The following proposition shows how the mean and variance of the mixture of posteriors from synthetic data behave as the size of the synthetic data grows.

**Proposition 16.** *If $\mu^*$ is a sample from $\bar{p}_n(\mu)$ in Gaussian mean estimation with synthetic data generation assuming unknown variance, but downstream analysis assuming known variance,*

$$\mathbb{E}_{\mu^*}(\mu^*) \to \bar{\mu}_{n_X}, \quad \text{Var}_{\mu^*}(\mu^*) \to \frac{\bar{\sigma}^2_0}{\bar{\kappa}_{n_X}} \tag{20}$$

*as $n_{X^*} \to \infty$.*

**Proof** See Appendix B.4. ■

This means that $\mu^*$ asymptotically has the same mean and variance as the marginal posterior $p(\mu|X, I_S)$ of $\mu$ in the synthetic data model, which is not the same as the downstream posterior distribution $p(\mu|X, I_A)$ on the real data.

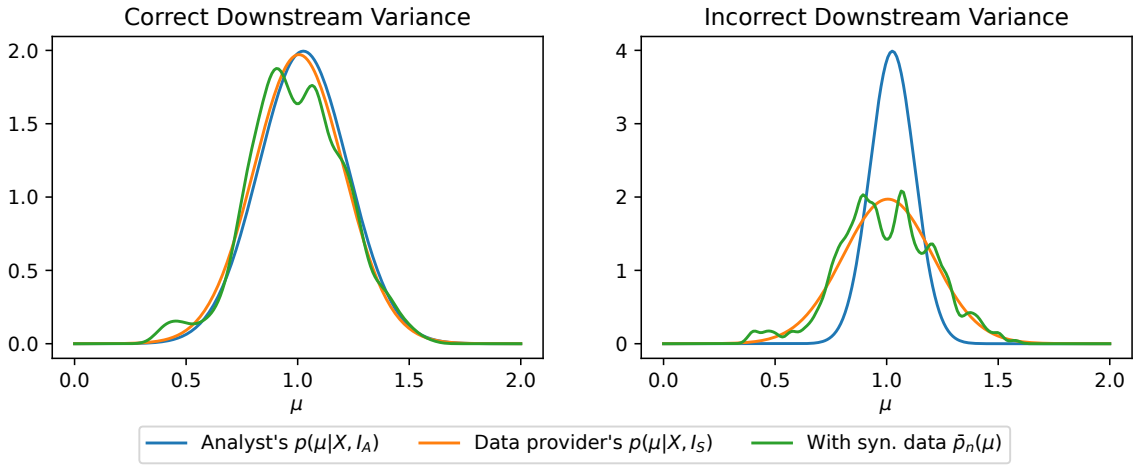



Figure 4: Results when the synthetic data is generated from the unknown variance Gaussian mean estimation model, and the analyst uses the model with known variance. On the left, the analyst's known variance is correct, on the right it is incorrect. In both cases, the mixture of synthetic data posteriors converges to the data provider's posterior. In both panels, $m = 400$ and $\frac{n_{X^*}}{n_X} = 20$.

We verify this with the simulation in Figure 4, where the synthetic data is generated from the model with unknown variance, while the analyst uses the known variance model. The setting is otherwise identical to the case where both used the known variance model in Figure 2. The mixture of synthetic data posteriors converges to the data provider's posterior, even when the analyst uses an incorrect value for the known variance $\hat{\sigma}_k^2$.

### 4.3 Size of the Synthetic Data set

In the preceding analysis, most of the approximations hold when $n_{X^*}$ is large, even when $n_{X^*} \approx n_X$. However, based on the experiment with different values of $n_{X^*}$ and $m$ in Figure 3, $n_{X^*} \gg n_X$ is needed for all of the approximations to hold.

This is explained by looking at $\mathrm{Var}_{X^*}(\bar{X}^*)$. In the case where both parties use the known variance model,

$$\begin{aligned}\mathrm{Var}_{X^*}(\bar{X}^*) &= \frac{1}{n_{X^*}}\mathbb{E}_{\bar{\mu}}(\mathrm{Var}_{x^*|\bar{\mu}}(x_i^*)) + \mathrm{Var}_{\bar{\mu}}(\bar{\mu}) = \frac{1}{n_{X^*}}(\bar{\sigma}_k^2 + \bar{\sigma}_{n_X}^2) + \bar{\sigma}_{n_X}^2 \\ &= \frac{1}{n_{X^*}}\bar{\sigma}_k^2 + \left(1 + \frac{1}{n_{X^*}}\right)\frac{1}{\frac{1}{\bar{\sigma}_0^2} + \frac{n_X}{\bar{\sigma}_k^2}}.\end{aligned}$$

If $n_X \approx n_{X^*}$ and both are large, $1 + \frac{1}{n_{X^*}} \approx 1$ and $\frac{1}{\bar{\sigma}_0^2} + \frac{n_X}{\bar{\sigma}_k^2} \approx \frac{n_X}{\bar{\sigma}_k^2}$, so

$$\mathrm{Var}_{X^*}(\bar{X}^*) \approx \frac{\bar{\sigma}_k^2}{n_{X^*}} + \frac{\bar{\sigma}_k^2}{n_X} \approx \frac{2\bar{\sigma}_k^2}{n_X}.$$

With these approximations,

$$\mathrm{Var}_{\bar{\mu}}(\bar{\mu}) \approx \frac{\bar{\sigma}_k^2}{n_X},$$

so

$$\mathrm{Var}_{X^*}(\bar{X}^*) \approx 2\mathrm{Var}_{\bar{\mu}}(\bar{\mu}),$$

while the $n_{X^*} \to \infty$ limit is $\mathrm{Var}_{X^*}(\bar{X}^*) \to \mathrm{Var}_{\bar{\mu}}(\bar{\mu})$. This means that $n_{X^*} \gg n_X$ is required.

The same happens when the synthetic data is generated from the unknown variance model:

$$\mathrm{Var}_{X^*}(\bar{X}^*) = \frac{1}{n_{X^*}}\mathbb{E}_{\bar{\sigma}^2}(\bar{\sigma}^2) + \mathrm{Var}_{\bar{\mu}}(\bar{\mu}) = \frac{1}{n_{X^*}}\frac{\bar{\nu}_0 + n_X}{\bar{\nu}_0 + n_X - 2}\bar{\sigma}_{n_X}^2 + \frac{\bar{\sigma}_{n_X}^2}{\bar{\kappa}_0 + n_X}.$$

If $n_X \approx n_{X^*}$ and both are large, $\frac{\bar{\nu}_0+n_X}{\bar{\nu}_0+n_X-2} \approx 1$ and $\bar{\kappa}_0 + n_X \approx n_X$, so

$$\mathrm{Var}_{X^*}(\bar{X}^*) \approx \frac{\bar{\sigma}_{n_X}^2}{n_{X^*}} + \frac{\bar{\sigma}_{n_X}^2}{n_X} \approx \frac{2\bar{\sigma}_{n_X}^2}{n_X}.$$

With these approximations,

$$\mathrm{Var}_{\bar{\mu}}(\bar{\mu}) \approx \frac{\bar{\sigma}_{n_X}^2}{n_X},$$

so

$$\mathrm{Var}_{X^*}(\bar{X}^*) \approx 2\mathrm{Var}_{\bar{\mu}}(\bar{\mu}).$$

This leads to the same conclusion as the previous case: $n_{X^*} \gg n_X$ is required.

### 4.4 Approximate Variance Correction

With similar approximations, we can find a correction term for the variance. When $n_{X^*}$ is large,

$$\mathrm{Var}_{\mu^*}(\mu^*) \approx \mathbb{E}_{X^*}(\hat{\sigma}_{n_{X^*}}^2) + \mathrm{Var}_{X^*}(\bar{X}^*).$$

If $n_{X^*} = cn_X$ for some $c > 0$, from the analyses in Section 4.3, we get

$$\mathrm{Var}_{X^*}(\bar{X}) \approx \left(1 + \frac{1}{c}\right)\mathrm{Var}_{\bar{\mu}}(\bar{\mu}),$$

so

$$\mathrm{Var}_{\mu^*}(\mu^*) \approx \mathbb{E}_{X^*}(\hat{\sigma}_{n_{X^*}}^2) + \left(1 + \frac{1}{c}\right)\mathrm{Var}_{\bar{\mu}}(\bar{\mu}).$$

Solving for $\mathrm{Var}_{\bar{\mu}}(\bar{\mu})$ gives

$$\mathrm{Var}_{\bar{\mu}}(\bar{\mu}) \approx \left(1 + \frac{1}{c}\right)^{-1}\left(\mathrm{Var}_{\mu^*}(\mu^*) - \mathbb{E}_{X^*}(\hat{\sigma}_{n_{X^*}}^2)\right). \tag{21}$$

which gives a Rubin's rules-like (Rubin, 1987) approximation of $\mathrm{Var}_{\bar{\mu}}(\bar{\mu})$ that can be computed from smaller synthetic data sets with $n_{X^*} \approx n_X$.

We validate this with the experiment in Figure 5, which shows that approximating $p(\mu|X, I_A)$ with a Gaussian with variance from (21) is closer to the real data posterior than the mixed posterior approximation from Section 3.

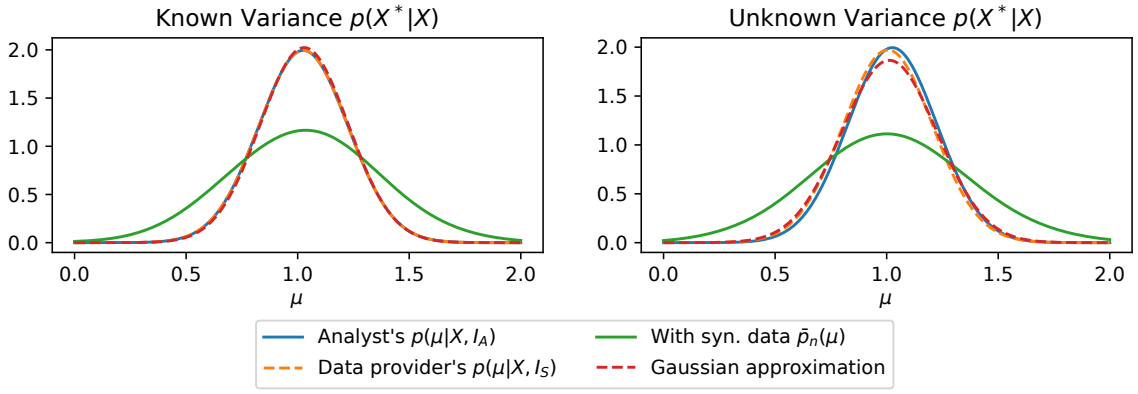



Figure 5: Results with the Gaussian approximation with $n_{X^*} = n_X$, showing that the Gaussian approximation is closer to the real data posterior than the mixture of synthetic data posteriors. On the left, the synthetic data is generated from the known variance model, and on the right, the synthetic data is generated from the unknown variance model. In both cases, the known variances for both parties are correct, and $m = 400$.

| | Data Provider | Analyst |
|---|---|---|
| Known Mean | $\bar{\mu}_k^2$ | $\hat{\mu}_k^2$ |
| Prior Parameters | $\bar{\nu}_0, \bar{\sigma}_0^2$ | $\hat{\nu}_0, \hat{\sigma}_0^2$ |
| Data | $X$ | $X^*$ |
| Single Datapoint | $x$ | $x^*$ |
| Data Size | $n_X$ | $n_{X^*}$ |
| Data Variance | $\bar{v}$ | $\hat{v}$ |
| Posterior Parameters | $\bar{\nu}_{n_X}, \bar{\sigma}_{n_X}^2$ | $\hat{\nu}_{n_{X^*}}, \hat{\sigma}_{n_{X^*}}^2$ |
| Posterior Sample | $\bar{\sigma}^2$ | $\hat{\sigma}^2$ |
| Mixture of Posteriors Sample | | $\sigma_*^2$ |

Table 3: Notation for known mean, unknown variance model in Section 4.5.

### 4.5 Gaussian with Known Mean, Unknown Variance

To asses the effects of uncongeniality when the downstream posterior is not Gaussian, we look at Bayesian estimation of the variance of a Gaussian, with known mean. The notation for this section is summarised in Table 3.

In this case, the data provider's conjugate prior is

$$\bar{\sigma}^2 \sim \text{Inv-}\chi^2(\bar{\nu}_0, \bar{\sigma}_0^2),$$

and their known mean is $\bar{\mu}_k$. The synthetic data is generated from

$$\begin{aligned}
x_i^* | \bar{\sigma}^2 &\sim \mathcal{N}(\bar{\mu}_k, \bar{\sigma}^2) \\
\bar{\sigma}^2 | X &\sim \text{Inv-}\chi^2(\bar{\nu}_{n_X}, \bar{\sigma}_{n_X}^2) \\
\bar{\sigma}_{n_X}^2 &= \frac{\bar{\nu}_0 \bar{\sigma}_0^2 + n_X \bar{v}}{\bar{\nu}_0 + n_X} \\
\bar{\nu}_{n_X} &= \bar{\nu}_0 + n_X \\
\bar{v} &= \frac{1}{n_X} \sum_{i=1}^{n_X} (x_i - \bar{\mu}_k)^2.
\end{aligned}$$

The analyst's conjugate prior is

$$\hat{\sigma}^2 \sim \text{Inv-}\chi^2(\hat{\nu}_0, \hat{\sigma}_0^2),$$

their known mean is $\hat{\mu}_k$, and the downstream posterior is

$$\begin{aligned}
\hat{v} &= \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} (x_i^* - \hat{\mu}_k)^2 \\
\hat{\nu}_{n_{X^*}} &= \hat{\nu}_0 + n_{X^*} \\
\hat{\sigma}_{n_{X^*}}^2 &= \frac{\hat{\nu}_0 \hat{\sigma}_0^2 + n_{X^*} \hat{v}}{\hat{\nu}_0 + n_{X^*}} \\
\hat{\sigma}^2 | X^* &\sim \text{Inv-}\chi^2(\hat{\nu}_{n_{X^*}}, \hat{\sigma}_{n_{X^*}}^2).
\end{aligned}$$

The behavior of the mixture of posteriors from synthetic data is given in the following proposition.

**Proposition 17.** *If $\sigma_*^2$ is a sample from $\bar{p}_n(\sigma)$ in Gaussian variance estimation with known mean,*

$$\mathbb{E}_{\sigma_*^2}(\sigma_*^2) \to \mathbb{E}_{\bar{\sigma}^2}(\bar{\sigma}^2) + (\bar{\mu}_k - \hat{\mu}_k)^2, \tag{22}$$

*as $n_{X^*} \to \infty$.*

**Proof** See Appendix B.4. ■

This means that mixing the downstream posteriors can only recover the data provider's posterior when both parties have equal known means.

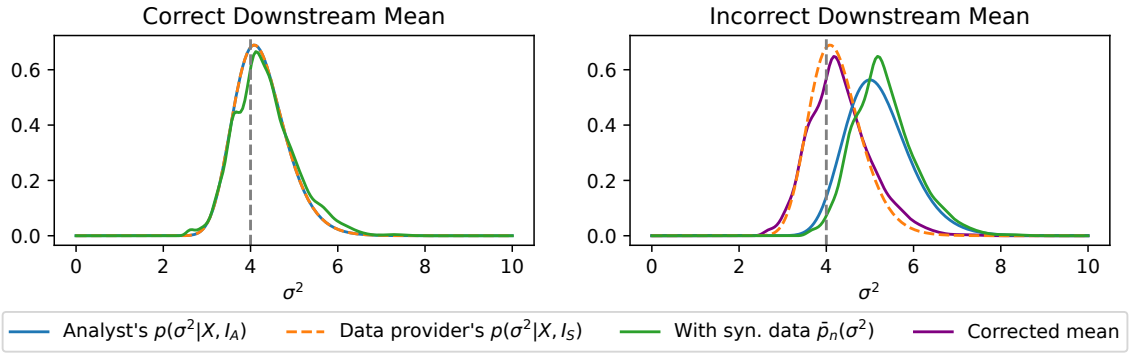



Figure 6: Posteriors from estimating a Gaussian variance $\sigma^2$ with known mean. On the right, both the data provider and the analyst use the correct known mean, so $\bar{p}_n(\sigma^2)$ converges as expected by our theory. On the right, the analyst mean is incorrect, so the model is not congenial. In this case, $\bar{p}_n(\sigma^2)$ does not converge to either the analyst's or the data provider's posterior. After correcting the mean of $\bar{p}_n(\sigma^2)$ as in (22), it appears to have the same variance and shape as the data provider's posterior. The gray line shows the true parameter value. In both panels, $m = 400$ and $\frac{n_{X^*}}{n_X} = 20$.

We verify this with a simulation shown in Figure 6. Both the data provider and analyst use the Gaussian with unknown variance and known mean as their model. Otherwise, the setting is identical with the other Gaussian examples. When both parties have the correct known mean, $\bar{p}_n(\sigma^2)$ converges as expected, but when the analyst has an incorrect known mean, $\bar{p}_n(\sigma^2)$ converges to neither party's posterior. However, after applying the mean correction from (22), $\bar{p}_n(\sigma^2)$ appears to have the same variance and shape as the data provider's posterior.

## 5. Differentially Private Logistic Regression

Our second example is logistic regression with DP synthetic data. We consider two settings used by Räisä et al. (2023) with frequentist logistic regression, and change the downstream task to Bayesian logistic regression. The first setting uses a simple toy dataset, and the second uses the UCI Adult dataset (Kohavi and Becker, 1996). We also consider a hierarchical logistic regression task on the Adult dataset.

Under DP, $Z$ is a noisy summary $\tilde{s}$ of the real data. We need synthetic data sampled from the posterior predictive $p(X^*|\tilde{s})$, which is exactly what the NAPSU-MQ algorithm of Räisä et al. (2023) provides. In NAPSU-MQ, $\tilde{s}$ contains the values of user-selected marginal queries with added Gaussian noise. We used the open-source implementation of NAPSU-MQ[3] by Räisä et al. (2023), and describe NAPSU-MQ in Section 2.3.

To demonstrate our theory outside of the NAPSU-MQ algorithm, we include the synthpop (Nowok et al., 2016) synthetic data genetator, which does not provide DP, in the Adult

3. `https://github.com/DPBayes/NAPSU-MQ-experiments`

logistic regression experiment. Like NAPSU-MQ, synthpop attempts to generate synthetic data from a posterior predictive distribution, which is done by sequentially training predictive models to predict one column from previous ones. Synthetic data is generated column-by-column by sampling the first column from the real data, and sampling the successive columns predicting them from the already sampled columns using the predictive models. Note that both synthpop and NAPSU-MQ generate both the covariates and the target variable jointly.

We asses the quality of posteriors from different methods by the coverages and widths of their *credible intervals*. Recall that a credible interval is an interval that contains a given amount of posterior probability mass, for example 95%, which we call the confidence level of the interval. The coverage is the number of times, in repeated experiments, the credible interval contained the true parameter value. We want the coverage to be above the confidence level for the interval to be valid. We also compare marginals of the posterior visually.

### 5.1 Posterior Sampling Background

For the DP experiments, we rely on Markov chain Monte Carlo (MCMC) algorithms (Gelman et al., 2014), in particular the No-U-Turn sampler (NUTS; Hoffman and Gelman, 2014), to sample the posterior in NAPSU-MQ. Recall that MCMC algorithms sample a distribution with the density only known up to a normalising constant by obtaining many samples from a Markov chain. We discard the initial elements of chain, where it has not yet converged (Gelman et al., 2014). We call these discarded elements warmup samples, and call the rest of elements kept samples. We run the sampler multiple times from different starting points, forming multiple chains, and mix the kept samples from each chain to form the final posterior sample. We check for convergence of the sampler with the $\hat{R}$ statistic (Gelman et al., 2014).

For downstream posteriors, we use Laplace approximations (Gelman et al., 2014), also known as Laplace's method. The Laplace approximation is simply a Gaussian distribution centered at the mode of the posterior, and the covariance is the negated inverse Hessian of the log posterior density at the mode.

### 5.2 Toy Data Logistic Regression

The first logistic regression setting we consider uses a simple toy data set of three binary variables, with $n_X = 2000$ samples. The first two variables are sampled with independent coinflips, and the third is sampled from logistic regression on the other two, with coefficients $(1, 0)$. The prior for the downstream logistic regression is $\mathcal{N}(0, 10I)$.

We generate synthetic data with the NAPSU-MQ algorithm (Räisä et al., 2023), instructing the algorithm to generate $m$ synthetic data sets of size $n_{X^*}$. For the privacy bounds, we vary $\epsilon$, and set $\delta = n_X^{-2} = 2.5 \cdot 10^{-7}$.

Because of the simplicity of this model, it is possible to use the exact posterior decomposition (4) as a baseline, by using $p(X|\tilde{s})$ instead of $p(X^*|\tilde{s})$ to generate synthetic data. We give a detailed description of this process in Appendix C. We have also included the DP-GLM algorithm (Kulkarni et al., 2021) that does not use synthetic data, and the non-DP posterior from the real data as baselines. We obtained the code for DP-GLM from Kulkarni et al. (2021) upon request.

#### 5.2.1 Hyperparameters

For NAPSU-MQ, we use the hyperparameters of Räisä et al. (2023), except we used NUTS (Hoffman and Gelman, 2014) with 200 warmup samples and 500 kept samples per chain for $\epsilon \in \{0.5, 1\}$, and 1500 kept samples per chain for $\epsilon = 0.1$, as the posterior sampling algorithm. The NAPSU-MQ prior is $\mathcal{N}(0, 10^2 I)$, and the summary is the single 3-way marginal query over all three variables.

The hyperparameters of DP-GLM are the $L_2$-norm upper bound $R$ for the covariates of the logistic regression, a coefficient norm upper bound $s$, and the parameters of the posterior sampling algorithm DP-GLM uses. We set $R = \sqrt{2}$ so that the covariates do not get clipped, and set $s = 5$ after some preliminary runs. The posterior sampling algorithm is NUTS (Hoffman and Gelman, 2014) with 1000 warmup samples and 1000 kept samples from 4 parallel chains.

#### 5.2.2 Results

Figure 7 compares the mixture of posteriors from synthetic data $\bar{p}_n(Q)$ from (11) that uses $p(Q|X^*)$, with $n_{X^*}/n_X = 20$ and $m = 400$ synthetic data sets, to the baselines. The mixture $\bar{p}_n(Q)$ is very close to the posterior $p(Q|\tilde{s})$ from (4). The DP-GLM posterior that does not use synthetic data is somewhat wider.

We ran the experiment 100 times and also with $\epsilon = 0.1$ and $\epsilon = 0.5$, and plot coverages and widths of credible intervals in Figure 8. With $\epsilon = 1$ and $\epsilon = 0.5$, the coverages are accurate and DP-GLM consistently produces wider intervals. With $\epsilon = 0.1$, the mixture of synthetic data posteriors likely needs more and larger synthetic data sets to converge, as it produced wider and slightly overconfident intervals for one coefficient.

Figures 9, 10 and 11 look at how $\bar{p}_n(Q)$ converges to $p(Q|\tilde{s})$ both visually and in terms of total variation distance as $n_{X^*}$ and $m$ increase. They show that both need to be increased together for $\bar{p}_n(Q)$ to converge to the target, otherwise the total variation distance plateaus at some point, and there is a clear visual difference.

#### 5.2.3 Plotting Details

The plotted density of DP-GLM in Figure 7 is a kernel density estimate from the posterior samples DP-GLM returns. The non-DP density is a Laplace approximation. Both synthetic data methods use Laplace approximations in the downstream analysis, so their posteriors are mixtures of these Laplace approximations for each synthetic data set. This was also used in Figure 9.

### 5.3 UCI Adult Logistic Regression

To test our theory on real data, we used the UCI Adult data set (Kohavi and Becker, 1996) setting that was used to test NAPSU-MQ (Räisä et al., 2023). We generate synthetic data with either NAPSU-MQ under DP, or with the synthpop algorithms (Nowok et al., 2016) without DP. In this setting, the synthetic data set is generated from a subset of 10 columns[4], with the continuous columns age and hours-per-week discretised to 5 categories,

4. age, workclass, education, marital-status, race, gender, capital-gain, capital-loss, hours-per-week and income

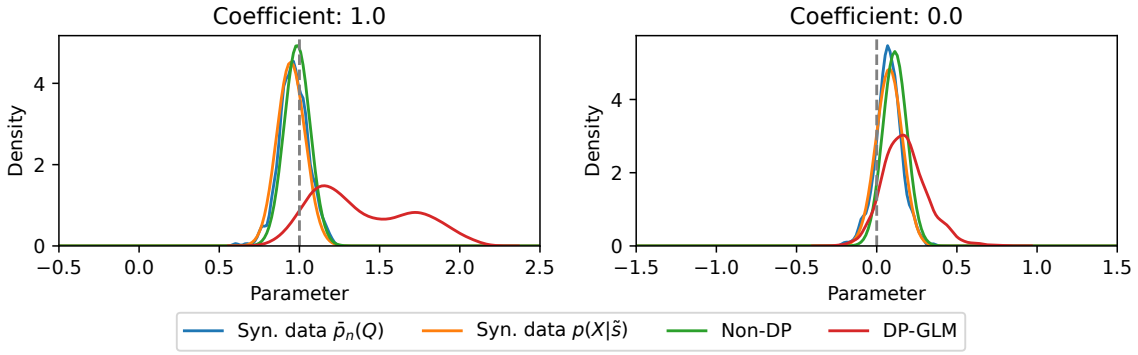


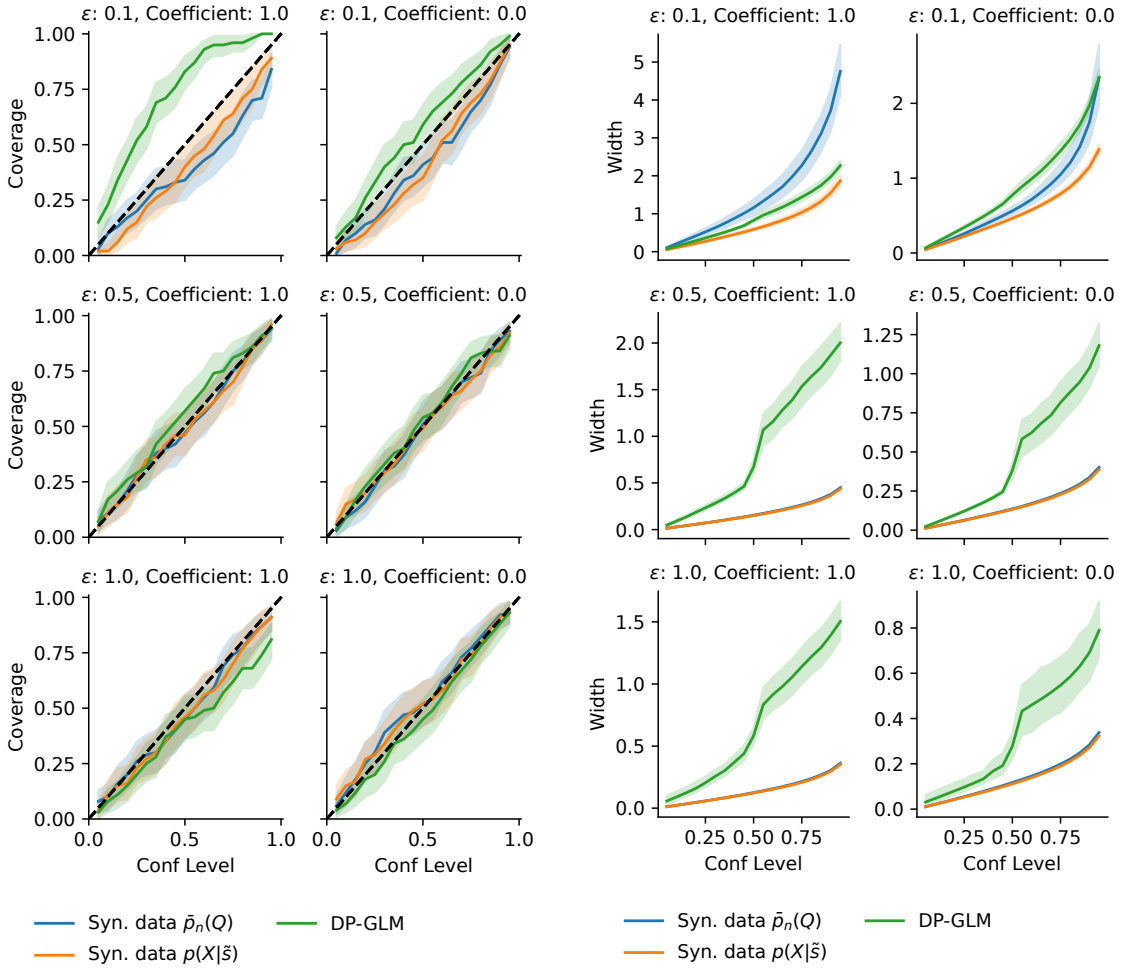

Figure 7: Posteriors in the DP logistic regression experiment, where $Q$ are the regression coefficients. The mixture of posteriors from synthetic data, $\bar{p}_n(Q)$, (with $n_{X^*}/n_X = 20$, $m = 400$) is very close the to the private posterior $p(Q|\tilde{s})$ computed using (4). Computing the posterior without synthetic data with DP-GLM gives a somewhat wider posterior. The true parameter values are highlighted by the grey dashed lines and shown in the panel titles. The privacy bounds are $\epsilon = 1$, $\delta = n_X^{-2} = 2.5 \cdot 10^{-7}$.

and capital-loss and capital-gain binarised according to whether they are greater than 0 or not. The income column is already binarised in the original data to denote whether it is over \$50000 or not. All rows with missing values in the original data set are deleted, which results in $n_X = 46043$ datapoints. The downstream task is logistic regression predicting income using age, race and gender, with age converted back to a continuous value by picking the midpoint of each category. The reference value for race is "white" and for gender is "female". These subsets were originally used to make the runtime of NAPSU-MQ manageable, and to make sure that enough relevant information for the downstream task can be included in the input queries for NAPSU-MQ (Räisä et al., 2023). We take bootstrap samples of the data to simulate draws from a population.

#### 5.3.1 Algorithms and Hyperparameters

The target distribution $p(Q|Z)$ is not tractable in this setting, so we used the non-DP Laplace approximation from the original data set as an approximate ground truth posterior, and the DP variational inference (DPVI) algorithm (Jälkö et al., 2017; Prediger et al., 2022) as a baselines. We also tried running DP-GLM (Kulkarni et al., 2021), but we were not able to get useful results out of it in this setting. We have also included the Gaussian approximation to the mixture of synthetic data posteriors discussed in Section 4.4, which is called "variance correction" in the figures. Unlike the mixture of posteriors, the variance correction approximation uses synthetic datasets of the same size as the real dataset.

The prior for the downstream Bayesian logistic regression is $\mathcal{N}(0, 10)$, i.i.d. for each coefficient. The privacy parameters are $\epsilon \in \{0.25, 0.5, 1\}$, and $\delta = n_X^{-2} \approx 4.7 \cdot 10^{-10}$. We repeat the experiment 20 times.




(a) (b)

Figure 8: (a) Coverages of credible intervals in the toy data experiment. The mixture of synthetic data posteriors is accurate, except with $\epsilon = 0.1$, where it may not have converged yet. (b) Widths of credible intervals in the toy data experiment. DP-GLM produces much wider intervals than other methods, except with $\epsilon = 0.1$.

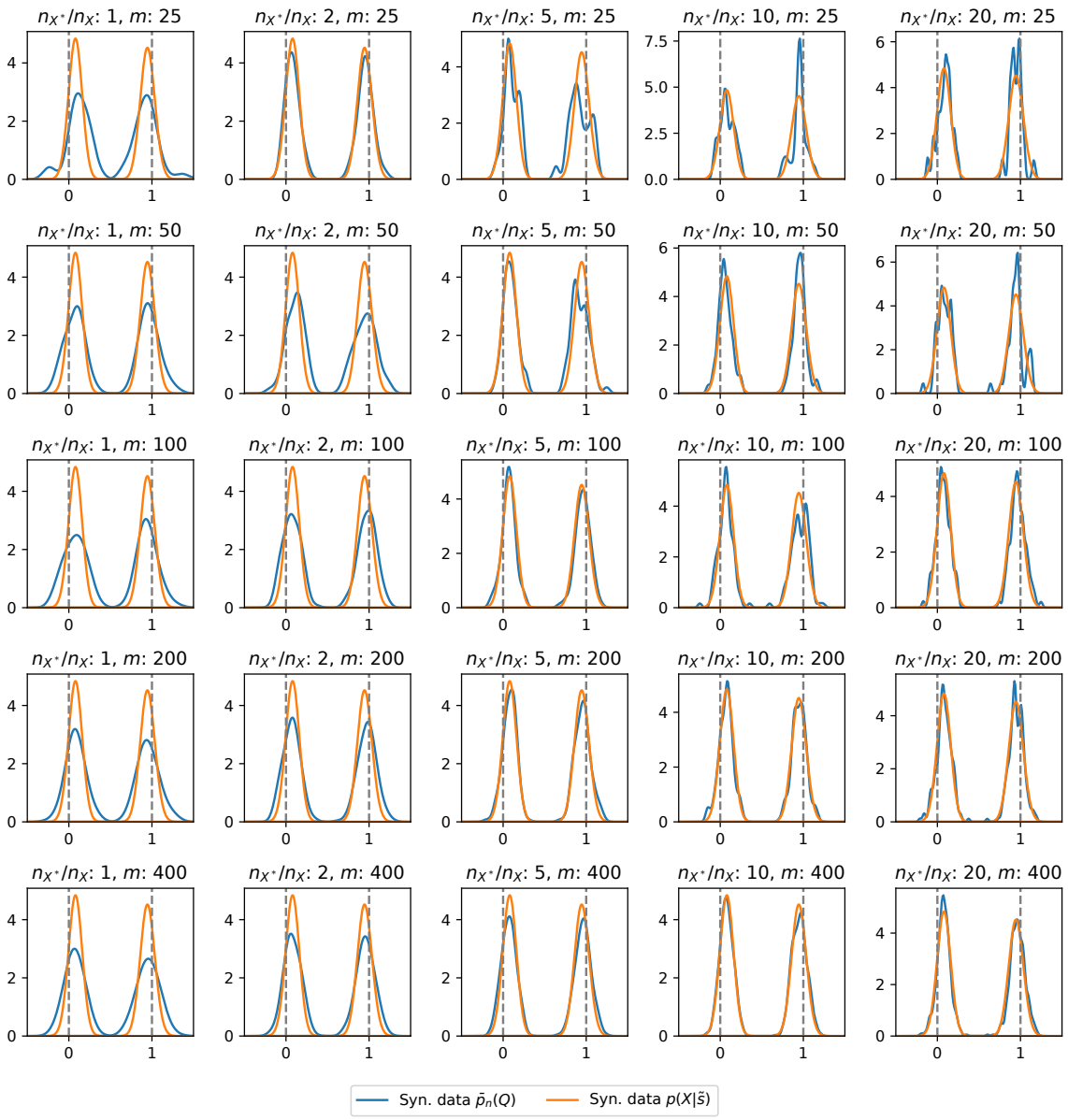



Figure 9: Convergence of the mixture of synthetic data posteriors (in blue) with different values of $m$ and $n_{X^*}$ in the toy data logistic regression experiment.

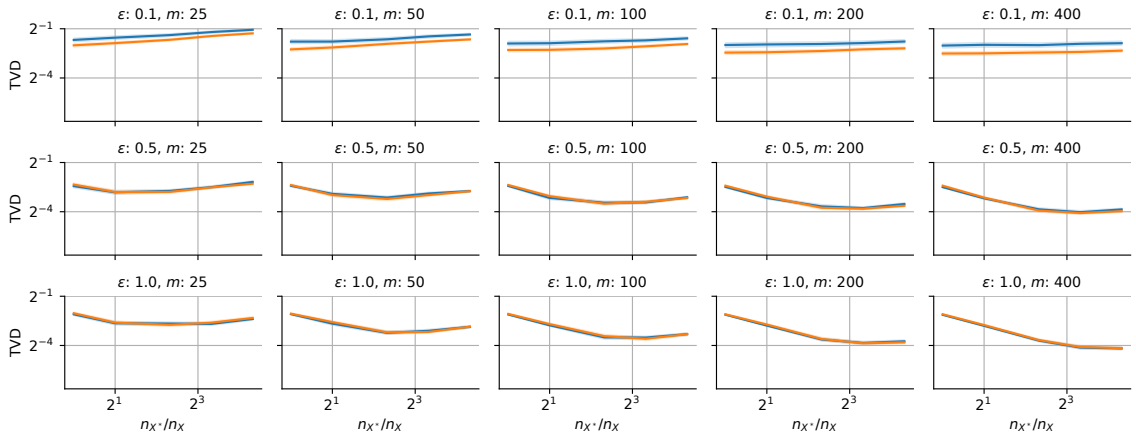



Figure 10: Total variation distance (TVD) between $\bar{p}_n(Q)$ and the target $p(Q|\tilde{s})$ for both 1D marginals (blue and orange) in the toy data experiment. For $\epsilon = \{0.5, 1\}$, increasing the size of the synthetic data sets $n_{X^*}$ decreases the total variation distance at a steady rate, until hitting a point where the decrease stops. This point moves further as number of synthetic data sets $m$ increases.

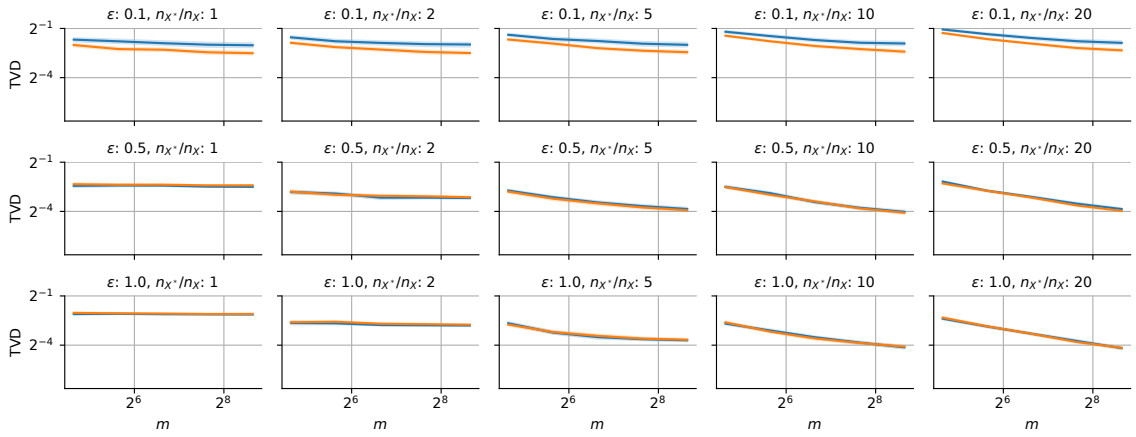



Figure 11: Total variation distance (TVD) between $\bar{p}_n(Q)$ and the target $p(Q|\tilde{s})$ for both 1D marginals (blue and orange) in the toy data experiment, with roles of $n_{X^*}$ and $m$ swapped from Figure 10. Increasing $m$ decreases the total variation distance at a rate which depends on $n_{X^*}$ and $\epsilon$.

The hyperparameters, prior, and selected queries for NAPSU-MQ are the same as in the original paper (Räisä et al., 2023). The synthetic data set size and number are $n_{X^*}/n_X = 10$, $m = 100$. For the variance correction approximation, $n_{X^*}/n_X = 1$.

For the predictive models used by synthpop, we used the "parametric" option, which chooses from various parametric models based on the types of the involved variables. We chose the "parametric" option after observing that the default of using decision trees as the predictor resulted in biased downstream estimates of several coefficients.

DPVI runs DP-SGD (Rajkumar and Agarwal, 2012; Song et al., 2013; Abadi et al., 2016), specifically DP-Adam, under the hood, so it inherits the clip bound, learning rate, number of iterations, and subsampling (without replacement) ratio hyperparameters from DP-SGD. We tuned these with the Optuna library (Akiba et al., 2019), using the bounds $[0.1, 50]$ for the clip bound, $[10^{-4}, 10^{-1}]$ for the learning rate, $[10^4, 10^5]$ for the number of iterations and $[0.001, 1]$ for the subsampling ratio. We used the distance of the DPVI posterior mean from the non-DP real data Laplace approximation as the optimisation criterion. We also tried using KL divergence, which gave hyperparameters that produced much wider posteriors. We used 100 trials for the tuning, and repeated it independently for all values of $\epsilon$. The privacy cost of the hyperparameter tuning is not reflected in the final results. As the variational posterior, we used a mean-field Gaussian.

#### 5.3.2 Results

Figure 12 compares posteriors from one of the 20 runs with $\epsilon = 1$. The mixture of synthetic data posteriors $\bar{p}_n(Q)$ from NAPSU-MQ is fairly close to the ground truth posterior from the real data set, with the exception of two coefficients. The Gaussian approximation to $\bar{p}_n(Q)$ from Section 4.4 is very close to $\bar{p}_n(Q)$. The mixture of posteriors from synthpop follows ground truth posterior fairly well. The posteriors from DPVI are close to ground truth posterior, but for some coefficients, they are too narrow to overlap the ground truth posterior.

The logistic regression coefficients for which $\bar{p}_n(Q)$ does not work well correspond to the two races with the smallest number of people in the original data set. These posteriors are very wide due to the fact that NAPSU-MQ adds noise uniformly to all queries, which means that the queries with small values, corresponding to minority groups in the data, get relatively larger amounts of noise. In contrast, synthpop is able to approximate these posteriors well, since it does not need to add noise for DP.

Figure 13 shows credible interval coverages from the experiment for the DP algorithms, computed from 20 runs, with different bootstrap samples each run to simulate draws from a population. The mixture $\bar{p}_n(Q)$ does not achieve perfect coverages, as some information is lost due to not running NAPSU-MQ with all marginal queries. The coverages are still much better than DPVI. The coverages of the variance correction approximation to $\bar{p}_n(Q)$ are close to the coverages of $\bar{p}_n(Q)$.

Figure 14 shows the widths of credible intervals from the same 20 runs. DPVI produces narrower posteriors than $\bar{p}_n(Q)$, but the width of $\bar{p}_n(Q)$ posteriors drops as $\epsilon$ increases, reflecting the reduced uncertainty from DP, which is not the case for DPVI. The variance correction approximation produces narrower posteriors than $\bar{p}_n(Q)$, especially with $\epsilon = 0.25$.

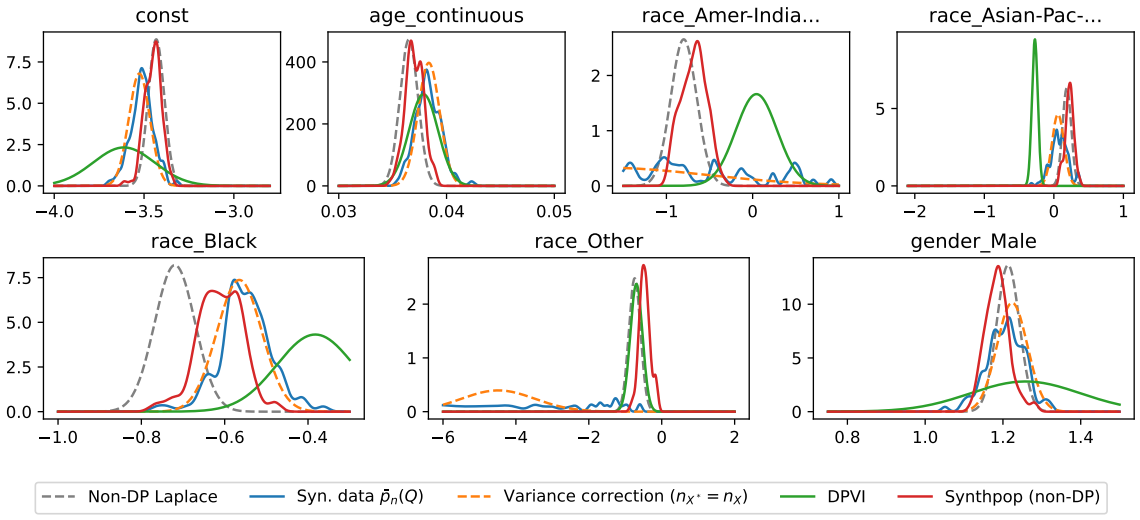



Figure 12: Posteriors from one run on the Adult logistic regression experiment with $\epsilon = 1$.

In summary, while DPVI is able to find the posterior mean fairly well, it fails to accurately reflect the additional uncertainty from DP, making the posteriors overconfident, which would lead to spurious findings if applied in practice. In contrast, $\bar{p}_n(Q)$ accounts for the DP noise very well, at the cost of producing very wide posteriors for the coefficients with a large amount of noise. Estimating uncertainty reliably is much more important than producing a narrow uncertainty estimate: a very wide posterior containing the correct value signals uncertainty, but a narrow posterior in the wrong place is confidently incorrect. The variance correction approximation performed well, despite using smaller synthetic datasets: it had valid coverage in most cases while producing somewhat narrower posteriors than the mixture of posteriors. The narrower posteriors are likely a result of the mixture of posteriors having heavier tails than the Gaussian that the variance correction uses.

Figure 15 shows the credible interval coverages and widths for synthpop, along with NAPSU-MQ results with $\epsilon = 1$ from Figures 13 and 14. Synthpop has similar coverages as NAPSU-MQ, but the credible interval widths demonstrate the price of DP: sythpop results in much narrower intervals due to not needing noise, especially for the two small minority races that NAPSU-MQ struggles with.

#### 5.3.3 Plotting Details

The plotted densities for the non-DP posterior and the mixture of synthetic data posteriors use Laplace approximations like in the toy data experiment. The posterior from DPVI is a multivariate Gaussian, which is plotted as is.

### 5.4 Hierarchical Logistic Regression

We test our theory on a downstream posterior with more complicated geometry with a hierarchical logistic regression model on the UCI Adult (Kohavi and Becker, 1996) dataset.

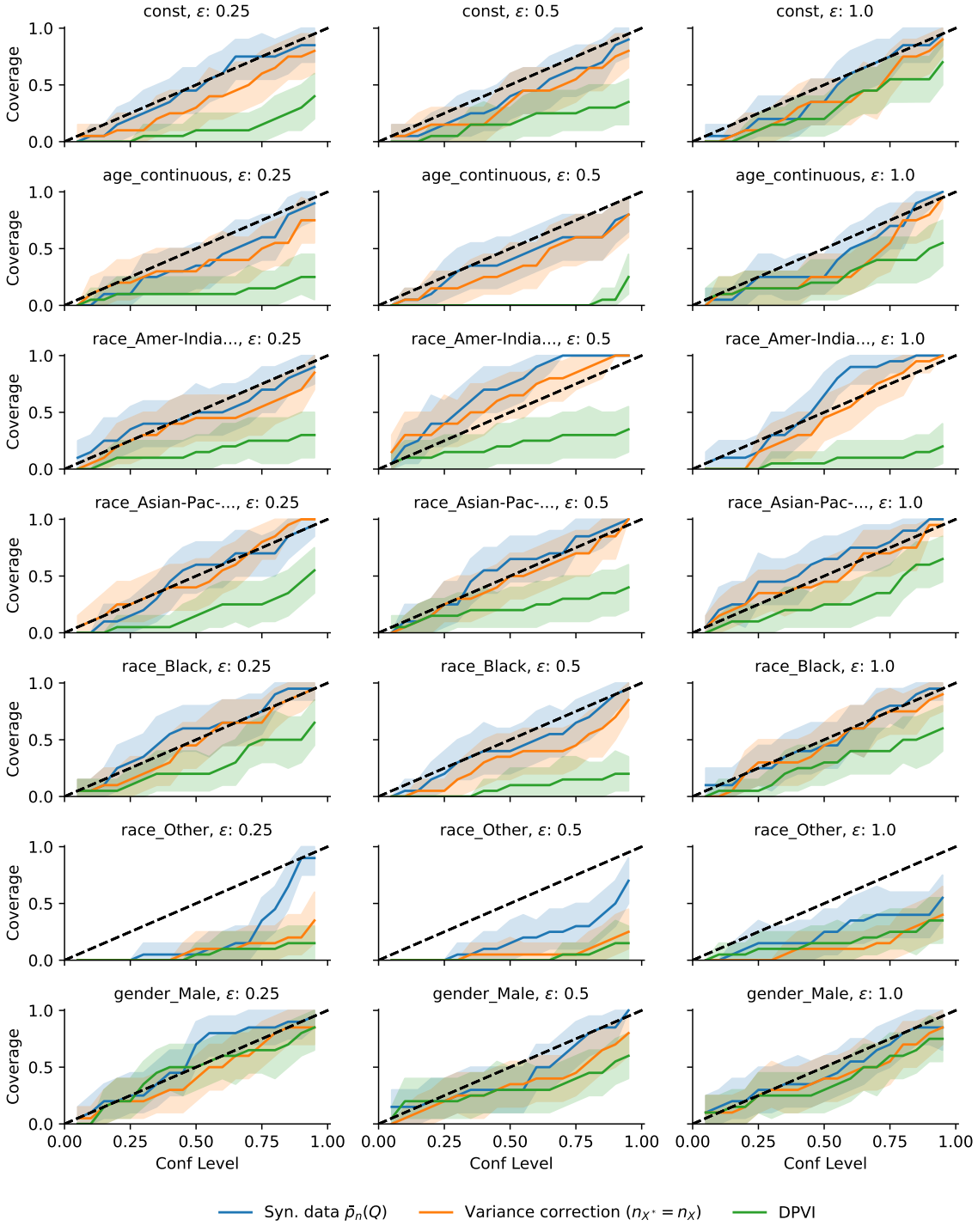



Figure 13: Credible interval coverages on the Adult logistic regression experiment.

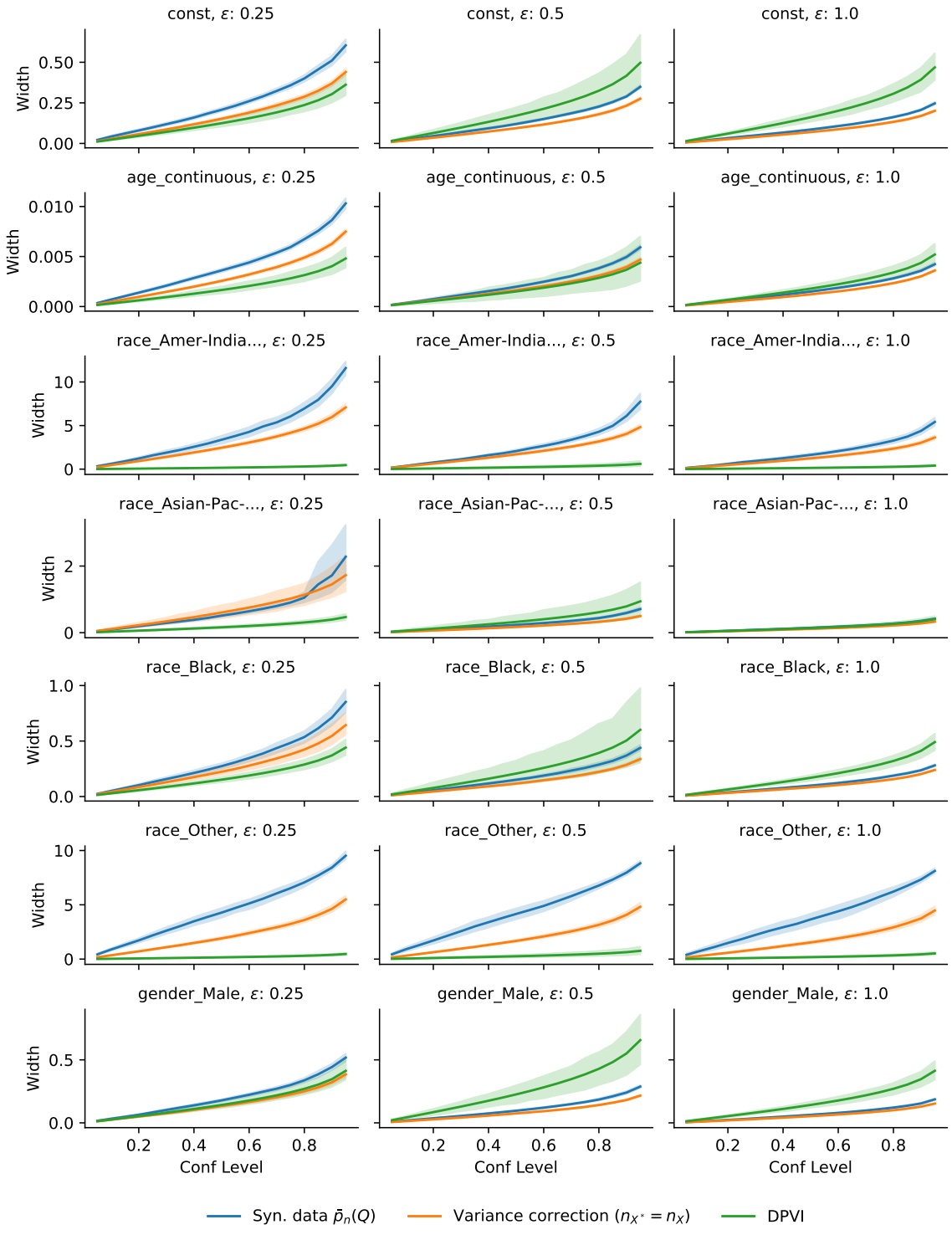



Figure 14: Credible interval widths on the Adult logistic regression experiment.

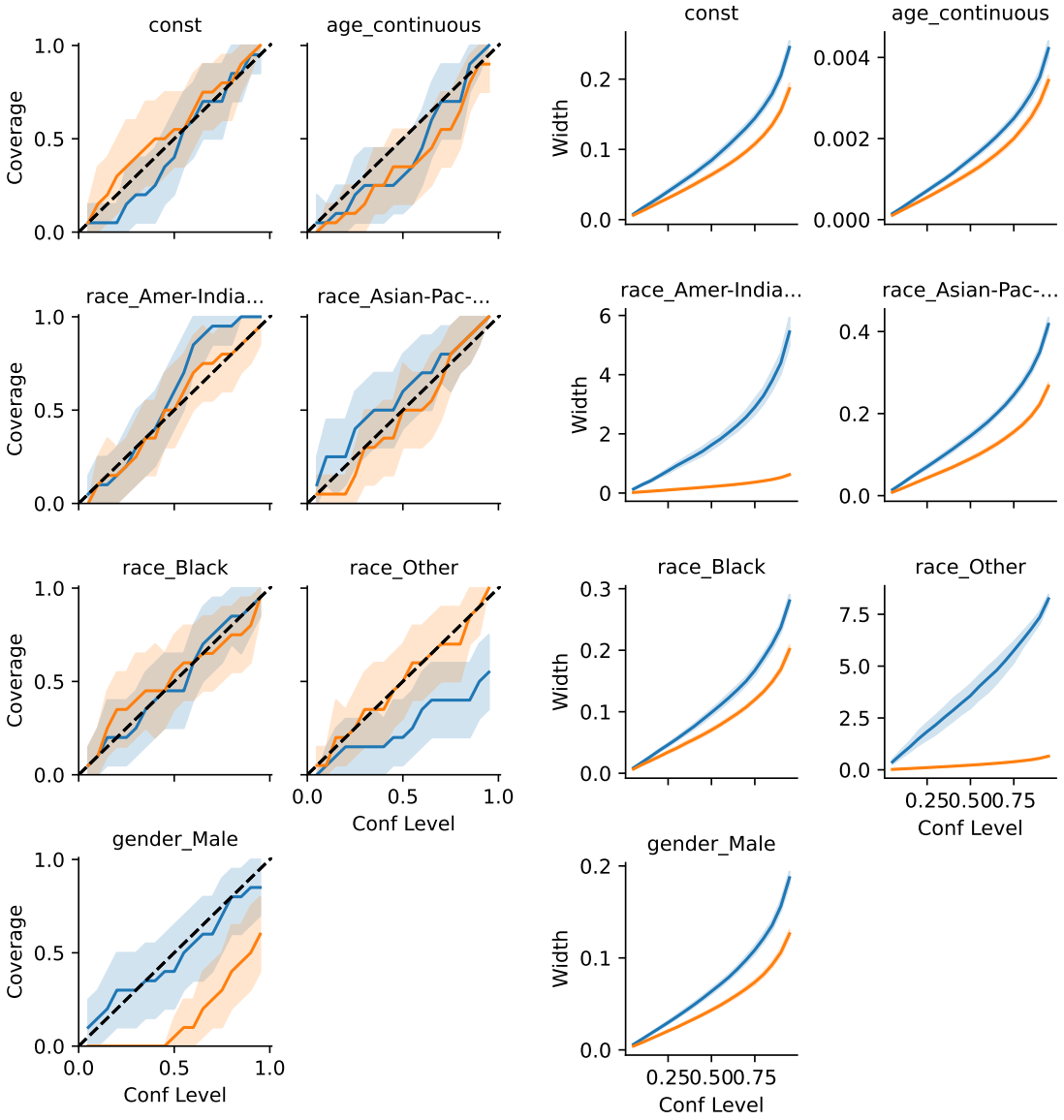



Figure 15: Results on the Adult dataset with NAPSU-MQ ($\epsilon = 1$) and synthpop (non-DP). The panels show (a) Credible interval coverages, (b) Credible interval widths.

We use the same preprocessing steps as in Section 5.3, except we drop two columns[5] from the synthetic data generation to reduce the runtime. The remaining columns are shown as the graph nodes in Figure 16. As in Section 5.3, we take bootstrap samples of the real data before generating synthetic data to simulate draws from a population. We repeat the whole experiment 5 times with different initial bootstrap samples.

The downstream task is logistic regression predicting the binary income variable from age, but with different coefficients for the two genders represented in the dataset. This type of model could be used if the analyst suspects the relationship between age and income to differ between men and women. Specifically, the model is

$$\begin{aligned}
\mu &\sim \mathcal{N}_2(0, 10I_2) \\
v &\sim \mathcal{N}_2(0, I_2) \\
\tau &= e^v \\
\theta_F &\sim \mathcal{N}_2(\mu, \mathrm{diag}(\tau^2)) \\
\theta_M &\sim \mathcal{N}_2(\mu, \mathrm{diag}(\tau^2)) \\
\alpha &= X^T(I_{G=F}\theta_F + I_{G=M}\theta_M) \\
y &\sim \mathrm{Bernoulli}\left(\frac{1}{1+e^{-\alpha}}\right)
\end{aligned}$$

where $X = (\mathrm{Age}, 1)$ is the covariate age and a constant, $y$ is the income, $G \in \{F, M\}$ is the gender, $\theta_F$ and $\theta_M$ are the coefficients for both genders, $\mu \in \mathbb{R}^2$ and $\tau \in \mathbb{R}^2_+$ are a common mean and standard deviation for the coefficients, and $v \in \mathbb{R}^2$ is an auxiliary variable that avoids to need to directly sample $\tau$ from a constrained space. In addition, $\mathrm{diag}(\tau^2)$ is a diagonal matrix with $\tau^2$ on the diagonal, $I_{G=g}$ is an indicator of the gender having a specific value $g$, and the operations $e^v$ and $\tau^2$ on vectors $v$ and $\tau$ are elementwise.

Hierarchical models are known to have difficult, non-Gaussian posterior geometry (Betancourt and Girolami, 2013). We see that this is the case for this hierarchical logistic regression model in Figure 18, which shows that the $(\mu_1, v_1)$ and $(\mu_2, v_2)$ marginal distributions have funnel-like geometry.

### 5.4.1 Algorithms and Hyperparameters

We ran a preliminary experiment with the same input queries to NAPSU-MQ that we used in Section 5.3, but we found those to give biased estimates of the coefficients in the hierarchical setting. To correct the bias, we added the 3-way marginal query (age, gender, income) to the input queries. We then removed two columns from the generated synthetic data to reduce runtime. Figure 16 shows the input queries and remaining columns. Otherwise, we used the same hyperparemeters for NAPSU-MQ as in Section 5.3. In particular, the privacy parameters are $\epsilon = 1$, $\delta = n_X^{-2} \approx 4.7 \cdot 10^{-10}$.

We keep the size of the synthetic datasets at $n_{X^*}/n_X = 10$, but increase the number of synthetic datasets $m = 200$ due to the increased complexity of the hierarchical downstream posterior.

We use NUTS (Hoffman and Gelman, 2014) to sample the downstream posterior, with 2 chains, 1000 kept samples, and 500 warmup samples.

5. These are workclass and marital-status.

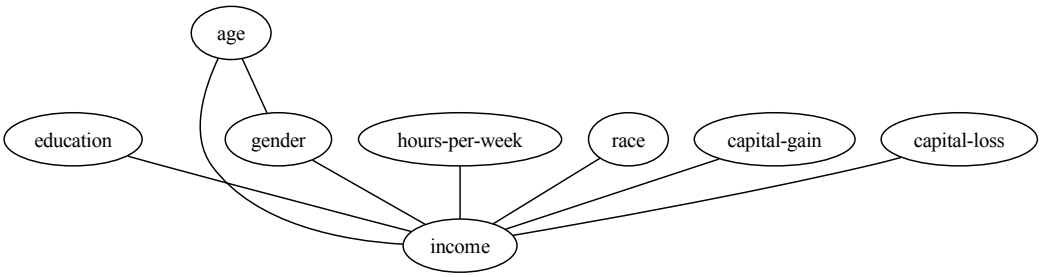


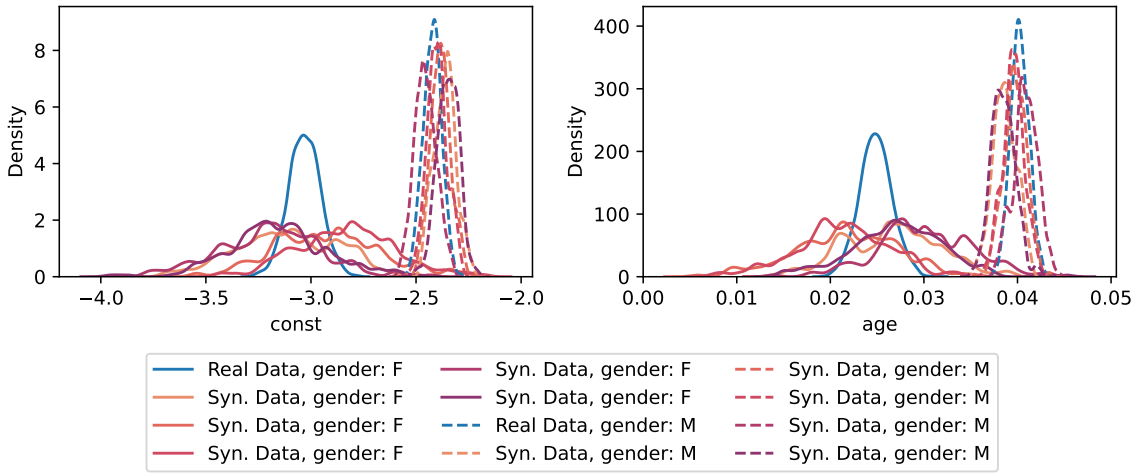

Figure 16: The graph representing the queries given to NAPSU-MQ in the hierarchical logistic regression experiment. The clique (age, gender, income) represents a 3-way marginal query, while the other edges represent 2-way marginal queries.

### 5.4.2 Results

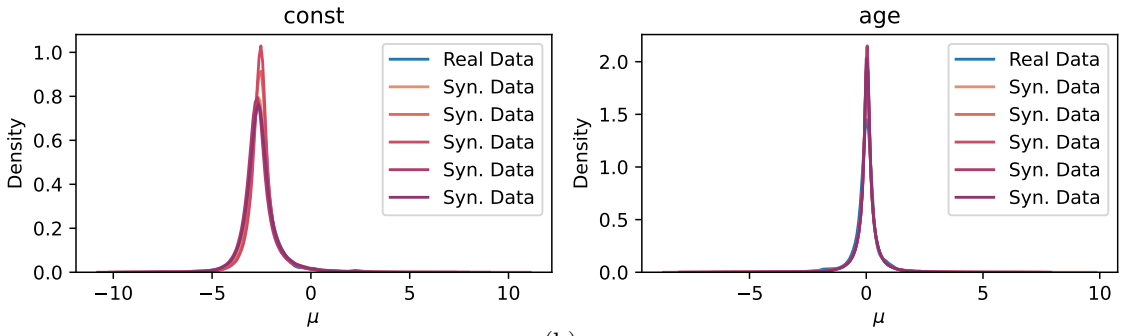
We compare the posterior from mixing the individual posterior from multiple DP synthetic datasets to the non-DP posterior from the real data. In Figure 17, we plot the one-dimensional marginals of all downstream posterior components with the exception of $v$, which is a deterministic transform of $\tau$. Each of the 5 repeats has the different line for synthetic data. In panel (a), we see that the female coefficient posteriors have a much larger variance in the synthetic data posterior compared to the real data posterior, which is to be expected due to DP noise and the fact that only about one third of the datapoints in the Adult dataset are women. The posteriors of the male coefficients are fairly similar between the synthetic and real data. In panels (b) and (c), we see that the posteriors for $\mu$ and $\tau$ are very similar between real and synthetic data.

In Figure 18, we plot selected 2-dimensional posterior marginals from the first repeat of the experiment. We selected 3 pairs of components between $\mu$ and $v$ for their non-Gaussian geometry, and the pairs comparing the same coefficient between men and women, which could be of interest to the analyst. In the 3 leftmost columns, we see that the posterior marginals with non-Gaussian geometry are very similar between real and synthetic data. In the rightmost 2 columns, we see the same result as with the one-dimensional marginals: the female coefficients have a larger variance in the synthetic data posterior due to the DP noise, but the synthetic data posterior still covers the bulk of the real data posterior.

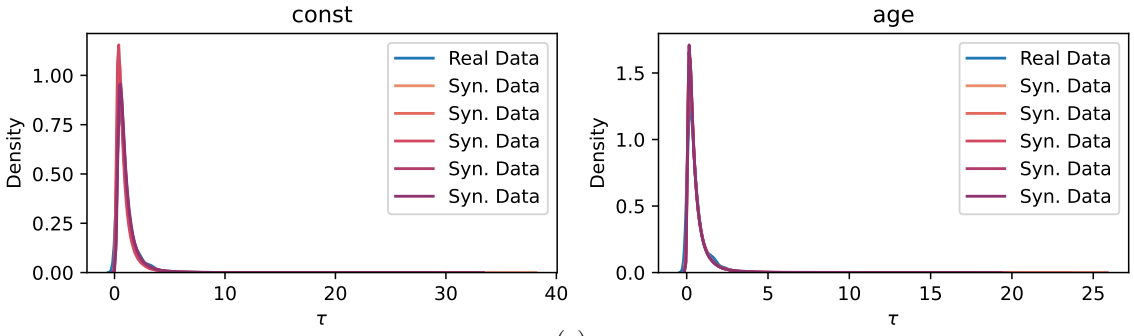
We checked for convergence of the downstream posterior sample wit the $\hat{R}$ statistic (Gelman et al., 2014). We find that in each of the 5 repeats, there are between 6 and 13 synthetic datasets out of $m = 200$ with $\hat{R} > 1.05$ for at least one coefficient, which indicates a convergence issue for these synthetic datasets. Since this is a very small fraction of the synthetic datasets, these convergence issues do not affect the results significantly.




(a)




(b)




(c)

Figure 17: Marginal posterior distributions from hierarchical logistic regression on the Adult dataset, with 5 repeats of generating synthetic data. The coefficients relating to the female gender have wider posteriors due to the DP noise and the lower number of women in the dataset. For the other coefficients, the DP synthetic data posterior is very close to the non-DP real data posterior. The privacy parameters are $\epsilon = 1$, $\delta \approx 4.7 \cdot 10^{-10}$.

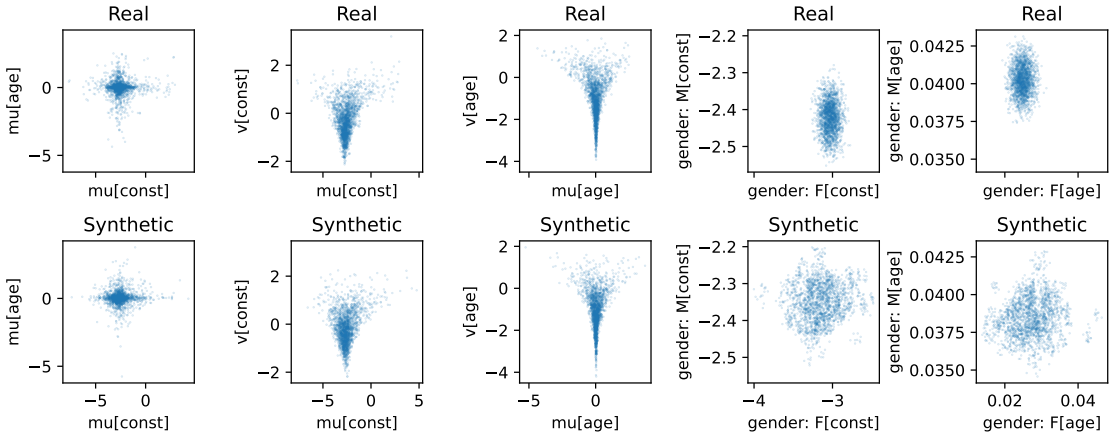



Figure 18: Selected pairwise posterior marginals from hierarchical logistic regression on the Adult dataset from the first run of synthetic data generation. The leftmost 3 pairs with non-Gaussian geometry are very similar between the DP synthetic data and non-DP real data. The gender-related coefficients have higher variance in the synthetic data posterior due to the DP noise. The privacy parameters are $\epsilon = 1$, $\delta \approx 4.7 \cdot 10^{-10}$.

#### 5.4.3 Plotting Details

The plots in Figure 17 use kernel density estimates obtained from posterior samples. The plots in Figure 18 are scatterplots of the posterior samples. The synthetic data scatterplots show a subsample of the posterior samples with an equal number of samples to the real data posterior.

## 6. Discussion

Synthetic data are often considered as a substitute for real data that are sensitive. Since the data generation process is based on having access to $Z$, one might ask why is the synthetic data needed in first place. Why cannot we simply perform the downstream posterior analysis directly using $Z$? Our analysis allows $Z$ to be an arbitrary, even noisy, representation of the data, and it might be difficult for the analyst to place a model for such generative process for $Q$. In most applications, the analyst does have a model for $Q$ arising from the data. Therefore using the synthetic data as a proxy for the $Z$ allows the analyst to use existing models and inference methods to perform the analysis.

### 6.1 Limitations

A clear limitation of mixing posteriors from multiple synthetic data sets is the computational cost of analysing many large synthetic data sets. This may be substantial for more complex Bayesian downstream models, where even a single analysis can be computationally expensive. However, the separate analyses can be run in parallel. We also expect that the information

gained from sampling the posteriors from a few synthetic data sets could be used to speed up sampling the others, for example by using importance sampling, as they likely won't be too far from the sampled ones.

Under DP, we need noise-aware synthetic data generation, which limits the settings in which the method can currently be applied. However, if new noise-aware methods are developed in the future, the method can immediately be used with them.

Condition 7 limits the applicability of our theory to downstream analyses where the prior's influence vanishes as the sample size grows. This does not always happen for some models, such as some infinite-dimensional models, models where the number of parameters increases with data set size, and models with a support that heavily depends on the parameters. The method also requires congeniality, which basically requires the analyst's prior to be compatible with the data provider's. Our Gaussian examples in Section 4 show that it is sometimes possible to recover useful inferences even without congeniality. This is not always the case, so an important direction for future research is separating these two cases, and finding out what can be done in the latter case.

### 6.2 Conclusion

We considered the problem of consistent Bayesian inference using multiple, potentially DP, synthetic data sets, and studied an inference method that mixes the posteriors from multiple large synthetic data sets while re-using existing analysis methods designed for real data. We proved, under congeniality and the general and well-understood regularity conditions of the Bernstein–von Mises theorem, that the method is asymptotically exact as the sizes of the synthetic data sets grow. We studied the method in two examples: non-private Gaussian mean or variance estimation and DP logistic regression. In the former, we were able to use the analytically tractable structure of the setting to derive additional properties of the method, in particular examining what can happen without congeniality. In both settings, we experimentally validated our theory, and showed that the method works in practice. When examining what can go wrong when our assumptions are not met, we found that the method can still give sensible results in some cases, but not all, showing that the method should be applied with care. This greatly expands the understanding of Bayesian inference from synthetic data, filling a major gap in the synthetic data analysis literature.

## Acknowledgments

This work was supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI and Grant 356499), the Strategic Research Council at the Research Council of Finland (Grant 358247) as well as the European Union (Project 101070617). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. The authors wish to thank the Finnish Computing Competence Infrastructure (FCCI) for supporting this project with computational and data storage resources. We thank Tejas Kulkarni for providing the DP-GLM code.

## Appendix A. Additional Background

In this section, we collected background material we use in the rest of the Appendix.

### A.1 Definition of Marginal Query

Let the data $X$ consist of $n$ points $x_1, \dots x_n$, where each $x_i$ is a $d$-tuple $(x_{i,1}, \dots, x_{i,d})$, and each $x_{i,j} \in \mathcal{X}_j$, with $\mathcal{X}_j$ finite. Let $q \subseteq \{1, \dots, d\}$ be a set of indices into $x_i$, and let the possible values the elements of $x_i$ indexed by $q$ can jointly take be $v_1, \dots, v_K$. A marginal query for variables $q$ is a vector $s \in \mathbb{N}^K$ with

$$s_j = \sum_{i=1}^{n} \mathbb{I}_{x_{i,q}=v_j}.$$

If $q$ has $k$-elements, the query is called a $k$-way marginal query. In other words, a marginal query counts, for given variables, how many data points have each of the possible values for those variables. Note that Räisä et al. (2023) use use the term "marginal query" for a single element of $s$, and use the term "full set of marginal queries" for what we call a marginal query.

### A.2 Total Variation Distance Properties

Recall the definition of total variation distance:

**Definition 1.** *The total variation distance between random variables (or distributions) $P_1$ and $P_2$ is*

$$\mathrm{TV}(P_1, P_2) = \sup_A |\Pr(P_1 \in A) - \Pr(P_2 \in A)|,$$

*where $A$ is any measurable set.*

**Lemma 18** (Kelbert, 2023)**.** *Properties of total variation distance:*

1. *For probability densities $p_1$ and $p_2$,*
$$\mathrm{TV}(p_1, p_2) = \frac{1}{2} \int |p_1(x) - p_2(x)| \, \mathrm{d}x.$$
2. *Total variation distance is a metric.*
3. *Pinsker's inequality: for distributions $P_1$ and $P_2$,*
$$\mathrm{TV}(P_1, P_2) \leq \sqrt{\frac{1}{2} \mathrm{KL}(P_1 \,||\, P_2)}.$$
4. *Invariance to bijections: if $f$ is a bijection and $P_1$ and $P_2$ are random variables,*
$$\mathrm{TV}(f(P_1), f(P_2)) = \mathrm{TV}(P_1, P_2).$$

We also occasionally write $\mathrm{TV}(p_1, p_2)$ for probability densities $p_1$ and $p_2$ as

$$\mathrm{TV}(p_1, p_2) = \sup_h \left| \int h(x) p_1(x) \, \mathrm{d}x - \int h(x) p_2(x) \, \mathrm{d}x \right|$$

where $h$ is an indicator function of some measurable set.

### A.3 Bernstein–von Mises Theorem Regularity Conditions

The version of the Bernstein–von Mises theorem we use is from van der Vaart (1998). To state the regularity conditions, we need two definitions:

**Definition 19.** *A parametric probability density $p_Q$ is differentiable in quadratic mean at $Q_0$ if there exists a measurable vector-valued function $\dot{\ell}_{Q_0}$ such that, as $Q \to Q_0$,*

$$\int \left( \sqrt{p_Q(x)} - \sqrt{p_{Q_0}(x)} - \frac{1}{2}(Q - Q_0)^T \dot{\ell}_{Q_0}(x) \sqrt{p_{Q_0}(x)} \right)^2 \mathrm{d}x = o(||Q - Q_0||_2^2).$$

**Definition 20.** *A randomised test is a function $\phi \colon \mathcal{X} \to [0, 1]$.*

The interepretation of $\phi(X)$ is the probability of rejecting some null hypothesis after observing data $X$.

Now we can state the regularity conditions of Theorem 2:

**Condition 21** (van der Vaart, 1998)**.** *For true parameter value $Q_0$ and observed data $X_n$:*

1. *The datapoints of $X_n$ are i.i.d.*
2. *The likelihood $p(x|Q)$ for a single datapoint $x$ is differentiable in quadratic mean at $Q_0$.*
3. *The Fisher information matrix of $p(x|Q)$ is nonsingular at $Q_0$.*
4. *For every $\beta > 0$, there exists a sequence of randomised tests $\phi_n$ such that*

$$p(X_n|Q_0)\phi_n(X_n) \to 0, \qquad \sup_{||Q - Q_0||_2 \geq \beta} p(X_n|Q)(1 - \phi_n(X_n)) \to 0.$$

5. *The prior $p(Q)$ is absolutely continuous (as a measure) in a neighbourhood of $Q_0$ with a continuous positive density at $Q_0$.*

### A.4 Bayesian Inference with Gaussian Models

In this section, we collect well-known results on Bayesian inference of a Gaussian mean. See Gelman et al. (2014) for proofs.

#### A.4.1 Scaled Inverse-Chi-Square Distribution

This parameterisation of the inverse gamma distribution is convenient in this setting.

$$\text{Inv-}\chi^2(\nu, s^2) = \text{Inv-Gamma}\left(\alpha = \frac{\nu}{2}, \beta = \frac{\nu}{2}s^2\right).$$

If $\theta \sim \text{Inv-}\chi^2(\nu, s^2)$, $\theta > 0$,

$$\begin{aligned} p(\theta) &= \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} s^\nu \theta^{-(\frac{\nu}{2}+1)} e^{-\frac{\nu s^2}{2\theta}} \\ \mathbb{E}(\theta) &= \frac{\nu}{\nu - 2} s^2, \quad \nu > 2 \\ \text{Var}(\theta) &= \frac{2\nu^2}{(\nu - 2)^2(\nu - 4)} s^4, \quad \nu > 4. \end{aligned}$$

### A.4.2 Gaussian Model with Known Variance

When the variance of the data is known to be $\sigma_k^2$, and only the mean is unknown, the conjugate prior is another Gaussian, and we get the following inference problem:

$$\begin{aligned}\mu &\sim \mathcal{N}(\mu_0, \sigma_0^2)\\ x_i|\mu &\sim \mathcal{N}(\mu, \sigma_k^2).\end{aligned}$$

The posterior with $n$ datapoints with sample mean $\bar{X}$ is:

$$\begin{aligned}\mu|X &\sim \mathcal{N}(\mu_n, \sigma_n^2)\\ \mu_n &= \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma_k^2}\bar{X}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma_k^2}}\\ \frac{1}{\sigma_n^2} &= \frac{1}{\sigma_0^2} + \frac{n}{\sigma_k^2}.\end{aligned}$$

### A.4.3 Gaussian Model with Unknown Variance

When the variance of the data is also unknown, the conjugate prior is a inverse-chi-squared for the variance, and Gaussian for the mean, which gives the following inference problem:

$$\begin{aligned}\sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)\\ \mu|\sigma^2 &\sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)\\ x_i|\mu, \sigma^2 &\sim \mathcal{N}(\mu, \sigma^2).\end{aligned}$$

The joint posterior of $\mu$ and $\sigma^2$ for $n$ datapoints is:

$$\begin{aligned}\sigma^2|X &\sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)\\ \mu|\sigma^2, X &\sim \mathcal{N}\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)\end{aligned}$$

with

$$\begin{aligned}\bar{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i\\ s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2\\ \mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{X}\\ \kappa_n &= \kappa_0 + n\\ \nu_n &= \nu_0 + n\\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{X} - \mu_0)^2.\end{aligned}$$

The marginal posterior of $\mu$ is

$$\mu|X \sim t_{\nu_n}\left(\mu_n, \frac{\sigma_n^2}{\kappa_n}\right).$$

## Appendix B. Missing Proofs

This sections contains the missing proofs from the main text. The proof of our main Theorem 9 is in Appendix B.1, and proofs of the convergence rate-related Theorems B.2 and 26 are in Appendix 13.

### B.1 Consistency Proof

For ease of reference, we repeat Theorem 2 and Condition 7:

**Theorem 2** (Bernstein–von Mises, van der Vaart, 1998)**.** *Let $n$ denote the size of the data set $X_n$. Under regularity conditions stated in Condition 21 in Appendix A.3, for true parameter value $Q_0$, the posterior $\bar{Q}(X_n) \sim p(Q|X_n)$ satisfies*

$$\mathrm{TV}\left(\sqrt{n}(\bar{Q}(X_n) - Q_0), \mathcal{N}(\mu(X_n), \Sigma)\right) \xrightarrow{P} 0$$

*as $n \to \infty$ for some $\mu(X_n)$ and $\Sigma$, that do not depend on the prior, where the convergence in probability is over sampling $X_n \sim p(X_n|Q_0)$.*

Recall that $\bar{Q}_n^+ \sim p(Q|Z, X_n^*)$, and $\bar{Q}_n \sim p(Q|X_n^*)$.

**Condition 7.** *For the observed $Z$ and all $Q$, there exist random variables $D_n$ such that*

$$\mathrm{TV}\left(\bar{Q}_n^+, D_n\right) \xrightarrow{P} 0 \quad \text{and} \quad \mathrm{TV}\left(\bar{Q}_n, D_n\right) \xrightarrow{P} 0$$

*as $n \to \infty$, where the convergence in probability is over sampling $X_n^* \sim p(X_n^*|Z, Q)$.*

**Lemma 8.** *Let the assumptions of Theorem 2 (stated in Condition 21) hold for the downstream analysis for all $Q_0$, and the following assumptions:*

*(1) $Z$ and $X^*$ are conditionally independent given $Q$; and*

*(2) $p(Z|Q) > 0$ for all $Q$,*

*hold. Then Condition 7 holds.*

**Proof** Under Assumption (1)

$$p(Q|Z, X_n^*) \propto p(X_n^*|Q)p(Z|Q)p(Q),$$

so we can view both $p(Q|Z, X_n^*)$ and $p(Q|X_n^*)$ as the posteriors for the same Bayesian inference problem with observed data $X_n^*$, and priors $p(Q|Z) \propto p(Z|Q)p(Q)$ and $p(Q)$, respectively. Due to Condition 21 (5) and Assumption (2), $p(Q|Z)$ has an everywhere positive density. Recall that $\bar{Q}_n^+ \sim p(Q|Z, X_n^*)$ and $\bar{Q}_n \sim p(Q|X_n^*)$. Now, Theorem 2 gives

$$\mathrm{TV}\left(\sqrt{n}(\bar{Q}_n^+ - Q_0), \mathcal{N}(\mu_n, \Sigma)\right) \xrightarrow{P} 0$$

and

$$\mathrm{TV}\left(\sqrt{n}(\bar{Q}_n - Q_0), \mathcal{N}(\mu_n, \Sigma)\right) \xrightarrow{P} 0$$

as $n \to \infty$, where $\mu_n, \Sigma$ are equal in the two cases because they do not depend on the prior. The probability is over $X_n^* \sim p(X_n^*|Q_0)$. Because of Assumption (1), $p(X_n^*|Q_0) = p(X_n^*|Z, Q_0)$, so the convergence also holds with probability over $X_n^* \sim p(X_n^*|Z, Q_0)$. These hold for any $Q_0$. Because the function $f_n(q) = \sqrt{n}(q - Q_0)$ is a bijection and total variation distance is invariant to bijections, Condition 7 holds with $D_n$ being the pushforward distribution $D_n = f_n^{-1} \circ \mathcal{N}(\mu_n, \Sigma)$, with the $Q$ of Condition 7 being $Q_0$. Note that $D_n$ is allowed to depend on $Q$ in Condition 7 due to the order of quantifiers. ■

**Lemma 22.** *Under Condition 7,*

$$\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) \xrightarrow{P} 0$$

*as $n \to \infty$, with the probability over $X_n^* \sim p(X_n^*|Z)$.*

**Proof** Total variation distance is a metric, so

$$\mathrm{TV}\left(\bar{Q}_n^+, \bar{Q}_n\right) \leq \mathrm{TV}\left(\bar{Q}_n^+, D_n\right) + \mathrm{TV}\left(\bar{Q}_n, D_n\right).$$

Now, by Condition 7

$$\mathrm{TV}\left(\bar{Q}_n^+, \bar{Q}_n\right) \xrightarrow{P} 0 \tag{23}$$

as $n \to \infty$, with the probability over $X_n^* \sim p(X_n^*|Z, Q)$.

It remains to show (23) with the probability over $X_n^* \sim p(X_n^*|Z)$ instead of $X_n^* \sim p(X_n^*|Z, Q)$. With $X_n^* \sim p(X_n^*|Z)$, for any $\epsilon > 0$,

$$\Pr_{X_n^*|Z}(\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) > \epsilon) = \int \Pr_{X_n^*|Z,Q}(\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) > \epsilon) p(Q|Z)\, \mathrm{d}Q.$$

(23) holds for any $Q$, so

$$\lim_{n\to\infty} \Pr_{X_n^*|Z,Q}(\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) > \epsilon) = 0.$$

The dominated convergence theorem then implies that

$$\lim_{n\to\infty} \Pr_{X_n^*|Z}(\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) > \epsilon) = 0,$$

so

$$\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) \xrightarrow{P} 0$$

as $n \to \infty$, with the probability over $X_n^* \sim p(X_n^*|Z)$. ■

**Lemma 23.** *Let $y_n \sim U_n$ be an arbitrary sequence of random variables with densities $f_{U_n}$ with regards to some measure $\mu$. Let $S(y_n)$, $T(y_n)$ be random variables that depend on $y_n$, with respective density functions $f_{S(y_n)}$ and $f_{T(y_n)}$, with regards to some measure $\nu$. If*

$$\mathrm{TV}(S(y_n), T(y_n)) \xrightarrow{P} 0$$

*as $n \to \infty$, where the probability is over $y_n \sim U_n$, then*

$$\mathrm{TV}\left(\int f_{S(y_n)}(x) f_{U_n}(y_n)\mu(\mathrm{d}y_n), \int f_{T(y_n)}(x) f_{U_n}(y_n)\mu(\mathrm{d}y_n)\right) \to 0$$

*as $n \to \infty$.*

**Proof** Let $h$ be an indicator function of $x$ over any measurable set and let $\epsilon > 0$. Then

$$\begin{aligned}
&\left|\int h(x) \int f_{S(y_n)}(x) f_{U_n}(y_n)\mu(\mathrm{d}y_n)\nu(\mathrm{d}x) - \int h(x) \int f_{T(y_n)}(x) f_{U_n}(y_n)\mu(\mathrm{d}y_n)\nu(\,\mathrm{d}x)\right| \\
&= \left|\int h(x) \int f_{U_n}(y_n)\big(f_{S(y_n)}(x) - f_{T(y_n)}(x)\big)\mu(\mathrm{d}y_n)\nu(\mathrm{d}x)\right| \\
&= \left|\int f_{U_n}(y_n) \int h(x)\big(f_{S(y_n)}(x) - f_{T(y_n)}(x)\big)\nu(\mathrm{d}x)\mu(\mathrm{d}y_n)\right| \\
&\leq \int f_{U_n}(y_n) \left|\int h(x)\big(f_{S(y_n)}(x) - f_{T(y_n)}(x)\big)\nu(\mathrm{d}x)\right| \mu(\mathrm{d}y_n) \\
&= \int f_{U_n}(y_n) \left|\int h(x) f_{S(y_n)}(x)\nu(\mathrm{d}x) - \int h(x) f_{T(y_n)}(x)\nu(\mathrm{d}x)\right| \mu(\mathrm{d}y_n).
\end{aligned}$$

Because $\mathrm{TV}(S(y_n), T(y_n)) \xrightarrow{P} 0$, for large enough $n$, there is a set $Y_n$ with

$$\mathrm{TV}(S(y_n), T(y_n)) < \frac{\epsilon}{2}$$

for all $y_n \in Y_n$, and $\Pr(y_n \in Y_n^C) < \frac{\epsilon}{2}$. As

$$\mathrm{TV}(S(y_n), T(y_n)) = \sup_h \left|\int h(x) f_{S(y_n)}(x)\nu(\mathrm{d}x) - \int h(x) f_{T(y_n)}(x)\nu(\mathrm{d}x)\right| \leq 1,$$

now

$$\begin{aligned}
&\int f_{U_n}(y_n) \left|\int h(x) f_{S(y_n)}(x)\nu(\mathrm{d}x) - \int h(x) f_{T(y_n)}(x)\nu(\mathrm{d}x)\right| \mu(\mathrm{d}y_n) \\
&= \int_{Y_n} f_{U_n}(y_n) \left|\int h(x) f_{S(y_n)}(x)\,\mathrm{d}x - \int h(x) f_{T(y_n)}(x)\nu(\mathrm{d}x)\right| \mu(\mathrm{d}y_n) \\
&+ \int_{Y_n^C} f_{U_n}(y_n) \left|\int h(x) f_{S(y_n)}(x)\,\mathrm{d}x - \int h(x) f_{T(y_n)}(x)\nu(\mathrm{d}x)\right| \mu(\mathrm{d}y_n) \\
&\leq \int_{Y_n} f_{U_n}(y_n)\frac{\epsilon}{2}\mu(\mathrm{d}y_n) + \int_{Y_n^C} f_{U_n}(y_n)\mu(\mathrm{d}y_n) \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon
\end{aligned}$$

for large enough $n$. Now

$$\begin{aligned}
&\mathrm{TV}\left(\int f_{S(y_n)}(x) f_{U_n}(y_n)\mu(\mathrm{d}y_n), \int f_{T(y_n)}(x) f_{U_n}(y_n)\mu(\mathrm{d}y_n)\right) \\
&= \sup_h \left|\int h(x) \int f_{S(y_n)}(x) f_{U_n}(y_n)\mu(\mathrm{d}y_n)\nu(\mathrm{d}x) - \int h(x) \int f_{T(y_n)}(x) f_{U_n}(y_n)\mu(\mathrm{d}y_n)\nu(\mathrm{d}x)\right| \\
&< \epsilon
\end{aligned}$$

for any $\epsilon > 0$ with large enough $n$. ■

**Theorem 9.** *Under congeniality and Condition 7,* $\mathrm{TV}\left(p(Q|Z), \bar{p}_n(Q)\right) \to 0$ *as* $n \to \infty$.

**Proof** The claim follows from Lemma 23 with $y_n = X_n^*$, $U_n = p(X_n^*|Z)$, $S(y_n) \sim p(Q|X_n^*)$ and $T(y_n) \sim p(Q|Z, X_n^*)$. These meet the condition for Lemma 23 due to Lemma 22. ■

### B.2 Convergence Rate

**Definition 10.** *A sequence of random variables* $X_n$ *is uniformly integrable if*

$$\lim_{M\to\infty} \sup_n \mathbb{E}(|X_n| \mathbb{I}_{|X_n|>M}) = 0.$$

**Lemma 24.** *If* $|X_n| \leq Y_n$ *almost surely and* $Y_n$ *is uniformly integrable,* $X_n$ *is uniformly integrable.*

**Proof**

$$0 \leq \lim_{M\to\infty} \sup_n \mathbb{E}(|X_n| \mathbb{I}_{|X_n|>M}) \leq \lim_{M\to\infty} \sup_n \mathbb{E}(Y_n \mathbb{I}_{Y_n>M}) = 0.$$

■

**Lemma 25** (Billingsley, 1995, Section 16)**.** *If* $X_n$ *and* $Y_n$ *are uniformly integrable,* $X_n + Y_n$ *is uniformly integrable.*

**Condition 11.** *For the observed* $Z$*, there exist random variables* $D_n$ *such that for a sequence* $R_1, R_2, \cdots > 0$*,* $R_n \to 0$ *as* $n \to \infty$*,*

$$\frac{1}{R_n} \mathrm{TV}\left(\bar{Q}_n^+, D_n\right) \quad \text{and} \quad \frac{1}{R_n} \mathrm{TV}\left(\bar{Q}_n, D_n\right)$$

*are uniformly integrable when* $X_n^* \sim p(X_n^*|Z)$.

**Theorem 13.** *Under congeniality and Condition 11,* $\mathrm{TV}\left(p(Q|Z), \bar{p}_n(Q)\right) = O(R_n)$.

**Proof** Total variation distance is a metric, so

$$\frac{1}{R_n} \mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) \leq \frac{1}{R_n} \mathrm{TV}\left(\bar{Q}_n^+, D_n\right) + \frac{1}{R_n} \mathrm{TV}\left(\bar{Q}_n, D_n\right).$$

Now Condition 11 and Lemmas 24 and 25 imply that

$$\frac{1}{R_n} \mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n)$$

is uniformly integrable with $X_n^* \sim p(X_n^*|Z)$.

Recall that

$$\frac{1}{R_n}\operatorname{TV}(\bar{Q}_n^+, \bar{Q}_n) = \frac{1}{R_n}\sup_h \left| \int h(Q)p(Q|Z, X_n^*)\,\mathrm{d}Q - \int h(Q)p(Q|X_n^*)\,\mathrm{d}Q \right|$$

and

$$\begin{aligned}
&\frac{1}{R_n}\operatorname{TV}(p(Q|Z), \bar{p}_n(Q)) \\
&= \frac{1}{R_n}\sup_h \left| \int h(Q) \int p(Q|Z, X_n^*)p(X_n^*|Z)\,\mathrm{d}X_n^*\,\mathrm{d}Q - \int h(Q) \int p(Q|X_n^*)p(X_n^*|Z)\,\mathrm{d}X_n^*\,\mathrm{d}Q \right|,
\end{aligned}$$

where $h$ is an indicator function of some measurable set.

For any indicator function $h$, using the start of the proof of Lemma 23 gives

$$\begin{aligned}
&\frac{1}{R_n}\left| \int h(Q) \int p(Q|Z, X_n^*)p(X_n^*|Z)\,\mathrm{d}X_n^*\,\mathrm{d}Q - \int h(Q) \int p(Q|X_n^*)p(X_n^*|Z)\,\mathrm{d}X_n^*\,\mathrm{d}Q \right| \\
&\leq \int p(X_n^*|Z)\frac{1}{R_n}\left| \int h(Q)p(Q|Z, X_n^*)\,\mathrm{d}Q - \int h(Q)p(Q|X_n^*)\,\mathrm{d}Q \right| \mathrm{d}X_n^* \\
&\leq \int p(X_n^*|Z)\frac{1}{R_n}\operatorname{TV}(\bar{Q}_n^+, \bar{Q}_n)\,\mathrm{d}X_n^*.
\end{aligned}$$

Because $R_n^{-1}\operatorname{TV}(\bar{Q}_n^+, \bar{Q}_n)$ is uniformly integrable when $X_n^* \sim p(X_n^*|Z)$, there exists an $M$ such that for all $n$,

$$\int_{Y_n} p(X_n^*|Z)\frac{1}{R_n}\left| \int h(Q)p(Q|Z, X_n^*)\,\mathrm{d}Q - \int h(Q)p(Q|X_n^*)\,\mathrm{d}Q \right| \mathrm{d}X_n^* \leq 1,$$

where $Y_n = \{X_n^* \mid \frac{1}{R_n}\operatorname{TV}(\bar{Q}_n^+, \bar{Q}_n) > M\}$. Now, for all $n$

$$\begin{aligned}
&\frac{1}{R_n}\left| \int h(Q) \int p(Q|Z, X_n^*)p(X_n^*|Z)\,\mathrm{d}X_n^*\,\mathrm{d}Q - \int h(Q) \int p(Q|X_n^*)p(X_n^*|Z)\,\mathrm{d}X_n^*\,\mathrm{d}Q \right| \\
&\leq \int p(X_n^*|Z)\frac{1}{R_n}\left| \int h(Q)p(Q|Z, X_n^*)\,\mathrm{d}Q - \int h(Q)p(Q|X_n^*)\,\mathrm{d}Q \right| \mathrm{d}X_n^* \\
&= \int_{Y_n} p(X_n^*|Z)\frac{1}{R_n}\left| \int h(Q)p(Q|Z, X_n^*)\,\mathrm{d}Q - \int h(Q)p(Q|X_n^*)\,\mathrm{d}Q \right| \mathrm{d}X_n^* \\
&+ \int_{Y_n^C} p(X_n^*|Z)\frac{1}{R_n}\left| \int h(Q)p(Q|Z, X_n^*)\,\mathrm{d}Q - \int h(Q)p(Q|X_n^*)\,\mathrm{d}Q \right| \mathrm{d}X_n^* \\
&\leq 1 + \int_{Y_n^C} p(X_n^*|Z)M\,\mathrm{d}X_n^* \\
&\leq 1 + M,
\end{aligned}$$

so $\operatorname{TV}(p(Q|Z), \bar{p}_n(Q)) = O(R_n)$. ■

### B.2.1 Convergence Rate in Gaussian Mean Estimation

We start by proving the one-dimensional case.

**Theorem 26.** *When the up- and downstream models are Gaussian mean estimations with known variance, and* $D_n = \mathcal{N}(\bar{X}, n^{-1}\sigma_k^2)$*,*

$$\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n^+, D_n\right) \quad \text{and} \quad \sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n, D_n\right)$$

*are uniformly integrable when* $X_n^* \sim p(X_n^*|X)$*.*

**Proof** When the downstream model is Gaussian mean estimation with known variance,

$$\bar{Q}_n = \mathcal{N}(\mu_n, \sigma_n^2)$$

$$\mu_n = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma_k^2}\bar{X}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma_k^2}}$$

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma_k^2}.$$

We start with the proof for $\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n, D_n\right)$. By Pinsker's equality and the formula for KL-divergence between Gaussians,

$$\begin{aligned}
\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n, D_n\right) &\leq \sqrt{\frac{1}{2}n\mathrm{KL}(\bar{Q}_n \,||\, D_n)} \\
&= \sqrt{\frac{1}{4}n\left(\frac{\sigma_k^2}{n\sigma_n^2} + \frac{(\mu_n - \bar{X})^2}{\sigma_n^2} - 1 + \ln\frac{n\sigma_n^2}{\sigma_k^2}\right)} \\
&\leq \sqrt{\left|\frac{1}{4}n\left(\frac{\sigma_k^2}{n\sigma_n^2} - 1\right)\right| + \frac{1}{4}n\frac{(\mu_n - \bar{X})^2}{\sigma_n^2} + \left|\frac{1}{4}n\ln\frac{n\sigma_n^2}{\sigma_k^2}\right|} \\
&\leq \sqrt{\left|\frac{1}{4}n\left(\frac{\sigma_k^2}{n\sigma_n^2} - 1\right)\right|} + \sqrt{\frac{1}{4}n\frac{(\mu_n - \bar{X})^2}{\sigma_n^2}} + \sqrt{\left|\frac{1}{4}n\ln\frac{n\sigma_n^2}{\sigma_k^2}\right|}.
\end{aligned}$$

The last inequality can be deduced from the fact that the $L_2$-norm is upper bounded by the $L_1$ norm. Denote

$$C_1(n) = n\left(\frac{\sigma_k^2}{n\sigma_n^2} - 1\right) = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma_k^2}\right)\sigma_k^2 - n = \frac{\sigma_k^2}{\sigma_0^2} + n - n = \frac{\sigma_k^2}{\sigma_0^2}$$

and

$$\begin{aligned}
C_2(n) &= n\ln\frac{n\sigma_n^2}{\sigma_k^2} = -n\ln\frac{\sigma_k^2}{n\sigma_n^2} = -n\ln\left(\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma_k^2}\right)\frac{\sigma_k^2}{n}\right) \\
&= -n\ln\left(\frac{\sigma_k^2}{n\sigma_0^2} + 1\right) = -\frac{u\sigma_k^2}{\sigma_0^2}\ln\left(\frac{1}{u} + 1\right) = -\frac{\sigma_k^2}{\sigma_0^2}\ln\left(\frac{1}{u} + 1\right)^u,
\end{aligned}$$

with $n = \frac{u\sigma_k^2}{\sigma_0^2}$. Because

$$\lim_{u\to\infty}\left(1+\frac{1}{u}\right)^u = e,$$

we have

$$\lim_{u\to\infty} -\frac{\sigma_k^2}{\sigma_0^2}\ln\left(\frac{1}{u}+1\right)^u = -\frac{\sigma_k^2}{\sigma_0^2},$$

which implies that $C_2(n)$ is bounded.

Futhermore,

$$\sqrt{\frac{1}{4}n\frac{(\mu_n-\bar{X})^2}{\sigma_n^2}} = \frac{1}{2}\sqrt{n\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}\right)}|\mu_n-\bar{X}|.$$

Denote

$$s_n = \frac{1}{2}\sqrt{n\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}\right)}.$$

Note that $s_n = O(n)$. Then

$$\sqrt{\frac{1}{4}n\frac{n(\mu_n-\bar{X})^2}{n\sigma_n^2}} = s_n|\mu_n-\bar{X}|,$$

and

$$\begin{aligned}
s_n|\mu_n-\bar{X}| &= s_n\left|\frac{\frac{1}{\sigma_0^2}\mu_0}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}}+\frac{\frac{n}{\sigma_k^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}}\bar{X}-\bar{X}\right| \\
&\le s_n\frac{\frac{1}{\sigma_0^2}|\mu_0|}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}}+s_n\left|\frac{\frac{n}{\sigma_k^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}}-1\right||\bar{X}| \\
&= s_n\frac{\frac{1}{\sigma_0^2}|\mu_0|}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}}+s_n\left|\frac{\frac{n}{\sigma_k^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}}-\frac{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}}\right||\bar{X}| \\
&= s_n\frac{\frac{1}{\sigma_0^2}|\mu_0|}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}}+s_n\frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}}|\bar{X}|.
\end{aligned}$$

Denote

$$C_3(n) = s_n\frac{\frac{1}{\sigma_0^2}|\mu_0|}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}},$$

and

$$C_4(n) = s_n\frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma_k^2}}.$$

Because $s_n = O(n)$, we have $C_3(n) = O(1)$ and $C_4(n) = O(1)$, so $C_3(n)$ and $C_4(n)$ are bounded.

We now have

$$\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n, D_n\right) \leq \sqrt{\frac{1}{4}|C_1(n)|} + \sqrt{\frac{1}{4}|C_2(n)|} + C_3(n) + C_4(n)|\bar{X}|.$$

By Lemmas 24 and 25, it suffices to show that each of the terms on the right is uniformly integrable. The terms containing $C_1, C_2$ and $C_3$ are non-random and bounded in $n$, so they are uniformly integrable. It remains to show that $C_4(n)|\bar{X}|$ is uniformly integrable. $C_4(n)$ is bounded, so we only need to show that $|\bar{X}|$ is uniformly integrable.

To bound the expectation in the definition of uniform integrability for $|\bar{X}|$, we need some background facts. For geometric series, with $a \in \mathbb{R}$ and $|r| < 1$,

$$\sum_{i=0}^{\infty} ar^i = \frac{a}{1-r},$$

and differentiating both sides with regards to $r$ gives

$$\sum_{i=0}^{\infty} a(i+1)r^i = \frac{a}{(1-r)^2}.$$

For a Gaussian random variable $Y$ with mean $\mu$ and variance $\sigma$, $\Pr(Y > \mu + t) \leq e^{-\frac{t^2}{2\sigma^2}}$. $\bar{X} \sim \mathcal{N}\left(\mu, \frac{1}{n}\sigma_k^2\right)$, so this tail bound gives

$$\Pr(\bar{X} > t + \mu) \leq 2e^{-\frac{nt^2}{2\sigma_k^2}}, \quad \Pr(\bar{X} < \mu - t) \leq 2e^{-\frac{nt^2}{2\sigma_k^2}}.$$

Now

$$\begin{aligned}
&\lim_{M\to\infty} \sup_n \mathbb{E}\big(|\bar{X}|\mathbb{I}_{|\bar{X}|>M}\big) \\
&= \lim_{M\to\infty} \sup_n \mathbb{E}_\mu\big(\mathbb{E}(|\bar{X}|\mathbb{I}_{|\bar{X}|>M}|\mu)\big) \\
&= \lim_{M\to\infty} \sup_n \mathbb{E}_\mu\left(\sum_{i=0}^{\infty} \mathbb{E}\big(|\bar{X}|\mathbb{I}_{M+i<|\bar{X}|\leq M+i+1}|\mu\big)\right) \\
&\leq \lim_{M\to\infty} \sup_n \mathbb{E}_\mu\left(\sum_{i=0}^{\infty} \mathbb{E}\big((M+i+1)\mathbb{I}_{|\bar{X}|>M+i}|\mu\big)\right) \\
&= \lim_{M\to\infty} \sup_n \mathbb{E}_\mu\left(\sum_{i=0}^{\infty} (M+i+1)\Pr\left(|\bar{X}| > M+i|\mu\right)\right) \\
&= \lim_{M\to\infty} \sup_n \mathbb{E}_\mu\left(\sum_{i=0}^{\infty} (M+i+1)\left(\Pr\left(\bar{X} > M+i|\mu\right) + \Pr\left(\bar{X} < -M-i|\mu\right)\right)\right) \\
&\leq \lim_{M\to\infty} \sup_n \mathbb{E}_\mu\left(\sum_{i=0}^{\infty} (M+i+1)\left(e^{-\frac{n(M+i-\mu)^2}{2\sigma_k^2}} + e^{-\frac{n(\mu+M+i)^2}{2\sigma_k^2}}\right)\right) \\
&\leq \lim_{M\to\infty} \mathbb{E}_\mu\left(\sum_{i=0}^{\infty} (M+i+1)\left(e^{-\frac{(M+i-\mu)^2}{2\sigma_k^2}} + e^{-\frac{(\mu+M+i)^2}{2\sigma_k^2}}\right)\right),
\end{aligned}$$

so

$$\begin{aligned}\lim_{M\to\infty}\sup_n \mathbb{E}\big(|\bar{X}|\mathbb{I}_{|\bar{X}|>M}\big) \leq \lim_{M\to\infty}\mathbb{E}_\mu\left(\sum_{i=0}^{\infty}(M+i+1)e^{-\frac{(M+i-\mu)^2}{2\sigma_k^2}}\right) \\ + \lim_{M\to\infty}\mathbb{E}_\mu\left(\sum_{i=0}^{\infty}(M+i+1)e^{-\frac{(\mu+M+i)^2}{2\sigma_k^2}}\right).\end{aligned} \tag{24}$$

Looking at the first term on the RHS of (24), when $|M+i-\mu|\geq 1$, $(M+i-\mu)^2 \geq M+i-\mu$. It is possible that $|M+i-\mu|<1$ for exactly two values of $i$ that depend on $\mu$. Let $i_{\mu 1}$ and $i_{\mu 2}$ be those values. We know that $i_{\mu j} < 1+\mu-M$ for $j\in\{1,2\}$. Now

$$\begin{aligned}\mathbb{E}_\mu\left(\sum_{i=0}^{\infty}(M+i+1)e^{-\frac{(M+i-\mu)^2}{2\sigma_k^2}}\right) &= \mathbb{E}_\mu\left(\sum_{i=0,i\neq i_{\mu1},i\neq i_{\mu2}}^{\infty}(M+i+1)e^{-\frac{(M+i-\mu)^2}{2\sigma_k^2}}\right) \\ &+ \mathbb{E}_\mu\left((M+i_{\mu1}+1)e^{-\frac{(M+i_{\mu1}-\mu)^2}{2\sigma_k^2}}\right) \\ &+ \mathbb{E}_\mu\left((M+i_{\mu2}+1)e^{-\frac{(M+i_{\mu2}-\mu)^2}{2\sigma_k^2}}\right).\end{aligned} \tag{25}$$

We can upper bound the series using $(M+i-\mu)^2 \geq M+i-\mu$ and the formulas for geometric series.

$$\begin{aligned}&\mathbb{E}_\mu\left(\sum_{i=0,i\neq i_{\mu1},i\neq i_{\mu2}}^{\infty}(M+i+1)e^{-\frac{(M+i-\mu)^2}{2\sigma_k^2}}\right) \\ &\leq \mathbb{E}_\mu\left(\sum_{i=0,i\neq i_{\mu1},i\neq i_{\mu2}}^{\infty}(M+i+1)e^{-\frac{(M+i-\mu)}{2\sigma_k^2}}\right) \\ &\leq \mathbb{E}_\mu\left(\sum_{i=0}^{\infty}(M+i+1)e^{-\frac{(M+i-\mu)}{2\sigma_k^2}}\right) \\ &= \mathbb{E}_\mu\left(\sum_{i=0}^{\infty}(M+i+1)e^{-\frac{(M-\mu)}{2\sigma_k^2}}\left(e^{-\frac{1}{2\sigma_k^2}}\right)^i\right) \\ &= \mathbb{E}_\mu\left(\sum_{i=0}^{\infty}Me^{-\frac{(M-\mu)}{2\sigma_k^2}}\left(e^{-\frac{1}{2\sigma_k^2}}\right)^i\right) + \mathbb{E}_\mu\left(\sum_{i=0}^{\infty}(i+1)e^{-\frac{(M-\mu)}{2\sigma_k^2}}\left(e^{-\frac{1}{2\sigma_k^2}}\right)^i\right) \\ &= \mathbb{E}_\mu\left(\frac{Me^{-\frac{(M-\mu)}{2\sigma_k^2}}}{1-e^{-\frac{1}{2\sigma_k^2}}}\right) + \mathbb{E}_\mu\left(\frac{Me^{-\frac{(M-\mu)}{2\sigma_k^2}}}{\left(1-e^{-\frac{1}{2\sigma_k^2}}\right)^2}\right) \\ &= \left(\frac{M}{1-e^{-\frac{1}{2\sigma_k^2}}} + \frac{M}{\left(1-e^{-\frac{1}{2\sigma_k^2}}\right)^2}\right)\mathbb{E}_\mu\left(e^{-\frac{(M-\mu)}{2\sigma_k^2}}\right).\end{aligned}$$

For the expectation, we have

$$\mathbb{E}_\mu\left(e^{-\frac{M-\mu}{2\sigma_k^2}}\right) = e^{-\frac{M}{2\sigma_k^2}}\mathbb{E}_\mu\left(e^{\frac{1}{2\sigma_k^2}\mu}\right).$$

$\mathbb{E}_\mu\left(e^{\frac{1}{2\sigma_k^2}\mu}\right)$ is finite, as it is an evaluation of the moment generating function of $\mu$, which means that

$$\lim_{M\to\infty}\left(\frac{M}{1-e^{-\frac{1}{2\sigma_k^2}}} + \frac{M}{\left(1-e^{-\frac{1}{2\sigma_k^2}}\right)^2}\right)\mathbb{E}_\mu\left(e^{-\frac{(M-\mu)}{2\sigma_k^2}}\right) = 0.$$

For the two other terms on the RHS of (25)

$$\begin{aligned}\mathbb{E}_\mu\left((M+i_{\mu j}+1)e^{-\frac{(M+i_{\mu j}-\mu)^2}{2\sigma_k^2}}\right) &\leq \mathbb{E}_\mu\left((M+1+\mu-M+1)e^{-\frac{(M+i_{\mu j}-\mu)^2}{2\sigma_k^2}}\right)\\ &= \mathbb{E}_\mu\left((\mu+2)e^{-\frac{(M+i_{\mu j}-\mu)^2}{2\sigma_k^2}}\right)\\ &\to 0\end{aligned}$$

as $M\to\infty$ by the dominated convergence theorem, as

$$(\mu+2)e^{-\frac{(M+i_{\mu j}-\mu)^2}{2\sigma_k^2}} \leq (\mu+2)e^{-\frac{0}{2\sigma_k^2}},$$

and the right-hand-side has a finite expectation.

We have now shown that the first limit on the RHS of (24) is 0. For the other limit, setting $\mu' = -\mu$, and using the reasoning above with $\mu$ replaced by $\mu'$, we have

$$\begin{aligned}&\lim_{M\to\infty}\mathbb{E}_\mu\left(\sum_{i=0}^{\infty}(M+i+1)\left(e^{-\frac{(\mu+M+i)^2}{2\sigma_k^2}}\right)\right)\\ &= \lim_{M\to\infty}\mathbb{E}_{\mu'}\left(\sum_{i=0}^{\infty}(M+i+1)\left(e^{-\frac{(M+i-\mu')^2}{2\sigma_k^2}}\right)\right)\\ &= 0.\end{aligned}$$

so the RHS of (24) is 0.

We have now shown that

$$\lim_{M\to\infty}\sup_n \mathbb{E}\left(|\bar{X}|\mathbb{I}_{|\bar{X}|>M}\right) = 0,$$

or, in other words, that $|\bar{X}|$ is uniformly integrable. As shown earlier, this concludes the proof that

$$\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n, D_n\right)$$

is uniformly integrable when $X_n^* \sim p(X_n^*|X)$.

To show that $\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n^+, D_n\right)$ is uniformly integrable, as in the proof of Lemma 8, we have

$$p(Q|X, X^*) \propto p(X^*|Q)p(X|Q)p(Q),$$

so we can view both $p(Q|X, X_n^*)$ and $p(Q|X_n^*)$ as the posteriors for the same Bayesian inference problem with observed data $X^*$, and priors $p(Q|X) \propto p(X|Q)p(Q)$ and $p(Q)$, respectively. $p(Q|X)$ is Gaussian, so the uniform integrability of

$$\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n^+, D_n\right)$$

follows from the previous case with different values for $\mu_0$ and $\sigma_0^2$. ■

**Lemma 27.** *Let $A, B$ be positive-definite matrices. Then $||(A + nB)^{-1}||_F = O(\frac{1}{n})$, where $||\cdot||_F$ is the Frobenious norm.*

**Proof** It suffices to show that $n||(A + nB)^{-1}||_F$ is bounded. Since $A$ and $B$ are positive-definite, the inverse $(A + nB)^{-1}$ exists for all $n \geq 0$. By the continuity of the matrix inverse and the Frobenius norm,

$$n||(A + nB)^{-1}||_F = \left|\left|\left(\frac{1}{n}A + B\right)^{-1}\right|\right|_F \to ||B^{-1}||_F$$

as $n \to \infty$. The required boundedness follows from continuity. ■

**Theorem 12.** *When the up- and downstream models are $d$-dimensional Gaussian mean estimations with known covariance $\Sigma_k$, and $D_n \sim \mathcal{N}(\bar{X}_n^*, n^{-1}\Sigma_k)$,*

$$\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n^+, D_n\right) \quad \text{and} \quad \sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n, D_n\right)$$

*are uniformly integrable when $X_n^* \sim p(X_n^*|X)$.*

**Proof** When the downstream model is Gaussian mean estimation with known covariance,

$$\begin{aligned}
\bar{Q}_n &= \mathcal{N}(\mu_n, \Sigma_n), \\
\mu_n &= \left(\Sigma_0^{-1} + n\Sigma_k^{-1}\right)^{-1}\left(\Sigma_0^{-1}\mu_0 + n\Sigma_k^{-1}\bar{X}_n^*\right), \\
\Sigma_n^{-1} &= \Sigma_0^{-1} + n\Sigma_k^{-1}.
\end{aligned}$$

We start with the proof for $\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n, D_n\right)$. By Pinsker's equality and the formula for KL-divergence between Gaussians, we have

$$\begin{aligned}
\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n, D_n\right) &\leq \sqrt{\frac{1}{2}n\mathrm{KL}(D_n \,||\, \bar{Q}_n)} \\
&= \sqrt{\frac{1}{4}n\left(\mathrm{tr}(n^{-1}\Sigma_n^{-1}\Sigma_k) + (\mu_n - \bar{X}_n^*)^T\Sigma_n^{-1}(\mu_n - \bar{X}_n^*) - d + \ln\frac{\det(\Sigma_n)}{\det(n^{-1}\Sigma_k)}\right)} \\
&\leq \sqrt{\left|\frac{1}{4}C_1(n)\right|} + \sqrt{\left|\frac{1}{4}n(\mu_n - \bar{X}_n^*)^T\Sigma_n^{-1}(\mu_n - \bar{X}_n^*)\right|} + \sqrt{\left|\frac{1}{4}C_2(n)\right|}
\end{aligned}$$

with

$$C_1(n) = n(\mathrm{tr}(n^{-1}\Sigma_n^{-1}\Sigma_k) - d)$$
$$C_2(n) = n \ln \frac{\det(\Sigma_n)}{\det(n^{-1}\Sigma_k)}.$$

For $C_1(n)$,

$$\begin{aligned} C_1(n) &= n(\mathrm{tr}(n^{-1}\Sigma_n^{-1}\Sigma_k) - d) \\ &= \mathrm{tr}((\Sigma_0^{-1} + n\Sigma_k^{-1})\Sigma_k) - nd \\ &= \mathrm{tr}(\Sigma_0^{-1}\Sigma_k) + n\mathrm{tr}(I_d) - nd \\ &= \mathrm{tr}(\Sigma_0^{-1}\Sigma_k), \end{aligned}$$

which does not depend on $n$, so $C_1(n)$ is bounded.

For $C_2(n)$,

$$\begin{aligned} C_2(n) &= n \ln \frac{\det(\Sigma_n)}{\det(n^{-1}\Sigma_k)} \\ &= -n \ln \frac{\det(n^{-1}\Sigma_k)}{\det(\Sigma_n)} \\ &= -n \ln \left( \frac{1}{n^d} \det(\Sigma_k) \det(\Sigma_0^{-1} + n\Sigma_k^{-1}) \right) \\ &= -n \ln \left( \frac{1}{n^d} \det(\Sigma_k\Sigma_0^{-1} + nI_d) \right) \end{aligned}$$

$\det(\Sigma_k\Sigma_0^{-1} + nI_d)$ is the characteristic polynomial of $\Sigma_k\Sigma_0^{-1}$ with $-n$ as the variable, so we have

$$\det(\Sigma_k\Sigma_0^{-1} + nI_d) = \prod_{j=1}^{d} (a_j + n)$$

where $a_1, \ldots, a_d$ are the eigenvalues of $\Sigma_k\Sigma_0^{-1}$. Plugging this into $C_2(n)$,

$$\begin{aligned} C_2(n) &= -n \ln \left( \frac{1}{n^d} \det(\Sigma_k\Sigma_0^{-1} + nI_d) \right) \\ &= -n \ln \left( \frac{1}{n^d} \prod_{j=1}^{d} (a_j + n) \right) \\ &= -n \sum_{j=1}^{d} \ln \left( \frac{a_j}{n} + 1 \right). \end{aligned}$$

Substituting $u_j = \frac{n}{a_j}$,

$$\begin{aligned} C_2(n) &= -n \sum_{j=1}^{d} \ln\left(\frac{a_j}{n} + 1\right). \\ &= -\sum_{j=1}^{d} u_j a_j \ln\left(\frac{1}{u_j} + 1\right). \\ &= -\sum_{j=1}^{d} a_j \ln\left(\frac{1}{u_j} + 1\right)^{u_j}. \end{aligned}$$

Since $\lim_{u_j \to \infty} \left(\frac{1}{u_j} + 1\right)^{u_j} = e$,

$$\lim_{n \to \infty} C_2(n) = -\sum_{j=1}^{d} a_j$$

so $C_2(n)$ is bounded.

For the remaining term, we have

$$\begin{aligned} (\mu_n - \bar{X}_n^*)^T \Sigma_n^{-1} (\mu_n - \bar{X}_n^*) &= (\Sigma_n^{-\frac{1}{2}} (\mu_n - \bar{X}_n^*))^T (\Sigma_n^{-\frac{1}{2}} (\mu_n - \bar{X}_n^*)) \\ &= ||\Sigma_n^{-\frac{1}{2}} (\mu_n - \bar{X}_n^*)||_2^2 \end{aligned}$$

where $\Sigma_n^{-\frac{1}{2}}$ is the upper triangular part of the Cholesky decomposition of $\Sigma_n^{-1}$. Now

$$\sqrt{\left|\frac{1}{4} n (\mu_n - \bar{X}_n^*)^T \Sigma_n^{-1} (\mu_n - \bar{X}_n^*)\right|} = \frac{1}{2}\sqrt{n} ||\Sigma_n^{-\frac{1}{2}} (\mu_n - \bar{X}_n^*)||_2.$$

Using $||M||_2$ for a matrix $M$ to denote the spectral 2-norm, we have

$$||\Sigma_n^{-\frac{1}{2}} (\mu_n - \bar{X}_n^*)||_2 \leq ||\Sigma_n^{-\frac{1}{2}}||_2 ||\mu_n - \bar{X}_n^*||_2$$

and using $||\cdot||_F$ to denote Frobenius norm, we have

$$||\Sigma_n^{-\frac{1}{2}}||_2 \leq ||\Sigma_n^{-\frac{1}{2}}||_F = \sqrt{\mathrm{tr}(\Sigma_n^{-1})} = \sqrt{\mathrm{tr}(\Sigma_0^{-1}) + n\mathrm{tr}(\Sigma_k^{-1})}.$$

Denote $s_n = \frac{1}{2}\sqrt{n}\sqrt{\mathrm{tr}(\Sigma_0^{-1}) + n\mathrm{tr}(\Sigma_k^{-1})}$. Clearly $s_n = O(n)$. Next,

$$\begin{aligned} \mu_n - \bar{X}_n^* &= (\Sigma_0^{-1} + n\Sigma_k^{-1})^{-1} \Sigma_0^{-1} \mu_0 + (\Sigma_0^{-1} + n\Sigma_k^{-1})^{-1} n\Sigma_k^{-1} \bar{X}_n^* - \bar{X}_n^* \\ &= (\Sigma_0^{-1} + n\Sigma_k^{-1})^{-1} \Sigma_0^{-1} \mu_0 + ((\Sigma_0^{-1} + n\Sigma_k^{-1})^{-1} n\Sigma_k^{-1} - I_d) \bar{X}_n^* \end{aligned}$$

and

$$\begin{aligned} (\Sigma_0^{-1} + n\Sigma_k^{-1})^{-1} n\Sigma_k^{-1} - I_d &= (\Sigma_0^{-1} + n\Sigma_k^{-1})^{-1} (n\Sigma_k^{-1} - (\Sigma_0^{-1} + n\Sigma_k^{-1})) \\ &= -(\Sigma_0^{-1} + n\Sigma_k^{-1})^{-1} \Sigma_0^{-1} \end{aligned}$$

so, denoting $M_n = (\Sigma_0^{-1} + n\Sigma_k^{-1})^{-1}\Sigma_0^{-1}$

$$||\mu_n - \bar{X}_n^*||_2 \leq ||M_n||_F||\mu_0||_2 + ||M_n||_F||\bar{X}_n^*||_2.$$

Denote

$$C_3(n) = s_n||M_n||_F||\mu_0||_2$$
$$C_4(n) = s_n||M_n||_F$$

By Lemma 27, $||M_n||_F = O(\frac{1}{n})$, so $C_3(n)$ and $C_4(n)$ are bounded. We now have

$$\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n, D_n\right) \leq \sqrt{\left|\frac{1}{4}C_1(n)\right|} + \sqrt{\left|\frac{1}{4}C_2(n)\right|} + C_3(n) + C_4(n)||\bar{X}_n^*||_2.$$

By Lemmas 24 and 25, it suffices to show that each of the terms on the right is uniformly integrable. The terms containing $C_1, C_2$ and $C_3$ are non-random and bounded in $n$, so they are uniformly integrable. Additionally, $C_4(n)$ is bounded, so we only need to show that $||\bar{X}_n^*||_2$ is uniformly integrable. We have

$$||\bar{X}_n^*||_2 \leq \sum_{j=1}^{d} |X_{n,j}^*| \tag{26}$$

so it suffices to show uniform integrability for a univariate Gaussian absolute value $|X_{n,j}^*|$, which is done in the proof of Theorem 26.

To show that $\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n^+, D_n\right)$ is uniformly integrable, as in the proof of Lemma 8, we have

$$p(Q|X, X^*) \propto p(X^*|Q)p(X|Q)p(Q),$$

so we can view both $p(Q|X, X_n^*)$ and $p(Q|X_n^*)$ as the posteriors for the same Bayesian inference problem with observed data $X^*$, and priors $p(Q|X) \propto p(X|Q)p(Q)$ and $p(Q)$, respectively. $p(Q|X)$ is Gaussian, so the uniform integrability of

$$\sqrt{n}\,\mathrm{TV}\left(\bar{Q}_n^+, D_n\right)$$

follows from the previous case with different values for $\mu_0$ and $\Sigma_0$. ■

### B.3 Convergence with Asymptotic Congeniality

**Theorem 14.** *If asymptotic congeniality and Condition 7 for all $Z_{n_Z}$, with the probabilities of Condition 7 taken conditional to $I_S$, hold,*

$$\lim_{n_Z\to\infty}\lim_{n_{X^*}\to\infty} \mathrm{TV}\left(p(Q|Z_{n_Z}, I_A), \bar{p}_{n_{X^*}}(Q)\right) = 0. \tag{16}$$

**Proof** By the triangle inequality of total variation distance,

$$\mathrm{TV}(p(Q|Z_{n_Z}, I_A), \bar{p}_{n_X^*}(Q)) \leq \mathrm{TV}(p(Q|Z_{n_Z}, I_A), p(Q|Z_{n_Z}, I_S)) + \mathrm{TV}(p(Q|Z_{n_Z}, I_S), \bar{p}_{n_X^*}(Q)).$$

Asymptotic congeniality implies

$$\lim_{n_Z \to \infty} \mathrm{TV}(p(Q|Z_{n_Z}, I_A), p(Q|Z_{n_Z}, I_S)) = 0.$$

Next, we examine $p(Q|Z, X^*_{n_{X^*}}, I_S)$ and $p(Q|X^*_{n_{X^*}}, I_A)$. We have

$$\begin{aligned}
&\mathrm{TV}(p(Q|Z_{n_Z}, X^*_{n_{X^*}}, I_S), p(Q|X^*_{n_{X^*}}, I_A)) \\
&\leq \mathrm{TV}(p(Q|Z_{n_Z}, X^*_{n_{X^*}}, I_S), p(Q|X^*_{n_{X^*}}, I_S)) + \mathrm{TV}(p(Q|X^*_{n_{X^*}}, I_S), p(Q|X^*_{n_{X^*}}, I_A)).
\end{aligned}$$

Asymptotic congeniality implies that

$$\lim_{n_{X^*} \to \infty} \mathrm{TV}(p(Q|X^*_{n_{X^*}}, I_S), p(Q|X^*_{n_{X^*}}, I_A)) = 0$$

for all $X^*_\infty$, which means that the limit holds in probability under all distributions for $X^*_{n_{X^*}}$. Combining this with Lemma 22, we get

$$\mathrm{TV}(p(Q|Z_{n_Z}, X^*_{n_{X^*}}, I_S), p(Q|X^*_{n_{X^*}}, I_A)) \xrightarrow{P} 0$$

as $n_{X^*} \to \infty$, with the probability over $X^*_{n_{X^*}} \sim p(X^*_{n_{X^*}}|Z_{n_Z}, I_S)$. Now we can apply Lemma 23 to $\mathrm{TV}(p(Q|Z_{n_Z}, I_S), \bar{p}_{n^*_X}(Q))$, since

$$\begin{aligned}
p(Q|Z_{n_Z}, I_S) &= \int p(Q|Z_{n_Z}, X^*_{n_{X^*}}, I_S) p(X^*_{n_{X^*}}|Z_{n_Z}, I_S) \,\mathrm{d}X^*_{n_{X^*}}, \\
\bar{p}_{n^*_X}(Q) &= \int p(Q|X^*_{n_{X^*}}, I_A) p(X^*_{n_{X^*}}|Z_{n_Z}, I_S) \,\mathrm{d}X^*_{n_{X^*}}.
\end{aligned}$$

This yields

$$\mathrm{TV}(p(Q|Z_{n_Z}, I_S), \bar{p}_{n^*_X}(Q)) \to 0$$

as $n_{X^*} \to \infty$ for all $Z_{n_Z}$, which concludes the proof. ■

### B.4 Non-private Gaussian Example Proofs

**Proposition 15.** *If $\mu^*$ is a sample from $\bar{p}_n(\mu)$ in Gaussian mean estimation with known variance, $\mu^*$ has a Gaussian distribution, and as $n_{X^*} \to \infty$,*

$$\mathbb{E}_{\mu^*}(\mu^*) \to \bar{\mu}_{n_X}, \quad \mathrm{Var}_{\mu^*}(\mu^*) \to \bar{\sigma}^2_{n_X}. \tag{19}$$

**Proof** We begin by checking where the mean and variance of $\mu^* \sim \bar{p}_n(\mu)$ converge when $n_{X^*} \to \infty$:

$$\begin{aligned}
\mathbb{E}_{\mu^*}(\mu^*) = \mathbb{E}_{X^*}(\mathbb{E}_{\mu^*|X^*}(\mu^*)) = \mathbb{E}_{X^*}\left(\hat{\mu}_{n_{X^*}}\right) &= \mathbb{E}_{X^*}\left(\frac{\frac{1}{\hat{\sigma}^2_0}\hat{\mu}_0 + \frac{n_{X^*}}{\hat{\sigma}^2_k}\bar{X}^*}{\frac{1}{\hat{\sigma}^2_0} + \frac{n_{X^*}}{\hat{\sigma}^2_k}}\right) \\
&= \frac{\frac{1}{\hat{\sigma}^2_0}\hat{\mu}_0 + \frac{n_{X^*}}{\hat{\sigma}^2_k}\mathbb{E}_{X^*}(\bar{X}^*)}{\frac{1}{\hat{\sigma}^2_0} + \frac{n_{X^*}}{\hat{\sigma}^2_k}} \to \mathbb{E}_{X^*}(\bar{X}^*) = \bar{\mu}_{n_X}
\end{aligned}$$

as $n_{X^*} \to \infty$.

For the variance,

$$\mathrm{Var}_{\mu^*}(\mu^*) = \mathbb{E}_{X^*}(\mathrm{Var}_{\mu^*|X^*}(\mu^*)) + \mathrm{Var}_{X^*}(\mathbb{E}_{\mu^*|X^*}(\mu^*)) = \mathbb{E}_{X^*}(\hat{\sigma}^2_{n_{X^*}}) + \mathrm{Var}_{X^*}(\hat{\mu}_{n_{X^*}}),$$

$$\mathbb{E}_{X^*}(\hat{\sigma}^2_{n_{X^*}}) = \mathbb{E}_{X^*}\left(\frac{1}{\frac{n_{X^*}}{\hat{\sigma}_k^2} + \frac{1}{\hat{\sigma}_0^2}}\right) \to 0, n_{X^*} \to \infty,$$

$$\mathrm{Var}_{X^*}\left(\hat{\mu}_{n_{X^*}}\right) = \mathrm{Var}_{X^*}\left(\frac{\frac{n_{X^*}}{\hat{\sigma}_k^2}\bar{X}^* + \frac{\hat{\mu}_0}{\hat{\sigma}_0^2}}{\frac{n_{X^*}}{\hat{\sigma}_k^2} + \frac{1}{\hat{\sigma}_0^2}}\right) = \left(\frac{\frac{n_{X^*}}{\hat{\sigma}_k^2}}{\frac{n_{X^*}}{\hat{\sigma}_k^2} + \frac{1}{\hat{\sigma}_0^2}}\right)^2 \mathrm{Var}_{X^*}(\bar{X}^*),$$

and

$$\begin{aligned}
\mathrm{Var}_{X^*}(\bar{X}^*) &= \mathbb{E}_{\bar{\mu}}(\mathrm{Var}_{X^*|\bar{\mu}}(\bar{X}^*)) + \mathrm{Var}_{\bar{\mu}}(\mathbb{E}_{X^*|\bar{\mu}}(\bar{X}^*)) \\
&= \frac{1}{n_{X^*}}\mathbb{E}_{\bar{\mu}}(\mathrm{Var}_{x^*|\bar{\mu}}(x^*)) + \mathrm{Var}_{\bar{\mu}}(\bar{\mu}) \\
&\to \mathrm{Var}_{\bar{\mu}}(\bar{\mu}) = \bar{\sigma}^2_{n_X}
\end{aligned}$$

as $n_{X^*} \to \infty$, where $x^*$ is a single element of $X^*$. Putting these together,

$$\mathbb{E}_{\mu^*}(\mu^*) \to \bar{\mu}_{n_X}, \quad \mathrm{Var}_{\mu^*}(\mu^*) \to \bar{\sigma}^2_{n_X}$$

as $n_{X^*} \to \infty$.

Next, we show that $\mu^*$ also has a Gaussian distribution. In

$$\mu^* \sim \int p(\mu|X_n^*)p(X_n^*|X)dX^*,$$

both $p(\mu|X_n^*) = \mathcal{N}(\hat{\mu}_{n_{X^*}}, \hat{\sigma}^2_{n_{X^*}})$ and $p(X_n^*|X)$ are Gaussian. Since $\hat{\mu}_{n_{X^*}}$ is a linear function of $X_n^*$, $p(\hat{\mu}_{n_{X^*}}|X)$ is also Gaussian. Since $\hat{\sigma}^2_{n_{X^*}}$ does not depend on $X^*$, $\mu^*$ is the sum of a random variable with distribution $\mathcal{N}(0, \hat{\sigma}^2_{n_{X^*}})$ and $\hat{\mu}_{n_{X^*}}$, which is also Gaussian, meaning that $\mu^*$ is Gaussian. ■

**Proposition 16.** *If $\mu^*$ is a sample from $\bar{p}_n(\mu)$ in Gaussian mean estimation with synthetic data generation assuming unknown variance, but downstream analysis assuming known variance,*

$$\mathbb{E}_{\mu^*}(\mu^*) \to \bar{\mu}_{n_X}, \quad \mathrm{Var}_{\mu^*}(\mu^*) \to \frac{\bar{\sigma}_0^2}{\bar{\kappa}_{n_X}} \tag{20}$$

*as $n_{X^*} \to \infty$.*

**Proof** Checking where the mean and variance of $\mu^* \sim \bar{p}_n(\mu)$ converge when $n_{X^*} \to \infty$:

$$\begin{aligned}
\mathbb{E}_{\mu^*}(\mu^*) = \mathbb{E}_{X^*}(\mathbb{E}_{\mu^*|X^*}(\mu^*)) = \mathbb{E}_{X^*}\left(\hat{\mu}_{n_{X^*}}\right) &= \mathbb{E}_{X^*}\left(\frac{\frac{1}{\hat{\sigma}_0^2}\hat{\mu}_0 + \frac{n_{X^*}}{\hat{\sigma}_k^2}\bar{X}^*}{\frac{1}{\hat{\sigma}_0^2} + \frac{n_{X^*}}{\hat{\sigma}_k^2}}\right) \\
&= \frac{\frac{1}{\hat{\sigma}_0^2}\hat{\mu}_0 + \frac{n_{X^*}}{\hat{\sigma}_k^2}\mathbb{E}_{X^*}(\bar{X}^*)}{\frac{1}{\hat{\sigma}_0^2} + \frac{n_{X^*}}{\hat{\sigma}_k^2}} \to \mathbb{E}_{X^*}(\bar{X}^*) = \bar{\mu}_{n_X}
\end{aligned}$$

as $n_{X^*} \to \infty$.

For the variance,

$$\begin{aligned}\mathrm{Var}_{\mu^*}(\mu^*) &= \mathbb{E}_{X^*}(\mathrm{Var}_{\mu^*|X^*}(\mu^*)) + \mathrm{Var}_{X^*}(\mathbb{E}_{\mu^*|X^*}(\mu^*)) \\ &= \mathbb{E}_{X^*}(\hat{\sigma}^2_{n_{X^*}}) + \mathrm{Var}_{X^*}(\hat{\mu}_{n_{X^*}}),\end{aligned}$$

$$\mathbb{E}_{X^*}(\hat{\sigma}^2_{n_{X^*}}) = \mathbb{E}_{X^*}\left(\frac{1}{\frac{n_{X^*}}{\hat{\sigma}_k^2} + \frac{1}{\hat{\sigma}_0^2}}\right) \to 0, n_{X^*} \to \infty,$$

$$\mathrm{Var}_{X^*}\left(\hat{\mu}_{n_{X^*}}\right) = \mathrm{Var}_{X^*}\left(\frac{\frac{n_{X^*}}{\hat{\sigma}_k^2}\bar{X}^* + \frac{\hat{\mu}_0}{\hat{\sigma}_0^2}}{\frac{n_{X^*}}{\hat{\sigma}_k^2} + \frac{1}{\hat{\sigma}_0^2}}\right) = \left(\frac{\frac{n_{X^*}}{\hat{\sigma}_k^2}}{\frac{n_{X^*}}{\hat{\sigma}_k^2} + \frac{1}{\hat{\sigma}_0^2}}\right)^2 \mathrm{Var}_{X^*}(\bar{X}^*),$$

and

$$\begin{aligned}\mathrm{Var}_{X^*}(\bar{X}^*) &= \mathbb{E}_{\bar{\mu},\bar{\sigma}^2}(\mathrm{Var}_{X^*|\bar{\mu},\bar{\sigma}^2}(\bar{X}^*)) + \mathrm{Var}_{\bar{\mu},\bar{\sigma}^2}(\mathbb{E}_{X^*|\bar{\mu},\bar{\sigma}^2}(\bar{X}^*)) \\ &= \frac{1}{n_{X^*}}\mathbb{E}_{\bar{\sigma}^2}(\bar{\sigma}^2) + \mathrm{Var}_{\bar{\mu}}(\bar{\mu}) \\ &\to \mathrm{Var}_{\bar{\mu}}(\bar{\mu}) = \frac{\bar{\sigma}_0^2}{\bar{\kappa}_{n_X}}\end{aligned}$$

as $n_{X^*} \to \infty$. Putting these together,

$$\mathbb{E}_{\mu^*}(\mu^*) \to \bar{\mu}_{n_X}, \quad \mathrm{Var}_{\mu^*}(\mu^*) \to \frac{\bar{\sigma}_0^2}{\bar{\kappa}_{n_X}}$$

as $n_{X^*} \to \infty$. ■

**Proposition 17.** *If $\sigma_*^2$ is a sample from $\bar{p}_n(\sigma)$ in Gaussian variance estimation with known mean,*

$$\mathbb{E}_{\sigma_*^2}(\sigma_*^2) \to \mathbb{E}_{\bar{\sigma}^2}(\bar{\sigma}^2) + (\bar{\mu}_k - \hat{\mu}_k)^2, \tag{22}$$

*as $n_{X^*} \to \infty$.*

**Proof** We have

$$\begin{aligned}\mathbb{E}_{\sigma_*^2}(\sigma_*^2) = \mathbb{E}_{X^*}(\mathbb{E}_{\sigma_*^2|X^*}(\sigma_*^2)) = \mathbb{E}_{X^*}\left(\frac{\hat{\nu}_{n_{X^*}}}{\hat{\nu}_{n_{X^*}} - 2}\hat{\sigma}^2_{n_{X^*}}\right) = \frac{\hat{\nu}_0 + n_{X^*}}{\hat{\nu}_0 + n_{X^*} - 2}\mathbb{E}_{X^*}\left(\hat{\sigma}^2_{n_{X^*}}\right) \\ = \frac{\hat{\nu}_0 + n_{X^*}}{\hat{\nu}_0 + n_{X^*} - 2}\frac{\hat{\nu}_0\hat{\sigma}_0^2 + n_{X^*}\mathbb{E}_{X^*}(\hat{v})}{\hat{\nu}_0 + n_{X^*}} = \frac{\hat{\nu}_0\hat{\sigma}_0^2 + n_{X^*}\mathbb{E}_{X^*}(\hat{v})}{\hat{\nu}_0 + n_{X^*} - 2},\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{X^*}(\hat{v}) &= \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} \mathbb{E}_{x_i^*}((x_i^* - \hat{\mu}_k)^2) = \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} \mathbb{E}_{x_i^*}((x_i^*)^2 - 2x_i^*\hat{\mu}_k + \hat{\mu}_k^2) \\
&= \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} \left(\mathbb{E}_{x_i^*}((x_i^*)^2) - 2\bar{\mu}_k\hat{\mu}_k + \hat{\mu}_k^2\right) = \hat{\mu}_k^2 - 2\bar{\mu}_k\hat{\mu}_k + \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} \mathbb{E}_{x_i^*}((x_i^*)^2) \\
&= \hat{\mu}_k^2 - 2\bar{\mu}_k\hat{\mu}_k + \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} (\mathbb{E}_{x_i^*}(x_i^*)^2 + \mathrm{Var}_{x_i^*}(x_i^*)) \\
&= \hat{\mu}_k^2 - 2\bar{\mu}_k\hat{\mu}_k + \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} (\bar{\mu}_k^2 + \mathrm{Var}_{x_i^*}(x_i^*)) \\
&= \bar{\mu}_k^2 + \hat{\mu}_k^2 - 2\bar{\mu}_k\hat{\mu}_k + \mathrm{Var}_{x_i^*}(x_i^*) = (\bar{\mu}_k - \hat{\mu}_k)^2 + \mathrm{Var}_{x_i^*}(x_i^*),
\end{aligned}$$

and

$$\mathrm{Var}_{x_i^*}(x_i^*) = \mathrm{Var}_{\bar{\sigma}^2}(\mathbb{E}_{x_i^*|\bar{\sigma}^2}(x_i^*)) + \mathbb{E}_{\bar{\sigma}^2}(\mathrm{Var}_{x_i^*|\bar{\sigma}^2}(x_i^*)) = \mathbb{E}_{\bar{\sigma}^2}(\bar{\sigma}^2).$$

Putting these together,

$$\mathbb{E}_{\sigma_*^2}(\sigma_*^2) \to \mathbb{E}_{\bar{\sigma}^2}(\bar{\sigma}^2) + (\bar{\mu}_k - \hat{\mu}_k)^2,$$

as $n_{X^*} \to \infty$. ■

## Appendix C. Sampling the Exact Posterior in the Toy Data Experiment

In order to sample the exact posterior $p(Q|\tilde{s})$, we use another decomposition:

$$p(Q|\tilde{s}) = \int p(Q|\tilde{s}, X)p(X|\tilde{s})\,\mathrm{d}X = \int p(Q|X)p(X|\tilde{s})\,\mathrm{d}X, \tag{27}$$

where $p(Q|\tilde{s}, X) = p(Q|X)$ due to the independencies of the graphical model in Figure 1. It remains to sample $p(X|\tilde{s})$. This is not tractable in general, but is possible in the toy data setting due to using the 3-way marginal query that covers all possible values of a datapoint, and the simplicity of the toy data.

We can decompose

$$p(X|\tilde{s}) = \int p(s|\tilde{s})p(X|s)\,\mathrm{d}\theta\,\mathrm{d}X = \int p(X|s) \int p(s, \theta|\tilde{s})\,\mathrm{d}\theta\,\mathrm{d}X,$$

so we can sample $(s, \theta) \sim p(s, \theta|s)$ and then sample $X \sim p(X|s)$ to obtain a sample from $p(X|\tilde{s})$. Due to the simplicity of the toy data setting, sampling both $p(s, \theta|s)$ and $p(X|s)$ is possible.

NAPSU-MQ uses the following Bayesian inference problem:

$$\begin{aligned}
\theta &\sim \text{Prior} \\
X &\sim \mathrm{MED}_\theta^n \\
s &= a(X) \\
\tilde{s} &\sim \mathcal{N}(s, \sigma_{DP}^2),
\end{aligned}$$

where $a$ are the marginal queries, $\sigma^2_{DP}$ is the noise variance of the Gaussian mechanism, and $\text{MED}^n_\theta$ is the maximum entropy distribution (Räisä et al., 2023) with point probability

$$p(x) = \exp(\theta^T a(x) - \theta_0(\theta)),$$

where $\theta_0$ is the log-normalising constant.

In the toy data setting, $a$ is the 3-way marginal query for all of the 3 variables. In other words, $a(x)$ is the one-hot coding of $x$, so $s = a(X)$ is a vector of counts of how many times each of the 8 possible values is repeated in $X$. This means that sampling $p(X|s)$ is simple:

1. For each possible value of a datapoint, find the corresponding count from $s$, and repeat that datapoint according the that count.

2. Shuffle the datapoints to a random order.

As the downstream analysis $p(Q|X)$ doesn't depend on the order of the datapoints, the second step is not actually needed.

To sample $p(s, \theta|\tilde{s})$, we use a Metropolis-within-Gibbs sampler (Gilks et al., 1995) that sequentially updates $s$ and $\theta$ while keeping the other fixed. The proposal for $\theta$ is obtained from Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 2011). The proposal for $s$ is obtained by repeatedly choosing a random index in $s$ to increment and another to decrement. It is possible to obtain negative values in $s$ from this proposal, but those will always be rejected by the acceptance test, as the likelihood for them is 0.

To initialise the sampler, we pick an initial value for $\theta$ from a Gaussian distribution, and pick the initial $s$ by rounding $\tilde{s}$ to integer values, changing the rounded values such that they sum to $n$ while ensuring that all values are non-negative.

The step size for the HMC we used is 0.05, and the number of steps is 20. In the $s$ proposal, we repeat the combination of an increment and a decrement 30 times. We take 20000 samples in total from 4 parallel chains, and drop the first 20% as warmup samples.

The method described in this section is similar to the noise-aware Bayesian inference method of Ju et al. (2022). The difference between the two is that Ju et al. (2022) use $X$ instead of $s$ as the auxiliary variable, and they sample the $X$ proposals from the model, changing one datapoint at a time. This makes their algorithm more generalisable.

## References

Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 308–318. ACM, 2016.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631. ACM, 2019.

Borja Balle and Yu-Xiang Wang. Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising. In *Proceedings of the 35th International*

*Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 394–403. PMLR, 2018.

Garrett Bernstein and Daniel Sheldon. Differentially Private Bayesian Inference for Exponential Families. In *Advances in Neural Information Processing Systems*, volume 31, pages 2924–2934, 2018.

Garrett Bernstein and Daniel Sheldon. Differentially Private Bayesian Linear Regression. In *Advances in Neural Information Processing Systems*, volume 32, pages 523–533, 2019.

Michael. J. Betancourt and Mark Girolami. Hamiltonian Monte Carlo for Hierarchical Models, 2013. arXiv:1312.0906.

Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, NY, 3rd edition, 1995.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 78 (5):1103–1130, 2016.

Peter Bühlmann. Discussion of Big Bayes Stories and BayesBag. *Statistical Science*, 29(1): 91–94, 2014.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians with the same mean, 2022. arXiv:1810.08693.

Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

Cynthia Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In *Advances in Cryptology - EUROCRYPT*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Third Theory of Cryptography Conference*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006b.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2. edition, 2004.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science Series. CRC Press, Boca Raton, FL, 3. edition, 2014.

Sahra Ghalebikesabi, Harry Wilde, Jack Jewson, Arnaud Doucet, Sebastian J. Vollmer, and Chris C. Holmes. Mitigating statistical bias within differentially private synthetic data. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 696–705. PMLR, 2022.

W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive Rejection Metropolis Sampling Within Gibbs Sampling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 44 (4):455–472, 1995.

C. Hipp and R. Michel. On the Bernstein-v. Mises Approximation of Posterior Distributions. *The Annals of Statistics*, 4(5):972–980, 1976.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Jonathan H. Huggins and Jeffrey W. Miller. Robust Inference and Model Criticism Using Bagged Posteriors, 2020. arXiv:1912.07104.

Jonathan H. Huggins and Jeffrey W. Miller. Reproducible Model Selection Using Bagged Posteriors. *Bayesian analysis*, 18(1):79–104, 2023.

Joonas Jälkö, Onur Dikmen, and Antti Honkela. Differentially Private Variational Inference for Non-conjugate Models. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017.

Robert I. Jennrich. Asymptotic Properties of Non-Linear Least Squares Estimators. *The Annals of Mathematical Statistics*, 40(2):633–643, 1969.

Nianqiao Ju, Jordan Awan, Ruobin Gong, and Vinayak Rao. Data Augmentation MCMC for Bayesian Inference from Privatized Data. In *Advances in Neural Information Processing Systems*, volume 35, pages 12732–12743, 2022.

Mark Kelbert. Survey of Distances between the Most Popular Distributions. *Analytics*, 2(1): 225–245, 2023.

Ronny Kohavi and Barry Becker. Adult. UCI Machine Learning Repository, 1996.

Tejas Kulkarni, Joonas Jälkö, Antti Koskela, Samuel Kaski, and Antti Honkela. Differentially Private Bayesian Inference for Generalized Linear Models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5838–5849. PMLR, 2021.

Chong K. Liew, Uinam J. Choi, and Chung J. Liew. A data distortion by probability distribution. *ACM Transactions on Database Systems*, 10(3):395–411, 1985.

Xiao-Li Meng. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, 9(4), 1994.

Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press, 2011.

Beata Nowok, Gillian M. Raab, and Chris Dibben. Synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, 74:1–26, 2016.

Lukas Prediger, Niki Loppi, Samuel Kaski, and Antti Honkela. D3p - A Python Package for Differentially-Private Probabilistic Programming. *Proceedings on Privacy Enhancing Technologies*, 2022(2):407–425, 2022.

Trivellore E. Raghunathan, Jerome P. Reiter, and Donald B. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1, 2003.

Ossi Räisä, Joonas Jälkö, Samuel Kaski, and Antti Honkela. Noise-aware statistical inference with differentially private synthetic data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 3620–3643. PMLR, 2023.

Arun Rajkumar and Shivani Agarwal. A differentially private stochastic gradient descent algorithm for multiparty classification. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *JMLR Proceedings*, pages 933–941. JMLR.org, 2012.

Jerome P. Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531, 2002.

Donald B. Rubin. The Bayesian Bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

Donald B. Rubin. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468, 1993.

Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE, 2013.

Boris van Breugel, Zhaozhi Qian, and Mihaela van der Schaar. Synthetic data, real errors: How (not) to publish and use synthetic data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34793–34808. PMLR, 2023a.

Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership Inference Attacks against Synthetic Data through Overfitting Detection. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 3493–3514. PMLR, 2023b.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, repr. 2000 edition, 1998.

Harrison Wilde, Jack Jewson, Sebastian J. Vollmer, and Chris Holmes. Foundations of Bayesian Learning from Synthetic Data. In *The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 541–549. PMLR, 2021.

Xianchao Xie and Xiao-Li Meng. Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? *Statistica Sinica*, 2016.