Noise-Aware Differentially Private Variational Inference

Talal Alrawajfeh
University of Helsinki
talal.alrawajfeh@helsinki.fi

Joonas Jälkö
University of Helsinki
joonas.jalko@helsinki.fi

Antti Honkela
University of Helsinki
antti.honkela@helsinki.fi

Abstract

Differential privacy (DP) provides robust privacy guarantees for statistical inference, but this can lead to unreliable results and biases in downstream applications. While several noise-aware approaches have been proposed which integrate DP perturbation into the inference, they are limited to specific types of simple probabilistic models. In this work, we propose a novel method for noise-aware approximate Bayesian inference based on stochastic gradient variational inference which can also be applied to high-dimensional and non-conjugate models. We also propose a more accurate evaluation method for noise-aware posteriors. Empirically, our inference method has similar performance to existing methods in the domain where they are applicable. Outside this domain, we obtain accurate coverages on high-dimensional Bayesian linear regression and well-calibrated predictive probabilities on Bayesian logistic regression with the UCI Adult dataset.

1 INTRODUCTION

When applying Bayesian inference on sensitive data, one needs to consider the privacy risks the released results might pose. Multiple methods combining the state-of-the-art privacy paradigm differential privacy (DP) (Dwork et al., 2006) with Bayesian inference have been proposed in the past. These include methods that are based on, e.g., Markov chain Monte Carlo (MCMC) (Wang et al., 2015; Heikkilä et al., 2019) and variational inference (VI) (Jälkö et al., 2017, 2023). Many of these methods do not consider how the additional noise due to DP affects the learned posteriors, which might lead to poor uncertainty quantification. As a solution, multiple works trying to explicitly model the DP noise have since been proposed (Bernstein and Sheldon, 2018, 2019; Ju et al., 2022), but these are limited to simple or small problems.

In this work, we propose a highly scalable noise-aware DP Bayesian inference approach that is applicable to a broad range of models. Our method uses DP variational inference (DPVI) (Jälkö et al., 2017, 2023) to obtain noisy gradients from an approximate Bayesian inference problem. Modelling the DP noise in the gradients, we form a probabilistic model connecting the noisy gradients to an optimal solution of the VI problem. Finally, learning the noise-aware posterior for the optimum, we can form a noise-aware posterior for the original Bayesian inference problem.

Related work: Several methods have been proposed to take the DP induced noise into account in private Bayesian inference. Earlier works include sufficient statistic based models such as noise-aware exponential family models (Bernstein and Sheldon, 2018) and Bayesian linear regression (Bernstein and Sheldon, 2019). In these works, the data is released through noisy sufficient statistics, and the perturbation noise is included as a part of the data generation process. Using approximate sufficient statistics, Kulkarni et al. (2021) extended the noise-awareness for Bayesian generalized linear models (Huggins et al., 2017). Gong (2022) proposed a general summary statistics–based approach using approximate Bayesian computation (ABC). Ju et al. (2022) proposed a Gibbs sampling based approach, where the latent confidential data is augmented into the inference model. Their approach is applicable for a wide range of models; however, the cost of the inference process scales by the number of samples in the data. Xiong et al. (2023) proposed a method which fits a normalizing flow as a surrogate for the true posterior, by iteratively drawing the model parameters from the surrogate, drawing the latent confidential data conditioned on the proposed parameter,

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

and finally updating the surrogate model based on the observed noisy summary statistics. As an important application of DP probabilistic modelling, Räisä et al. (2023) proposed a synthetic data generation method, which is based on learning a noise-aware posterior for a discrete marginal query based summary statistic.

Contributions:

1. We propose a theoretical framework for noise-aware inference and amend the theory of Lemos et al. (2023) to assess approximate noise-aware posteriors in Section 3.1.
2. We propose Noise-Aware DP VI (NA-DPVI) for approximate noise-aware inference in Section 3.3. The method is based on post-processing the gradient trace from DPVI using a Bayesian linear model to capture the uncertainty from DP, and combining this with data-modelling uncertainty using the VI posterior approximation.
3. We provide a theoretical analysis of the conditions under which our approach performs well, focusing on how the hyperparameters, including the learning-rate, affect the noise-aware posterior approximation in Appendix D.
4. We employ an accurate evaluation method for approximate noise-aware posteriors by modifying the Test of Accuracy with Random Points (TARP) method in (Lemos et al., 2023). We apply this algorithm to evaluate our method and compare it against existing baselines in Section 3.3. Additionally, we demonstrate the real-world applicability of our method by applying it to a Bayesian logistic regression model on the UCI Adult dataset.

2 BACKGROUND

2.1 Bayesian Inference

Assume we have a model $p\left(\mathbf{D} \mid \boldsymbol{\theta}\right)$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^n$ denotes the unobserved model parameters; and $\mathbf{D} \in \mathcal{D}$ denotes the data. Given a prior $p(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$, Bayes' Theorem states that the posterior can be written as:

$$p(\boldsymbol{\theta} \mid \mathbf{D}) = \frac{p\left(\mathbf{D} \mid \boldsymbol{\theta}\right) p(\boldsymbol{\theta})}{\int_{\Theta} p\left(\mathbf{D} \mid \boldsymbol{\theta}\right) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}. \tag{1}$$

In many cases, the denominator in Eq. (1) is intractable. For such cases, the true posterior can only be approximated with some other distribution $\widetilde{p}(\boldsymbol{\theta} \mid \mathbf{D})$ using methods such as Markov chain Monte Carlo (MCMC) (Neal, 1993) or Variational inference (VI) Jordan et al. (1999).

2.2 Validating Approximate Bayesian Inference

After obtaining $\widetilde{p}$, we want to test how well $\widetilde{p}$ approximates p. In order to devise such a test, let us first assume that $\widetilde{p}$ has support anywhere p has support. Next, we define a $(1-\alpha)$ credible region for $\tilde{p}$ as a mapping $\widetilde{\mathcal{R}}_\alpha : \mathcal{D} \to \mathcal{P}(\Theta)$, where $\mathcal{P}$ denotes a power set, such that for any $\mathbf{D} \in \mathcal{D}$

$$\int_{\Theta} \mathbb{1}_{\widetilde{\mathcal{R}}_\alpha(\mathbf{D})}\left(\boldsymbol{\theta}\right) \widetilde{p}\left(\boldsymbol{\theta} \mid \mathbf{D}\right) \mathrm{d}\boldsymbol{\theta} = 1 - \alpha. \tag{2}$$

Following Lemos et al. (2023), we define the Expected Coverage Probability (ECP) of $\widetilde{\mathcal{R}}_\alpha\left(\mathbf{D}\right)$ as

$$\mathrm{ECP}\left[\widetilde{\mathcal{R}}_\alpha\left(\mathbf{D}\right)\right] = \mathop{\mathbb{E}}_{\boldsymbol{\theta}, \mathbf{D} \sim p(\boldsymbol{\theta}, \mathbf{D})}\left[\mathbb{1}_{\widetilde{\mathcal{R}}_\alpha(\mathbf{D})}\left(\boldsymbol{\theta}\right)\right]. \tag{3}$$

Lemos et al. (2023) showed, that if $\mathrm{ECP}\left[\widetilde{\mathcal{R}}_\alpha\left(\mathbf{D}\right)\right] = 1-\alpha$ for every mapping $\widetilde{\mathcal{R}}_\alpha$ that satisfied Eq. (2) and every $\alpha \in (0,1)$, we have $\tilde{p}(\boldsymbol{\theta} \mid \mathbf{D}) = p(\boldsymbol{\theta} \mid \mathbf{D})$. Unfortunately, testing this is not practically feasible, as we would need to enumerate over every possible credible region. In practice, we need to perform the test over a finite selection of credible regions, and see if the expected coverage over this set is close to $1-\alpha$. However, the way the set of credible regions is constructed is critical as shown by Lemos et al. (2023). For example, we can obtain perfect $1-\alpha$ coverage by choosing $\widetilde{\mathcal{R}}_\alpha\left(\mathbf{D}\right)$ to be the $(1-\alpha)$ Highest Posterior Density (HPD) region and then setting $\widetilde{p}\left(\boldsymbol{\theta} \mid \mathbf{D}\right) = p(\boldsymbol{\theta})$, i.e. using the prior as the posterior approximation. To solve this problem, Lemos et al. (2023) introduced the concept of a **positionable credible region**.

Definition 1 (Lemos et al. (2023))**.** *A positionable credible region is a mapping*

$$\widetilde{\mathcal{C}}_\alpha : \mathcal{D} \times \Theta \to \mathcal{P}\left(\Theta\right), \tag{4}$$

for which the following conditions hold for all $\boldsymbol{\theta}_{ref} \in \Theta$*:*

1. *for any* $\alpha \in (0,1)$ *and* $\mathbf{D} \in \mathcal{D}$*, the set* $\widetilde{\mathcal{C}}_\alpha\left(\mathbf{D}, \boldsymbol{\theta}_{ref}\right)$ *is a* $(1-\alpha)$ *credible region for* $\widetilde{p}\left(\boldsymbol{\theta} \mid \mathbf{D}\right)$
2. *for all* $\mathbf{D} \in \mathcal{D}$*,* $\lim_{\alpha \to 1} \widetilde{\mathcal{C}}_\alpha\left(\mathbf{D}, \boldsymbol{\theta}_{ref}\right) = \{\boldsymbol{\theta}_{ref}\}$*.*

Intuitively, this means that we can position credible regions around any point in Θ. In addition, we want to choose $\boldsymbol{\theta}_{\mathrm{ref}}$ as a function of $\mathbf{D}$; that is, $\boldsymbol{\theta}_{\mathrm{ref}} : \mathcal{D} \to \Theta$. Based on Theorem 3 in Lemos et al. (2023), if

$$\mathrm{ECP}\left[\widetilde{\mathcal{C}}_\alpha\left(\mathbf{D}, \boldsymbol{\theta}_{\mathrm{ref}}\left(\mathbf{D}\right)\right)\right] = 1 - \alpha, \tag{5}$$

for all $\alpha \in (0,1)$ and any function $\boldsymbol{\theta}_{\mathrm{ref}} : \mathcal{D} \to \Theta$, then $\widetilde{p}\left(\boldsymbol{\theta} \mid \mathbf{D}\right) = p\left(\boldsymbol{\theta} \mid \mathbf{D}\right)$ for all $\boldsymbol{\theta} \in \Theta$ and $\mathbf{D} \in \mathcal{D}$. This is an important property of the ECP with the positionable-credible regions, as it establishes the equality of $\widetilde{p}$ and p if and only if $\widetilde{p}$ has perfect coverages. For the rest of the paper we will abbreviate $\widetilde{\mathcal{C}}_\alpha\left(\mathbf{D}, \boldsymbol{\theta}_{\text{ref}}\left(\mathbf{D}\right)\right)$ as $\widetilde{\mathcal{C}}_\alpha\left(\mathbf{D}\right)$.

The resulting coverage test by Lemos et al. (2023) called TARP can be implemented with the following steps.

Algorithm 1 TARP method. Modified from Lemos et al. (2023).

```
$S_\alpha \leftarrow 0$
**for** $k \leq K$ **do**
    Sample $\boldsymbol{\theta}, \mathbf{D}$ from the model $p(\boldsymbol{\theta}, \mathbf{D})$;
    Compute the approximate posterior $\tilde{p}(\boldsymbol{\theta} \mid \mathbf{D})$;
    Sample reference point $\boldsymbol{\theta}_{\text{ref}}^{(k)}$
    Set $I_k = \mathbb{1}_{\widetilde{\mathcal{C}}_\alpha(\mathbf{D})}(\boldsymbol{\theta})$;
    $S_\alpha = S_\alpha + I_k$
**end for**
**return** $\frac{1}{K} S_\alpha$
```

2.3 Differential Privacy (DP) and DP Stochastic Gradient Descent

DP (Dwork et al., 2006) is the standard definition for privacy in modern computer science. DP is defined over so called neighbouring data sets, i.e. data sets that differ in only single element. The formal definition is given as:

Definition 2. *Dwork et al. (2006) A randomized algorithm $\mathcal{A}$ is said to be (ϵ, δ)-DP if for all neighbouring data sets D, D' and all sets $S \subset im(A)$*

$$\Pr(\mathcal{A}(D) \in S) \leq e^{\epsilon} \Pr(\mathcal{A}(D') \in S) + \delta. \tag{6}$$

DP has a number of appealing theoretical properties, including compositionality, whereby repeated accesses to the data weaken the guarantees in a predictable manner, and post-processing immunity, whereby the guarantees can never be weakened by post-processing.

DP Stochastic Gradient Descent (DP-SGD) (Rajkumar and Agarwal, 2012; Song et al., 2013; Abadi et al., 2016) modifies the traditional SGD algorithm to satisfy DP. In order to apply DP-SGD, we need to assume that the loss function, parametrized by $\boldsymbol{\phi}$, can be represented as a sum over individuals $\boldsymbol{x}_i \in \mathbf{D}$:

$$\mathcal{L}\left(\boldsymbol{\phi}; \mathbf{D}\right) = \sum_{i=1}^{N} \ell(\boldsymbol{\phi}; \boldsymbol{x}_i). \tag{7}$$

DP-SGD makes three modifications to the SGD algorithm: 1) gradients are computed for each $\boldsymbol{x}_i$ in the batch 2) these *per-example* gradients are clipped to a bounded norm C and 3) Gaussian noise is added to the summed per-example gradients. The DP-SGD update rule can be written as

$$\begin{aligned} \boldsymbol{g}_{t+1} &= \sum_{i \in \mathcal{B}_{t+1}} \operatorname{clip}\left(\nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}_t; \boldsymbol{x}_i), C\right), \\ \boldsymbol{\phi}_{t+1} &= \boldsymbol{\phi}_t - \lambda_t \left[\boldsymbol{g}_{t+1} + \sigma_{\text{DP}} C \boldsymbol{\eta}_{t+1}\right], \end{aligned} \tag{8}$$

where λ_t is the learning rate, $\mathcal{B}_t$ is a Poisson subsampled minibatch and $\boldsymbol{\eta}_{t+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The Gaussian noise-level σ_{DP} is chosen so that the (ϵ, δ)-DP guarantees hold for T iterations of the DP-SGD algorithm. The general privacy proofs for DP-SGD allow releasing all of the intermediate steps and the noisy gradients of the algorithm under the same privacy guarantee.

2.4 DP Variational Inference

Variational inference (Jordan et al., 1999) is a method used to approximate a posterior $p\left(\boldsymbol{\theta} \mid \mathbf{D}\right)$ using another distribution called the variational distribution $q_{\text{VI}}\left(\boldsymbol{\theta}; \boldsymbol{\phi}\right)$ parameterized by a vector $\boldsymbol{\phi} \in \Phi \subseteq \mathbb{R}^d$. The idea is to find $\boldsymbol{\phi}$ that maximizes the Evidence Lower Bound (ELBO):

$$\mathcal{L}\left(\boldsymbol{\phi}; \mathbf{D}\right) = \mathbb{E}_{q_{\text{VI}}(\boldsymbol{\theta}; \boldsymbol{\phi})}\left[\log \frac{p(\boldsymbol{\theta}, \mathbf{D})}{q_{\text{VI}}(\boldsymbol{\theta}; \boldsymbol{\phi})}\right]. \tag{9}$$

Furthermore, when the observations are i.i.d. we can decompose the ELBO as

$$\begin{aligned} \mathcal{L}\left(\boldsymbol{\phi}; \mathbf{D}\right) = &\, \mathbb{E}_{q_{\text{VI}}(\boldsymbol{\theta}; \boldsymbol{\phi})}\left[\log \frac{p(\boldsymbol{\theta})}{q_{\text{VI}}(\boldsymbol{\theta}; \boldsymbol{\phi})}\right] \\ &+ \sum_{i=1}^{N} \mathbb{E}_{q_{\text{VI}}(\boldsymbol{\theta}; \boldsymbol{\phi})}\left[\log p(\boldsymbol{x}_i \mid \boldsymbol{\theta})\right]. \end{aligned} \tag{10}$$

The expectations in ELBO are typically intractable. To solve this, Kingma and Welling (2014) proposed a solution called Stochastic Gradient Variational Bayes (SGVB), which approximates the expectations with Monte-Carlo estimates and parametrizes q_{VI} s.t. the Monte-Carlo approximator of the ELBO becomes a differentiable function that can be optimized with gradient based methods. For more details on the Monte-Carlo estimation and the differentiable parametrization of q_{VI} see Appendix H.

Jälkö et al. (2017) proposed an DP variant of the SGVB method called DPVI, by replacing the gradient optimizer with DP-SGD. Later Jälkö et al. (2023) improved the DPVI by using certain structural knowledge of the gradients in order to improve the convergence.

3 NOISE-AWARE INFERENCE

3.1 Formalism

The distributions in this paper are assumed to be continuous unless explicitly stated otherwise. Optimizing the ELBO using DP-SGD introduces uncertainty into the final VI approximation. Our goal is to integrate the noise from the DP mechanism into the posterior, this motivates the following definition.

Definition 3. *Given any prior distribution $p(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$, let $\mathcal{M}(\mathbf{D};\epsilon,\delta)$ be randomized algorithm that outputs a random vector $\boldsymbol{\xi} \in \Xi \subseteq \mathbb{R}^w$ according to some distribution $p(\boldsymbol{\xi} \mid \mathbf{D}, \boldsymbol{\theta})$ and provides (ϵ,δ) differential privacy, then the following posterior distribution given according to Bayes' Theorem,*

$$p(\boldsymbol{\theta} \mid \boldsymbol{\xi}) = \frac{\int_{\mathcal{D}} p(\boldsymbol{\xi} \mid \mathbf{D}, \boldsymbol{\theta})\, p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \mathrm{d}\mathbf{D}}{\int_{\Theta}\int_{\mathcal{D}} p(\boldsymbol{\xi} \mid \mathbf{D}, \boldsymbol{\theta})\, p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \mathrm{d}\mathbf{D}\mathrm{d}\boldsymbol{\theta}}, \tag{11}$$

is called the (true) ***noise-aware posterior****. The following joint distribution which is found in Eq. (11) is very important and will be used repeatedly throughout this paper,*

$$p(\boldsymbol{\theta}, \mathbf{D}, \boldsymbol{\xi}) = p(\boldsymbol{\xi} \mid \mathbf{D}, \boldsymbol{\theta})\, p(\mathbf{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}). \tag{12}$$

Similar to $\mathbf{D}$, the vector $\boldsymbol{\xi}$ can be thought of as either one point or a list of points concatenated into one large vector. For instance, DP-SGD provides the iterates $(\boldsymbol{\phi}_0, \ldots, \boldsymbol{\phi}_T)$ as a list of vectors. Eq. (11) is obtained by applying Bayes' theorem using the joint distribution in Eq. (12) with additionally marginalizing out the data $\mathbf{D}$, as it cannot be observed in the DP setting.

Computing Eq. (11) analytically is often intractable, and instead it is approximated by another distribution. Given any approximation of Eq. (11), say $\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{\xi})$, we want to evaluate this approximation using the ECP. However, compared to the non-private Bayesian inference, we now have three random vectors $\boldsymbol{\theta}$, $\mathbf{D}$, and $\boldsymbol{\xi}$ in the data generating process. Thus, we propose a slight modification to the ECP as follows. Note that $p(\boldsymbol{\theta}, \boldsymbol{\xi}) = p(\boldsymbol{\xi} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$, so we can define the ECP equivalently using the joint distribution in Eq. (12) as,

$$\mathrm{ECP}\left[\widetilde{\mathcal{C}}_\alpha(\boldsymbol{\xi})\right] = \underset{\boldsymbol{\theta},\mathbf{D},\boldsymbol{\xi}\sim p(\boldsymbol{\theta},\mathbf{D},\boldsymbol{\xi})}{\mathbb{E}}\left[\mathbb{1}_{\widetilde{\mathcal{C}}_\alpha(\boldsymbol{\xi})}(\boldsymbol{\theta})\right]. \tag{13}$$

The reason why Eq. (13) and Eq. (3) are equivalent is that the indicator function $\mathbb{1}_{\widetilde{\mathcal{C}}_\alpha(\boldsymbol{\xi})}(\boldsymbol{\theta})$ does not depend on $\mathbf{D}$, so by applying Fubini's Theorem (Folland, 2013), the expectation in Eq. (13) becomes

$$\underset{\boldsymbol{\theta},\mathbf{D},\boldsymbol{\xi}\sim p(\boldsymbol{\theta},\mathbf{D},\boldsymbol{\xi})}{\mathbb{E}}\left[\mathbb{1}_{\widetilde{\mathcal{C}}_\alpha(\boldsymbol{\xi})}(\boldsymbol{\theta})\right] = \mathbb{1}_{\widetilde{\mathcal{C}}_\alpha(\boldsymbol{\xi})}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}, \boldsymbol{\xi}). \tag{14}$$

Similarly, the ECP in Eq. (13) with the positionable-credible regions establishes the equality of $\widetilde{p}$ and p if and only if $\widetilde{p}$ has perfect coverages.

To evaluate the approximate noise-aware posterior based on Eq. (13), the coverage test in Algorithm 1 needs to be modified to include the data generation process for $\boldsymbol{\xi}$. We first draw samples from the joint distribution Eq. (12), where the samples are drawn in the following order,

$$\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}) \to \mathbf{D}_i \sim p(\mathbf{D} \mid \boldsymbol{\theta}_i) \to \boldsymbol{\xi}_i \sim p(\boldsymbol{\xi}_i \mid \mathbf{D}_i, \boldsymbol{\theta}_i). \tag{15}$$

Second, we replace I_k in Algorithm 1 with $I_k = \mathbb{1}_{\widetilde{\mathcal{C}}_\alpha(\boldsymbol{\xi}_i)}(\boldsymbol{\theta}_i)$.

3.2 Approximate Inference with DPVI

We build our noise-aware approximate Bayesian inference method on top of DPVI. Recall that DPVI is based on optimizing the ELBO with DP-SGD. The final posterior approximation returned by DPVI is the $q(\boldsymbol{\theta}; \boldsymbol{\phi}_T)$, where $\boldsymbol{\phi}_T$ denote the last parameters returned by the DP-SGD optimization. The DPVI treats these parameters as the true optima, and hence disregards all the stochasticity that was introduces by the Gaussian perturbation. Therefore, the DPVI as such is completely unaware of the noise.

However, due to the privacy accounting of DP-SGD, we do not need to limit our considerations to the last iterate. In fact, we can use the full parameter and noisy gradient traces, respectively $\mathcal{T} = \{\boldsymbol{\phi}_t\}_{t=1}^T$ and $\widetilde{\mathcal{G}} = \{\widetilde{\boldsymbol{g}}_t\}_{t=1}^T$, for arbitrary post-processing. In the next Section, we will introduce our post-processing model, which allows us to model the DP-induced noise, and make the approximate inference noise-aware.

3.3 Post-processing Model

Assume for now that our gradients norms are bounded by C. Hence $\mathrm{clip}(\nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi}_t; x_i), C) = \nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi}_t; x_i)$. From the DP-SGD update equations, we have that the noisy gradient at iteration t is

$$\widetilde{\boldsymbol{g}}_{t+1} = \sum_{i \in \mathcal{B}_t} \nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi}_t; x_i) + \sigma_{\mathrm{DP}} C \boldsymbol{\eta}, \tag{16}$$

where $\boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I}_d)$ with $\mathbf{I}_d$ denoting a d-dimensional identity matrix. Next, we assume that the subsampled gradients could be approximated with a Gaussian distribution through the central limit theorem. This is possible since, in DP-SGD, a large sampling rate or batch size is usually used (e.g., $\kappa = 0.1$). Given the full-data gradient at iteration t, $\nabla\mathcal{L}(\boldsymbol{\phi}_t; \mathbf{D})$, we have

$$\sum_{i \in \mathcal{B}_t} \nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi}_t; x_i) \sim \mathcal{N}\left(\kappa \nabla\mathcal{L}(\boldsymbol{\phi}_t; \mathbf{D}), \Sigma_{\mathrm{sub}}(\boldsymbol{\phi}_t)\right). \tag{17}$$

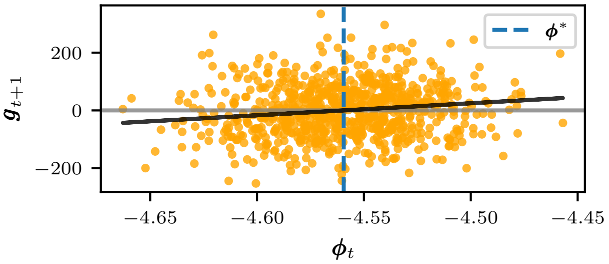

Figure 1: An example of the linear model of the perturbed gradients $\kappa\mathbf{A}(\boldsymbol{\phi}_t-\boldsymbol{\phi}^*)$ (black line) based on M1.

Combining Eqs. (16) and (17) yields

$$\widetilde{\boldsymbol{g}}_{t+1} \sim \mathcal{N}(\kappa\nabla\mathcal{L}(\boldsymbol{\phi}_t;\mathbf{D}), \sigma^2_{\mathrm{DP}}C^2\mathbf{I}_d + \Sigma_{\mathrm{sub}}(\boldsymbol{\phi}_t)). \quad (18)$$

Since the true gradient, as a data-dependent quantity is clearly unknown, we need to approximate it. We will approximate the true gradient as a linear function parameterized with a matrix $\mathbf{A}$ and a vector $\boldsymbol{\phi}^*$ through the second-order Taylor approximation of $\mathcal{L}$:

$$\nabla\mathcal{L}(\boldsymbol{\phi}_t;\mathbf{D}) \approx \mathbf{A}(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*). \quad (19)$$

This parametrization is convenient, as it allows us to interpret the $\boldsymbol{\phi}^*$ as the *true optimum* for the VI problem. If we further assume that Σ_{sub} is approximately constant around $\boldsymbol{\phi}^*$, then denoting the matrix $\Sigma_{\mathrm{total}} = \sigma^2_{\mathrm{DP}}C^2\mathbf{I}_d + \Sigma_{\mathrm{sub}}$, we can write

$$\widetilde{\boldsymbol{g}}_{t+1} \mid \boldsymbol{\phi}_t, \mathbf{A}, \boldsymbol{\phi}^*, \Sigma_{\mathrm{sub}} \sim \mathcal{N}\left(\kappa\mathbf{A}\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\right), \Sigma_{\mathrm{total}}\right). \quad (20)$$

Figure 1 shows an example of this model based on the exponential distribution M1, with the noisy gradients $\widetilde{\boldsymbol{g}}_t$ and a MAP estimate for $\boldsymbol{\phi}^*$ together with the fitted gradient. We now have a model for the noisy gradients, and can perform Bayesian inference on the unknown variables $\mathbf{A}, \boldsymbol{\phi}^*$ and Σ_{sub}. Note that the parameters $\mathbf{A}$ and $\boldsymbol{\phi}^*$ encapsulate all the information about the latent sensitive data $\mathbf{D}$. This avoids the costly marginalization of individual samples over the latent data, and instead our models scales for arbitrary number of data samples.

After obtaining the posterior for $\mathbf{A}, \boldsymbol{\phi}^*$ and Σ_{sub}, we can form the final noise-aware posterior approximation. We start by marginalizing out $\mathbf{A}$ and Σ_{sub} from the post-processing model,

$$p\left(\boldsymbol{\phi}^* \mid \mathcal{T}\right) = \int p\left(\mathbf{A}, \boldsymbol{\phi}^*, \Sigma_{\mathrm{sub}} \mid \mathcal{T}\right) \mathrm{d}\mathbf{A}\mathrm{d}\Sigma_{\mathrm{sub}}. \quad (21)$$

Then by observing that the joint distribution of $\boldsymbol{\theta}, \boldsymbol{\phi}^*$ conditioned on $\mathcal{T}$ is given by,

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\phi}^* \mid \mathcal{T}) &= p(\boldsymbol{\theta} \mid \boldsymbol{\phi}^*, \mathcal{T})p(\boldsymbol{\phi}^* \mid \mathcal{T}) \\ &= q_{\mathrm{VI}}(\boldsymbol{\theta}; \boldsymbol{\phi}^*)p(\boldsymbol{\phi}^* \mid \mathcal{T}), \end{aligned} \quad (22)$$

we obtain the final noise-aware approximate posterior $\widetilde{p}\left(\boldsymbol{\theta} \mid \mathcal{T}\right)$ by marginalizing out $\boldsymbol{\phi}^*$ in Eq. (22),

$$\widetilde{p}(\boldsymbol{\theta} \mid \mathcal{T}) = \int q_{\mathrm{VI}}(\boldsymbol{\theta}; \boldsymbol{\phi}^*)p(\boldsymbol{\phi}^* \mid \mathcal{T})\mathrm{d}\boldsymbol{\phi}^*. \quad (23)$$

To provide theoretical justification for our gradient-based model, we first need to state the following assumptions.

Assumption 1. *There exists a stationary point $\boldsymbol{\phi}^*$ for $\mathcal{L}$ such that $\nabla_{\boldsymbol{\phi}}\mathcal{L}\left(\boldsymbol{\phi}^*;\mathbf{D}\right) = \mathbf{0}$.*

Assumption 1 is a reasonable assumption for the VI problem to be solvable via DP-SGD.

Assumption 2. *We assume the loss function can be well-approximated by the second-order Taylor expansion on a neighborhood around $\boldsymbol{\phi}^*$ that includes the stationary part of the trace.*

Motivation for the second-order Taylor approximation can be found in Appendix A which implies that $\mathbf{A}$ is the Hessian. We can rigorously formalize Assumption 2 through the Fréchet derivative of $\nabla_{\boldsymbol{\phi}}\mathcal{L}\left(\boldsymbol{\phi};\mathbf{D}\right)$ at $\boldsymbol{\phi}^*$. For some tolerance $e_{\mathrm{tay}} > 0$, that is problem-dependent, there exists $1 < T^* < T$, and an open ball $\mathrm{B}_{r^*}(\boldsymbol{\phi}^*)$ around $\boldsymbol{\phi}^*$ with radius r^* that contains $\boldsymbol{\phi}_t$ for all $t \geq T^*$, such that, for all $\boldsymbol{\phi} \in \mathrm{B}_{r^*}(\boldsymbol{\phi}^*)$

$$\frac{\left\|\nabla_{\boldsymbol{\phi}}\mathcal{L}\left(\boldsymbol{\phi};\mathbf{D}\right) - \nabla^2_{\boldsymbol{\phi}}\mathcal{L}\left(\boldsymbol{\phi}^*;\mathbf{D}\right)\left(\boldsymbol{\phi}-\boldsymbol{\phi}^*\right)\right\|}{\left\|\boldsymbol{\phi}-\boldsymbol{\phi}^*\right\|} < e_{\mathrm{tay}}. \quad (24)$$

Assumption 3. *We assume that the clipping threshold C is high enough such that $C \geq \|\nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi}_t, \boldsymbol{x}_i)\|$ for all $1 \leq t \leq T$ and for all $1 \leq i \leq N$.*

Assumption 3 is required to make the clipping operation redundant, allowing us to write the noisy gradient models as in Eq. (20).

Assumption 4. *We assume that for all $1 \leq t \leq T$ the subsampling error is bounded, i.e. there exists $e_{sub} > 0$ such that*

$$\mathop{\mathbb{E}}_{Ber(\kappa)}\left[\left\|\boldsymbol{g}_{t+1} - \kappa\nabla_{\boldsymbol{\phi}}\mathcal{L}\left(\boldsymbol{\phi}_t;\mathbf{D}\right)\right\|^2\right] \leq e^2_{sub}. \quad (25)$$

We verify this assumption experimentally through the application of our method. We can approximate Σ_{sub} in principle; however, most of our experiments show that this is not necessary due to the relatively large magnitude of the DP noise, so we can set $\Sigma_{\mathrm{sub}} = \mathbf{0}$. Now we will introduce the following theorem which theoretically justifies our post-processing model.

Theorem 1. *Under Assumptions 1, 2, 3, and 4, there exists a matrix $\mathbf{A}$ such that*

$$\mathop{\mathbb{E}}_{Ber(\kappa)}\left[\left\|\boldsymbol{g}_{t+1} - \kappa\mathbf{A}\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\right)\right\|^2\right] < e^2_{approx}, \quad (26)$$

where $e^2_{approx} = e^2_{sub} + (\kappa \times e_{tay} \times r^*)^2$.

The proof of Theorem 1 is found in Appendix B. If some of these assumptions do not hold in practice, such as Assumption 3, then it is sufficient that the approximation given by Eq. (20) holds for our method to work. These assumptions mainly aid the theoretical analysis and the derivation of our method. We verify that the approximation in Eq. (20) is reasonable in the following section experimentally, by applying the method to various models. Furthermore, we find that the method is sensitive to the learning rate of DP-SGD and the priors. Thus, we analyze this further in Appendix D. We obtain a heuristic for the learning rate based on that analysis, which works well across our experiments.

4 EXPERIMENTS

4.1 Setup

In our experiments, we use a diagonal Gaussian distribution as the variational distribution, i.e. $q_{\mathrm{VI}}(\boldsymbol{\theta}; \boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{s}))$, where $\boldsymbol{\mu}$ are the means and $\boldsymbol{s}$ are the variances. Instead of directly optimizing the variances $\boldsymbol{s} \in \mathbb{R}^n_+$, we parametrize the variational distribution with $\boldsymbol{u} \in \mathbb{R}^n$ such that $\boldsymbol{s}$ is an element-wise softplus transformation of $\boldsymbol{u}$ as $s_j = \log(1 + e^{u_j})$. Now the variational distribution is parameterized with $\boldsymbol{\phi} = (\boldsymbol{\mu}, \boldsymbol{u})$ leading to $d = 2n$ variational parameters to optimize in total.

Jälkö et al. (2023) demonstrated that the gradients w.r.t. $\boldsymbol{u}$ are usually a lot smaller in magnitude than the gradients w.r.t. $\boldsymbol{\mu}$. Since DP-SGD adds identically distributed noise to each dimension of the gradient, the gradients w.r.t. $\boldsymbol{u}$ will be affected disproportionally by the noise. To avoid this, we scale up gradients w.r.t. $\boldsymbol{u}$ before clipping and noise addition, and revert this as a post-processing before we take the update step. For more details about this *preconditioning* see Appendix H.

We infer the post-processing model in Eq. (22) through two methods, the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014) and using Laplace's approximation which we obtain using the Adam optimizer (Kingma and Ba, 2015). More details about these methods can be found in Appendix C. Additionally we compare our method against the last iterate DPVI.

We approximate $\mathbf{A}$ with a diagonal matrix, that is $\mathbf{A} = \mathrm{diag}(a_1, \ldots, a_d)$. Further, from the fact that the Hessian of the loss function $\mathcal{L}$ is positive-definite around its optimum, we know$\mathbf{A}$ to be positive-definite. Therefore, we constrain $\mathbf{A}$ to be element-wise positive by parameterizing the post-processing model with $\boldsymbol{v} \in \mathbb{R}^d$ and mapping the v_i to a_i with softplus transformation, i.e. $a_i = \mathrm{softplus}(v_i)$. The algorithm blocks for NA-DPVI are provided in Appendix H.7.

We observed empirically that non-informative priors worked poorly for our post-processing model which motivated us to do further analysis (see Appendix D). Instead, we chose the priors for $\boldsymbol{v}$ and $\boldsymbol{\phi}^*$ based on $\mathcal{T}$ and $\widetilde{\mathcal{G}}$ which we discuss in Appendix H.3.

For evaluating the coverages, we apply a slightly modified version of the TARP algorithm (Algorithm 1) according to Section 3.1 which we elaborate on further in Appendix H.4. We repeat TARP 20 times in each experiment with $K = 500$ to obtain error estimates for the TARP. We estimate the error as the difference between the coverages at each credible level and the perfect coverages ($C(\alpha) - (1 - \alpha)$). We summarize these coverage errors with a single number using the RMSE (Root Mean Square Error) metric for N_α values of α,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_\alpha} \sum_{i=1}^{N_\alpha} \left(C\left(\alpha_i\right) - \left(1 - \alpha_i\right)\right)^2}. \qquad (27)$$

Across all the coverage experiments we use $\delta = 10^{-5}$, $T = 10^4$, $N = 5000$, $\kappa = 0.1$, and we use the NUTS method with 1000 warmup iterations and then run it for 4000 iterations.

4.2 Exponential Families

We compare our approach to the method of Bernstein and Sheldon (2018) which only works for exponential families, and with DPVIm (Jälkö et al., 2023), to show that our method works well for similar models. We implement the following conjugate models,

$$\begin{aligned} &\mathrm{M1}\ : \boldsymbol{\theta} \sim \mathrm{Gamma}\left(\alpha, \beta\right), \boldsymbol{x} \mid \boldsymbol{\theta} \sim \mathrm{Exp}(\boldsymbol{\theta}), \\ &\mathrm{M2}\ : \boldsymbol{\theta} \sim \mathrm{Beta}\left(\alpha, \beta\right), \boldsymbol{x} \mid \boldsymbol{\theta} \sim \mathrm{Ber}(\boldsymbol{\theta}), \\ &\mathrm{M3}\ : \boldsymbol{\theta} \sim \mathrm{Dir}\left(\alpha_1, \alpha_2, \alpha_3\right), \boldsymbol{x} \mid \boldsymbol{\theta} \sim \mathrm{Cat}(\boldsymbol{\theta}). \end{aligned}$$

For details on the specific parameter transformations we use for these models see, Appendix H.5. The coverages for both our method and Bernstein & Sheldon's are shown in Fig. 2 including the coverages for the naive baseline $q_{\mathrm{VI}}(\boldsymbol{\theta} \mid \boldsymbol{\phi}_T)$. We use $\epsilon = 0.1$ for all the models. We also show the errors between the coverages and the perfect coverages in Fig. 2). Table 1 summarizes the errors using the RMSE in Eq. (27). From the coverages and coverage errors, we can see that both methods work similarly compared to the naive baseline. Our method performs better for M1 and M2; however, Bernstein & Sheldon's method works better for M3. The marginal coverages and the marginal coverage errors for M3, see Appendix H.5.

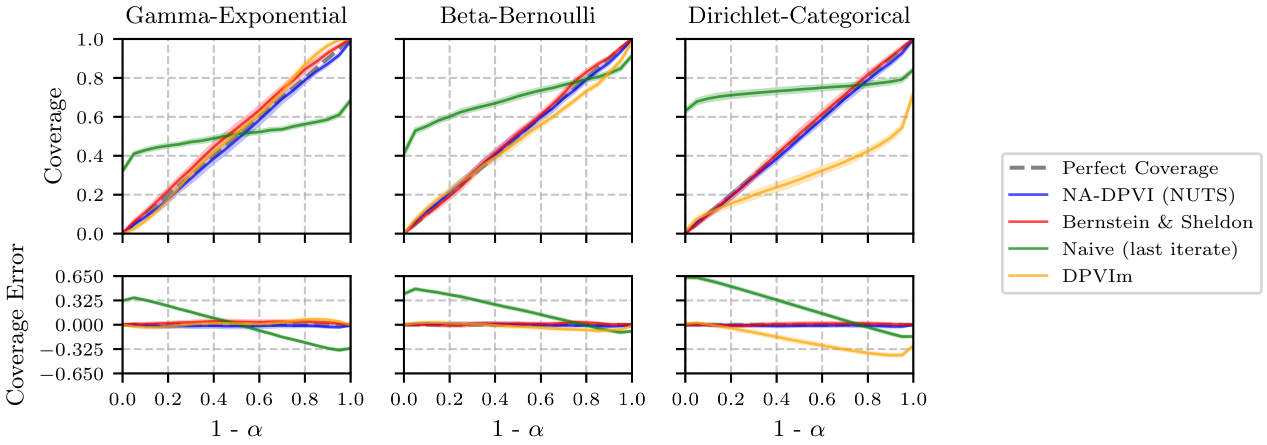

Figure 2: The first row in the figure shows the TARP coverages for the exponential families experiment for NA-DPVI (NUTS), Bernstein & Sheldon's method (Bernstein and Sheldon, 2018), last iterate DPVI, and DPVIm (Jälkö et al., 2023). The second row shows the error for the coverages $(C(\alpha) - (1 - \alpha))$. The solid lines show the average performance over 20 independent TARP repetitions and the error bars show the corresponding std. The parameters for NA-DPVI are $\delta = 10^{-5}$, $N = 5000$, $\kappa = 0.1$ and $T = 10^4$.

Table 1: The RMSE errors corresponding to the exponential family coverages in Fig. 2. Average RMSE ± std. $\delta = 10^{-5}$ in all experiments.

Method	**Gamma-Exponential**	**Beta-Bernoulli**	**Dirichlet-Categorical**
NA-DPVI (NUTS)	**0.023 ± 0.008**	**0.016 ± 0.006**	0.020 ± 0.004
Bernstein & Sheldon	0.034 ± 0.011	0.018 ± 0.006	**0.017 ± 0.007**
Naive (last iterate)	0.232 ± 0.003	0.273 ± 0.007	0.355 ± 0.009
DPVIm	0.038 ± 0.005	0.044 ± 0.004	0.251 ± 0.011

4.3 10D Bayesian linear regression

Noise-aware Bayesian linear regression was done before by (Bernstein and Sheldon, 2019); however their approach does not scale to many dimensions. The baselines we provide are Gibbs-SS-Noisy (Bernstein and Sheldon, 2019), DPVIm (Jälkö et al., 2023), and by using the last iterate $\boldsymbol{\phi}_T$, i.e. constructing credible regions from the distribution $q_{\text{VI}}(\boldsymbol{\theta}; \boldsymbol{\phi}_T)$. We apply our method and compare it with the baselines over the following model:

$$\begin{aligned}&\left(\theta_i, \theta_\sigma^2\right) \sim \text{N-}\Gamma^{-1}\left(0, \frac{1}{4}, 20, \frac{1}{2}\right), \\ &x_i \sim \mathcal{N}\left(0, 1\right), \quad y \mid \boldsymbol{x}, \boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{w}^\top \boldsymbol{x}, \theta_\sigma^2\right),\end{aligned} \tag{28}$$

where the parameters are $\boldsymbol{\theta} = \left(\theta_1, \dots, \theta_{10}, \theta_{11}, \theta_\sigma^2\right)$ and N-Γ^{-1} is the normal-inverse-gamma distribution. We define $\boldsymbol{w} = (\theta_1, \dots, \theta_{11})$ and $\boldsymbol{x} = (x_1, \dots, x_{10}, 1)$. The bias term is θ_{11}. For the constrained optimization problem, we only condition θ_σ^2 and use the same function as Eq. (A64).

We compute the coverages for our method using both NUTS and Laplace's approximation for $\epsilon \in \{0.1, 0.3, 1.0\}$. The coverages for our method is shown in Fig. 3 including the coverages for the naive baseline $q_{\text{VI}}(\boldsymbol{\theta} \mid \boldsymbol{\phi}_T)$. We also show the errors between the coverages and the perfect coverages in 3. Table 2 summarizes the errors using the RMSE Eq. (27). From the coverages and coverage errors, we can see that both methods work similarly compared to the naive baseline. However, NUTS performs better compared to Laplace's approximation.

Table 2: The RMSE errors corresponding to the Bayesian linear regression coverages in Fig. 3. Average RMSE ± std, with all the values scaled by 10^3. $\delta = 10^{-5}$ in all experiments.

Method	$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 1.0$
NA-DPVI (NUTS)	**36 ± 11**	**35 ± 11**	**27 ± 9**
NA-DPVI (Laplace)	78 ± 11	98 ± 6	60 ± 7
Naive (last iterate)	512 ± 3	307 ± 3	360 ± 3
Bernstein & Sheldon	301 ± 1	299 ± 1	323 ± 3
DPVIm	584 ± 0	584 ± 0	584 ± 0

4.4 UCI Adult Bayesian Logistic Regression

We perform an experiment on the UCI Adult dataset (Becker and Kohavi, 1996) with the Bayesian logistic

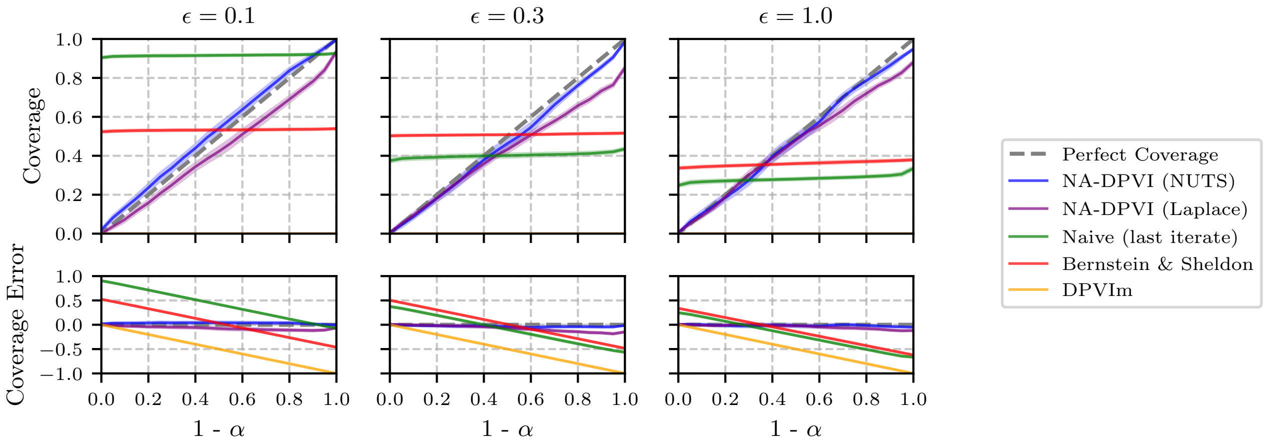

Figure 3: The first row in the figure shows the TARP coverages for the 10D Bayesian linear regression experiment for NA-DPVI (NUTS, Laplace), last iterate DPVI, DPVIm (Jälkö et al., 2023), and Gibbs-SS-Noisy (Bernstein and Sheldon, 2019). The second row shows the error for the coverages ($C(\alpha) - (1 - \alpha)$). The solid lines show the average performance over 20 independent TARP repetitions and the error bars show the corresponding std. The parameters for NA-DPVI are $\delta = 10^{-5}$, $N = 5000$, $\kappa = 0.1$ and $T = 10^4$.

regression model with standard normal prior. No baseline method can scale to this problem other than the naive (last iterate) baseline and DPVIm (Jälkö et al., 2023). After learning the posterior for $\boldsymbol{\theta}$, we test the posterior predictive distribution given as

$$\widetilde{p}(y = 1 \mid \mathcal{T}, \boldsymbol{x}) = \int p(y = 1 \mid \boldsymbol{\theta}) \widetilde{p}(\boldsymbol{\theta} \mid \mathcal{T}) \mathrm{d}\boldsymbol{\theta}, \qquad (29)$$

where we replace the integral with its Monte-Carlo estimator. We do 20 repeats for each experiment and we plot the calibration curves for NA-DPVI (NUTS & Laplace) and for the other baselines with $\epsilon \in \{0.1, 0.3, 1.0\}$. The calibration curves and calibration errors are found in Fig. 4. We also compute the RMSE for the calibration errors in Table 3, this is the same as the square root of the Brier score (Brier, 1950). From the calibrations and calibration errors, we can see that both NUTS and Laplace outperform the naive baseline and DPVIm. However, NUTS performs slightly better than Laplace's approximation. For more details about the experiment, see Appendix H.6.

Table 3: The RMSE errors corresponding to the logistic regression calibration in Fig. 4. Average RMSE ± std, with all the values scaled by 10^3. $\delta = 10^{-5}$ in all experiments.

Method	$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 1.0$
NA-DPVI (NUTS)	**61 ± 31**	**26 ± 5**	**24 ± 7**
NA-DPVI (Laplace)	84 ± 43	44 ± 11	46 ± 14
Naive (last iterate)	120 ± 65	67 ± 24	54 ± 17
DPVIm	101 ± 52	65 ± 25	38 ± 14

5 DISCUSSION

Bayesian inference should be a very natural companion to DP because it naturally deals with uncertain information and should thus be able to easily tolerate the noise added for DP. The fact that noise-aware inference is so difficult reminds us that commonly used general-purpose inference methods such as MCMC are not really Bayesian since they do not incorporate a mechanism for reasoning about the accuracy of their results. In this sense, algorithms required for noise-aware inference share a similarity with probabilistic numerics (Hennig et al., 2022).

We cite DPVIm (Jälkö et al., 2023) in our work regarding DPVI, which they extensively investigate; however, the noise-awareness concept in their work is not exactly equivalent to ours which we generalized from Bernstein and Sheldon (2019). Although our proposed NA-DPVI greatly expands the domain of noise-aware inference, it is still restricted by its reliance on potentially inaccurate VI to capture the data-modelling uncertainty. Additionally, our approximations only work under certain conditions, which may not always hold. An explicit privacy-utility-computation trade-off is very difficult to obtain for NA-DPVI, because there are several factors that are challenging to account for, theoretically speaking. These factors include DP-SGD, VI, the post-processing model, and the final approximate posterior obtained either in closed form or through an approximate inference method (MCMC, Laplace).

Another limitation in this work is that we did not account for the privacy leakage from hyper-parameter

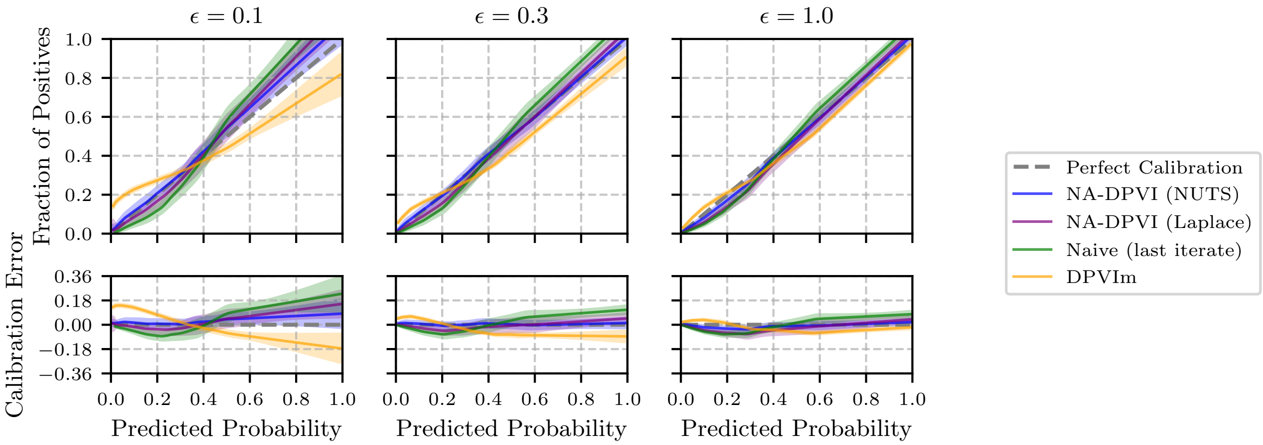

Figure 4: The first row in the figure shows the predictive calibration for the UCI Adult logistic regression experiment for NA-DPVI (NUTS), NA-DPVI (Laplace), last iterate DPVI, and DPVIm (Jälkö et al., 2023). The second row shows the calibration error (Fraction of Positives - Predicted Probability). The solid lines show the average performance over 20 independent repetitions, and the error bars show the corresponding std. The parameters for NA-DPVI are $\delta = 10^{-5}$, $\kappa = 0.1$ and $T = 10^4$.

selection, which was done manually due to the high computational cost. Currently, many papers that employ DP-SGD ignore the privacy accounting of hyper-parameter tuning. For example, see Section 5.2 of Sander et al. (2023). We applied heuristics to choose some of the hyper-parameters (e.g., the learning rate), and other hyper-parameters were shared among most experiments which reduces the privacy leakage. The state of privacy accounting for hyper-parameter tuning is still at its infancy, see Section 5.4 of Ponomareva et al. (2023), and this presents an area for future research and improvement, especially for noise-aware inference.

The search for the ultimate noise-aware inference algorithm that could accurately capture both data-modelling uncertainty as well as uncertainty due to DP inference for arbitrary models remains an important goal for future research.

Acknowledgments

This work was supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI, Grant 356499 and Grant 359111), 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. This work has been performed using resources provided by the CSC – IT Center for Science, Finland (Project 2003275).

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, 2016.

B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996.

G. Bernstein and D. Sheldon. Differentially private Bayesian inference for exponential families. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pages 2924–2934, 2018.

G. Bernstein and D. Sheldon. Differentially private Bayesian linear regression. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 523–533, 2019.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518): 859–877, 2017.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78 (1):1 – 3, 1950.

C. Dwork, F. McSherry, K. Nissim, and A. D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

G. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013.

R. Gong. Exact inference with approximate computation for differentially private data via perturbations. *J. Priv. Confidentiality*, 12(2), 2022.

S. Gopi, Y. T. Lee, and L. Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 11631–11642, 2021.

M. A. Heikkilä, J. Jälkö, O. Dikmen, and A. Honkela. Differentially private Markov chain Monte Carlo. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 4115–4125, 2019.

P. Hennig, M. A. Osborne, and H. P. Kersting. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, 2022.

M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

J. H. Huggins, R. P. Adams, and T. Broderick. PASS-GLM: polynomial approximate sufficient statistics for scalable Bayesian GLM inference. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 3611–3621, 2017.

J. Jälkö, O. Dikmen, and A. Honkela. Differentially private variational inference for non-conjugate models. In *Uncertainty in Artificial Intelligence 2017*. The Association for Uncertainty in Artificial Intelligence, 2017.

J. Jälkö, L. Prediger, A. Honkela, and S. Kaski. DPVIm: Differentially private variational inference improved. *Transactions on Machine Learning Research*, 2023.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, 1999.

N. Ju, J. Awan, R. Gong, and V. Rao. Data augmentation MCMC for Bayesian inference from privatized data. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings*, 2014.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

T. Kulkarni, J. Jälkö, A. Koskela, S. Kaski, and A. Honkela. Differentially private Bayesian inference for generalized linear models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 5838–5849. PMLR, 2021.

P. Lemos, A. Coogan, Y. Hezaveh, and L. P. Levasseur. Sampling-based accuracy testing of posterior estimators for general inference. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 19256–19273. PMLR, 2023.

Q. Li, C. Tai, and W. E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: mathematical foundations. *J. Mach. Learn. Res.*, 20:40:1–40:47, 2019.

Z. Li, S. Malladi, and S. Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 12712–12725, 2021.

S. Mandt, M. D. Hoffman, and D. M. Blei. A variational analysis of stochastic gradient algorithms. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 354–363. JMLR.org, 2016.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical report, Department of Computer Science, University of Toronto Toronto, ON, Canada, 1993.

N. Ponomareva, H. Hazimeh, A. Kurakin, Z. Xu, C. Denison, H. B. McMahan, S. Vassilvitskii, S. Chien, and A. G. Thakurta. How to dp-fy ML: A practical guide to machine learning with differential privacy. *J. Artif. Intell. Res.*, 77:1113–1201, 2023.

O. Räisä, J. Jälkö, S. Kaski, and A. Honkela. Noise-aware statistical inference with differentially private synthetic data. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, volume 206 of *Proceedings of Machine Learning Research*, pages 3620–3643. PMLR, 2023.

A. Rajkumar and S. 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 *Proceedings of Machine Learning Research*, pages 933–941. PMLR, 2012.

L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes and Martingales.* Cambridge Mathematical Library. Cambridge University Press, 2 edition, 2000.

T. Sander, P. Stock, and A. Sablayrolles. TAN without a burn: Scaling laws of DP-SGD. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 29937–29949. PMLR, 2023.

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

Y. Wang, S. E. Fienberg, and A. J. Smola. Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2493–2502, 2015.

Y. Xiong, N. P. Ju, and S. Zhang. Conditional density estimations from privacy-protected data. *CoRR*, abs/2310.12781, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] Refer to Section 3, Section 4.1, and Appendix H.7.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] Refer to Appendix H.7 for the time complexity of NA-DPVI.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] Refer to Section 3 and Appendix D.
   (b) Complete proofs of all theoretical results. [Yes] Refer to Appendix E, Appendix F, and Appendix G.
   (c) Clear explanations of any assumptions. [Yes] Refer to Section 3 and Appendix D.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] Refer to Section 3.3 and Appendix H.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] Refer to Section 3.3 and Appendix H.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] Refer to Section 5 and Appendix H.8.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Second-Order Taylor Approximation of the Loss Function

For any function $\mathcal{L}(\boldsymbol{\phi})$, if $\nabla \mathcal{L}(\boldsymbol{\phi}^*) = 0$, then by using the second-order Taylor approximation around $\boldsymbol{\phi}^*$, we can write $\mathcal{L}(\boldsymbol{\phi})$ as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}) &\approx \mathcal{L}(\boldsymbol{\phi}^*) + \nabla \mathcal{L}(\boldsymbol{\phi}^*)^T (\boldsymbol{\phi} - \boldsymbol{\phi}^*) + \frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\phi}^*)^T \nabla^2 \mathcal{L}(\boldsymbol{\phi}^*)(\boldsymbol{\phi} - \boldsymbol{\phi}^*) \\ &\approx \mathcal{L}(\boldsymbol{\phi}^*) + \frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\phi}^*)^T \nabla^2 \mathcal{L}(\boldsymbol{\phi}^*)(\boldsymbol{\phi} - \boldsymbol{\phi}^*),\end{aligned} \tag{A1}$$

where $\nabla^2 \mathcal{L}(\boldsymbol{\phi}^*)$ is the hessian of $\mathcal{L}$ at $\boldsymbol{\phi}^*$. Taking the gradient of both sides of (A1) with respect to $\boldsymbol{\phi}$, we obtain,

$$\nabla \mathcal{L}(\boldsymbol{\phi}) \approx \nabla^2 \mathcal{L}(\boldsymbol{\phi}^*)(\boldsymbol{\phi} - \boldsymbol{\phi}^*). \tag{A2}$$

B Proof of Theorem 1

Proof. Let $\mathbf{A} = \nabla^2_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}^*; \mathbf{D})$, then for all $t \geq T^*$,

$$\begin{aligned}&\|\boldsymbol{g}_{t+1} - \kappa \nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}_t; \mathbf{D}) + \kappa \nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}_t; \mathbf{D}) - \kappa \mathbf{A}(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*)\|^2 \\ &\leq \|\boldsymbol{g}_{t+1} - \kappa \nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}_t; \mathbf{D})\|^2 + \kappa^2 \|\nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}_t; \mathbf{D}) - \mathbf{A}(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*)\|^2 \\ &\leq \|\boldsymbol{g}_{t+1} - \kappa \nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}_t; \mathbf{D})\|^2 + \kappa^2 e^2_{\text{tay}} \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2,\end{aligned}$$

taking the expectation of both sides with respect to the subsampling distribution and using the fact that for all $t \geq T^*$, $\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2 < (r^*)^2$, we obtain

$$\mathop{\mathbb{E}}_{\text{Ber}(\kappa)} \left[\|\boldsymbol{g}_{t+1} - \kappa \mathbf{A}(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*)\|^2 \right] < e^2_{\text{approx}}. \tag{A3}$$

where $\text{Ber}(\kappa)$ is the Bernoulli distribution with $p = \kappa$, which is the subsampling distribution. □

C Noise-Aware Posterior Approximation

Deriving the closed form of $p(\boldsymbol{\phi}^*, \mathbf{A} \mid \mathbf{T})$ might not always be feasible so we obtain the following approximations. The first Laplace's approximation. Since the Hessian is positive-definite, we want $\mathbf{A}$ to be positive-definite as well, so we need a bijective differentiable function (diffeomorphism) F, such that $F(\mathbf{V}) = \mathbf{A}$ where $\mathbf{V}$ is an unconstrained version of $\mathbf{A}$. Sometimes, F is referred to as a conditioning transformation (Blei et al., 2017). Therefore,

$$p(\mathbf{V}, \boldsymbol{\phi}^* \mid \mathcal{T}) = p(F(\mathbf{V}), \boldsymbol{\phi}^* \mid \mathcal{T}) |J_F(\mathbf{V})|,$$

where J_F is the Jacobian of F. To obtain Laplace's approximation first we need the MAP (Maximum a posteriori) estimate for both $\mathbf{V}$ and $\boldsymbol{\phi}^*$,

$$\left(\widehat{\mathbf{V}}_{\text{map}}, \widehat{\boldsymbol{\phi}}^*_{\text{map}}\right) = \underset{(\boldsymbol{\phi}^*, \mathbf{V})}{\text{argmax}} \log p(\mathbf{V}, \boldsymbol{\phi}^* \mid \mathcal{T}). \tag{A4}$$

Let $p_{\text{gauss}}(\mathbf{V}, \boldsymbol{\phi}^*)$ be the distribution

$$\mathcal{N}\left(\left(\text{vec}\left(\widehat{\mathbf{V}}_{\text{map}}\right), \widehat{\boldsymbol{\phi}}^*_{\text{map}}\right), \Sigma_{\text{lap}}\right), \tag{A5}$$

where $\text{vec}(\cdot)$ is the vector representation of a matrix. We can easily marginalize out $\widehat{\boldsymbol{\phi}}^*_{\text{map}}$ from $p_{\text{gauss}}(\mathbf{V}, \boldsymbol{\phi}^*)$ to approximate (21) since it is a Guassian distribution. We then use this marginalized distribution to approximate (23) and obtain a new distribution that we will denote $\widetilde{p}_{\text{lap}}(\boldsymbol{\theta} \mid \mathcal{T})$. The second method to approximate (23) is by using any Markov Chain Monte Carlo (MCMC) method to approximately sample from the post-processing model,

$$(\mathbf{A}_i, \boldsymbol{\phi}^*_i) \sim p(\mathbf{A}, \boldsymbol{\phi}^* \mid \mathcal{T}), \ 1 \leq i \leq N_{\text{mc}}. \tag{A6}$$

We can use these samples to approximate (23) using the following mixture model,

$$\widetilde{p}(\boldsymbol{\theta} \mid \mathcal{T}) \approx \frac{1}{N_{\text{mc}}} \sum_{i=1}^{N_{\text{mc}}} q_{\text{VI}}(\boldsymbol{\theta}; \boldsymbol{\phi}^*_i) \equiv_{\text{def}} \widetilde{p}_{\text{mc}}(\boldsymbol{\theta} \mid \mathcal{T}). \tag{A7}$$

D Estimation of the Hessian Matrix and the Influence of the Learning Rate

From our experiments, we noticed that the uncertainty estimates of the approximate noise-aware posterior are sensitive to the estimation of the Hessian matrix $\mathbf{A}$. Through observation, we found that the accuracy of the estimation of $\mathbf{A}$ is highly correlated with the choice of the constant learning rate λ. This motivated us to conduct further analysis which we present in this section.

There is also an equivalent model to (20) which we call parameter-based model of the trace that is useful in this analysis,

$$\boldsymbol{\phi}_{t+1} \mid \boldsymbol{\phi}_t, \mathbf{A}, \boldsymbol{\phi}^* \sim \mathcal{N}\left(\boldsymbol{\phi}_t - \lambda\kappa\mathbf{A}\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\right), \lambda^2\Sigma_{total}\right), \tag{A8}$$

where we assume that $\Sigma_{\text{sub}} = \mathbf{0}$.

For Assumption 2 to hold, the part of the trace $\mathcal{T}$ that is stationary around the optimum $\boldsymbol{\phi}^*$ (after $t = T^*$) needs to be close enough to $\boldsymbol{\phi}^*$. On the other hand, the trace should have enough variability around $\boldsymbol{\phi}^*$ for the inference to be accurate. Intuitively speaking, we want to sample parameters $\boldsymbol{\phi}_t$ and gradients $\boldsymbol{g}_t$ around $\boldsymbol{\phi}^*$ to use that information for the post-processing model; and in particular, to estimate the Hessian that describes the curvature of the loss function around $\boldsymbol{\phi}^*$. For a rigorous analysis, we present the following Theorem with the full proof in Supplement E.

Theorem 2. *Under assumptions 1, 2, 3, and 4 with the additional assumptions that $\mathbf{A}$ is a diagonal matrix, i.e. $\mathbf{A} = \mathrm{diag}(a_1, \ldots, a_d)$, and that $\boldsymbol{\phi}^*$ is known, if we set the prior for $\mathbf{A}$ to any uniform distribution $a_i \sim U(\cdot, \cdot)$, then the MAP estimate of $\mathbf{A}$, $\widehat{\mathbf{A}}_{map} = \mathrm{diag}(\hat{a}_1, \ldots, \hat{a}_d)$ has to satisfy the following inequality,*

$$r^* \times \max_{1\leq i\leq d} Var\left[\hat{a}_i\right] \geq \frac{1}{\kappa}\sqrt{\frac{d}{T - T^*}}\sigma_{DP}C. \tag{A9}$$

The analysis in Theorem 2 can be done without the assumption that $\mathbf{A}$ is diagonal; however, this assumption makes the analysis easier and the relation more clear. We also perform our experiments with a diagonal $\mathbf{A}$ matrix so it makes sense to analyze this particular case. Inequality (A9) entails that there is a trade-off between $\max_i \mathrm{Var}\left[\hat{a}_i\right]$ and r^* (the radius of the open ball that contains the iterates in $\mathcal{T}^*$), and this puts a limit on how well we can estimate the Hessian with a non-informative prior even if we assume perfect knowledge about $\boldsymbol{\phi}^*$.

Also, according to (A8), other than Σ_{DP}, it is clear that the value of λ affects the variance of the trace around $\boldsymbol{\phi}^*$. We show the relation between λ and $\mathbb{E}\left[\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2\right]$ through the Ornstein–Uhlenbeck (OU) SDE which is a known tool for analyzing variants of SGD (Mandt et al., 2016; Li et al., 2019, 2021). This is usually done by modeling the noise in the discrete-time SGD process through the Wiener process and then approximating it by a continuous-time process using the Itô integral (Rogers and Williams, 2000).

Theorem 3. *Under assumptions 1, 2, 3, and 4, if the DP-SGD process can be well-approximated by the OU SDE, then the stationary mean of $\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2$ is proportional to $\lambda\sigma_{DP}^2C^2$. Moreover, for the special case when $\mathbf{A}$ is a non-singular diagonal matrix,*

$$\lambda < \frac{2\kappa(r^*)^2}{\sigma_{DP}^2C^2\mathbf{Tr}\left(\mathbf{A}^{-1}\right)}. \tag{A10}$$

What Theorem 3 entails is that

$$\lambda\sigma_{\text{DP}}^2C^2 \propto \mathbb{E}\left[\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2\right] < (r^*)^2,$$

so if we make λ large enough, then $\mathbb{E}\left[\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2\right]$ will increase proportional to λ which will invalidate Assumption 2 (the second-order Taylor approximation of the loss function) by contraposition. The full proof of Theorem 3 and a further discussion is found in Appendix F.

To find a good learning rate, we analyze the convergence of DP-SGD with some additional Assumptions and then find the learning rate that optimizes the convergence bound given in the following theorem. Also, we want to write the optimization problem as a minimization problem. If the optimization problem is initially a maximization problem, then we change the direction of optimization by re-defining $\mathcal{L}$ as $-\mathcal{L}$.

Theorem 4. *Under assumptions 3 and 4, further assume that $\mathcal{L}$ has a lower bound, namely M, and the gradient $\nabla_{\boldsymbol{\phi}}\mathcal{L}$ is Lipschitz continuous with Lipschitz constant H (so $\mathcal{L}$ is Lipschitz smooth), and that the per-examples* *losses are Lipschitz continuous, then the following inequality holds,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\phi}_t)\right\|^2\right] \leq \frac{\mathcal{L}\left(\boldsymbol{\phi}_0\right)-M}{T\lambda\kappa} + \frac{H}{2\kappa}\left(e_{sub}^2+\kappa^2G^2+\sigma_{DP}^2C^2d\right)\lambda. \tag{A11}$$

We obtain a heuristic based on finding λ that optimizes the right-hand side of (A11) by taking the derivative, setting it to zero, and then solving for λ from which we obtain a quantity proportional to the following,

$$\lambda_{\text{heur}} = \frac{\sqrt{2}\lambda_c}{\sigma_{\text{DP}}C\sqrt{Td}},$$

where $\lambda_c > 0$ is a hyper-parameter. See supplement G for the full proof and other details. For our experiments, we mostly use $\lambda_c = 1$ or $\lambda_c = \sqrt{\frac{d}{2}}$. This heuristic makes it easier to tune the learning-rate because it sets λ to the appropriate scale that is required by convergence results. We show experimentally through our experiments that this is a good choice for the learning rate.

E Proof of Theorem 2

As said before, an intuitive explanation why the variance of the trace around $\boldsymbol{\phi}^*$ affects the estimation of the hessian is that assuming no to little variability, i.e. $\boldsymbol{\phi}_t \approx \boldsymbol{\phi}^*$ for $t = T^*, \ldots, T$, then $\mathbf{A}\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\right) \approx \nabla\mathcal{L}\left(\boldsymbol{\phi}_t; \mathbf{D}\right)$ becomes $\mathbf{A}\left(\boldsymbol{\phi}^* - \boldsymbol{\phi}^*\right) \approx \mathcal{L}\left(\boldsymbol{\phi}^*; \mathbf{D}\right) = 0$. This implies that $\mathbf{A} \times 0 \approx 0$, and thus, in this case, it is impossible to estimate $\mathbf{A}$ from this equation. The following is the proof of Theorem 2.

Proof. The MAP estimate of $\mathbf{A}$ under a inform prior and with perfect knowledge about $\boldsymbol{\phi}^*$ is the same as the maximum likelihood estimate (MLE). Using the gradient-based model of the trace, and assuming that the subsampling noise Σ_{sub} is $\mathbf{0}$, the likelihood of $\mathbf{A}$ could be written as:

$$L\left(\mathbf{A}; \widetilde{\boldsymbol{g}}_{T^*+1}, \ldots, \widetilde{\boldsymbol{g}}_T, \boldsymbol{\phi}_{T^*}, \ldots, \boldsymbol{\phi}_{T-1}\right) \propto \prod_{t=T^*}^{T-1} p(\widetilde{\boldsymbol{g}}_{t+1} \mid \boldsymbol{\phi}_t, \mathbf{A}),$$

taking the logarithm of both sides:

$$\log L\left(\mathbf{A}; \widetilde{\boldsymbol{g}}_{T^*+1}, \ldots, \widetilde{\boldsymbol{g}}_T, \boldsymbol{\phi}_{T^*}, \ldots, \boldsymbol{\phi}_{T-1}\right) \underset{c}{=} \sum_{t=T^*}^{T-1} \log p\left(\widetilde{\boldsymbol{g}}_{t+1} \mid \boldsymbol{\phi}_t, \mathbf{A}\right),$$

where $\underset{c}{=}$ means equals with a constant difference. For each t, we have:

$$\log p\left(\widetilde{\boldsymbol{g}}_{t+1} \mid \boldsymbol{\phi}_t, \mathbf{A}\right) = -\frac{d}{2}\log 2\pi - \frac{1}{2}\,\log\det\left(\Sigma_{\text{DP}}\right) - \frac{1}{2}\left(\widetilde{\boldsymbol{g}}_{t+1} - \kappa\mathbf{A}(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*)\right)^\top \Sigma_{\text{DP}}^{-1}\left(\widetilde{\boldsymbol{g}}_{t+1} - \kappa\mathbf{A}(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*)\right),$$

therefore,

$$-\log L\left(\mathbf{A}\right) \underset{c}{=} \frac{1}{2}\sum_{t=T^*}^{T-1}\left(\widetilde{\boldsymbol{g}}_{t+1} - \kappa\mathbf{A}(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*)\right)^\top \Sigma_{\text{DP}}^{-1}\left(\widetilde{\boldsymbol{g}}_{t+1} - \kappa\mathbf{A}(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*)\right). \tag{A12}$$

If we want to find $\widehat{\mathbf{A}}$ that maximizes the likelihood $L(\mathbf{A})$, then we need to minimize $-\log L\left(\mathbf{A}\right)$ in (A12). Thus, we have to minimize the sum:

$$\sum_{t=T^*}^{T-1}\left\|\widetilde{\boldsymbol{g}}_{t+1} - \kappa\mathbf{A}(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*)\right\|^2. \tag{A13}$$

Let $\mathbf{a} = (a_1, \ldots, a_d)$, then we can also re-write the sum (A13) as:

$$\sum_{t=T^*}^{T-1}\left\|\widetilde{\boldsymbol{g}}_{t+1} - \kappa\mathbf{a}\odot(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*)\right\|^2, \tag{A14}$$

where the operation $\odot$ is element-wise multiplication. Hence, we want to find $\hat{\mathbf{a}}$ such that for all $1 \leq i \leq d$:

$$\frac{\partial}{\partial \hat{a}_i} \sum_{t=T^*}^{T-1} \left(\widetilde{\boldsymbol{g}}_{t+1}^{(i)} - \kappa \hat{a}_i \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right) \right)^2 = 2\kappa \sum_{t=T^*}^{T-1} \left(\kappa \hat{a}_i \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right) - \widetilde{\boldsymbol{g}}_{t+1}^{(i)} \right) \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right) = 0,$$

where $\boldsymbol{v}^{(i)}$ denotes the ith component of a vector $\boldsymbol{v}$, in other words,

$$\kappa \hat{a}_i \sum_{t=T^*}^{T-1} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)^2 = \sum_{t=T^*}^{T-1} \widetilde{\boldsymbol{g}}_{t+1}^{(i)} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right),$$

$$\hat{a}_i = \frac{\sum_{t=T^*}^{T-1} \widetilde{\boldsymbol{g}}_{t+1}^{(i)} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)}{\kappa \sum_{t=T^*}^{T-1} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)^2}. \tag{A15}$$

We are interested in the distribution of each $\hat{a}_i \mid \mathcal{T}$. The gradient-based model implies that:

$$\begin{aligned}
\widetilde{\boldsymbol{g}}_t^{(i)} \mid \mathcal{T} &\sim \mathcal{N}\left(\kappa a_i \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right), \sigma_{\mathrm{DP}}^2 C^2 \right), \\
\widetilde{\boldsymbol{g}}_t^{(i)} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right) \mid \mathcal{T} &\sim \mathcal{N}\left(\kappa a_i \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)^2, \sigma_{\mathrm{DP}}^2 C^2 \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)^2 \right), \\
\sum_{t=T^*}^{T-1} \widetilde{\boldsymbol{g}}_t^{(i)} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right) \mid \mathcal{T} &\sim \mathcal{N}\left(\kappa a_i \sum_{t=T^*}^{T-1} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)^2, \sigma_{\mathrm{DP}}^2 C^2 \sum_{t=T^*}^{T-1} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)^2 \right), \qquad \text{(A16)} \\
\hat{a}_i \mid \mathcal{T} &\sim \mathcal{N}\left(a_i, \frac{\sigma_{\mathrm{DP}}^2 C^2}{\kappa^2 \sum_{t=T^*}^{T-1} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)^2} \right).
\end{aligned}$$

From this, we can clearly see that the variance of $\hat{a}_i \mid \mathcal{T}$ is inversely proportional to the sum:

$$\sum_{t=T^*}^{T-1} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)^2.$$

Let $e_{\mathrm{std}} = \max_{1 \leq i \leq d} \mathrm{Var}\left[\hat{a}_i \right]$, then for all i,

$$\frac{\sigma_{\mathrm{DP}}^2 C^2}{\kappa^2 \sum_{t=T^*}^{T-1} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)^2} \leq e_{\mathrm{std}}^2,$$

re-arranging, we obtain

$$\sum_{t=T^*}^{T-1} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)^2 \geq \frac{\sigma_{\mathrm{DP}}^2 C^2}{\kappa^2 e_{\mathrm{std}}^2}. \tag{A17}$$

Moreover,

$$\sum_{t=T^*}^{T-1} \left\| \boldsymbol{\phi}_t - \boldsymbol{\phi}^* \right\|^2 = \sum_{t=T^*}^{T-1} \sum_{i=1}^{d} \left(\boldsymbol{\phi}_t^{(i)} - \boldsymbol{\phi}^{*(i)} \right)^2 \geq \frac{d \sigma_{\mathrm{DP}}^2 C^2}{\kappa^2 e_{\mathrm{std}}^2}. \tag{A18}$$

On the other hand, each $\left\| \boldsymbol{\phi}_t - \boldsymbol{\phi}^* \right\|$ cannot be too large for the second-order Taylor approximation of $\mathcal{L}\left(\boldsymbol{\phi}; \mathbf{D} \right)$ to be valid. We know that $\mathbf{A}$ is the hessian of $\mathcal{L}\left(\boldsymbol{\phi}; \mathbf{D} \right)$ at $\boldsymbol{\phi}^*$, so:

$$\lim_{\boldsymbol{\phi} \to \boldsymbol{\phi}^*} \frac{\left\| \nabla_{\boldsymbol{\phi}} \mathcal{L}\left(\boldsymbol{\phi} \right) - \nabla_{\boldsymbol{\phi}} \mathcal{L}\left(\boldsymbol{\phi}^* \right) - \mathbf{A}(\boldsymbol{\phi} - \boldsymbol{\phi}^*) \right\|}{\left\| \boldsymbol{\phi} - \boldsymbol{\phi}^* \right\|} = 0,$$

since $\nabla_{\boldsymbol{\phi}} \mathcal{L}\left(\boldsymbol{\phi}^*; \mathbf{D} \right) = 0$,

$$\lim_{\boldsymbol{\phi} \to \boldsymbol{\phi}^*} \frac{\left\| \nabla_{\boldsymbol{\phi}} \mathcal{L}\left(\boldsymbol{\phi} \right) - \mathbf{A}(\boldsymbol{\phi} - \boldsymbol{\phi}^*) \right\|}{\left\| \boldsymbol{\phi} - \boldsymbol{\phi}^* \right\|} = 0.$$

In other words, for all $e_{\text{tay}} > 0$, there exists $e_{\text{trace}} > 0$ such that

$$0 < \|\boldsymbol{\phi} - \boldsymbol{\phi}^*\| < e_{\text{trace}} \Longrightarrow \frac{\|\nabla \mathcal{L}(\boldsymbol{\phi}) - \mathbf{A}(\boldsymbol{\phi} - \boldsymbol{\phi}^*)\|}{\|\boldsymbol{\phi} - \boldsymbol{\phi}^*\|} < e_{\text{tay}}.$$

Let e_{tay} be as in Assumption 2, then $e_{\text{trace}} = r^*$, and

$$0 < \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\| < r^*, \tag{A19}$$

for all $t = T^*, \ldots, T-1$. This implies that:

$$0 < \sum_{t=T^*}^{T-1} \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2 < (T - T^*)(r^*)^2. \tag{A20}$$

Combining (A18) and (A20), we obtain:

$$\frac{d\sigma_{dp}^2 C^2}{\kappa^2 e_{\text{std}}^2} \le \sum_{t=T^*}^{T-1} \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2 < (T - T^*)(r^*)^2. \tag{A21}$$

Finally, (A21) implies that:

$$e_{\text{std}} \times r^* \ge \frac{1}{\kappa}\sqrt{\frac{d}{T - T^*}}\sigma_{\text{DP}} C. \tag{A22}$$

□

F Proof of Theorem 3

For the second-order Taylor approximation to be accurate, certain values of the learning rate are valid. To show this, we will first approximate the discrete DP-SGD process using a stochastic differential equation (SDE). We provide the following proof for Theorem 3.

Proof. Let $\mathbf{X}^{\frac{1}{2}}$ denote the square root of a positive semidefinite matrix $\mathbf{X}$. According to the parameter-based model, assuming $\Sigma_{\text{sub}} = \mathbf{0}$,

$$\boldsymbol{\phi}_{t+1} = \boldsymbol{\phi}_t - \lambda\kappa\mathbf{A}\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\right) + \lambda\Sigma_{\text{DP}}^{\frac{1}{2}}\boldsymbol{\eta}_{t+1}, \tag{A23}$$

where $\boldsymbol{\eta}_{t+1} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_d\right)$, and $\boldsymbol{\eta}_0 = \mathbf{0}$. The random variable $\boldsymbol{\eta}_t$ can be written using the Wiener process $\mathbf{W}_t$:

$$\boldsymbol{\eta}_{t+1} = \mathbf{W}_{t+1} - \mathbf{W}_t \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_d\right),$$

this way, we can re-write equation (A23) as:

$$\Delta\boldsymbol{\phi}_{t+1} = -\lambda\kappa\mathbf{A}\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\right) + \lambda\Sigma_{\text{DP}}^{\frac{1}{2}}\Delta\mathbf{W}_{t+1}. \tag{A24}$$

In continuous time, we can approximate (A24) as:

$$\mathbf{d}\boldsymbol{\phi}_t = -\lambda\kappa\mathbf{A}\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\right)\mathbf{d}t + \lambda\Sigma_{\text{DP}}^{\frac{1}{2}}\mathbf{d}\mathbf{W}_t, \tag{A25}$$

This SDE is a special case of the multi-dimensional Ornstein-Uhlenbeck (OU) process. It is equivalent to the following integral form:

$$\boldsymbol{\phi}_t = \boldsymbol{\phi}_0 - \int_0^t \lambda\kappa\mathbf{A}\left(\boldsymbol{\phi}_s - \boldsymbol{\phi}^*\right)\mathbf{d}s + \int_0^t \lambda\Sigma_{\text{DP}}^{\frac{1}{2}}\mathbf{d}\mathbf{W}_s, \tag{A26}$$

To write equation (A25) in the more common form, define $\boldsymbol{\psi}_t = (\boldsymbol{\phi}_t - \boldsymbol{\phi}^*)$, then

$$\mathbf{d}\boldsymbol{\psi}_t = -\lambda\kappa\mathbf{A}\boldsymbol{\psi}_t\mathbf{d}t + \lambda\Sigma_{\text{DP}}^{\frac{1}{2}}\mathbf{d}\mathbf{W}_t. \tag{A27}$$

from the fact that

$$\boldsymbol{\psi}_t - \boldsymbol{\psi}_0 = (\boldsymbol{\phi}_t - \boldsymbol{\phi}^*) - (\boldsymbol{\phi}_0 - \boldsymbol{\phi}^*) = \boldsymbol{\phi}_t - \boldsymbol{\phi}_0,$$

The solution of equation (A27) is given by,

$$\boldsymbol{\psi}_t = \boldsymbol{e}^{-\lambda\kappa\mathbf{A}t}\boldsymbol{\psi}_0 + \int_0^t \lambda\Sigma_{\mathrm{DP}}^{\frac{1}{2}}\boldsymbol{e}^{-\lambda\kappa\mathbf{A}(t-s)}\mathbf{dW}_s \tag{A28}$$

$$\iff \boldsymbol{\phi}_t = \boldsymbol{\phi}^* + \boldsymbol{e}^{-\lambda\kappa\mathbf{A}t}\left(\boldsymbol{\phi}_0 - \boldsymbol{\phi}^*\right) + \int_0^t \lambda\Sigma_{\mathrm{DP}}^{\frac{1}{2}}\boldsymbol{e}^{-\lambda\kappa\mathbf{A}(t-s)}\mathbf{dW}_s, \tag{A29}$$

where $\boldsymbol{e}^{-\lambda\kappa\mathbf{A}t}$ is the matrix exponent of $-\lambda\kappa\mathbf{A}t$,

$$\boldsymbol{e}^{-\lambda\kappa\mathbf{A}t} = \sum_{k=0}^{\infty}\frac{(-1)^k\lambda^k t^k}{k}\mathbf{A}^k = \mathbf{I}_d - \lambda\kappa t\mathbf{A} + \frac{\lambda^2 t^2}{2}\kappa^2\mathbf{A}^2 + \dots.$$

Note that

$$\mathbb{E}\left[\boldsymbol{\phi}_t\right] = \boldsymbol{\phi}^* + \boldsymbol{e}^{-\lambda\kappa\mathbf{A}t}\left(\boldsymbol{\phi}_0 - \boldsymbol{\phi}^*\right).$$

We are interested in the stationary distribution of $\boldsymbol{\phi}_t$, so from (A28, A29):

$$\mathbb{E}\left[\boldsymbol{\phi}_t\right] \underset{t\to\infty}{\longrightarrow} \boldsymbol{\phi}^*,$$

$$\mathrm{Cov}\left[\boldsymbol{\phi}_t\right] \underset{t\to\infty}{\longrightarrow} \Sigma_{\mathrm{sde}},$$

where Σ_{sde} is the solution of the following Lyapunov equation:

$$\begin{aligned} \lambda\kappa\mathbf{A}\Sigma_{\mathrm{sde}} + \lambda\kappa\Sigma_{\mathrm{sde}}\mathbf{A}^\top &= \lambda^2\sigma_{\mathrm{DP}}^2 C^2\mathbf{I}_d \\ \iff \mathbf{A}\Sigma_{\mathrm{sde}} + \Sigma_{\mathrm{sde}}\mathbf{A} &= \frac{\lambda}{\kappa}\sigma_{\mathrm{DP}}^2 C^2\mathbf{I}_d. \end{aligned} \tag{A30}$$

Hence, given that $\mathbf{A}$ is constant with respect to λ (since it only depends on the $\mathcal{L}$, $\mathbf{D}$, and $\boldsymbol{\phi}^*$), then Σ_{sde} is proportional to λ. Now inequality (A19), implies that:

$$(r^*)^2 > \mathbb{E}\left[\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2\right] = \mathrm{Tr}\left(\Sigma_{\mathrm{sde}}\right) \propto \frac{\lambda}{\kappa}\sigma_{\mathrm{DP}}^2 C^2. \tag{A31}$$

For the special case when $\mathbf{A}$ is diagonal, then Σ_{sde} is a diagonal matrix since it has to satisfy (A30) and the right hand side is a diagonal matrix.

If $\mathbf{A} = \mathrm{diag}\left(a_1, a_2, \dots, a_d\right)$, then $\Sigma_{\mathrm{sde}} = \mathrm{diag}\left(\sigma_{\mathrm{sde}_1}^2, \sigma_{\mathrm{sde}_2}^2, \dots, \sigma_{\mathrm{sde}_d}^2\right)$, and (A30) implies that:

$$\sigma_{\mathrm{sde}_k}^2 = \frac{\lambda\sigma_{\mathrm{DP}}^2 C^2}{2a_k\kappa}$$

or

$$\sigma_{\mathrm{sde}_k} = \sqrt{\frac{\lambda}{2a_k\kappa}}\sigma_{\mathrm{DP}}C.$$

Hence, $\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2$ has the following stationary mean:

$$\sum_{k=1}^{d}\frac{\lambda\sigma_{\mathrm{DP}}^2 C^2}{2a_k\kappa} = \frac{\lambda\sigma_{\mathrm{DP}}^2 C^2}{2\kappa}\sum_{k=1}^{d}\frac{1}{a_k} = \frac{\lambda\sigma_{\mathrm{DP}}^2 C^2}{2\kappa}\mathbf{Tr}\left(\mathbf{A}^{-1}\right),$$

thus,

$$\mathbb{E}\left[\|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\|^2\right] = \frac{\lambda\sigma_{\mathrm{DP}}^2 C^2}{2\kappa}\mathbf{Tr}\left(\mathbf{A}^{-1}\right). \tag{A32}$$

Now inequality (A19), implies that:

$$(r^*)^2 > \frac{\lambda\sigma_{\mathrm{DP}}^2 C^2}{2\kappa}\mathbf{Tr}\left(\mathbf{A}^{-1}\right).$$

Therefore,

$$\lambda < \frac{2\kappa(r^*)^2}{\sigma_{\mathrm{DP}}^2 C^2\mathbf{Tr}\left(\mathbf{A}^{-1}\right)}. \tag{A33}$$

□

Inequality (A33) nor inequality (A31) help us choose the learning rate since the relation between them and inequality (A19) is not a logical equivalence and we typically don't have any information about $\mathbf{A}$ in advance. The reason why we approximated the DP-SGD process using the OU SDE is that it is easier to work with. That said, it is possible to do a similar analysis with the discrete process but with some additional assumptions. To see this, observe that (A23) can be written as (when $t = T - 1$):

$$\boldsymbol{\phi}_T = (\mathbf{I}_d - \lambda\kappa\mathbf{A})\,\boldsymbol{\phi}_{T-1} + \lambda\kappa\mathbf{A}\boldsymbol{\phi}^* + \lambda\Sigma_{\mathrm{DP}}^{\frac{1}{2}}\boldsymbol{\eta}_T,$$

and that this could be expanded as:

$$\boldsymbol{\phi}_T = (\mathbf{I}_d - \lambda\kappa\mathbf{A})^T\boldsymbol{\phi}_0 + \lambda\kappa\sum_{t=0}^{T-1}(\mathbf{I}_d - \lambda\kappa\mathbf{A})^t\,\mathbf{A}\boldsymbol{\phi}^* + \lambda\sum_{t=0}^{T-1}(\mathbf{I}_d - \lambda\kappa\mathbf{A})^t\,\Sigma_{\mathrm{DP}}^{\frac{1}{2}}\boldsymbol{\eta}_{t+1}.$$

Since $\mathbf{A}$ is a symmetric matrix and thus is diagonalizable, there is a diagonal matrix $\mathbf{L}$ and an orthogonal matrix $\mathbf{U}$ such that $\mathbf{U}\mathbf{U}^\top = \mathbf{I}_d$ and $\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{U}^\top$, then $\mathbf{I}_d - \lambda\mathbf{L}$ is also a diagonal matrix, Let l be the absolute value of the largest entry in the diagonal of $\mathbf{L}$, if $\lambda < \frac{1}{l\kappa}$, then

$$(\mathbf{I}_d - \lambda\kappa\mathbf{A})^T\boldsymbol{\phi}_0 = \left(\mathbf{I}_d - \lambda\kappa\mathbf{U}\mathbf{L}\mathbf{U}^\top\right)^T\boldsymbol{\phi}_0 = \left(\mathbf{U}\left(\mathbf{I}_d - \lambda\kappa\mathbf{L}\right)\mathbf{U}^\top\right)^T\boldsymbol{\phi}_0 = \mathbf{U}\left(\mathbf{I}_d - \lambda\kappa\mathbf{L}\right)^T\mathbf{U}^\top\boldsymbol{\phi}_0 \overset{T\to\infty}{\longrightarrow} \mathbf{0}. \tag{A34}$$

Further assume that the diagonal entries of $\mathbf{L}$ are all non-zero so that the matrix $\mathbf{A}$ is invertible, let $\mathbf{Q} = \mathbf{I}_d - \lambda\kappa\mathbf{A}$ then, applying the geometric series sum for matrices,

$$\begin{aligned}
\lambda\kappa\sum_{t=0}^{T-1}\mathbf{Q}^t\mathbf{A}\boldsymbol{\phi}^* &= \lambda\kappa\left(\sum_{t=0}^{T-1}\mathbf{Q}^t\right)\mathbf{A}\boldsymbol{\phi}^* \\
&= \lambda\kappa\left(\left(\mathbf{I}_d - \mathbf{Q}^T\right)\left(\mathbf{I}_d - \mathbf{I}_d + \lambda\kappa\mathbf{A}\right)^{-1}\right)\mathbf{A}\boldsymbol{\phi}^* \\
&= \lambda\kappa\left(\left(\mathbf{I}_d - \mathbf{Q}^T\right)\left(\lambda\kappa\mathbf{A}\right)^{-1}\right)\mathbf{A}\boldsymbol{\phi}^* \\
&= \frac{\lambda\kappa}{\lambda\kappa}\left(\mathbf{I}_d - \mathbf{Q}^T\right)\boldsymbol{\phi}^* \\
&= \left(\mathbf{I}_d - \left(\mathbf{I}_d - \lambda\kappa\mathbf{U}\mathbf{L}\mathbf{U}^\top\right)^T\right)\boldsymbol{\phi}^* \\
&= \left(\mathbf{I}_d - \left(\mathbf{U}\left(\mathbf{I}_d - \lambda\kappa\mathbf{L}\right)\mathbf{U}^\top\right)^T\right)\boldsymbol{\phi}^* \\
&= \left(\mathbf{I}_d - \mathbf{U}\left(\mathbf{I}_d - \lambda\kappa\mathbf{L}\right)^T\mathbf{U}^\top\right)\boldsymbol{\phi}^* \\
&= \mathbf{U}\left(\mathbf{I}_d - \left(\mathbf{I}_d - \lambda\kappa\mathbf{L}\right)^T\right)\mathbf{U}^\top\boldsymbol{\phi}^* \overset{T\to\infty}{\longrightarrow} \boldsymbol{\phi}^*.
\end{aligned} \tag{A35}$$

Since Σ_{DP} is $\sigma_{\mathrm{DP}}^2 C^2\mathbf{I}_d$, it commutes with $\mathbf{Q}^t$, thus from the fact that the noises $\boldsymbol{\eta}_t$ are i.i.d.,

$$\begin{aligned}
\lambda\sum_{t=0}^{T-1}\mathbf{Q}^t\Sigma_{\mathrm{DP}}^{\frac{1}{2}}\boldsymbol{\eta}_{t+1}\,\Big|\,\boldsymbol{\phi}_0 &\sim \mathcal{N}\left(\mathbf{0}, \lambda^2\sigma_{\mathrm{DP}}^2 C^2\sum_{t=0}^{T-1}\mathbf{Q}^{2t}\right) \\
&\sim \mathcal{N}\left(\mathbf{0}, \lambda^2\sigma_{\mathrm{DP}}^2 C^2\left(\mathbf{I}_d - \mathbf{Q}^{2T}\right)\left(\mathbf{I}_d - \mathbf{Q}^2\right)^{-1}\right).
\end{aligned} \tag{A36}$$

Hence, from (A34, (A35), and (A36),

$$\mathbb{E}\left[\boldsymbol{\phi}_T\right] \overset{T\to\infty}{\longrightarrow} \boldsymbol{\phi}^*,$$

$$\mathrm{Cov}\left[\boldsymbol{\phi}_T\right] \overset{T\to\infty}{\longrightarrow} \lambda^2\sigma_{\mathrm{DP}}^2 C^2\left(\mathbf{I}_d - \mathbf{Q}^2\right)^{-1}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\phi}_T-\boldsymbol{\phi}^*\|^2\right] &\overset{T\to\infty}{\longrightarrow} \lambda^2\sigma_{\mathrm{DP}}^2C^2\mathrm{Tr}\left(\mathbf{U}^\top\left(\mathbf{I}_d-\left(\mathbf{I}_d-\lambda\kappa\mathbf{L}\right)^2\right)^{-1}\mathbf{U}\right)\\
&=\lambda^2\sigma_{\mathrm{DP}}^2C^2\mathrm{Tr}\left(\left(\mathbf{I}_d-\left(\mathbf{I}_d-\lambda\kappa\mathbf{L}\right)^2\right)^{-1}\right)\\
&=\lambda^2\sigma_{\mathrm{DP}}^2C^2\sum_{i=1}^{d}\frac{1}{1-(1-\lambda\kappa l_i)^2}\\
&=\lambda^2\sigma_{\mathrm{DP}}^2C^2\sum_{i=1}^{d}\frac{1}{1-1+2\lambda\kappa l_i-\lambda^2\kappa^2 l_i^2}\\
&=\frac{\lambda\sigma_{\mathrm{DP}}^2C^2}{\kappa}\sum_{i=1}^{d}\frac{1}{2l_i-\lambda l_i^2}
\end{aligned}$$

where l_i $(1\le i\le d)$ are the diagonal entries of $\mathbf{L}$.

G Proof of Theorem 4

We state the additional assumptions of Theorem 4 again,

Assumption 5. *$\mathcal{L}(\boldsymbol{\phi};\mathbf{D})$ is bounded below, i.e. $M=\inf_{\boldsymbol{\phi}}\mathcal{L}(\boldsymbol{\phi};\mathbf{D})$ exists.*

Assumption 6. *The per-example losses are Lipschitz continuous $\ell(\boldsymbol{\phi};\boldsymbol{x}_i)$ in $\boldsymbol{\phi}$. Further, assume that the clipping threshold is set to*

$$C=\max_i\sup_{\boldsymbol{\phi}}\|\nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi};\boldsymbol{x}_i)\|.$$

Assumption 7. *The loss $\mathcal{L}$ is Lipschitz smooth, i.e. that the gradients $\nabla\mathcal{L}(\boldsymbol{\phi})$ are Lipschitz continuous with Lipschitz constant H.*

Since the per-example losses are assumed to be Lipschitz continuous, then

$$\|\nabla_{\boldsymbol{\phi}}\mathcal{L}(\boldsymbol{\phi};\mathbf{D})\|\le\sum_{i=1}^{N}\|\nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi};\boldsymbol{x}_i)\|\le NC,$$

so $\mathcal{L}$ is Lipschitz continuous, also

$$\sup_{\boldsymbol{\phi}}\|\nabla_{\boldsymbol{\phi}}\mathcal{L}(\boldsymbol{\phi};\mathbf{D})\|=G\le NC.$$

We will denote $\mathcal{L}(\boldsymbol{\phi};\mathbf{D})$ as $\mathcal{L}(\boldsymbol{\phi})$ for short. Also, we want to write the optimization problem as a minimization problem. If the optimization problem is initially a maximization problem, then we change the direction of optimization by re-defining $\mathcal{L}$ as $-\mathcal{L}$. The following lemma is a well-known result; however we prove it within this context.

Lemma 1. *If $\mathcal{L}(\boldsymbol{\phi})$ is Lipschitz smooth, then for any $\boldsymbol{\phi}_1,\boldsymbol{\phi}_2\in\boldsymbol{\phi}$, the following inequality holds:*

$$\mathcal{L}(\boldsymbol{\phi}_2)\le\mathcal{L}(\boldsymbol{\phi}_1)+\nabla\mathcal{L}(\boldsymbol{\phi}_1)^\top(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1)+\frac{H}{2}\|\boldsymbol{\phi}_2-\boldsymbol{\phi}_1\|^2.$$

Proof. Define $\gamma(t)=\mathcal{L}(\boldsymbol{\phi}_1+t(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1))$. By applying the fundamental theorem of calculus,

$$\mathcal{L}(\boldsymbol{\phi}_2)-\mathcal{L}(\boldsymbol{\phi}_1)=\int_{t=0}^{1}\gamma'(t)\mathrm{d}t=\int_{t=0}^{1}\nabla\mathcal{L}(\boldsymbol{\phi}_1+t(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1))^\top(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1)\mathrm{d}t.$$

Rearranging the terms,

$$\mathcal{L}(\boldsymbol{\phi}_2)=\mathcal{L}(\boldsymbol{\phi}_1)+\int_{t=0}^{1}\nabla\mathcal{L}(\boldsymbol{\phi}_1+t(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1))^\top(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1)\mathrm{d}t,$$

then adding and subtracting $\nabla \mathcal{L}\left(\boldsymbol{\phi}_1\right)^\top\left(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1\right)$ to the right side we obtain

$$\mathcal{L}\left(\boldsymbol{\phi}_2\right)=\mathcal{L}\left(\boldsymbol{\phi}_1\right)+\nabla \mathcal{L}\left(\boldsymbol{\phi}_1\right)^\top\left(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1\right)+\int_{t=0}^{1}\left(\nabla \mathcal{L}\left(\boldsymbol{\phi}_1+t(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1)\right)-\nabla \mathcal{L}(\boldsymbol{\phi}_1)\right)^\top\left(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1\right)\mathrm{d}t. \tag{A37}$$

Applying the Cauchy-Schwarz inequality,

$$\left(\nabla \mathcal{L}\left(\boldsymbol{\phi}_1+t(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1)\right)-\nabla \mathcal{L}(\boldsymbol{\phi}_1)\right)^\top\left(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1\right)\le \|\nabla \mathcal{L}\left(\boldsymbol{\phi}_1+t(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1)\right)-\nabla \mathcal{L}(\boldsymbol{\phi}_1)\|\,\|\boldsymbol{\phi}_2-\boldsymbol{\phi}_1\|. \tag{A38}$$

From the Lipschitz smoothness condition,

$$\|\nabla \mathcal{L}\left(\boldsymbol{\phi}_1+t(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1)\right)-\nabla \mathcal{L}(\boldsymbol{\phi}_1)\|\le Ht\|\boldsymbol{\phi}_2-\boldsymbol{\phi}_1\|. \tag{A39}$$

From (A37), (A38), and (A39),

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\phi}_2\right)&\le \mathcal{L}\left(\boldsymbol{\phi}_1\right)+\nabla \mathcal{L}(\boldsymbol{\phi}_1)^\top(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1)+\int_{t=0}^{1}Ht\|\boldsymbol{\phi}_2-\boldsymbol{\phi}_1\|^2\mathrm{d}t\\ &\le \mathcal{L}\left(\boldsymbol{\phi}_1\right)+\nabla \mathcal{L}(\boldsymbol{\phi}_1)^\top(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1)+\frac{H}{2}\|\boldsymbol{\phi}_2-\boldsymbol{\phi}_1\|^2.\end{aligned}$$

□

Now the following is the proof of Theorem 4:

Proof. From lemma [1],

$$\mathcal{L}(\boldsymbol{\phi}_{t+1})-\mathcal{L}(\boldsymbol{\phi}_t)\le \nabla \mathcal{L}(\boldsymbol{\phi}_t)^\top(\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t)+\frac{H}{2}\left\|\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t\right\|^2. \tag{A40}$$

Since for any t, we have:

$$\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t=-\lambda\left[\boldsymbol{g}_{t+1}+\sigma_{\mathrm{DP}}C\boldsymbol{\eta}_{t+1}\right],$$

then we can write:

$$\begin{aligned}\nabla \mathcal{L}(\boldsymbol{\phi}_t)^\top(\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t)&=-\lambda\nabla \mathcal{L}(\boldsymbol{\phi}_t)^\top\left[\boldsymbol{g}_{t+1}+\sigma_{\mathrm{DP}}C\boldsymbol{\eta}_{t+1}\right]\\ &=-\lambda\nabla \mathcal{L}(\boldsymbol{\phi}_t)^\top\boldsymbol{g}_{t+1}-\lambda\sigma_{\mathrm{DP}}C\nabla \mathcal{L}(\boldsymbol{\phi}_t)^\top\boldsymbol{\eta}_{t+1}.\end{aligned} \tag{A41}$$

Observe that $\nabla \mathcal{L}(\boldsymbol{\phi}_t)^\top\boldsymbol{\eta}_{t+1}\sim\mathcal{N}\left(\mathbf{0},\nabla \mathcal{L}(\boldsymbol{\phi}_t)^\top\mathbf{I}_d\nabla \mathcal{L}(\boldsymbol{\phi}_t)\right)\equiv\mathcal{N}\left(\mathbf{0},\left\|\nabla \mathcal{L}(\boldsymbol{\phi}_t)\right\|^2\mathbf{I}_d\right)$, and $\mathbb{E}\left[\boldsymbol{g}_{t+1}\mid\boldsymbol{\phi}_t\right]=\kappa\nabla \mathcal{L}(\boldsymbol{\phi}_t)$, and hence,

$$\mathbb{E}\left[\nabla \mathcal{L}(\boldsymbol{\phi}_t)^\top(\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t)\mid\boldsymbol{\phi}_t\right]=-\lambda\kappa\left\|\nabla \mathcal{L}(\boldsymbol{\phi}_t)\right\|^2. \tag{A42}$$

Also since $\left\|\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t\right\|^2=\langle\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t,\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t\rangle$, then

$$\begin{aligned}\left\|\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t\right\|^2&=\langle-\lambda\left[\boldsymbol{g}_{t+1}+\sigma_{\mathrm{DP}}C\boldsymbol{\eta}_{t+1}\right],-\lambda\left[\boldsymbol{g}_{t+1}+\sigma_{\mathrm{DP}}C\boldsymbol{\eta}_{t+1}\right]\rangle\\ &=\lambda^2\langle\boldsymbol{g}_{t+1}+\sigma_{\mathrm{DP}}C\boldsymbol{\eta}_{t+1},\boldsymbol{g}_{t+1}+\sigma_{\mathrm{DP}}C\boldsymbol{\eta}_{t+1}\rangle\\ &=\lambda^2\left(\left\|\boldsymbol{g}_{t+1}\right\|^2+2\sigma_{\mathrm{DP}}C\boldsymbol{g}_{t+1}^\top\boldsymbol{\eta}_{t+1}+\sigma_{\mathrm{DP}}^2C^2\left\|\boldsymbol{\eta}_{t+1}\right\|^2\right)\\ &=\lambda^2\left(\left\|\boldsymbol{g}_{t+1}-\kappa\nabla \mathcal{L}(\boldsymbol{\phi}_t)+\kappa\nabla \mathcal{L}(\boldsymbol{\phi}_t)\right\|^2+2\sigma_{\mathrm{DP}}C\boldsymbol{g}_{t+1}^\top\boldsymbol{\eta}_{t+1}+\sigma_{\mathrm{DP}}^2C^2\left\|\boldsymbol{\eta}_{t+1}\right\|^2\right)\\ &\le\lambda^2\left(\left\|\boldsymbol{g}_{t+1}-\kappa\nabla \mathcal{L}(\boldsymbol{\phi}_t)\right\|^2+\left\|\kappa\nabla \mathcal{L}(\boldsymbol{\phi}_t)\right\|^2+2\sigma_{\mathrm{DP}}C\boldsymbol{g}_{t+1}^\top\boldsymbol{\eta}_{t+1}+\sigma_{\mathrm{DP}}^2C^2\left\|\boldsymbol{\eta}_{t+1}\right\|^2\right),\end{aligned} \tag{A43}$$

and hence, from assumption Assumption 7, and the fact that $\mathcal{L}$ is Lipschitz continuous with Lipschitz constant G,

$$\mathbb{E}\left[\left\|\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t\right\|^2\mid\boldsymbol{\phi}_t\right]\le\lambda^2\left(e_{\mathrm{sub}}^2+\kappa^2G^2+\sigma_{\mathrm{DP}}^2C^2d\right). \tag{A44}$$

Taking the expectation of both sides of the inequality (A40) conditioned on $\boldsymbol{\phi}_t$, and plugging in both (A42) and (A44), we obtain:

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\phi}_{t+1})\mid\boldsymbol{\phi}_t\right]-\mathcal{L}(\boldsymbol{\phi}_t)\le-\lambda\kappa\left\|\nabla \mathcal{L}(\boldsymbol{\phi}_t)\right\|^2+\frac{H}{2}\left(e_{\mathrm{sub}}^2+\kappa^2G^2+\sigma_{\mathrm{DP}}^2C^2d\right)\lambda^2. \tag{A45}$$

Taking the expectation of both sides again, then from the law of total expectation:

$$\mathbb{E}\left[\mathbb{E}\left[\mathcal{L}(\boldsymbol{\phi}_{t+1}) \mid \boldsymbol{\phi}_t\right]\right] = \mathbb{E}\left[\mathcal{L}(\boldsymbol{\phi}_{t+1})\right],$$

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\phi}_{t+1})\right] - \mathbb{E}\left[\mathcal{L}(\boldsymbol{\phi}_t)\right] \leq -\lambda\kappa\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\phi}_t)\right\|^2\right] + \frac{H}{2}\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right)\lambda^2. \tag{A46}$$

Re-arranging the terms, we obtain:

$$\lambda\kappa\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\phi}_t)\right\|^2\right] \leq \mathbb{E}\left[\mathcal{L}(\boldsymbol{\phi}_t)\right] - \mathbb{E}\left[\mathcal{L}(\boldsymbol{\phi}_{t+1})\right] + \frac{H}{2}\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right)\lambda^2. \tag{A47}$$

Now by taking the sum of both side from $t = 0$ to $R - 1$ and telescoping, we obtain:

$$\sum_{t=0}^{T-1}\lambda\kappa\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\phi}_t)\right\|^2\right] \leq \mathbb{E}\left[\mathcal{L}(\boldsymbol{\phi}_0)\right] - \mathbb{E}\left[\mathcal{L}(\boldsymbol{\phi}_T)\right] + \frac{H}{2}\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right)\sum_{t=0}^{T-1}\lambda^2. \tag{A48}$$

from assumption Assumption 5,

$$M \leq \mathbb{E}\left[\mathcal{L}(\boldsymbol{\phi}_T)\right] \Longrightarrow -\mathbb{E}\left[\mathcal{L}(\boldsymbol{\phi}_T)\right] \leq -M, \tag{A49}$$

and since $\boldsymbol{\phi}_0$ is a constant,

$$\sum_{t=0}^{T-1}\lambda\kappa\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\phi}_t)\right\|^2\right] \leq \mathcal{L}\left(\boldsymbol{\phi}_0\right) - M + \frac{H}{2}\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right)\sum_{t=0}^{T-1}\lambda^2. \tag{A50}$$

Since $\min\limits_{0\leq t\leq T-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\phi}_t)\right\|^2\right] \leq \mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\phi}_t)\right\|^2\right]$, then the inequality becomes:

$$\min_{0\leq t\leq T-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\phi}_t)\right\|^2\right]T\lambda\kappa \leq \mathcal{L}\left(\boldsymbol{\phi}_0\right) - M + \frac{H}{2}\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right)T\lambda^2, \tag{A51}$$

therefore,

$$\min_{0\leq t\leq T-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\phi}_t)\right\|^2\right] \leq \frac{\mathcal{L}\left(\boldsymbol{\phi}_0\right) - M}{T\lambda\kappa} + \frac{H}{2\kappa}\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right)\lambda. \tag{A52}$$

We can also get the same result as (A52) by dividing both sides of (A50) by T and re-arranging to get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\phi}_t)\right\|^2\right] \leq \frac{\mathcal{L}\left(\boldsymbol{\phi}_0\right) - M}{T\lambda\kappa} + \frac{H}{2\kappa}\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right)\lambda, \tag{A53}$$

so we have the same upper bound for the average expected gradient norms and the minimum expected gradient norm. Now we want to find λ that minimizes the right-hand side of inequality (A52), in other words, we want λ such that

$$\begin{aligned}
&\frac{d}{d\lambda}\left(\frac{\mathcal{L}\left(\boldsymbol{\phi}_0\right) - M}{T\kappa\lambda} + \frac{H}{2\kappa}\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right)\lambda\right) = 0 \\
\Longleftrightarrow\ & -\frac{\mathcal{L}\left(\boldsymbol{\phi}_0\right) - M}{T\lambda^2} + \frac{H}{2}\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right) = 0 \\
\Longleftrightarrow\ & \frac{H}{2}\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right) = \frac{\mathcal{L}\left(\boldsymbol{\phi}_0\right) - M}{T\lambda^2} \\
\Longleftrightarrow\ & \lambda^2 = \frac{2\left(\mathcal{L}\left(\boldsymbol{\phi}_0\right) - M\right)}{HT\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right)}.
\end{aligned}$$

□

Define λ_{opt} as:

$$\lambda_{\text{opt}} \equiv_{\text{def}} \frac{\sqrt{2\left(\mathcal{L}\left(\boldsymbol{\phi}_0\right) - M\right)}}{\sqrt{HT\left(e_{\text{sub}}^2 + \kappa^2G^2 + \sigma_{\text{DP}}^2C^2d\right)}}.$$

If e_{sub}^2 is insignificant compared to $\kappa^2 G^2 + \sigma_{\text{DP}}^2 C^2 d$, then we can clearly see that

$$\lambda_{\text{opt}} \propto \frac{\sqrt{2}}{\sqrt{HT\left(\kappa^2 G^2 + \sigma_{\text{DP}}^2 C^2 d\right)}}. \tag{A54}$$

We choose our λ according to (A54); however, if we don't know about H and G, then we can either estimate them or simply choose λ_{heur} such that:

$$\lambda_{\text{heur}} = \frac{\sqrt{2}\lambda_c}{\sigma_{\text{DP}} C\sqrt{Td}}, \tag{A55}$$

where $\lambda_c > 0$ is a hyper-parameter.

H Experiments Additional Details

H.1 Per-Example Loss Function Approximation

The per-example losses based on the ELBO might be difficult to compute in closed form for some problems, so generally we approximate ℓ by sampling from the variational distribution q_{VI} and then replacing the expected values with averages. So let $\{\boldsymbol{\theta}_i\}_{i=1}^{N_{\text{VI}}}$, $N_{\text{VI}} > 0$ be samples from q_{VI}, then

$$\ell\left(\boldsymbol{\phi}; \boldsymbol{x}\right) \approx \frac{1}{N_{\text{VI}}} \sum_{i=1}^{N_{\text{VI}}} \left[\log p\left(\boldsymbol{x} \mid \boldsymbol{\theta}_i\right)\right] - \frac{1}{N_{\text{VI}} N} \sum_{i=1}^{N_{\text{VI}}} \left[\log q_{\text{VI}}(\boldsymbol{\theta}_i \mid \boldsymbol{\phi}) - \log p\left(\boldsymbol{\theta}_i\right)\right]. \tag{A56}$$

Throughout our experiments we use $N_{\text{VI}} = 10$.

If $p\left(\mathbf{D} \mid \boldsymbol{\theta}_{\text{con}}\right)$ and $p(\boldsymbol{\theta}_{\text{con}})$ require some constraints on $\boldsymbol{\theta}$, then we would need to transform $\boldsymbol{\theta}_{\text{con}}$ to an unconstrained domain to work with. From (Kucukelbir et al., 2017), we can define any diffeomorphism $\mathcal{U} : \Theta \to \mathbb{R}^n$ such that it transforms $\boldsymbol{\theta}_{\text{con}}$ from the constrained domain to the unconstrained domain $\mathbb{R}^n$. Because $\boldsymbol{\theta}$ is unconstrained with respect to the variational distribution, this requires an additional adjustment to (A56),

$$\begin{aligned}\ell\left(\boldsymbol{\phi}; \boldsymbol{x}\right) \approx & \frac{1}{N_{\text{VI}}} \sum_{i=1}^{N_{\text{VI}}} \left[\log p\left(\boldsymbol{x} \mid \mathcal{U}^{-1}\left(\boldsymbol{\theta}_i\right)\right)\right] \\ & - \frac{1}{N_{\text{VI}} N} \sum_{i=1}^{N_{\text{VI}}} \left[\log q_{\text{VI}}(\boldsymbol{\theta}_i \mid \boldsymbol{\phi}) - \log p\left(\mathcal{U}^{-1}\left(\boldsymbol{\theta}_i\right)\right)\right] \\ & + \frac{1}{N_{\text{VI}} N} \sum_{i=1}^{N_{\text{VI}}} \left|\det J_{\mathcal{U}^{-1}}\left(\boldsymbol{\theta}_i\right)\right|.\end{aligned} \tag{A57}$$

where $J_{\mathcal{U}^{-1}}\left(\boldsymbol{\theta}_i\right)$ is the Jacobian matrix of $\mathcal{U}^{-1}$ evaluated at $\boldsymbol{\theta}_i$ and det denotes the determinant.

The estimation of $\boldsymbol{\phi}^*$ and $\mathbf{A}$ is affected by the choice of the learning-rate, especially the estimation of $\mathbf{A}$ which we fully expand on in Appendix D. We also establish a heuristic for the learning-rate in Appendix E which we use extensively in our experiments,

$$\lambda_{\text{heur}} = \frac{\sqrt{2}\lambda_c}{\sigma_{\text{DP}} C\sqrt{Td}}. \tag{A58}$$

It makes it easier to get the learning rate to the right scale and the only decision to be made is the choice of λ_c which is a hyper-parameter. Usually, the default value ($\lambda_c = 1$) yields good results. Another value for λ_c that we found to yield good results for other models is $\lambda_c = \sqrt{\frac{d}{2}}$.

H.2 Gradients Preconditioning

We apply a simple preconditioning technique by modifying (8) so that for a vector $\boldsymbol{\beta} = (\boldsymbol{\beta_\mu}, \boldsymbol{\beta_u})$ which is the concatenation of two vectors $\boldsymbol{\beta_\mu}$ and $\boldsymbol{\beta_u}$ each of dimension n, the update rule becomes,

$$\begin{aligned} \boldsymbol{g}_{t+1} &= \sum_{i \in \mathcal{B}_{t+1}} \text{clip}\left(\boldsymbol{\beta} \odot \nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}_t; \boldsymbol{x}_i), C\right), \\ \widetilde{\boldsymbol{g}}_{t+1} &= \frac{1}{\boldsymbol{\beta}} \odot \left[\boldsymbol{g}_{t+1} + \sigma_{\text{DP}} C \boldsymbol{\eta}_{t+1}\right], \\ \boldsymbol{\phi}_{t+1} &= \boldsymbol{\phi}_t - \lambda \widetilde{\boldsymbol{g}}_{t+1}, \end{aligned} \tag{A59}$$

where $\odot$ is the element-wise multiplication operation, and $\frac{1}{\boldsymbol{\beta}}$ denotes the vector obtained from the reciprocals of the components of $\boldsymbol{\beta}$. In our experiments, $\boldsymbol{\beta_\mu}$ is filled with 1s and we only choose specific values for $\boldsymbol{\beta_u}$. We also need to modify the gradient-based (20) model so that

$$\widetilde{\boldsymbol{g}}_{t+1} \mid \boldsymbol{\phi}_t, \mathbf{A}, \boldsymbol{\phi}^* \sim \mathcal{N}\left(\kappa \mathbf{A}\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\right), \frac{1}{\boldsymbol{\beta}} \odot \left(\sigma_{\text{DP}}^2 C^2 \mathbf{I}_d + \Sigma_{\text{sub}}\right)\right), \tag{A60}$$

where $\frac{1}{\boldsymbol{\beta}} \odot \Sigma_{\text{DP}}$ is the element-wise multiplication of each row-vector of Σ_{DP} by $\frac{1}{\boldsymbol{\beta}}$. Regarding the learning rate λ, we use our heuristic but multiply the heuristic with $\boldsymbol{\beta}$ to ensure also that all the parameters converge at a similar rate,

$$\lambda = \lambda_{\text{heur}} \boldsymbol{\beta},$$

and this makes lambda into a vector so that the product in the last equation in (A59), i.e. $\lambda \widetilde{g}_{t+1}$ becomes an element-wise product $\lambda \odot \widetilde{g}_{t+1}$.

H.3 Setting the Priors

We choose Gaussian distributions as priors for both $\boldsymbol{\phi}^*$ and each v_i. For $\boldsymbol{\phi}^*$ we set,

$$\boldsymbol{\phi}^* \sim \mathcal{N}\left(\frac{1}{T - T^*} \sum_{t=T^*}^{T-1} \boldsymbol{\phi}_t, \mathbf{I}_d\right). \tag{A61}$$

The prior for each v_i is set based on the MLE estimate for $\mathbf{A} = \text{diag}(v_1, \dots, v_d)$ in the proof of Theorem 2,

$$\begin{aligned} v_i &\sim \mathcal{N}\left(\mu_{v_i}, \sigma_{v_i}^2\right), \\ \mu_{v_i} &= \frac{\left|\sum_{t=T^*}^{T-1} \widetilde{\boldsymbol{g}}_{t+1}^{(i)} \left(\boldsymbol{\phi}_t^{(i)} - \overline{\boldsymbol{\phi}}^{(i)}\right)\right|}{\kappa \sum_{t=T^*}^{T-1} \left(\boldsymbol{\phi}_t^{(i)} - \overline{\boldsymbol{\phi}}^{(i)}\right)^2}, \\ \sigma_{v_i} &= \frac{\sigma_{\text{DP}}^2 C^2}{\kappa^2 \left(\boldsymbol{\beta}^{(i)}\right)^2 \sum_{t=T^*}^{T-1} \left(\boldsymbol{\phi}_t^{(i)} - \overline{\boldsymbol{\phi}}^{(i)}\right)^2}. \end{aligned}$$

where $\overline{\boldsymbol{\phi}}$ is the average of the trace after the burn-in index $\boldsymbol{\phi}^*$,

$$\overline{\boldsymbol{\phi}} = \frac{1}{T - T^*} \sum_{t=T^*}^{T-1} \boldsymbol{\phi}_t.$$

H.4 TARP Evaluation Method

We can further approximate the average coverages:

$$C(\alpha) = \frac{1}{K} \sum_{i=1}^{K} \mathbb{1}_{i,\alpha}\left(\boldsymbol{\theta}_i\right),$$

by applying the TARP algorithm from (Lemos et al., 2023). This is done by sampling from the approximate posterior $\widetilde{p}\left(\boldsymbol{\theta} \mid \boldsymbol{\xi}\right)$ and assuming that the credible regions are balls centered around $\boldsymbol{\theta}_{\mathrm{ref}}\left(\boldsymbol{\xi}_i\right)$. Denote $\boldsymbol{\theta}_{\mathrm{ref}}\left(\boldsymbol{\xi}_i\right)$ as $\boldsymbol{\theta}_{\mathrm{ref}(i)}$ and let $\left\{\widetilde{\boldsymbol{\theta}}_{j,i} : 1 \leq j \leq N_{\mathrm{tarp}}\right\}$ be samples from the approximate posterior $\widetilde{p}\left(\boldsymbol{\theta} \mid \boldsymbol{\xi}_i\right)$ for each i, then

$$C(\alpha) \approx \frac{1}{K} \sum_{i=1}^{K} \mathbb{1}\left[f_i < 1 - \alpha\right], \tag{A62}$$

where

$$f_i = \frac{1}{N_{\mathrm{tarp}}} \sum_{j=1}^{N_{\mathrm{tarp}}} \mathbb{1}\left[d\left(\widetilde{\boldsymbol{\theta}}_{j,i}, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right) < d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right)\right]. \tag{A63}$$

The motivation for why this is a valid approximation is that:

$$\boldsymbol{\theta}_i \in \widetilde{\mathcal{C}}_\alpha\left(\boldsymbol{\xi}_i, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right) \iff d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right) < r\left(\alpha, \boldsymbol{\xi}_i\right),$$

where $r\left(\alpha, \boldsymbol{\xi}_i\right)$ is the radius of the credible region. Also, this is equivalent to

$$\mathrm{Ball}\left(\boldsymbol{\theta}_{\mathrm{ref}(i)}, d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right)\right) \subset \widetilde{\mathcal{C}}_\alpha\left(\boldsymbol{\xi}_i, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right),$$

i.e. the ball centered around $\boldsymbol{\theta}_{\mathrm{ref}(i)}$ with radius $d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right)$ is contained in the credible region. Therefore, it is also equivalent to

$$\int_{\Theta} \mathbb{1}\left[\widetilde{\boldsymbol{\theta}} \in \mathrm{Ball}\left(\boldsymbol{\theta}_{\mathrm{ref}(i)}, d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right)\right)\right] \widetilde{p}\left(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{\xi}_i\right) < 1 - \alpha.$$

This integral can be approximated by sampling from $\widetilde{p}\left(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{\xi}_i\right)$ and then using these samples to calculate the average of $\mathbb{1}\left[\widetilde{\boldsymbol{\theta}} \in \mathrm{Ball}\left(\boldsymbol{\theta}_{\mathrm{ref}(i)}, d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right)\right)\right]$. Also,

$$\begin{aligned} &\widetilde{\boldsymbol{\theta}} \in \mathrm{Ball}\left(\boldsymbol{\theta}_{\mathrm{ref}(i)}, d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right)\right) \\ \iff &\mathbb{1}\left[d\left(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right) < d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathrm{ref}(i)}\right)\right], \end{aligned}$$

and thus the approximation holds. One important thing to consider here, is that if $\boldsymbol{\theta} \in \Theta$ is constrained, we must have a diffeomorphism $\mathcal{U} : \Theta \to \mathbb{R}^n$ that transforms the prior samples $\boldsymbol{\theta}_i$ to the unconstrained space $\mathcal{U}\left(\boldsymbol{\theta}_i\right)$. Since $\mathcal{U}$ is a diffeomorphism and thus both injective and surjective, it must preserve the relation $\in$. Hence,

$$\boldsymbol{\theta} \in X \iff \mathcal{U}\left(\boldsymbol{\theta}\right) \in \mathcal{U}\left(X\right)$$

similarly if $\boldsymbol{\theta}' \in \mathbb{R}^n$, then

$$\boldsymbol{\theta}' \in Y \iff \mathcal{U}^{-1}\left(\boldsymbol{\theta}'\right) \in \mathcal{U}^{-1}\left(Y\right)$$

According to Theorem 3 of (Lemos et al., 2023), if $\widetilde{p}\left(\boldsymbol{\theta} \mid \mathcal{T}\right)$ had the correct coverages, then the marginals $\widetilde{p}\left(\boldsymbol{\theta}^{(i)} \mid \mathcal{T}\right)$ should also have correct coverages. We can also use the TARP algorithm to calculate the coverages for each dimension of the parameters vector $\boldsymbol{\theta}$, i.e., the marginal coverages $\widetilde{p}\left(\boldsymbol{\theta}^{(i)} \mid \mathcal{T}\right)$ where $\boldsymbol{v}^{(i)}$ denotes the ith component of a vector $\boldsymbol{v}$.

H.5 Exponential Families Experiment Details

For the constrained optimization problem of M1, $\boldsymbol{\theta}_{\mathrm{con}} > 0$, so we define

$$\mathcal{U}_1\left(\boldsymbol{\theta}_{\mathrm{con}}\right) = \log\left(e^{\boldsymbol{\theta}_{\mathrm{con}}} - 1\right) \implies \mathcal{U}_1^{-1}\left(\boldsymbol{\theta}\right) = \mathrm{softplus}\left(\boldsymbol{\theta}\right). \tag{A64}$$

For Model M2, $\boldsymbol{\theta}_{\mathrm{con}} \in (0, 1)$, so we define

$$\mathcal{U}_2\left(\boldsymbol{\theta}_{\mathrm{con}}\right) = \log\left(\frac{\boldsymbol{\theta}_{\mathrm{con}}}{1 - \boldsymbol{\theta}_{\mathrm{con}}}\right) \implies \mathcal{U}_2^{-1}\left(\boldsymbol{\theta}\right) = \frac{1}{1 + e^{-\boldsymbol{\theta}}}.$$

Finally, for Model M3, let $\boldsymbol{\theta}_{\text{con}} = (\theta_1, \theta_2, \theta_3)$, then $\theta_1 + \theta_2 + \theta_3 = 1$ so we only have two degrees of freedom and can let $\theta_3 = 1 - \theta_1 - \theta_2$. We define,

$$\mathcal{U}_3\left(\boldsymbol{\theta}_{\text{con}}\right) = \left(\log\left(\frac{\theta_1}{\theta_3}\right), \log\left(\frac{\theta_2}{\theta_3}\right), 0\right).$$

For the inverse, let $\boldsymbol{\theta} = (\theta_1', \theta_2')$, then

$$\mathcal{U}_3^{-1}\left(\boldsymbol{\theta}\right) = \operatorname{softmax}\left(\theta_1', \theta_2', 0\right).$$

H.6 UCI Adult Experiment Details

For data pre-processing, we removed the columns “education-num”, “native-country”, and “relationship”, and converted categorical values to one-hot encoded vectors. The continuous values were normalized to be within $(0, 1)$. Moreover, we evaluated our method for $\epsilon \in \{0.1, 0.3, 1.0\}$. Across these different values of ϵ, we used the same number of iterations $T = 10^4$ and a sampling rate of $\kappa = 0.1$.

H.7 NA-DPVI Algorithm

The variables that are used in all the algorithms are found in Algorithm 2. The DPVI algorithm that we used can be found in Algorithm 3 from which the parameter and gradient traces ($\mathcal{T}$, $\widetilde{\mathcal{G}}$) are obtained. Finally, our NA-DPVI algorithm can be found in Algorithm 4.

The time complexity of DPVI (Algorithm 3) is $\mathcal{O}(T \times B \times d)$ where T is the number of steps of DP-SGD, $B = \kappa N$ is the expected batch size (κ is the subsampling ratio and N is the number of samples), and d is dimensionality of $\boldsymbol{\phi}$ (i.e. parameters).

The time complexity of NA-DPVI (Algorithm 4) has an additional cost of sampling M from the posterior $p\left(\boldsymbol{\phi}^*, \mathbf{A}, \Sigma_{\text{sub}} \mid \widetilde{\mathcal{G}}, \mathcal{T}\right)$, which depends on the method of approximating the noise-aware posterior Eq. (23) (e.g., NUTS and Laplace). The complexity of the noise-aware posterior approximation method is a function M, d, and T since the approximation method takes as its input the trace $\mathcal{T}$ and perturbed gradients $\widetilde{\mathcal{G}}$.

Algorithm 2 **Global variables**

```
$(\boldsymbol{x}_i)_{i=1}^N$ : data;
$(\epsilon, \delta)$ : DP privacy level;
$C$ : clipping threshold;
$T$ : training iterations;
$\lambda$ : learning rate;
$\kappa$ : Poisson sampling rate;
$\boldsymbol{\beta}$ : gradients preconditioning vector;        ▷ see Appendix H.2, for more details about preconditioning.
$\boldsymbol{\phi}_0$ : initial values for $\boldsymbol{\phi}$;
$\sigma_{\text{DP}}$ : PRVAccountant$(\epsilon, \delta, T, \kappa)$;        ▷ see Gopi et al. (2021).
```

Algorithm 3 DPVI

```
for $t = 1, \ldots, T-1$ do
    $\boldsymbol{g}_{t+1} \leftarrow \sum_{i \in \mathcal{B}_{t+1}} \operatorname{clip}\left(\boldsymbol{\beta} \odot \nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}_t; \boldsymbol{x}_i), C\right)$;        ▷ see Appendix H.2, for more details about preconditioning.
    Sample $\boldsymbol{\eta}_{t+1} \sim \mathcal{N}(0, \mathbf{I}_d)$;
    $\widetilde{\boldsymbol{g}}_{t+1} \leftarrow \frac{1}{\boldsymbol{\beta}} \odot \left[\boldsymbol{g}_{t+1} + \sigma_{\text{DP}} C \boldsymbol{\eta}_{t+1}\right]$;
    $\boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t - \lambda \widetilde{\boldsymbol{g}}_{t+1}$;
end for
return $\mathcal{T} = (\boldsymbol{\phi}_t)_{t=0}^T$, $\widetilde{\mathcal{G}} = (\widetilde{\boldsymbol{g}}_t)_{t=1}^T$;
```

H.8 Computational Resources

We used the computational resources offered by CSC – IT Center for Science, Finland. In particular, the Puhti supercomputer was used to run experiments on CPU nodes in parallel. Each Puhti node is equipped with two Intel Xeon processors, code name Cascade Lake, with 20 cores each running at 2.1 GHz.

Algorithm 4 NA-DPVI

```
Input:
    $M$ : number of samples;
    $\boldsymbol{\mu}_{\boldsymbol{\phi}^*}$ : $\boldsymbol{\phi}^*$ prior distribution mean;
    $\boldsymbol{\Sigma}_{\boldsymbol{\phi}^*}$ : $\boldsymbol{\phi}^*$ prior distribution covariance matrix;
    $\boldsymbol{\mu}_{\mathbf{A}}$ : $\mathbf{A}$ entries prior distribution mean;
    $\boldsymbol{\Sigma}_{\mathbf{A}}$ : $\mathbf{A}$ entries prior distribution covariance matrix;
    $\boldsymbol{\mu}_{\Sigma_{\text{sub}}}$ : $\Sigma_{\text{sub}}$ entries prior distribution mean;
    $\boldsymbol{\Sigma}_{\Sigma_{\text{sub}}}$ : $\Sigma_{\text{sub}}$ entries prior distribution covariance matrix;
    $\mathcal{T}, \widetilde{\mathcal{G}}$ : DPVI output;
model-definition $p\left(\widetilde{\mathcal{G}} \mid \mathcal{T}, \boldsymbol{\phi}^*, \mathbf{A}, \Sigma_{\text{sub}}\right)$
    $\boldsymbol{\phi}^* \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\phi}^*}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}^*})$;
    $\mathbf{A} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{A}}, \boldsymbol{\Sigma}_{\mathbf{A}})$;
    $\Sigma_{\text{sub}} \sim \mathcal{N}(\boldsymbol{\mu}_{\Sigma_{\text{sub}}}, \boldsymbol{\Sigma}_{\Sigma_{\text{sub}}})$;
    $\widetilde{\boldsymbol{g}}_{t+1} \mid \boldsymbol{\phi}_t, \mathbf{A}, \boldsymbol{\phi}^*, \Sigma_{\text{sub}} \sim \mathcal{N}\left(\kappa \mathbf{A}\left(\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\right), \frac{1}{\boldsymbol{\beta}} \odot \left(\sigma_{\text{DP}}^2 C^2 \mathbf{I}_d + \Sigma_{\text{sub}}\right)\right)$;    ▷ see Eq. (A60).
end model-definition
Sample $(\boldsymbol{\phi}_i^*, \mathbf{A}_i, \Sigma_{\text{sub},i})_{i=1}^M \sim p\left(\boldsymbol{\phi}^*, \mathbf{A}, \Sigma_{\text{sub}} \mid \widetilde{\mathcal{G}}, \mathcal{T}\right)$;    ▷ sampling using any approximate inference method.
return $\widetilde{p}(\boldsymbol{\theta} \mid \mathcal{T}) = \frac{1}{M} \sum_{i=1}^M q_{\text{VI}}(\boldsymbol{\theta}; \boldsymbol{\phi}_i^*)$;    ▷ approximate noise-aware posterior mixture model.
```