1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–21

Fast Bayesian Inference in GLMs with Low Rank Data Approximations

Brian L. Trippe BTRIPPE@MIT.EDU and **Jonathan H. Huggins** JHUGGINS@MIT.EDU
Tamara Broderick TBRODERICK@CSAIL.MIT.EDU

1. Introduction

Scientists, engineers, and social scientists are often interested in characterizing the relationship between a response and a set of features. For example, a biologist may wish to understand the effect of certain genes on the presence of a disease or a medical practitioner may wish to understand the effect of a patient's history on their future health. In these applications and countless others, the relative ease of modern data collection methods often lead to particularly large collections of features for data analysts to study. While this rich data should ultimately aid understanding, it poses a number of practical challenges. One challenge is how to discover interpretable relationships between features and a response. Generalized linear models (GLMs) are widely used in part because they provide such interpretability – as well as the flexibility to accommodate many different types of responses (including binary, count, and heavy-tailed responses). A second challenge is that, unless the number of data points is substantially larger than the number of features, there is likely to be non-trivial uncertainty about the relationships between various features and the response. A Bayesian approach to GLM inference provides exactly the desired coherent uncertainty quantification – as well as the ability to improve inference by incorporating expert information and sharing power across experiments. A final challenge is computational. Naive approaches to point inference in GLMs require at least linear time in both the number of features and data points, but capturing Bayesian uncertainty is typically either super-quadratic in the number of features or super-quadratic in the number of data points. For data and feature cardinalities in the tens of thousands or larger – as is often achieved in biological, medical, and other applications – this latter cost can be prohibitive.

In what follows, we propose to reduce the dimensionality of the feature set as a preprocessing step to speed up Bayesian inference; in particular, we show that low rank descriptions of the data permit fast MCMC routines and Laplace approximations of the Bayesian posterior for the full feature set. We motivate our proposal with a conjugate linear regression analysis in the case where the data is exactly low rank. When the data are merely approximately low rank, our proposal is an approximation. We provide preliminary theory and experiments to assess the quality of this approximation both for conjugate linear regression as well as non-conjugate GLMs.

Related work. Geppert et al. (2017) and Lee and Oh (2013) focus on conjugate Bayesian regression, respectively using random projections and PCA to reduce covariate dimension. Spantini et al. (2015) use conjugate Bayesian regression as stepping-off point to derive a point estimator for Bayesian inverse problems. Guhaniyogi and Dunson (2015) use random projections for Bayesian GLMs but focus on the predictive performance rather than parameter estimation. Outside the Bayesian context, Zhang et al. (2014), Wang et al. (2017), and many others have analyzed random projections for regression and classification

using, for example, an M-estimation framework. Some additional related work is discussed in Appendix A.

2. Approach

Background. Suppose we have N data points, each with dimension D. We collect our covariates, or features, in the design matrix $X \in \mathbb{R}^{N\times D}$ and our responses in the column vector $Y \in \mathbb{R}^N$. Let $\beta \in \mathbb{R}^D$ be an unknown parameter characterizing the relationship between the covariates and the response for each data point. In particular, we take β to parameterize a likelihood $p(Y|X,\beta)$, which we assume takes a particular GLM form: $p(Y|X\beta)$. To complete our Bayesian model specification, we assume a prior $p(\beta)$ expressing our knowledge of β before seeing any data. Bayes' theorem gives the Bayesian posterior: $p(\beta|Y,X) = p(\beta)p(Y|X\beta)/Z$, where $Z := \int p(\beta)p(Y|X\beta)d\beta$ is the normalizing constant. Notably, when both N and D are large, computing the likelihood up to this constant requires an expensive $O(ND)$ matrix-vector multiplication.

Proposal. Our proposal is simple. Choose some integer M with $0 < M < D$. For any real matrix X, its singular value decomposition exists and may be written as:

$$X^T = U\mathrm{diag}(\lambda)V^T + \bar{U}\mathrm{diag}(\bar{\lambda})\bar{V}^T.$$

Here $U \in \mathbb{R}^{D\times M}, \bar{U} \in \mathbb{R}^{D\times(D-M)}, V \in \mathbb{R}^{N\times M}$, and $\bar{V} \in \mathbb{R}^{N\times(D-M)}$ are matrices of orthonormal rows, and $\lambda \in \mathbb{R}^M$ and $\bar{\lambda} \in \mathbb{R}^{D-M}$ are vectors of decreasing singular values $\lambda_1 \geq \lambda_2 \geq \lambda_M \geq \bar{\lambda}_1 \geq \cdots \geq \bar{\lambda}_{D-M} \geq 0$.

We propose to replace X with the low-rank approximation XUU^T. Note that the resulting posterior approximation $\tilde{p}(\beta)$ is still a distribution over the original D-dimensional β vector: $p(\beta|Y,X) \approx \tilde{p}(\beta) := p(\beta)p(Y|XUU^T\beta)/\tilde{Z}$, where $\tilde{Z} = \int p(\beta)p(Y|XUU^T\beta)d\beta$ is the normalizing constant. We evaluate this approximation theoretically and empirically.

3. Evaluation of the approximation

Conjugate linear regression, low rank data. The classic linear regression likelihood fits our desired framework above: $p(Y|X,\beta) = \mathcal{N}(Y|X\beta,(\tau I_N)^{-1})$, where $\tau > 0$ is the precision, and I_N is the identity matrix of size N. Consider the conjugate prior $p(\beta) = \mathcal{N}(\beta|0,\Sigma_\beta)$. Due to conjugacy, we can write the posterior in closed form: $p(\beta|Y,X) = \mathcal{N}(\beta|\mu_N,\Sigma_N)$ with $\Sigma_N := \left(\Sigma_\beta^{-1} + \tau X^T X\right)^{-1}$ and $\mu_N := \tau\Sigma_N X^T Y$ (Bishop, 2006). Notably, while conjugacy avoids the computational expense of approximate Bayesian inference, it does not avoid the prohibitive $O(D^3)$ cost of calculating Σ_N by inverting Σ_N^{-1}.

By contrast, suppose that X is low rank and can therefore be written as $X = XUU^T$ exactly for some $M \ll D$. Then, if $\Sigma_\beta = \sigma_\beta^2 I_D$, we can write (see Appendix B.1 for the derivation):

$$\Sigma_N = \sigma_\beta^2\left\{I - U\mathrm{diag}\left(\frac{\tau\lambda^2}{\sigma_\beta^{-2}+\tau\lambda^2}\right)U^T\right\} \quad \text{and} \quad \mu_N = U\frac{\tau\lambda}{\sigma_\beta^{-2}+\tau\lambda^2}V^T Y. \tag{1}$$

The partial SVD and posterior mean and covariance can altogether be computed in $O(NDM + DM^2)$ time and with $O(MD)$ memory, an improvement over the $O(D^3 + ND^2)$ time and $O(D^2)$ memory demands of computing the exact solution (see Appendix B.2).

Conjugate linear regression, approximately low rank data. While the case above is illustrative about the time and memory savings achievable with our approach, real data are rarely exactly low rank. So, more generally, our approach will yield an approximation $\mathcal{N}(\beta|\tilde{\mu}_N, \tilde{\Sigma}_N)$ to the posterior $\mathcal{N}(\beta|\mu_N, \Sigma_N)$ – rather than the exact posterior, as in the previous case. We next upper bound the error in our approximation and then interpret our bounds. In particular, we note that practitioners typically report posterior means and variances, so we focus on how well we approximate these functionals. In what follows we rewrite λ_m as $\lambda_{N,m}$ to emphasize the N dependence.

Theorem 1 *Consider any conjugate Bayesian linear regression and our low rank approximation in Section 2. We also assume here that each element of Y is bounded by b and that the exact posterior is α_N-strongly log concave. Then $\|\tilde{\mu}_N - \mu_N\|_2$ is bounded above by*

$$\alpha_N^{-1}\tau\big(\lambda_{N,M+1}^2\|\bar{U}^T\mu_N\|_2 + \lambda_{N,M+1}\|\bar{V}^T Y\|_2\big) \leq \alpha_N^{-1}\tau\big(\lambda_{N,M+1}^2\|\mu_N\|_2 + \lambda_{N,M+1}\sqrt{N}b\big). \quad (2)$$

$$\textit{Also,} \quad \Sigma_N^{-1} - \tilde{\Sigma}_N^{-1} = \tau(X^TX - UU^TX^TXUU^T), \qquad \textit{hence} \qquad \|\Sigma_N^{-1} - \tilde{\Sigma}_N^{-1}\|_2 = \tau\lambda_{N,M+1}^2. \quad (3)$$

We examine the implications of this result via a number of corollaries. All proofs may be found in Appendix C.1.

Corollary 2 *When the data is generated i.i.d. from the likelihood with true parameter β, then under mild assumptions on the distribution of the covariates, $\tilde{\mu}_N$ converges in probability to the maximum a priori vector $\tilde{\mu}$ satisfying $U^T\tilde{\mu} = U^T\beta$.*

This corollary captures that the posterior mean estimate is not, in general, consistent for a true data-generating parameter but demonstrates its reasonable asymptotic behavior. In particular, $\tilde{\mu}_N$ is consistent within the span of U and converges to the most probable vector (a priori) with this characteristic. In the special case that Σ_β is diagonal, $\tilde{\mu}_N \to UU^T\beta$. As expected, we see that we are not learning anything about the relation between the response and covariates in the data directions we truncate away with our truncated SVD approach. If the response has little dependence on these directions, $\bar{U}\bar{U}^T\beta = \tilde{\mu} - \beta$ will be small and the error in our approximation will be low. If the response depends heavily on these directions, our error will be higher. This challenge is ubiquitous in dealing with high-dimensional covariates. Indeed, when theory is developed in these cases, we often see explicit assumptions encoding the notion that high-variance directions in X are also highly predictive of the response (see, e.g., Zhang et al. (2014) Theorem 2).

Corollary 3 *The approximate posterior uncertainty in any linear combination of parameters is no smaller than the exact posterior uncertainty, or equivalently, $\tilde{\Sigma} - \Sigma$ is positive semi-definite.*

This corollary demonstrates that our approximation never underestimates uncertainty. From an approximation perspective, overestimating uncertainty can be seen as preferable to underestimation as it leads to more conservative decision-making. An alternative perspective is that we actually engender additional uncertainty simply by making an approximation, with more uncertainty for coarser approximations, and we should express that in reporting our inferences. We see similar behavior in a corollary showing that our approximate posterior never has lower entropy than the exact posterior (see Appendix C.4).

Non-conjugate GLMs, approximately low rank data. The conjugate linear setting above facilitates intuition and theory. But many models are non-conjugate and require posterior approximation, such as: (1) Markov Chain Monte Carlo (MCMC), which has theoretical guarantees asymptotic in running time but may be relatively slow in practice and (2) the Laplace approximation, which is typically faster but does not become arbitrarily accurate given enough time. In these cases, we still obtain large time savings from our low-rank approximation. Each likelihood computation in MCMC takes $O(ND)$ without our low-rank approximation and takes $O(NM + DM)$ time with our approximation. Likewise, the Laplace approximation in general takes $O(D^3)$ time and $O(D^2)$ memory. With our low-rank approximation, Laplace takes $O(DMN + DM^2 + NM^2)$ time and $O(DM)$ memory. See Appendix D.2 for proofs and for further theory on the quality of Laplace together with our low-rank approximation.

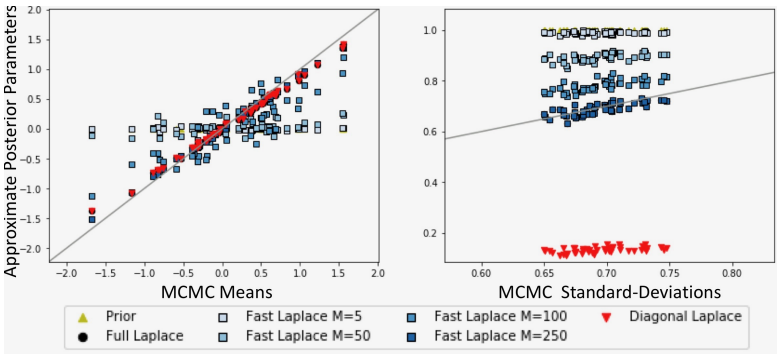

Figure 1: Approximate posterior mean and standard deviation across a parameter subset as M varies. X-axis represents ground truth from running MCMC.

We test our approximation on a simulated logistic regression task (details in Appendix E.2). The Laplace results are in Figures 1 and 2, and the MCMC results (using HMC in Stan (Carpenter et al., 2017)) are in Figures 4 and 5. We see how the truncation level M provides a knob to turn for trading off statistical accuracy for computational cost. We are able to obtain statistical performance essentially as good as the full Laplace approximation at a fraction of the computational budget.

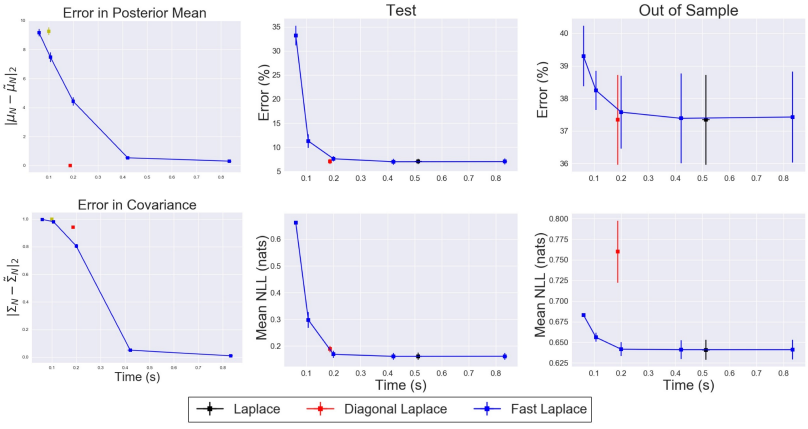

Figure 2: Trade-off between approximation quality and runtime in logistic regression, found by varying M. In left panels, error is relative to the full Laplace approximation.

Appendix A. Additional Related Work

An additional related line of work on speeding up Bayesian inference for high dimensional linear models has pursued approaches to reduce the time required to draw samples from Gaussian conditionals within Gibbs samplers. In particular, Brown et al. (2016) takes advantage of low rank structure in the Hessian of the log likelihood to draw approximate samples from Gaussian conditionals by approximating this matrix with a truncated eigenvalue decomposition; this then allows for an analytic factorization of the associated approximate covariance matrix, with which samples may be generated transforming i.i.d. Gaussian noise. More recently, Nishimura and Suchard (2018) similarly took advantage of such low rank structure to efficiently draw samples from Gaussian conditionals within a Gibbs sampler by solving a linear system using conjugate gradient methods. In both cases the sampling operations take on order $O\big(\min(DNM, D^2M)\big)$ time. In the case of Brown et al. (2016), this cost arises from the truncated eigenvalue decomposition and in the case of Nishimura and Suchard (2018) this cost come from performing M iterations of conjugate gradient descent.

Appendix B. Supplementary material on conjugate Gaussian regression with exactly low rank design

B.1. Derivation of Equation (1)

We here derive the expressions for the mean and covariance of the posterior over parameters in conjugate Bayesian provided in Equation (1). Our starting point for these expressions is:

$$\Sigma_N^{-1} = \sigma_\beta^{-2} I + \tau X^T X \text{ and } \mu_N = \tau \Sigma_N X^T$$

When $X = V\text{diag}(\boldsymbol{\lambda})U^T$ for orthonormal U, we may write $X = V\text{diag}(\boldsymbol{\lambda})I_D U^T = V\text{diag}(\boldsymbol{\lambda})U^T UU^T = XUU^T$, this allows us to write Σ_N as:

$$\begin{aligned}
\Sigma_N &= \big(\sigma_\beta^{-2} I + \tau X^T X\big)^{-1} \\
&= \big(\sigma_\beta^{-2} I + U(U^T X^T \tau X U)U^T\big)^{-1} \\
&= \sigma_\beta^2 I - \sigma_\beta^2 U\big(\tau^{-1}(U^T X^T X U)^{-1} + \sigma_\beta^2 I\big)^{-1} U^T \sigma_\beta^2 \\
&= \sigma_\beta^2 I - \sigma_\beta^2 U\big(\tau^{-1}(\text{diag}(\boldsymbol{\lambda}^2))^{-1} + \sigma_\beta^2 I\big)^{-1} U^T \sigma_\beta^2 \\
&= \sigma_\beta^2 I - \sigma_\beta^2 U\text{diag}\big((\tau^{-1}\boldsymbol{\lambda}^{-2} + \sigma_\beta^2)^{-1}\big) U^T \sigma_\beta^2 \\
&= \sigma_\beta^2 I - \sigma_\beta^2 U\text{diag}\Big(\frac{1}{\tau^{-1}\boldsymbol{\lambda}^{-2} + \sigma_\beta^2}\Big) U^T \sigma_\beta^2 \\
&= \sigma_\beta^2 (I - U\text{diag}\Big(\frac{\sigma_\beta^2}{\tau^{-1}\boldsymbol{\lambda}^{-2} + \sigma_\beta^2}\Big) U^T) \\
&= \sigma_\beta^2 (I - U\text{diag}\Big(\frac{\tau\boldsymbol{\lambda}^2}{\sigma_\beta^{-2} + \tau\boldsymbol{\lambda}^2}\Big) U^T)
\end{aligned} \tag{4}$$

Where in the third line we rely on the Woodbury matrix lemma.

We next derive the expression for the means as:

$$\begin{aligned}
\mu_N &= (\sigma_\beta^{-2} I + \tau X^T X)^{-1} \tau X^T Y \\
&= (\sigma_\beta^{-2} I + X^T V(\tau I) V^T X)^{-1} X^T V(\tau I) V^T Y \\
&= \sigma_\beta^2 X^T V(\sigma_\beta^2 V^T X X^T V + \tau^{-1} I)^{-1} V^T Y \\
&= \sigma_\beta^2 U \mathrm{diag}(\boldsymbol{\lambda})(\sigma_\beta^2 \mathrm{diag}(\boldsymbol{\lambda})^2 + \tau^{-1} I)^{-1} V^T Y \\
&= \sigma_\beta^2 U \frac{\boldsymbol{\lambda}}{\sigma_\beta^2 \boldsymbol{\lambda}^2 + \tau^{-1}} V^T Y \\
&= U \frac{\tau \boldsymbol{\lambda}}{\sigma_\beta^{-2} + \tau \boldsymbol{\lambda}^2} V^T Y
\end{aligned} \tag{5}$$

The second line uses the $V^T V = I_D$ and the third line applies the matrix identity $RW^T(WRW^T + Q)^{-1} = (W^T Q^{-1} W + R^{-1})^{-1} W^T Q^{-1}$ (Petersen et al., 2008; Luenberger, 1969).

B.2. Time complexity of low rank inference

As discussed in the main text, Gaussian-Gaussian conjugacy allows us to express the exact posterior as:

$$\log p(\beta|X,Y) = \log \mathcal{N}(\beta|\mu_N, \Sigma_N) \text{ where } \Sigma_N^{-1} = \sigma_\beta^{-2} I + \tau X^T X \text{ and } \mu_N = \tau \Sigma_N X^T \tag{6}$$

where for simplicity of exposition we here restrict to the case of an isotropic prior.

While these expressions are illuminating from a theoretical perspective, they provide no more of an answer to the inference problem (e.g. computing the posterior mean and marginal variances) than our initial symbolic expression of the posterior as $p(\beta|X,Y) \propto p(\beta)p(Y|\beta,X)$. In particular, when the dimensionality of β is large (say $D > 30,000$), the computation of the posterior mean and parameter covariances is challenging, requiring a $O(D^2)$ space and $O(D^3)$ time to store and invert Σ_N^{-1}, additionally, this requires computing $X^T X$, which takes $O(ND^2)$ time in general.

A notable exception when inference can be done more efficiently is when the design X is low rank such that we may write its singular value decomposition (SVD) as $X^T = U\mathrm{diag}(\boldsymbol{\lambda})V^T$ for orthonormal $U \in \mathbb{R}^{D,M}$and $V^T \in \mathbb{R}^{N,M}$, and some $\boldsymbol{\lambda} \in \mathbb{R}^M$ with decreasing entries. We can now do inference even faster by calculating Σ_N and μ_N as:

$$\Sigma_N = \sigma_\beta^2 \big(I - U\mathrm{diag}(\frac{\tau \boldsymbol{\lambda}^2}{\sigma_\beta^{-2} + \tau \boldsymbol{\lambda^2}}) U^T\big) \text{ and } \mu_N = U \frac{\tau \boldsymbol{\lambda}}{\sigma_\beta^{-2} + \tau \boldsymbol{\lambda}^2} V^T Y \tag{7}$$

The singular vectors U may be obtained in $O(ND \log M)$ via a randomized SVD (Halko et al., 2010). Next, $\boldsymbol{\lambda^2}$ may be computed as $\boldsymbol{\lambda^2} = \mathrm{diag}(\mathbf{U^T X X U})$ with two $O(NDM)$ and $O(DM^2)$ matrix multiplications, and $V^T Y$ as an $O(MN)$ matrix vector multiplication, thereby providing the necessary ingredients of the posterior mean and covariance in $O(NDM)$ time. As for storage, this approach demands keeping only U, λ and $V^T Y$, which comes to just $O(MD)$. In total this demands $O(NDM + DM^2)$ time and $O(DM)$ memory as opposed to $O(ND^2 + D^3)$ time and $O(D^2)$ memory using the naive approach. In the large D, large N regime this can lead to significant savings.

Appendix C. Proofs of Theorem 1, corollaries and supporting lemmas for conjugate Bayesian linear regression with low rank data approximations

We begin with the proof of Theorem 1.

C.1. The proof of theorem 1

Recall that for conjugate Gaussian Bayesian linear regression the exact posterior is available analytically as:

$$p(\beta|X,Y) = \mathcal{N}(\mu_N, \Sigma_N)$$

where

$$\Sigma_N^{-1} = \Sigma_\beta^{-1} + \tau X^T X \;\&\; \mu_N = \Sigma_N \tau X^T Y$$

Where Σ_β is the prior covariance for $p(\beta) = \mathcal{N}(\beta|0, \Sigma_\beta)$, and we have assumed without loss of generality that the prior is zero-mean.

Using an orthonormal projection, U, leads to a normal approximate posterior $\tilde{p}(\beta|X,Y) = \mathcal{N}(\tilde{\mu}_N, \tilde{\Sigma}_N)$ where

$$\tilde{\Sigma}_N^{-1} = \Sigma_\beta^{-1} + \tau U U^T X^T X U U^T \;\&\; \tilde{\mu}_N = \tilde{\Sigma}_N \tau U U^T X^T Y \tag{8}$$

As before, we considering the case in which the projection U is obtained by performing a truncated svd of X. In this setting, we can obtain upper bounds on the approximation error induced which are informative with respect to the quality of the resulting approximation more generally. In the following two subsections, we provide derivations of these bounds for the posterior mean and precision.

C.1.1. Upper bound on the error in the means

We first consider the error in the means when using the approximate likelihood. To do this we consider the error in the gradient.

The gradients of the exact log likelihood and the approximate log likelihood are given by:

$$\begin{aligned} \nabla_\beta \log p(Y|X,\beta) &= -\nabla_\beta \frac{\tau}{2}(X\beta - Y)^T(X\beta - Y) \\ &= -\tau(X^T X\beta + X^T Y) \end{aligned} \tag{9}$$

and

$$\begin{aligned} \nabla_\beta \log \tilde{p}(Y|X,\beta) &= -\nabla_\beta \frac{\tau}{2}(XUU^T\beta - Y)^T(XUU^T\beta - Y) \\ &= -\tau(UU^T X^T X U U^T \beta + UU^T X^T Y) \end{aligned} \tag{10}$$

where τ is the precision of the Gaussian likelihood.

We can thus rewrite and upper bound the difference between the gradients as follows:

$$\begin{aligned}
E(\beta) :&= \|\nabla_\beta \log \tilde{p}(Y|X,\beta) - \nabla_\beta \log p(Y|X,\beta)\|_2 \\
&= \|\tau(-UU^TX^TXUU^T\beta + UU^TX^TY) - \tau(-X^TX\beta + X^TY)\|_2 \\
&= \tau\|(X^TX - UU^TX^TXUU^T)\beta + UU^TX^TY - X^TY\|_2 \\
&= \tau\|\bar{U}\bar{U}^TX^TX\bar{U}\bar{U}^T\beta - \bar{U}\bar{U}^TX^TY\|_2 \\
&= \tau\|\bar{U}\bar{\boldsymbol{\lambda}}^2\bar{U}^T\beta - \bar{U}\bar{\boldsymbol{\lambda}}\bar{V}^TY\|_2 \\
&\leq \tau\|\bar{U}\bar{\boldsymbol{\lambda}}^2\bar{U}^T\beta\| + \|\bar{U}\bar{\boldsymbol{\lambda}}\bar{V}^TY\|_2 \\
&\leq \tau\big(\|\bar{U}\|_{\text{op}}\|\bar{\boldsymbol{\lambda}}^2\|_2\|\bar{U}^T\beta\|_2 + \|\bar{U}\|_{\text{op}}\|\bar{\boldsymbol{\lambda}}\|_2\|\bar{V}^TY\|_2\big) \\
&= \tau\big(\lambda_{M+1}^2\|\bar{U}^T\beta\|_2 + \lambda_{M+1}\|\bar{V}^TY\|_2\big)
\end{aligned} \tag{11}$$

Where the fifth line relies on that $X^TX = U\boldsymbol{\lambda}U^T + \bar{U}\bar{\boldsymbol{\lambda}}\bar{U}^T$, and in the sixth line we use the triangle inequality. Here we see that the error will be small for β when it is captured primarily within the span of U, and when Y is captured within the span of V.

If the responses, Y, and parameters β are bounded by scalars a and b such that $\forall d, |\beta_d| < a$ and $\forall n, |y_n| < b$ we additionally have that:

$$\begin{aligned}
E(\beta) &\leq \tau\big(\lambda_{M+1}^2\|\bar{U}^T\beta\|_2 + \lambda_{M+1}\|\bar{V}^TY\|_2\big) \\
&\leq \tau\big(\lambda_{M+1}^2\sqrt{D}a + \lambda_{M+1}\sqrt{N}b\big)
\end{aligned} \tag{12}$$

Noting that μ_1 and μ_2 are the maximum a posteriori values of β under $p(\beta|X,Y,\alpha)$ and $\tilde{p}(\beta|X,Y,\alpha)$ respectively, and given the strong convexity parameter α on the negative log posterior (e.g from a Gaussian prior precision of αI) this gradient error upper bound constrains the error in means as:

$$\|\mu_1 - \mu_2\|_2 \leq \tau\frac{\lambda_{M+1}^2\sqrt{D}a + \lambda_{M+1}\sqrt{N}b}{\alpha} \tag{13}$$

Which follows from lemma 4, given below. It is worth noting though that when X is not exactly low rank, α will grow as $\Omega(N)$.

Lemma 4 *If the norm of the difference between the gradients of two function f, g mapping $\mathbb{R}^D \to \mathbb{R}$ is bounded by some $c \in \mathbb{R}$ such that $\forall\beta \in \mathbb{R}^D \|\nabla_\beta f(\beta) - \nabla_\beta g(\beta)\|_2 \leq c$ and if both functions are strongly convex with parameter $\alpha > 0$ such that $\forall\beta, v \in \mathbb{R}^D$ with $\|v\|_2 = 1$ $v^T\nabla_\beta^2 h(\beta)v \geq \alpha$ for each $h \in \{f, g\}$ then*

$$\|\beta_f^* - \beta_g^*\|_2 \leq \frac{c}{\alpha}$$

where $\beta_f^* = \arg\min_\beta f(\beta)$ and $\beta_g^* = \arg\min_\beta g(\beta)$.

Proof Taylor's theorem provides that we may write that

$$\begin{aligned}
f(\beta) = f(\beta_g^*) &+ (\beta - \beta_g^*)^T\nabla_\beta f(\beta_g^*) \\
&+ \frac{1}{2}(\beta - \beta_g^*)^T\nabla_\beta^2 f(\hat{\beta})(\beta - \beta_g^*)
\end{aligned} \tag{14}$$

for some $\hat{\beta} \in \{\beta_g^* + (\beta - \beta_g^*)t | t \in [0,1]\}$.

We can next note that

$$\nabla f(\beta) = \nabla_\beta f(\beta_g^*) + (\beta - \beta_g^*)^T \nabla_\beta^2 f(\hat{\beta}) \tag{15}$$

This then allows us to say that

$$\begin{aligned} \|\nabla f(\beta)\|_2 &= \|\nabla_\beta f(\beta_g^*) + (\beta - \beta_g^*)^T \nabla_\beta^2 f(\hat{\beta})\|_2 \\ &= \|\nabla_\beta f(\beta_g^*) - \nabla_\beta g(\beta_g^*) + (\beta - \beta_g^*)^T \nabla_\beta^2 f(\hat{\beta})\|_2 \\ &\geq \|\beta - \beta_g^*\|_2 \|\nabla_\beta^2 f(\hat{\beta})\|_2 - \|\nabla_\beta f(\beta_g^*) - \nabla_\beta g(\beta_g^*)\|_2 \\ &\geq \|\beta - \beta_g^*\|_2 \alpha - c \end{aligned} \tag{16}$$

Where in the second line we use the fact that gradient of g is zero at β_g^*, in the third line we use the triangle inequality and in the fourth line we use the strong convexity of f and the supposed upper bound on the error in the gradients.

Concluding the proof, we note that for $\beta = \beta_f^*$ (where $\|\nabla f(\beta)\|_2 = 0$) in Equation (16) implies that $\|\beta_f^* - \beta_g^*\|_2 \leq \frac{c}{\alpha}$. ■

C.1.2. Error in Posterior Precision

The error in the precision matrices for the approximate and exact posteriors in linear regression are particularly straightforward since they do not depend on the responses, Y. In particular, we have:

$$\begin{aligned} \Sigma_1^{-1} - \Sigma_2^{-1} &= (\Sigma_\beta^{-1} + \tau X^T X) - (\Sigma_\beta^{-1} + \tau U U^T X^T X U U^T) \\ &= \tau X^T X - \tau U U^T X^T X U U^T \\ &= \tau \bar{U} \bar{U}^T X^T X \bar{U} \bar{U}^T \\ &= \tau \bar{U} \bar{\Lambda}^2 \bar{U}^T \end{aligned} \tag{17}$$

Thus the spectral norm of the error in the precisions is precisely $\|\Sigma_1^{-1} - \Sigma_2^{-1}\|_2 = \tau \lambda_{M+1}^2$.

C.2. Low rank data approximations and asymptotic consistency

Perhaps unsurprisingly, by neglecting the information which the data provides outside of the subspace defined by the low rank approximation, some irreducible error is introduced, and as a result the inferences are not asymptotically consistent. In particular, $\tilde{\mu}_N$ does not converge to μ_N. Much of the intuition behind the behavior of $\tilde{\mu}_N$ is captured by the toy example in Figure 3.

The asymptotic behavior of μ_N is characterized by Corollary 2, which we now prove.

Proof We find below that $\tilde{\mu}_N \xrightarrow{p} \Sigma_\beta U (U^T \Sigma_\beta U)^{-1} U^T \beta$. We then appeal to Theorem 5, which shows that this is the vector of minimum Σ_β^{-1}-norm satisfying $U^T \tilde{\mu} = U^T \beta$. Because for any closed $S \subset \mathbb{R}^D$, $\tilde{\mu} = \arg\min_{v \in S} \|v\|_{\Sigma_\beta^{-1}} = \arg\max_{v \in S} -\frac{1}{2} v^T \Sigma_\beta^{-1} v = \arg\max_{v \in S} \mathcal{N}(0, \Sigma_\beta)$, this shows that $\tilde{\mu}$ is the maximum a priori vector satisfying the constraints.

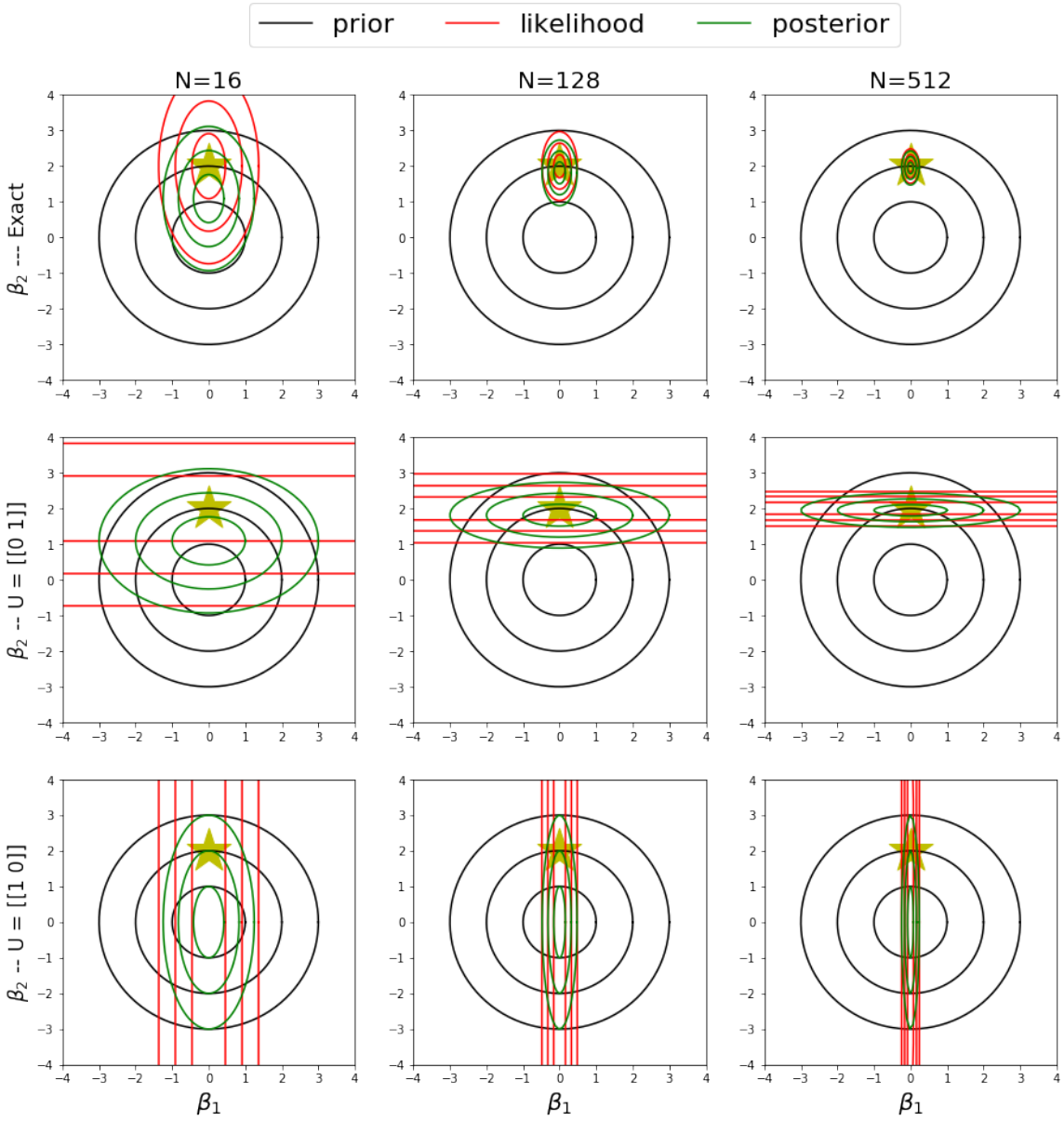

Figure 3: Example of posterior approximations with different projections, U for increasing sample sizes. In all plots the contours of the densities of the prior, likelihood and posterior (or approximations thereof). The top row represents the exact posterior. The middle row is the approximation found by using the best rank 1 approximation to X. The bottom row is the approximation obtained using the orthogonal rank 1 approximation.

From Equation (8) we have that $\tilde{\mu}_N = \tau\tilde{\Sigma}_N UU^T X^T Y$. Noting that $Y = X\beta + \frac{1}{\tau}\epsilon$ for some $\epsilon \in \mathbb{R}^N$ with $\epsilon_i \sim \mathcal{N}(0,1)$, we may expand this out and write:

$$\begin{aligned}
\tilde{\mu}_N &= \tau(\Sigma_\beta^{-1} + UU^T X^T \tau X UU^T)^{-1} UU^T X^T (X\beta + \frac{1}{\tau}\epsilon) \\
&= \tau\left(\Sigma_\beta^{-1} + U(\tau\Lambda^2)U^T\right)^{-1} U\Lambda V^T (V\Lambda U^T\beta + \frac{1}{\tau}\epsilon) \\
&= \left(\Sigma_\beta^{-1} + U(\tau\Lambda^2)U^T\right)^{-1} U(\tau\Lambda^2)(U^T\beta + \Lambda^{-1}V^T\epsilon) \\
&= \Sigma_\beta U(U^T\Sigma_\beta U + \tau^{-1}\Lambda^{-2})^{-1}(U^T\beta + \Lambda^{-1}V^T\epsilon) \\
&\xrightarrow{p} \Sigma_\beta U(U^T\Sigma_\beta U)U^T\beta
\end{aligned} \tag{18}$$

Where in the third line we use the matrix identity, $(W^TQW + R^{-1})^{-1}W^TQ = RW^T(WRW^T + Q^{-1})^{-1}$ (Petersen et al., 2008). Weak convergence in the last line follows from the weak convergence of Λ (Vershynin, 2012) to zero and Slutsky's theorem. ■

Lemma 5 *$\tilde{\mu} := \Sigma_\beta U(U^T\Sigma_\beta U)^{-1}U^T\beta$ is the vector of minimum Σ_β^{-1}-norm satisfying $U^T\tilde{\mu} = U^T\beta$.*

Proof We show that $\beta^* = \Sigma_\beta U(U^T\Sigma_\beta U)^{-1}U^T\beta$ is the vector of minimum norm satisfying the above constraints in the Hilbert space $\mathbb{R}^D$ with inner product $\langle v_1, v_2\rangle = v_1^T\Sigma_\beta^{-1}v_2$ for vectors $v_1, v_2 \in \mathbb{R}^D$.

Define β^* as:

$$\beta^* = \underset{v\in\mathbb{R}^D}{\arg\min} \|v\|_{\Sigma_\beta^{-1}} \text{ subject to } U^Tv = U^T\beta \tag{19}$$

First note that the condition $U^T\beta^* = U^T\beta$ may be expressed as a set the M linear constraints:

$$\langle \Sigma_\beta^{-1}U[:,i], \beta^*\rangle = U[:,i]^T\beta \tag{20}$$

for $i = 1, 2, \ldots, M$. We thereby see that the constraint restricts β^* to the linear variety $\beta + \left[\left\{\Sigma_\beta U[:,i]\right\}_{i=1}^M\right]^\perp$. Where $\left[A\right]$ denotes the subspace generated by the vectors of the set A, and $\left[A\right]^\perp$ denotes the set of all vectors orthogonal to all vectors in $\left[A\right]$ (i.e. the orthogonal complement of $\left[A\right]$).

By the projection theorem (Luenberger, 1969), β^* is orthogonal to $\left[\left\{\Sigma_\beta U[:,i]\right\}_{i=1}^M\right]^\perp$, or $\beta^* \in \left[\left\{\Sigma_\beta U[:,i]\right\}_{i=1}^M\right]^{\perp\perp} = \left[\left\{\Sigma_\beta U[:,i]\right\}_{i=1}^M\right]$. We can therefore write β^* as a linear combination of vectors in $\left[\left\{\Sigma_\beta U[:,i]\right\}_{i=1}^M\right]$:

$$\beta^* = \Sigma_\beta Uc \tag{21}$$

for some $c \in \mathbb{R}^M$.

Our constraints in Equation (20) then demand that $\langle \Sigma_\beta U[:,i], \Sigma_\beta Uc\rangle = U[:,i]^T\beta$ for each i, or equivalently that $U^T\Sigma_\beta\Sigma_\beta^{-1}\Sigma_\beta Uc = U^T\beta$. This implies that $c = (U^T\Sigma_\beta U)^{-1}U^T\beta$. Plugging this into Equation (21) provides that $\beta^* = \Sigma_\beta U(U^T\Sigma_\beta U)^{-1}U^T\beta$, as desired. ■

C.3. Proof of corollary on more conservative posterior uncertainties

We here show for conjugate Bayesian regression, that under $\tilde{p}$ the uncertainty (i.e. posterior variance) for any linear combination of parameters, $\mathrm{Var}_{\tilde{p}}[v^T\beta]$ is no smaller than the exact posterior variance, thereby proving Corollary 3.

Proof

First, we note that this statement is formally equivalent to stating that $v^T\tilde{\Sigma}_N v \geq v^T\Sigma_N v$, or that $E := \tilde{\Sigma}_N - \Sigma_N \succeq 0$ (where $\succeq$ denotes positive definiteness). By Theorem 1, $\Sigma_N - \tilde{\Sigma}_N = \bar{U}\mathrm{diag}(\hat{\boldsymbol{\lambda}}^2)\bar{U}^T \succeq 0$. Since this implies that the inverse of the difference of these matrices is positive definite, we can then see that $(\Sigma_N^{-1} - \tilde{\Sigma}_N^{-1})^{-1} = \tilde{\Sigma}_N(\tilde{\Sigma}_N - \Sigma_N)^{-1}\Sigma_N \succeq 0$. Because, as valid covariance matrices, Σ_N and $\tilde{\Sigma}_N$ are both positive definite, and because inverses and product of positive definite matrices are positive definite, this implies that $\tilde{\Sigma}_N^{-1}\tilde{\Sigma}_N(\tilde{\Sigma}_N - \Sigma_N)^{-1}\Sigma_N\Sigma_N^{-1} = (\tilde{\Sigma}_N - \Sigma_N)^{-1} \succeq 0$. Finally, this implies that $\tilde{\Sigma}_N - \Sigma_N \succeq 0$ as desired. ■

C.4. A quantitative assessment of information loss due to our approximation

We know take an information theoretic perspective to assess the loss of information due to the approximation. We find that our approximation reduces the information we extract about parameters from our data. Concretely, we look add the reduction of entropy in the approximate posterior relative to the exact posterior (MacKay, 2003), where entropy is defined as:

$$H\big(p(\beta)\big) := \mathbb{E}_p[-\log_2 p(\beta)]$$

Corollary 6 *When using an isotropic Gaussian prior $\Sigma_\beta = \sigma_\beta^2 I$ the information loss relative to the exact posterior (in nats) is upper bounded as $H\big(\tilde{p}(\beta|X)\big) - H\big(p(\beta|X)\big) \leq \frac{\tau}{2\sigma_\beta^2}\sum_{i=M+1}^{D}\lambda_{N,i}^2$. In other words, when U is obtained via an M truncated SVD, less than $\frac{\tau}{2\sigma_\beta^2}\lambda_{N,M+1}^2$ additional nats of information would have been provided by using the $M+1-$ truncated SVD.*

Proof The entropy of the exact and approximate posteriors are given as:

$$\begin{aligned}
H(p) &= -\frac{1}{2}\log|2\pi e\Sigma_N^{-1}| \\
&= -\frac{1}{2}(D\log 2\pi e - \sum_{i=1}^{D}\log\sigma_\beta^{-2} + \tau\lambda_{N,i}^2) \text{ and} \\
H(\tilde{p}) &= -\frac{1}{2}\log|2\pi e\tilde{\Sigma}_N^{-1}| \\
&= -\frac{1}{2}(D\log 2\pi e - \sum_{i=1}^{M}\log\sigma_\beta^{-2} + \tau\lambda_{N,i}^2 - (D-M)\log\sigma_\beta^{-2})
\end{aligned} \tag{22}$$

Therefore

$$
\begin{aligned}
&H\big(\tilde{p}(\beta|X)\big) - H\big(p(\beta|X)\big) \\
&= -\frac{1}{2}\sum_{i=M+1}^{D} \log \sigma_\beta^{-2} + \frac{1}{2}\sum_{i=M+1}^{D} \log \sigma_\beta^{-2} + \tau\lambda_{N,i}^2 \\
&= \frac{1}{2}\sum_{i=M+1}^{D} \log 1 + \frac{\tau}{\sigma_\beta^{-2}}\lambda_{N,i}^2 \\
&\leq \frac{1}{2}\sum_{i=M+1}^{D} \frac{\tau}{\sigma_\beta^{-2}}\lambda_{N,i}^2
\end{aligned}
\tag{23}
$$

■

Appendix D. Additional details & theory on low rank approximations in non-conjugate linear models

Low rank approximations can provide computational gains in non-conjugate linear models. In what follows we provide some theoretical justification for this approach which is demonstrated empirically in Figures 1, 2, 4 and 5. Our results and analysis are in the context of logistic regression, but the approaches are general to other Bayesian generalized linear models with log concave posteriors (as is the case for several widely used GLMs and priors). These results above show that the error induced by the proposed approximation is controlled to some extent.

D.1. Logistic regression

Before diving in, we briefly review logistic regression. In particular the logistic regression mapping function (Huggins et al., 2017) is given as

$$
\phi(y_n, x_n \cdot \beta) = -\log\big(1 + \exp\{-y_n x_n \cdot \beta\}\big) \tag{24}
$$

Where each $y_n \in \{-1, 1\}$.

In our bounds for the approximation quality in Bayesian logistic regression below we will take advantage of the first and second derivatives of the mapping function and bounds on their absolute values:

$$
\frac{d}{da}\phi(y, a) = y\phi'(ya) = y\frac{\exp\{-ya\}}{1 + \exp\{-ya\}} \tag{25}
$$

Notably, $\forall a \in \mathbb{R}, y \in \{0, 1\}$, $|\phi'(y, a)| < 1$.

$$
\frac{d^2}{da^2}\phi(y, a) = \phi''(ya) = \frac{-\exp\{a\}}{(1 + \exp\{a\})^2} \tag{26}
$$

Notably, $\forall a \in \mathbb{R}, y \in \{0,1\}$, $-\frac{1}{4} \leq \phi''(y,a) < 0$. This implies that the Hessian of the negative log likelihood will be positive semi-definite everywhere.

Additionally, if we have a zero-mean, isotropic Gaussian prior such that $p(\beta) = \mathcal{N}\big(\beta|\mu = \mathbf{0}, \Sigma = \text{diag}(\boldsymbol{\sigma_\beta})\big)$ for some $\sigma_\beta \in \mathbb{R}^+$, then we have that:

$$-\big(\nabla^2_\beta \log p(\beta) + \nabla^2_\beta \log p(Y|X,\beta)\big) \succeq \frac{1}{\sigma^2_\beta} I + \frac{1}{4} X^T X$$

$$\frac{d^3}{da^3}\phi(y,a) = \phi'''(a) = \frac{\big(\exp\{a\}(\exp(-a)-1)\big)}{(1+\exp\{a\})^3} \tag{27}$$

Notably, $\forall a \in \mathbb{R}, -\frac{1}{6\sqrt{3}} \leq \phi'''(a) \leq \frac{1}{6\sqrt{3}}$.

D.2. Fast Laplace approximations for logistic regression

The Laplace approximation refers to a Gaussian approximation to a distribution defined by a 2nd order Taylor approximation of the log density. When applied to Bayesian inference this local approximation is typically formed at the maximum a posteriori (MAP) parameter. As such a Laplace approximation applied to our proposed likelihood approximation is given by:

$$\log \hat{p}(\beta|X,Y) = \log \mathcal{N}\big(\beta\big|\mu = \beta^*, \Sigma = -H^{-1}\big) \tag{28}$$

Where $\beta^* = \arg\max_\beta \log \tilde{p}(\beta|X,Y)$, $H = \nabla^2_\beta \log \tilde{p}(\beta|X,Y)\Big|_{\beta=\beta^*}$, and can we neglect the first order term since $\nabla \log \tilde{p}(\beta|X,Y)\Big|_{\beta=\beta^*} = 0$.

While typically much computationally cheaper than MCMC, Laplace approximations become expensive or intractable for high dimensional problems as they demand inverting a D by D Hessian, which is in general an $O(D^3)$ operation. In contrast, with Theorem 7, we show that the Laplace approximation of $\tilde{p}(\beta|X,Y)$ may be computed more efficiently.

Theorem 7 *Given projected data $XU \in \mathbb{R}^{N,M}$ and a diagonal Gaussian prior $p(\beta) = \mathcal{N}(\mu_\beta, \Sigma_\beta)$ and $\tilde{\beta}^* = \arg\max_\beta \tilde{p}(\beta|XU,Y)$, forming a Laplace approximation to $\tilde{p}(\beta|X,Y)$ requires only $O(DM^2 + NM^2)$ time and $O(DM)$ memory.*

The resulting concise representation of the covariance is then easier to store, query, and to use to make predictions.

Proof The Laplace approximation is defined by a mean and covariance, with the mean given ($\tilde{\beta}^*$) all that remains is calculating $\tilde{\Sigma}_N$. This is taken to be the inverse of the Hessian of the negative log posterior, H^{-1}, where H is given as:

$$\begin{aligned} H :&= \nabla^2_\beta - \log \tilde{p}(\beta|X,Y) \\ &= \Sigma^{-1}_\beta - \sum_{i=1}^{N} \frac{d^2}{da^2} \log p(y_i|x^T\beta = a)|_{a=x_i^T UU^T\beta} UU^T x_i x_i^T UU^T \\ &= \Sigma^{-1}_\beta - UU^T X^T \phi'' X UU^T \end{aligned} \tag{29}$$

Where $\phi''_i = \frac{d^2}{da^2} \log p(y_i|x^T\beta = a)|_{a=x_i^T UU^T\beta}$.

The Woodbury matrix lemma provides that we may compute $\tilde{\Sigma}_N := H^{-1}$ as

$$\tilde{\Sigma}_N = \Sigma_\beta - \Sigma_\beta U\big(U^T\Sigma_\beta U - (U^T X^T \phi'' XU)^{-1}\big)U^T\Sigma_\beta$$

Computing $U^T\Sigma_\beta U$ requires the multiplication of 2, M by D matrices which requires $O(DM^2)$ time, and inverting H requires M^3 time. Additional matrix products require $O(DM^2)$ time.

■

We additionally find in Theorem 8 that the result approximation has bounded errors in estimated posterior means and precision relative to the usual Laplace approximation.

Theorem 8 *The Laplace approximation applied to logistic regression with a low rank approximation the design induces an error in estimate of the posterior means and covariances relative to the exact Laplace approximation which is upper bounded as:*

$$\|\mathbb{E}_{\tilde{p}}[\beta] - \mathbb{E}_p[\beta]\|_2 \leq \alpha^{-1}\lambda_{M+1}\big(\sqrt{N} + \frac{\lambda_1\sqrt{D}a}{4}\big) \tag{30}$$

Where $\lambda_1 := \|X\|_2$ *and* $\lambda_{M+1} = \|X - XUU^T\|_2$ *We additionally assume that* $\forall d\,|\beta_d| < a$ *and that the exact posterior is strongly* $\alpha-$*log concave.*

Additionally, we have that the error in the posterior precision is upper bounded as:

$$\|\nabla^2_\beta \log p(\beta|X,Y) - \nabla^2_\beta \log \tilde{p}(\beta|X,Y)\|_2 \leq \lambda^2_{M+1}\frac{\sqrt{N}}{4} + \lambda_{M+1}\lambda_1^2\|\beta\|_2\frac{1}{6\sqrt{3}}\sqrt{N} \tag{31}$$

Theorem 8 extends the worst case bounds on approximation error from the case of linear regression to logistic regression. Mirroring our approach to linear regression taken earlier, we will derive upper bounds on $\|\mu_1 - \mu_2\|_2$ and $\|\Sigma_1^{-1} - \Sigma_2^{-1}\|_2$.

Bound on error of the mean To begin, recall that the gradient of the log likelihood of the model and its approximation are given as:

$$\nabla_\beta \log p(Y|X,\beta) = \sum_{i=1}^N \phi'(y_i, x_i^T\beta)x_i \tag{32}$$

and

$$\nabla_\beta \log \tilde{p}(Y|X,\beta) = \sum_{i=1}^N \phi'(y_i, x_i^T UU^T\beta)UU^T x_i \tag{33}$$

As such, we may write the norm of error in the gradient as

$$\begin{aligned} E(\beta) :&= \|\nabla_\beta \log p(Y|X,\beta) - \nabla_\beta \log \tilde{p}(Y|X,\beta)\|_2 \\ &= \|\sum_{i=1}^N \phi'(y_i, x_i^T\beta)x_i - \phi'(y_i, x_i^T UU^T\beta)UU^T x_i\| \end{aligned} \tag{34}$$

We note that using Taylor's theorem we may rewrite for each x_i:

$$\phi'(y_i, x_i^T UU^T\beta) = \phi'(y_i, x_i^T\beta) + (x_i^T\beta - x_i UU^T\beta)\phi''(a_i) \tag{35}$$

For some $a_i \in [x_i^T\beta, x_i^T UU^T\beta]$.

For convenience, we now introduce vectorized notation for the above:

$$\phi' = \begin{bmatrix} y_1\phi'(y_1 x_1^T\beta) \\ y_2\phi'(y_2 x_2^T\beta) \\ \cdots \\ y_N\phi'(y_N x_N^T\beta) \end{bmatrix} \quad , \phi'' = \begin{bmatrix} \phi''(a_1) \\ \phi''(a_2) \\ \cdots \\ \phi''(a_N) \end{bmatrix}$$

And now can rewrite the gradient error as

$$\begin{aligned}
E(\beta) &= \|X^T\phi' - UU^T X^T\big(\phi' + (X\beta - XUU^T\beta)\circ\phi''\big)\|_2 \\
&= \|(X^T - UU^T X^T)\phi' - UU^T X^T(X\beta - XUU^T\beta)\circ\phi''\|_2 \\
&\leq \|(X^T - UU^T X^T)\phi'\|_2 + \|UU^T X^T(X\beta - XUU^T\beta)\circ\phi''\|_2 \\
&= \|\bar{U}\bar{U}^T X^T\phi'\|_2 + \|UU^T X^T(\phi''\circ X\bar{U}\bar{U}^T\beta)\|_2 \\
&= \|\bar{U}\bar{\Lambda}\bar{V}^T\phi'\|_2 + \|U\Lambda V^T(\phi''\circ\bar{V}\bar{\Lambda}\bar{U}^T\beta)\|_2 \\
&\leq \|\bar{U}\bar{\Lambda}\|_{\text{op}}\|\bar{V}^T\phi'\|_2 + \|U\Lambda V^T\|_{\text{op}}\|\phi''\circ\bar{V}\bar{\Lambda}\bar{U}^T\beta\|_2 \\
&= \lambda_{M+1}\|\phi'\|_2 + \lambda_1\sqrt{\langle\phi''\circ\bar{V}\bar{\Lambda}\bar{U}^T\beta, \phi''\circ\bar{V}\bar{\Lambda}\bar{U}^T\beta\rangle} \\
&= \lambda_{M+1}\|\phi'\|_2 + \lambda_1\sqrt{\sum_{i=1}^N {\phi''_i}^2(\bar{V}_i^T\bar{\Lambda}\bar{U}^T\beta)^2}
\end{aligned} \tag{36}$$

Where in the third line we use the triangle inequality. Similar to our analysis of the approximation error induced to linear regression, the $\|\bar{V}^T\phi'\|_2$ term in line six indicates that our error is smaller when the components of ϕ' outside of the span of V^T is small.

Taking additional knowledge of constraints into account, in particular that $\forall i\,|\phi'_i| < 1$ and $|\phi''| \leq \frac{1}{4}$ and that $\forall d\,|\beta_d| < a$, we can further bound the error as:

$$\begin{aligned}
E(\beta) &\leq \lambda_{M+1}\|\phi'\|_2 + \lambda_1\sqrt{\sum_{i=1}^N {\phi''_i}^2(\bar{V}_i^T\bar{\Lambda}\bar{U}^T\beta)^2} \\
&\leq \lambda_{M+1}\sqrt{N} + \lambda_1\sqrt{\sum_{i=1}^N(-\frac{1}{4})^2(\bar{V}_i^T\bar{\Lambda}\bar{U}^T\beta)^2} \\
&= \lambda_{M+1}\sqrt{N} + \lambda_1\frac{1}{4}\sqrt{\sum_{i=1}^N(\bar{V}_i^T\bar{\Lambda}\bar{U}^T\beta)^2} \\
&= \lambda_{M+1}\sqrt{N} + \lambda_1\frac{1}{4}\|\bar{V}\bar{\Lambda}\bar{U}^T\beta\|_2 \\
&= \lambda_{M+1}\sqrt{N} + \lambda_1\frac{1}{4}\|\bar{\Lambda}\|_2\|\bar{U}^T\beta\|_2 \\
&= \lambda_{M+1}\big(\sqrt{N} + \frac{\lambda_1\sqrt{D}a}{4}\big)
\end{aligned} \tag{37}$$

Where in the second line we rely on the monotonicity of $\sqrt{x}$. In the fifth line we see that, as in the case of linear regression, our error will be small when most of β is captured within the span of U as this reduces the magnitude of the second term. Additionally we see that when the data is modeled well (i.e. each ϕ_i'' and ϕ_i' will be small and so the approximation error will be small as well.

Error in Precisions We first recall that the Hessian of the log posterior with and without the approximation are given as:

$$\nabla_\beta^2 \log p(\beta|X,Y) = \Sigma_\beta^{-1} + \sum_{i=1}^N \phi''(x_i^T\beta)x_i x_i^T$$

and

$$\nabla_\beta^2 \log \tilde{p}(\beta|X,Y) = \Sigma_\beta^{-1} + \sum_{i=1}^N \phi''(x_i^T\beta)UU^T x_i x_i^T UU^T$$

We note that using Taylor's theorem we can write for each x_i:

$$\begin{aligned}\phi''(y_i, x_i^T UU^T\beta) = \phi''(y_i, x_i^T\beta) \\ + (x_i^T\beta - x_i UU^T\beta)\phi'''(a_i)\end{aligned} \tag{38}$$

For some $a_i \in [x_i^T\beta, x_i^T UU^T\beta]$.

For convenience, we now introduce vectorized notation for the above:

$$\phi'' = \begin{bmatrix}\phi'(x_1^T\beta)\\ \phi'(x_2^T\beta)\\ \cdots \\ \phi'(x_N^T\beta)\end{bmatrix}, \phi''' = \begin{bmatrix}\phi'''(a_1)\\ \phi'''(a_2)\\ \cdots \\ \phi'''(a_N)\end{bmatrix}$$

$$\begin{aligned}E(\beta) &= \|\nabla_\beta^2 \log p(\beta|X,Y) - \nabla_\beta^2 \log \tilde{p}(\beta|X,Y)\|_2 \\ &= \|X^T\phi''X - UU^TX\big(\phi'' + (X\beta - XUU^T\beta)\phi'''\big)XUU^T\|_2 \\ &\le \|X^T\phi''X - UU^TX\phi''XUU^T\|_2 + \|UU^TX(X\beta - XUU^T\beta)\phi'''XUU^T\|_2 \\ &= \|[U,\bar{U}]^T\bar{U}\bar{U}^TX\phi''X\bar{U}\bar{U}^T)[U,\bar{U}]\|_2 + \|UU^TX(X\bar{U}\bar{U}^T\beta\phi''')XUU^T\|_2 \\ &= \|\bar{\Lambda}\bar{V}\phi''\bar{V}\bar{\Lambda}\|_2 + \|U\Lambda V^T(\bar{V}\bar{\Lambda}\bar{U}^T\beta\phi''')V\Lambda U^T\|_2 \\ &\le \lambda_{M+1}^2\|\bar{V}\phi''\bar{V}\|_2 + \lambda_1^2\|\bar{V}\bar{\Lambda}\bar{U}^T\beta\|_2\|\phi'''\|_2 \\ &\le \lambda_{M+1}^2\|\phi''\|_2 + \lambda_{M+1}\lambda_1^2\|\beta\|_2\|\phi'''\|_2\end{aligned} \tag{39}$$

Where in the third line we use the triangle inequality. In the fourth line we use that a rotation does not change the spectral norm of a matrix.

Incorporating additional constraints we can provide a more complete upper bound. In particular, noting that equations 26 and 27 implies that each $|\phi''(x_i^T\beta)| < \frac{1}{4}$ and $|\phi'''(x_i^T\beta)| < \frac{1}{6\sqrt{3}}$, respectively, and supposing that $\forall d\, |\beta_d| < a$, we can further refine the bound on the error as:

$$\begin{aligned}E(\beta) &\le \lambda_{M+1}^2\|\phi''\|_2 + \lambda_{M+1}\lambda_1^2\|\beta\|_2\|\phi'''\|_2 \\ &\le \lambda_{M+1}^2\frac{\sqrt{N}}{4} + \lambda_{M+1}\lambda_1^2 a\frac{1}{6\sqrt{3}}\sqrt{ND}\end{aligned} \tag{40}$$

D.3. Factorized Laplace approximations underestimate marginal variances

We here illustrate that the factorized Laplace approximation underestimates marginal variances. Consider a toy case of a bivariate Gaussian with

$$\Sigma = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix},$$

for which the Hessian evaluated anywhere is

$$\Sigma^{-1} = \begin{bmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{bmatrix},$$

Ignoring off diagonal terms and inverting to approximate Σ_N, as is done by a diagonal Laplace approximation, yields:

$$\tilde{\Sigma} = \begin{bmatrix} \frac{3}{4} & 0 \\ 0 & \frac{3}{4} \end{bmatrix}$$

This approximation underestimates marginal variances.

Appendix E. Additional Experimental Details and Results

E.1. Additional Figures for fast MCMC

We here include results analogous to those in the main text for Laplace approximations using low rank data approximations to perform faster MCMC using HMC with (Carpenter et al., 2017), in Figures 4 and 5.

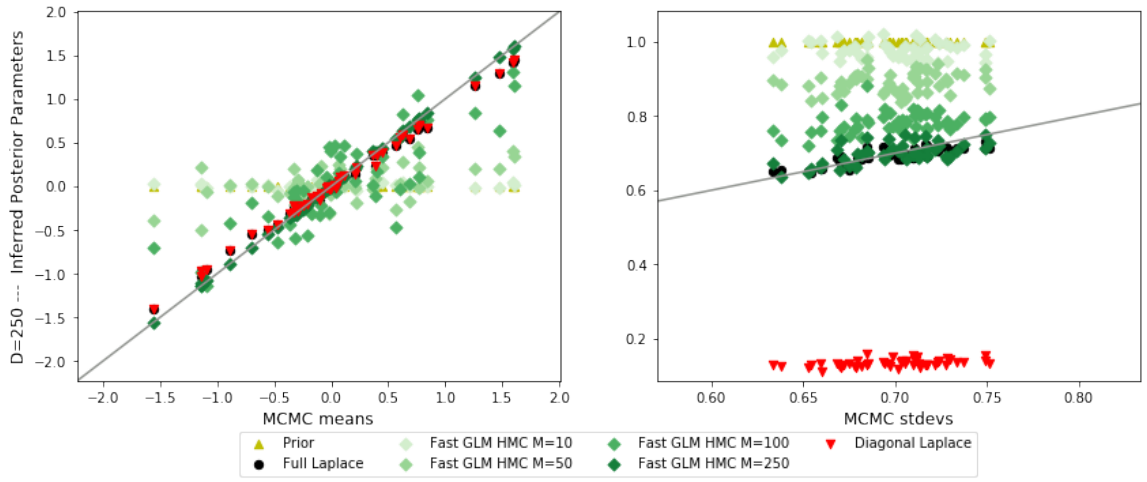

Figure 4: This figure mirrors Figure 1, but examines the trade-off between computation and accuracy for HMC with the low rank approximations rather than Laplace.

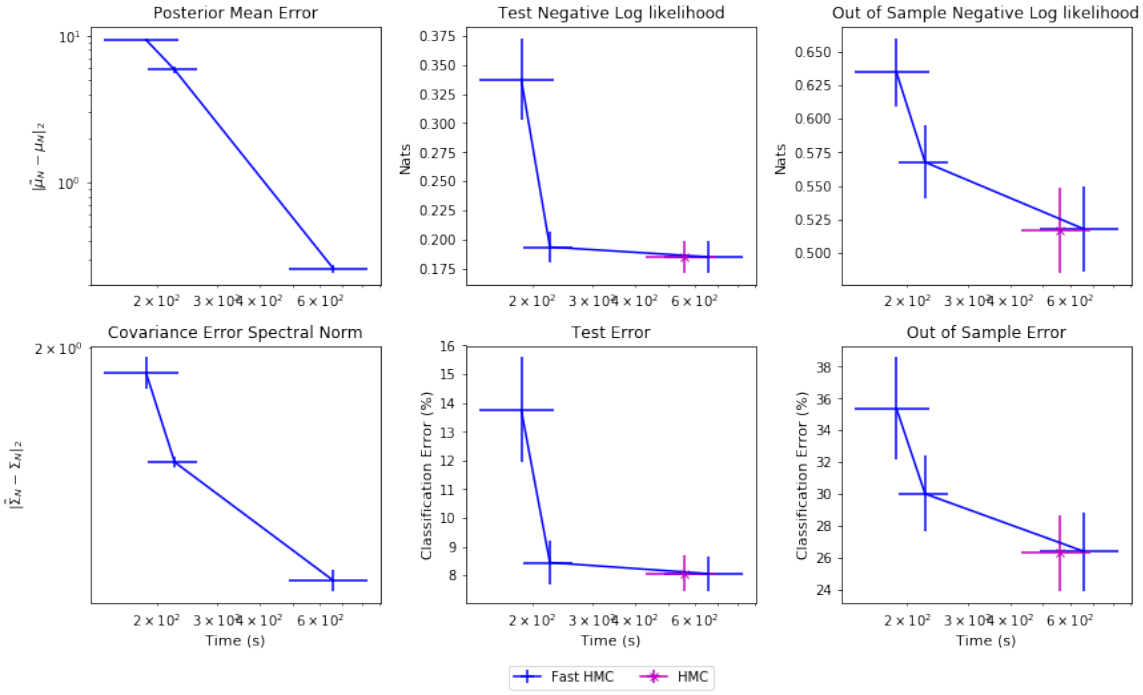

Figure 5: This figure mirrors Figure 2 but with HMC rather than Laplace. Additionally, D=250 instead of D=2000

E.2. Experimental Details

For all experiments we sampled β from an isotropic Gaussian prior with unit variance. For all synthetic data results we first generated a design matrix by sampling from a zero-mean Gaussian with diagonal covariance Σ^2 with each $\Sigma^2_{i,i} = 5 * 1.05^{-i}$. We then used a Scikit-Learn (Pedregosa et al., 2011) implementation of a randomized SVD algorithm due to (Halko et al., 2010), computed from two iterations (i.e. passes through X).

To assess the robustness, in all experiments we used 3 or more replicate experiments, defined by independently generated synthetic datasets or train/test splits as well as re-rerun the randomized truncated SVD.

Because the performance of the diagonal Laplace approximation is dependent upon the shape the exact posterior at β^{MAP}, using a dataset with axis aligned covariance structure gives this method an unrealistic advantage given that in most real applications we do not believe that our low rank structure will be axis aligned. As such, for all synthetic data experiments presented, we randomly generated a basis of orthonormal vectors and used this to rotate our the design matrix. This preserves the eigenspectrum of the data but eliminates the axis alignment of the synthetic data.

All experiments fit to $N = 2500$ training examples. Results on "Out of Sample Data" (in Figures 2 and 5) were obtained by sampling X from an alternative distribution over covariates. Specifically, these out of sample covariates were generated in the manner described above, but with a different random rotation matrix.

We found MAP estimation using $L - BFBS - B$ to be the most efficient of several available options in the scipy optimize library, and used this method in all MAP estimation and Laplace approximation experiments.

For all Bayesian predictions, we use the probit approximation to the logistic function to enable fast approximation (Bishop, 2006).

Acknowledgments

The authors thank Raj Agrawal for helpful comments and discussion. This research is supported in part by an NSF CAREER Award, an ARO YIP Award, a Google Faculty Research Award, and ONR and an NSF GRFP grant.

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

D Andrew Brown, Arvind Saibaba, and Sarah Vallélian. Low rank independence samplers in bayesian inverse problems. *arXiv preprint arXiv:1609.07180*, 2016.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

Leo N Geppert, Katja Ickstadt, Alexander Munteanu, Jens Quedenfeld, and Christian Sohler. Random projections for bayesian regression. *Statistics and Computing*, 27(1): 79–101, 2017.

Rajarshi Guhaniyogi and David B Dunson. Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512):1500–1514, 2015.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *arXiv*, 2010. ISSN 0036-1445. doi: 10.1137/090771806. URL http://arxiv.org/abs/0909.4061.

Jonathan Huggins, Ryan P Adams, and Tamara Broderick. Pass-glm: polynomial approximate sufficient statistics for scalable bayesian glm inference. In *Advances in Neural Information Processing Systems*, pages 3611–3621, 2017.

Jaeyong Lee and Hee-Seok Oh. Bayesian regression based on principal components for high-dimensional data. *Journal of Multivariate Analysis*, 117:175–192, 2013.

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1969.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Akihiko Nishimura and Marc A Suchard. Prior-preconditioned conjugate gradient for accelerated gibbs sampling in" large n & large p" sparse bayesian logistic regression models. *arXiv preprint arXiv:1810.12437*, 2018.

Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and douard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning*, 12:2825–2830, 2011. ISSN ISSN 1533-7928. doi: 10.1007/s13398-014-0173-7.2.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Alessio Spantini, Antti Solonen, Tiangang Cui, James Martin, Luis Tenorio, and Youssef Marzouk. Optimal low-rank approximations of bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487, 2015.

Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Jialei Wang, Jason D Lee, Mehrdad Mahdavi, Mladen Kolar, Nathan Srebro, et al. Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data. *Electronic Journal of Statistics*, 11(2):4896–4944, 2017.

Lijun Zhang, Mehrdad Mahdavi, Rong Jin, Tianbao Yang, and Shenghuo Zhu. Random projections for classification: A recovery approach. *IEEE Transactions on Information Theory*, 60(11):7300–7316, 2014.