A. Additional Experimental Details and Empirical Results

A.1. 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 with each $\Sigma_{i,i} = 5 * 1.05^{-i}$. We then used a scikit-learn (Pedregosa et al., 2011) implementation of a randomized SVD algorithm due to Halko et al. (2011), computed from two iterations (i.e., passes through X).

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

The performance of the Diagonal Laplace approximation is dependent upon the shape the exact posterior at β^{MAP}. In particular, using a dataset with axis aligned covariance structure gives Diagonal Laplace an unrealistic advantage given that in most real applications we do not believe that 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 basis to rotate our the design matrix. This rotation preserves the spectral decay of the data but eliminates the axis alignment of the synthetic data.

In all experiments we consider $N = 2{,}500$ training examples. We obtained results on "Out of Sample Data" (in Figures A.1 and A.5) by sampling X from an alternative distribution over covariates. Specifically, we generated these out-of-sample covariates 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, Chap. 4.5).

A.2. Additional Figures

In Figure A.1 we present results on prediction performance, in term of classification error, as well as negative log likelihood, reported for "Training", "Test", and "Out of Sample Data". In Figure A.2 we report the error of LR-Laplace and Random-Laplace relative to NUTS for estimation of posterior means and variances. We see here that the estimates exhibit behavior increasingly similar to that of the prior as the rank of the approximation, M, decreases. Next, Figure A.3 depicts the same error trends for LR-MCMC using NUTS in `Stan`. We report calibration performance of the approximations of interest for credible sets of parameters (Figure A.4) as well as for prediction (Figure A.5).

We additionally include results analogous to those in the main text for Laplace approximations using low-rank data approximations to perform faster MCMC using NUTS with `Stan` (Carpenter et al., 2017), in Figure A.6. Finally, we also here provide the relative error of posterior mean and standard deviation estimation for logistic regression with a regularized horseshoe prior using the LR-MCMC approximation in Figure A.7. This experiment uses `Stan` for inference as well.

A.2.1. Horseshoe logistic regression experiment

For the logistic regression experiment using a regularized horseshoe prior we used $N = 1{,}000$ data points of dimension $D = 200$. We used ten non-zero effects, each of size 10. Our implementation of the regularized horseshoe and inference in `Stan` closely followed M. Betancourt's "Bayes Sparse Regression" case study.[6] We generated covariates as described in the previous section.

A.3. Stan Model Code

First we show `Stan` code for Bayesian logistic regression.

```
data {
 int<lower=1> N; // # of data
 int<lower=1> D; // # of covariates
 matrix[N, D] X; // Design matrix
```

[6] https://betanalpha.github.io/assets/case_studies/bayes_sparse_regression.html

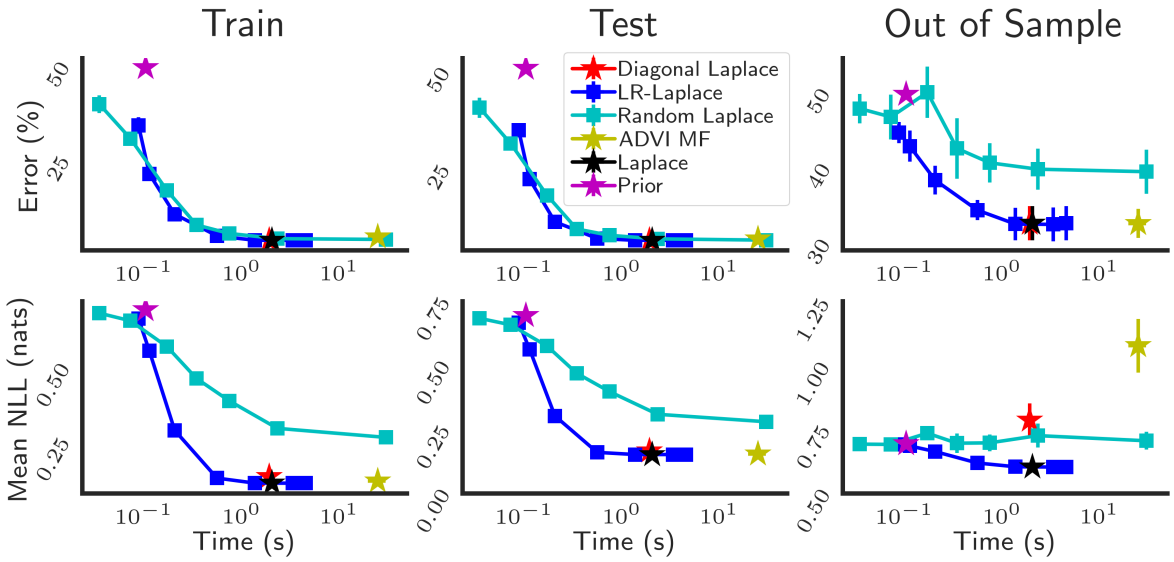

Figure A.1. Predictive performance of posterior approximations in Bayesian logistic regression in terms of (Top) classification error and (Bottom) average negative log likelihood (NLL) of responses under approximate posterior predictive distributions on (Left) *train*, (Center) *test* and (Right) *out of sample* datasets. Lower is better.

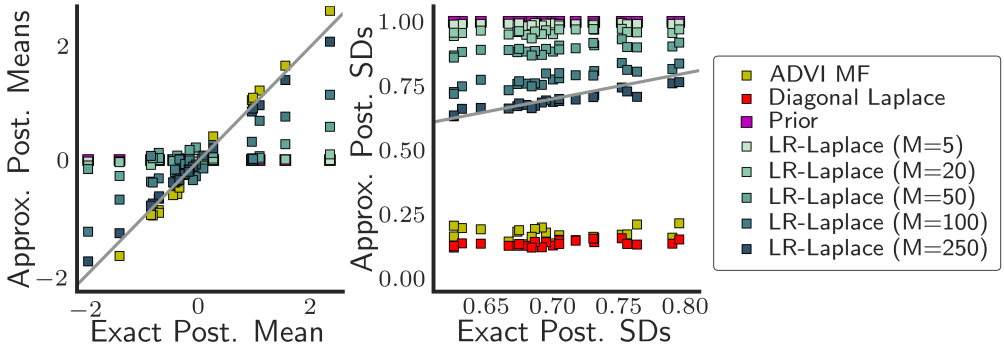

Figure A.2. Approximate posterior mean and standard deviation across a parameter subset as M varies. Horizontal axis represents ground truth from running NUTS using `Stan` without the LR-GLM approximation. $D = 250$.

```
 int<lower=0> y[N]; // labels
 real<lower=0> sigma;
}
parameters {
 vector[D] beta;
}
model {
 beta ~ normal(0, sigma);
 y ~ bernoulli_logit(X * beta);
}
```

Second, we show `Stan` code for logistic regression with our low-rank approximation.

```
data {
 int<lower=1> N; // # of data
 int<lower=1> D; // # of covariates
 int<lower=1> M; // Projected dimension
```

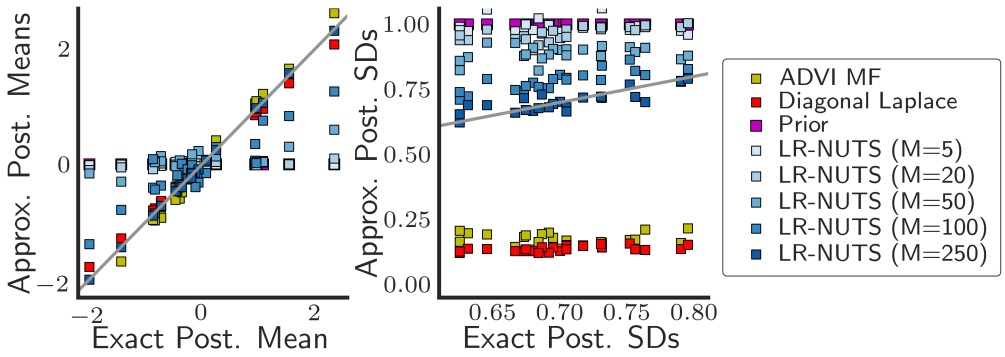

Figure A.3. This figure is analogous to Figure A.2 but examines the trade-off between computation and accuracy of LR-MCMC using NUTS in `Stan`. $D = 250$.

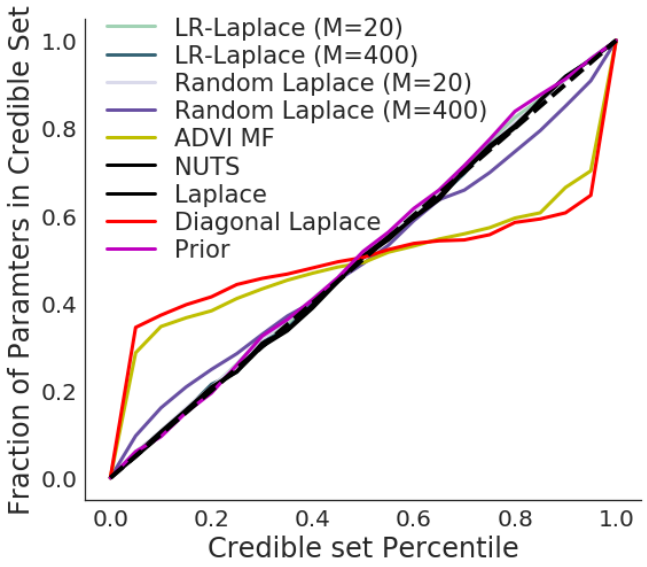

Figure A.4. Credible set calibration. The fraction of parameters in the credible sets defined by different lower tail intervals as a function of the approximate posterior probability of parameters taking values in that interval. The black dotted line (on the diagonal) reflects perfect calibration.

```
 matrix[D, M] U; // Projection matrix
 matrix[N, M] barX; // Projected design matrix
 int<lower=0> y[N]; // labels
 real<lower=0> sigma;
}
parameters {
 vector[D] beta;
}

transformed parameters {
 vector[M] bar_beta = U' * beta;
}
model {
 beta ~ normal(0, sigma);
 y ~ bernoulli_logit(barX * bar_beta);
}
```

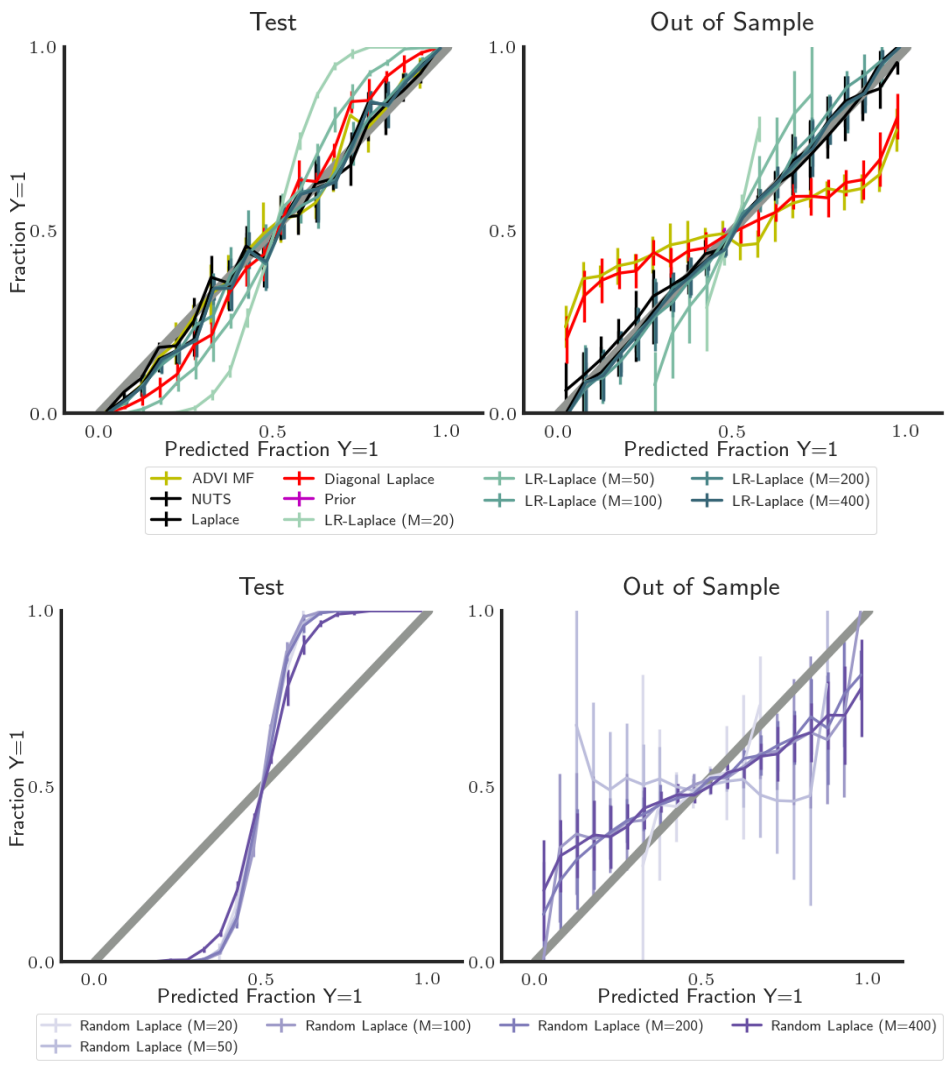

Figure A.5. Prediction calibration.

B. Related Work on Scalable Bayesian Inference

Developing scalable approximate Bayesian inference for models with many parameters (large D) and many data points (large N) has been active area of research for decades, and researchers have developed a large variety of methods applicable to GLMs. Historically, Markov chain Monte Carlo (MCMC) methods based on the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) have been dominant. However MCMC is computationally expensive on large-scale problems in which both D and N are very large. In particular, each likelihood evaluation requires $O(DN)$ time, due to the matrix vector product $X\beta$. Further, estimating posterior covariances uniformly well requires $O(\log D)$ samples (Cai et al., 2010). Therefore, the total cost of collecting those samples is $O(ND \log D)$ time in the case of perfect, independent Monte Carlo samples. In practice, though, mixing times may also have unfavorable scaling with dimensionality and sample size; these issues can lead to even worse scaling in N and D. Several lines of research have explored the use of subsampling methods to reduce the dependence on N. But these methods either lose the asymptotic guarantees of exact MCMC or fail to provide faster inference in practice due to poor mixing behavior (Bardenet et al., 2017).

Other work has pursued deterministic approximations to the Bayesian posterior. Some of the most widely used of these approximations include (1) the Laplace approximation, which is a Gaussian approximation of the posterior defined locally at the posterior mode, (2) extensions of the Laplace approximation such as the integrated nested Laplace approximation (INLA)

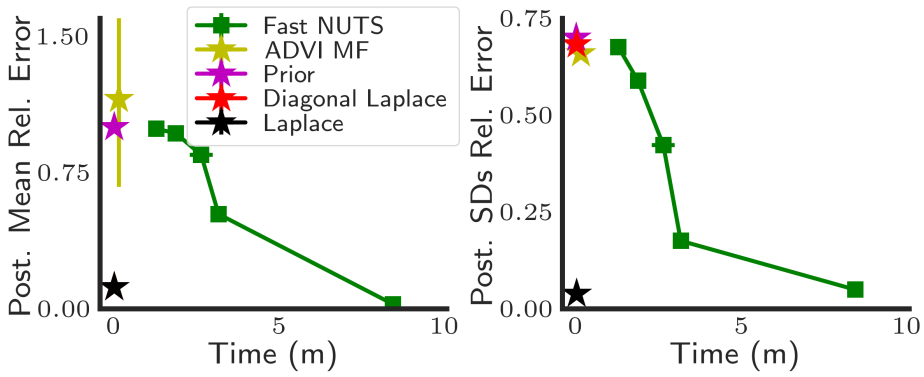

Figure A.6. This figure is analogous to Figure 2A but assesses LR-MCMC using NUTS in `Stan` rather than LR-Laplace. $D = 250$.

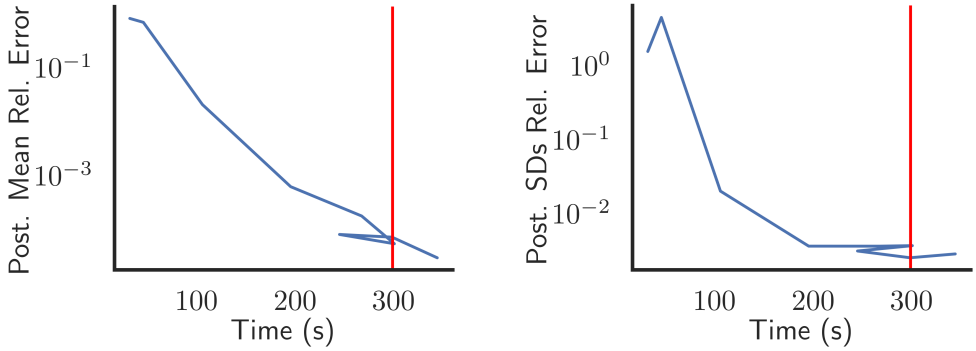

Figure A.7. Bayesian logistic regression with a regularized Horseshoe prior using NUTS in `Stan`. The red vertical line indicates the runtime of inference with `Stan` using the exact likelihood.

(Rue et al., 2009), and (3) variational Bayes; see, e.g., (Bishop, 2006, Chap. 10) and (Blei et al., 2017). However, these approaches also scale poorly with dimension in general. The Laplace approximation requires computing and inverting the Hessian of the log posterior which demand $O(ND^2)$ and $O(D^3)$ time respectively, in order to compute approximate posterior means and variances. In the $N \ll D$ setting, this cost can be reduced to $O(N^2D)$ time (Appendix C). However, in large-N settings of interest, the $O(N^2D)$ cost can be prohibitive as well. The cost of inference is further compounded when we give a fully Bayesian treatment to model hyperparameters as well as parameters; e.g., INLA requires this heavy computation for each nested approximation. In the face of difficulties posed by high dimensionality, practitioners frequently turn to factorized (or "mean-field") approximations. In the case of VB, the mean-field approach can yield biased approximations that underestimate uncertainty (MacKay, 2003; Turner & Sahani, 2011). Likewise, factorized Laplace approximations, which approximate the Hessian with only its diagonal elements, similarly underestimate uncertainty (Appendix F.8).

Some more recent work has approached scalable approximate inference in generalized linear models with theoretical guarantees on quality in the large-N regime by using likelihood approximations that are cheap to evaluate (Huggins et al., 2017; Campbell & Broderick, 2019; 2018; Huggins et al., 2016). But these methods fail to scale well to the large-D case.

More closely related to the present work, Geppert et al. (2017) and Lee & Oh (2013) focus on conjugate Bayesian regression, respectively using random projections and principle component analysis to define low-rank descriptions of the design. Lee & Oh (2013) restrict their consideration to the exactly low-rank case and primarily discuss the asymptotic consistency of the resulting posterior mean without discussing computational considerations. Spantini et al. (2015) use conjugate Bayesian regression as stepping-off point to derive a point estimator for Bayesian inverse problems. Guhaniyogi & Dunson (2015) use random projections for Bayesian GLMs but focus on 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.

C. Fast matrix inversions in the $N \ll D$ setting

In this section we focus on Gaussian conjugate linear regression with $N \ll D$. In this case, we can detail formulas for more efficient computation of the posterior mean and covariance. We start from the standard expressions for the posterior mean μ_N and covariance Σ_N when the prior is mean zero with covariance Σ_β; see Section 3 and Section 4.1 for further notation and setup of the model. These expressions are:

$$\Sigma_N^{-1} = \Sigma_\beta^{-1} + \tau X^\top X \tag{C.1}$$

$$\mu_N = \tau \Sigma_N X^\top Y. \tag{C.2}$$

Using these formulas naively in the $D \gg N$ setting is computationally expensive due to the $O(D^3)$ time cost of matrix inversion and $O(D^2)$ storage cost.

Using the Woodbury matrix identity, $(A^{-1} + UCV)^{-1} = A - AU(C^{-1} + VAU)^{-1}VA$, allows us to write $\Sigma_N = (\Sigma_\beta^{-1} + X^\top(\tau I_N)X)^{-1}$ as

$$\Sigma_N = \Sigma_\beta - \Sigma_\beta X^\top(\tau^{-1} I_N + X\Sigma_\beta X^\top)^{-1} X\Sigma_\beta. \tag{C.3}$$

Computing Σ_N via Eq. (C.3) requires only $O(DN^2)$ cost for the matrix multiplications and an $O(N^3)$ cost for the matrix inversion. The posterior mean μ_N may then be computed in $O(ND)$ time by multiplying through by $X^\top Y$. These time costs can be significant reductions over the naive $O(D^3)$ cost when $N \ll D$.

Fast inversions for the Laplace approximation to the GLM posterior

We here show that the same approach described above may be used for the Laplace approximation in the context of Bayesian GLMs. We say that we have a GLM likelihood if we can write

$$p(Y \mid \beta, X) = \sum_{n=1}^{N} \phi(y_n, x_n^\top \beta)$$

for some *mapping function* $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. The Bayesian posterior then becomes

$$\log p(\beta \mid X, Y) = \log p(\beta) + \sum_{n=1}^{N} \phi(y_n, x_n^\top \beta) + Z, \tag{C.4}$$

where Z is a typically-intractable log normalizing constant.

Due to the analytic intractability of posterior inference in many common GLMs, approximations are necessary; the Laplace approximation is a particularly widely used approximation and takes the form

$$\bar{p}(\beta) = \mathcal{N}(\beta \mid \bar{\mu}, \bar{\Sigma}), \tag{C.5}$$

where $\bar{\mu} := \arg\max_\beta \log p(\beta \mid X, Y)$ and $\bar{\Sigma} := \left(-\nabla_\beta^2 \log p(\beta|X,Y)|_{\beta=\bar{\mu}}\right)^{-1}$. However, as in the conjugate case, computing this matrix inverse naively can be expensive in the high-dimensional setting, and we are motivated to consider more computationally efficient routes to evaluate it. In settings when $N \ll D$ and when we have a Gaussian prior $p(\beta) = \mathcal{N}(\beta \mid \mu_\beta, \Sigma_\beta)$, we may take an approach similar to our approach in the conjugate case. We first note

$$\nabla_\beta^2 \log p(\beta \mid X, Y)|_{\beta=\bar{\mu}} = -\Sigma_\beta^{-1} + X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu}))X, \tag{C.6}$$

where $\vec{\phi}''(Y, A)$ is a vector in $\mathbb{R}^N$ defined such that for any n in $1, 2, \ldots, N$, $\vec{\phi}''(Y, A)_n := \frac{d^2}{da^2}\phi(y_i, a)|_{a=A_n}$. Applying the same trick to this expression as before, we obtain

$$\bar{\Sigma}_N = \left(-\nabla_\beta^2 \log p(\beta \mid X, Y)|_{\beta=\bar{\mu}}\right)^{-1} = \Sigma_\beta - \Sigma_\beta X^\top \left(\mathrm{diag}[-\vec{\phi}''(Y, X\bar{\mu})]^{-1} + X\Sigma_\beta X^\top\right)^{-1} X\Sigma_\beta, \tag{C.7}$$

which again can yield computational gains.

It is worth noting however that this route is more computationally efficient only when the prior covariance matrix is structured in some way that allows for fast matrix-vector and matrix-matrix multiplications. This will be the case, for example, if Σ_β is diagonal, block-diagonal, banded diagonal, or diagonal plus a low-rank matrix.

D. Conjugate Gaussian regression with exactly low rank design

D.1. Derivation of Eq. (2)

Here we consider the setting of conjugate Bayesian linear regression, with X exactly low rank and $\Sigma_\beta = \sigma_\beta^2 I_D$, as detailed in Section 4.1. We now derive the expressions (Eq. (2)) for the mean and covariance of the Gaussian posterior for β in this case. We suppose $X = V\mathrm{diag}(\lambda)U^\top$ for U, V matrices of orthonormal rows and λ a vector. The preceding equation for X will capture low rank structure when $U \in \mathbb{R}^{D\times M}$ for some M with $M \ll \min(D, N)$.

For the covariance, we start from Eq. (C.1). Then we can rewrite Σ_N as follows.

$$
\begin{aligned}
\Sigma_N &= \left(\sigma_\beta^{-2} I_D + \tau X^\top X\right)^{-1} \\
&= \left(\sigma_\beta^{-2} I_D + \tau U\mathrm{diag}(\lambda)V^\top V\mathrm{diag}(\lambda)U^\top\right)^{-1} \\
&= \left(\sigma_\beta^{-2} I_D + U\mathrm{diag}(\tau\lambda\odot\lambda)U^\top\right)^{-1} \\
&\quad \text{where } \odot \text{ denotes component-wise multiplication, in this case across the components of the vector } \lambda \\
&= \sigma_\beta^2 I - \sigma_\beta^2 U(\mathrm{diag}(\tau\lambda\odot\lambda)^{-1} + \sigma_\beta^2 I_M)^{-1}U^\top\sigma_\beta^2 \\
&\quad \text{by the Woodbury matrix identity and } U^\top U = I_M \\
&= \sigma_\beta^2 I - \sigma_\beta^2 U\mathrm{diag}\left\{\left(\frac{1}{\tau\lambda\odot\lambda} + \sigma_\beta^2 1_M\right)^{-1}\sigma_\beta^2\right\}U^\top \\
&\quad \text{where division within the diag input is component-wise and } 1_M \text{ is the all-ones vector of length } M \\
&= \sigma_\beta^2\left(I_D - U\mathrm{diag}\left\{\frac{\tau\lambda\odot\lambda}{\sigma_\beta^{-2}1_M + \tau\lambda\odot\lambda}\right\}U^\top\right).
\end{aligned}
$$

Starting from Eq. (C.2), we can rewrite the posterior mean as follows.

$$
\begin{aligned}
\mu_N &= \tau\Sigma_N X^\top Y \\
&= \tau\sigma_\beta^2\left(I_D - U\mathrm{diag}\left\{\frac{\tau\lambda\odot\lambda}{\sigma_\beta^{-2}1_M + \tau\lambda\odot\lambda}\right\}U^\top\right)U\mathrm{diag}(\lambda)V^\top Y \\
&\quad \text{from the derivation above and substituting for } X \\
&= \tau\sigma_\beta^2\left(U - U\mathrm{diag}\left\{\frac{\tau\lambda\odot\lambda}{\sigma_\beta^{-2}1_M + \tau\lambda\odot\lambda}\right\}\right)\mathrm{diag}(\lambda)V^\top Y \\
&\quad \text{since } U^\top U = I_M \\
&= \tau\sigma_\beta^2 U\left(I_M - \mathrm{diag}\left\{\frac{\tau\lambda\odot\lambda}{\sigma_\beta^{-2}1_M + \tau\lambda\odot\lambda}\right\}\right)\mathrm{diag}(\lambda)V^\top Y \\
&= U\mathrm{diag}\left\{\frac{\tau\lambda}{\sigma_\beta^{-2}1_M + \tau\lambda\odot\lambda}\right\}V^\top Y.
\end{aligned}
$$

E. Proofs and further results for conjugate Bayesian linear regression with low-rank data approximations

E.1. Proof of Theorem 4.1

Recall that for conjugate Gaussian Bayesian linear regression, the exact posterior is $p(\beta \mid X, Y) = \mathcal{N}(\beta \mid \mu_N, \Sigma_N)$, where μ_N and Σ_N are given in Eqs. (C.1) and (C.2).

Using an orthonormal projection U yields a Gaussian approximate posterior $\tilde{p}(\beta \mid X, Y) = \mathcal{N}(\beta \mid \tilde{\mu}_N, \tilde{\Sigma}_N)$. Recall from Section 3 that we obtain this approximate posterior by replacing X with $XUU^\top$. Thus, we can find $\tilde{\mu}_N$ and $\tilde{\Sigma}_N$ by

consulting Eqs. (C.1) and (C.2):

$$\tilde{\Sigma}_N^{-1} = \Sigma_\beta^{-1} + \tau U U^\top X^\top X U U^\top \tag{E.1}$$

$$\tilde{\mu}_N = \tilde{\tau} \Sigma_N U U^\top X^\top Y. \tag{E.2}$$

Upper bound on the posterior mean approximation error

We will obtain our upper bound on the error of the approximate posterior mean relative to the exact posterior mean by upper bounding the norm of the difference between the gradient of the log posterior with respect to β at the approximate posterior mean, $\tilde{\mu}_N$, and the exact posterior mean, μ_N. Together with the strong convexity of the negative log posterior, this bound will allow us to arrive at the desired upper bound on $\|\mu_N - \tilde{\mu}_N\|_2$.

First, we bound the norm of the gradient difference. To that end, the gradients of the exact log likelihood and the approximate log likelihood are given by

$$\begin{aligned}\nabla_\beta \log p(Y \mid X, \beta) &= \nabla_\beta \left[-\frac{\tau}{2} (X\beta - Y)^\top (X\beta - Y) \right] \\ &= -\tau (X^\top X \beta - X^\top Y)\end{aligned}$$

and

$$\begin{aligned}\nabla_\beta \log \tilde{p}(Y \mid X, \beta) &= \nabla_\beta \left[-\frac{\tau}{2} (XUU^\top \beta - Y)^\top (XUU^\top \beta - Y) \right] \\ &= -\tau (UU^\top X^\top X U U^\top \beta - UU^\top X^\top Y).\end{aligned}$$

We can thus upper bound the norm of the difference between the two log posteriors as follows.

$$\begin{aligned}&\|\nabla_\beta \log \tilde{p}(\beta \mid X, Y) - \nabla_\beta \log p(\beta \mid X, Y)\|_2 \\ &= \|\nabla_\beta \log \tilde{p}(Y \mid X, \beta) - \nabla_\beta \log p(Y \mid X, \beta)\|_2 \\ &\text{since the prior is the same in both the exact and approximate model} \\ &\quad \text{and since the normalizing constant has no } \beta \text{ dependence} \\ &= \left\| -\tau \left(UU^\top X^\top X U U^\top \beta - UU^\top X^\top Y \right) + \tau \left(X^\top X \beta - X^\top Y \right) \right\|_2 \\ &= \tau \left\| \left(X^\top X - UU^\top X^\top X U U^\top \right) \beta + UU^\top X^\top Y - X^\top Y \right\|_2 \\ &= \tau \left\| \bar{U}\bar{U}^\top X^\top X \bar{U}\bar{U}^\top \beta - \bar{U}\bar{U}^\top X^\top Y \right\|_2 \\ &\text{where } \bar{U} \text{ (above) as well as } \bar{\lambda} \text{ and } \bar{V} \text{ (below) are defined in Section 3} \\ &= \tau \left\| \bar{U} \mathrm{diag}(\bar{\lambda} \odot \bar{\lambda}) \bar{U}^\top \beta - \bar{U} \mathrm{diag}(\bar{\lambda}) \bar{V}^\top Y \right\|_2 \\ &\leq \tau \left(\left\| \bar{U} \mathrm{diag}(\bar{\lambda} \odot \bar{\lambda}) \bar{U}^\top \beta \right\|_2 + \left\| \bar{U} \mathrm{diag}(\bar{\lambda}) \bar{V}^\top Y \right\|_2 \right) \\ &\text{by the triangle inequality} \\ &= \tau \left(\left\| \mathrm{diag}(\bar{\lambda} \odot \bar{\lambda}) \bar{U}^\top \beta \right\|_2 + \left\| \mathrm{diag}(\bar{\lambda}) \bar{V}^\top Y \right\|_2 \right) \\ &\text{since } \|v\|_2^2 = v^\top v \text{ for a vector } v \text{ and } U^\top U = I_M \\ &\leq \tau \left(\left\| \mathrm{diag}(\bar{\lambda} \odot \bar{\lambda}) \right\|_{\mathrm{op}} \left\| \bar{U}^\top \beta \right\|_2 + \left\| \mathrm{diag}(\bar{\lambda}) \right\|_{\mathrm{op}} \left\| \bar{V}^\top Y \right\|_2 \right) \\ &\text{by definition of the operator norm in this space} \\ &= \tau \left(\bar{\lambda}_1^2 \|\bar{U}^\top \beta\|_2 + \bar{\lambda}_1 \|\bar{V}^\top Y\|_2 \right)\end{aligned} \tag{E.3}$$

Second, we need a result that will let us use the strong convexity of the negative log posterior. We prove the following result in Appendix E.2.

Lemma E.1. *Let f, g be twice differentiable functions mapping $\mathbb{R}^D \to \mathbb{R}$ and attaining minima at $\beta_f = \arg\min_\beta f(\beta)$ and $\beta_g = \arg\min_\beta g(\beta)$, respectively. Additionally, assume that f is α–strongly convex for some $\alpha > 0$ on the set $\{t\beta_f + (1-t)\beta_g | t \in [0,1]\}$ and that $\|\nabla_\beta f(\beta_g) - \nabla_\beta g(\beta_g)\|_2 = \|\nabla_\beta f(\beta_g)\|_2 \leq c$. Then*

$$\|\beta_f - \beta_g\|_2 \leq \frac{c}{\alpha}. \tag{E.4}$$

To use the preceding result, we need a lower bound on the strong convexity constant of the negative log posterior; we now calculate such a bound. We have that μ_N and $\tilde{\mu}_N$ are the maximum a posteriori values of β under $p(\beta|X, Y, \alpha)$ and $\tilde{p}(\beta|X, Y, \alpha)$, respectively; equivalently they minimize the respective negative log of these distributions. For a matrix A, let $\lambda_{\min}(A)$ denote its minimum eigenvalue. The Hessian of the negative log posterior with respect to β is precisely $\Sigma_\beta^{-1} + \tau X^\top X$ everywhere. So the negative log posterior is α–strongly convex, where

$$\alpha = \lambda_{\min}(\Sigma_\beta^{-1} + \tau X^\top X) \geq \lambda_{\min}(\Sigma_\beta^{-1}) + \tau\lambda_{\min}(X^\top X) = \|\Sigma_\beta\|_2^{-1} + \tau\bar{\lambda}_{D-M}^2. \tag{E.5}$$

In the first part of the final equality above, we use that the spectral norm of a matrix inverse is equal to the reciprocal of the minimum eigenvalue of the matrix.

Now we have an upper bound on the norm of the difference in gradients of the negative log posteriors (the same as for the log posteriors, in Eq. (E.3)) and a lower bound on the strong convexity constant from Eq. (E.5). So we can apply these together with Lemma E.1 to find

$$\begin{aligned}
\|\mu_N - \tilde{\mu}_N\|_2 &\leq \frac{\tau\left(\bar{\lambda}_1^2\|\bar{U}^\top\tilde{\mu}_N\|_2 + \bar{\lambda}_1\|\bar{V}^\top Y\|_2\right)}{\alpha} \\
&\quad \text{by Lemma E.1 taking } \log p(\beta|X, Y) \text{ and } \log\tilde{p}(\beta|X, Y) \\
&\quad \text{as } f \text{ and } g \text{ respectively, with } c \text{ given by Eq. (E.3)} \\
&\leq \frac{\tau\left(\bar{\lambda}_1^2\|\bar{U}^\top\tilde{\mu}_N\|_2 + \bar{\lambda}_1\|\bar{V}^\top Y\|_2\right)}{\|\Sigma_\beta\|_2^{-1} + \tau\bar{\lambda}_{D-M}^2} \\
&\quad \text{by Eq. (E.5)} \\
&= \frac{\bar{\lambda}_1\left(\bar{\lambda}_1\|\bar{U}^\top\tilde{\mu}_N\|_2 + \|\bar{V}^\top Y\|_2\right)}{\|\tau\Sigma_\beta\|_2^{-1} + \bar{\lambda}_{D-M}^2}.
\end{aligned}$$

Notably, in the common special case that Σ_β is diagonal, as we saw in Section 4.1, $\tilde{\mu}_N$ will be in the span of U, and we will have that $\|\bar{U}^\top\tilde{\mu}_N\|_2 = 0$.

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

$$\begin{aligned}
\Sigma_N^{-1} - \tilde{\Sigma}_N^{-1} &= (\Sigma_\beta^{-1} + \tau X^\top X) - (\Sigma_\beta^{-1} + \tau UU^\top X^\top XUU^\top) && \text{(E.6)} \\
&= \tau X^\top X - \tau UU^\top X^\top XUU^\top && \text{(E.7)} \\
&= \tau\bar{U}\bar{U}^\top X^\top X\bar{U}\bar{U}^\top && \text{(E.8)} \\
&= \tau\bar{U}\mathrm{diag}(\bar{\lambda}\odot\bar{\lambda})\bar{U}^\top. && \text{(E.9)}
\end{aligned}$$

Thus, since it is equal to the maximum eigenvalue, the spectral norm of the error in the precisions is precisely $\|\Sigma_N^{-1} - \tilde{\Sigma}_N^{-1}\|_2 = \tau\bar{\lambda}_1^2$.

E.2. Proof of Lemma E.1

By the fundamental theorem of calculus, we may write

$$\nabla_\beta f(\beta) = \nabla_\beta f(\beta_g) + \int_{t=0}^1 (\beta - \beta_g)^\top\nabla_\beta^2 f(t\beta + (1-t)\beta_g)dt.$$

Considering the norm of $\nabla_\beta f(\beta)$ and applying the triangle inequality provides that for any β in $\{t\beta_f + (1-t)\beta_g \mid t \in [0, 1]\}$,

$$\begin{aligned}
\|\nabla_\beta f(\beta)\|_2 &\geq \left\|\int_{t=0}^1 (\beta - \beta_g)^\top\nabla_\beta^2 f(t\beta + (1-t)\beta_g)dt\right\|_2 - \|\nabla_\beta f(\beta_g)\|_2 && \text{(E.10)} \\
&\geq \|\beta - \beta_g\|_2\left\|\int_{t=0}^1 \nabla_\beta^2 f(t\beta + (1-t)\beta_g)dt\right\|_2 - \|\nabla_\beta f(\beta_g)\|_2 && \text{(E.11)} \\
&\geq \|\beta - \beta_g\|_2\alpha - \|\nabla_\beta f(\beta_g)\|_2. && \text{(E.12)}
\end{aligned}$$

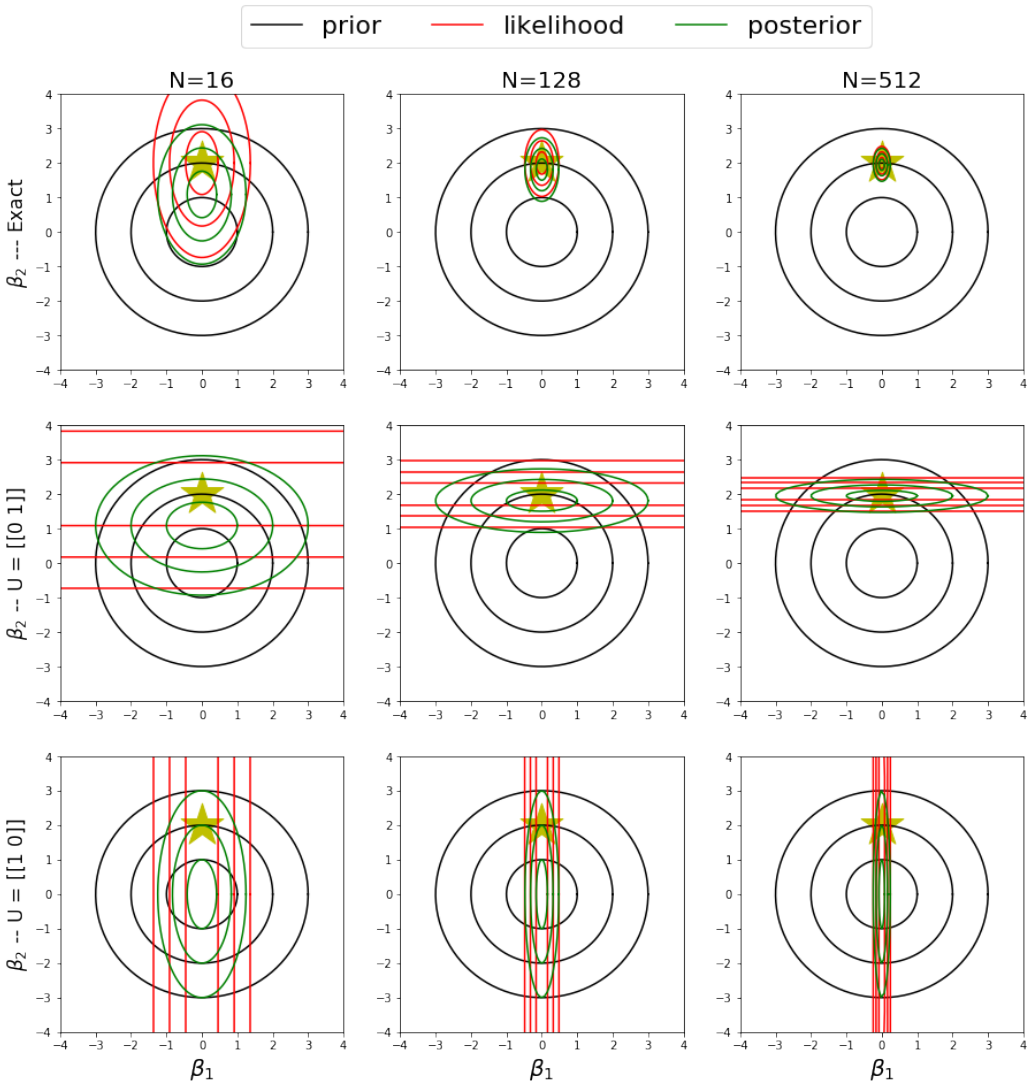

Figure E.1. Example of posterior approximations with different projections (characterized by U) for increasing sample sizes. Each plot shows the contours of three densities: the prior, likelihood, and posterior (or approximations thereof). The top row shows the exact posterior. The middle row shows the approximations found by using the best rank-1 approximation to X. The bottom row shows the approximations found using the orthogonal rank-1 approximation. The star is at the parameter value used to generate simulated data for these plots.

We consider this bound at β_f. Recall we assume that $\|\nabla_\beta f(\beta_g)\|_2 \leq c$. And $\|\nabla_\beta f(\beta_f)\|_2 = 0$ since f is twice differentiable. Therefore, we have that $0 \geq \|\beta_f - \beta_g\|_2 \alpha - c$, and the result follows.

E.3. Proof of Corollary 4.2

Our approach is to show that

$$\tilde{\mu}_N \xrightarrow{p} \Sigma_\beta U_* (U_*^\top \Sigma_\beta U_*)^{-1} U_*^\top \beta. \tag{E.13}$$

We then appeal to the following result, which we prove in Appendix E.4:

Lemma E.2. *$\tilde{\mu} := \Sigma_\beta U (U^\top \Sigma_\beta U)^{-1} U^\top \beta$ is the vector of minimum Σ_β^{-1}-norm satisfying $U^\top \tilde{\mu} = U^\top \beta$.*

Finally, for any closed $S \subset \mathbb{R}^D$, $\tilde{\mu} = \arg\min_{v \in S} \|v\|_{\Sigma_\beta^{-1}} = \arg\max_{v \in S} -\frac{1}{2} v^\top \Sigma_\beta^{-1} v = \arg\max_{v \in S} \mathcal{N}(0, \Sigma_\beta)$. Therefore, the $\tilde{\mu}$ in Lemma E.2 is the maximum a priori vector satisfying the constraint in Lemma E.2.

We first turn to proving Eq. (E.13). Let $U_N \text{diag}(\lambda^{(N)}) V_N^\top$ denote the M-truncated SVD of the design matrix consisting of N samples $X = (x_1, x_2, \ldots, x_N)$ where $x_i \overset{\text{i.i.d.}}{\sim} p_*$. When the low rank approximation is defined by this SVD, from Eq. (E.1) we have that $\tilde{\mu}_N = \tau \tilde{\Sigma}_N U_N U_N^\top X^\top Y$. Noting that $Y = X\beta + \frac{1}{\tau}\epsilon$ for some $\epsilon \in \mathbb{R}^N$ with $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, we may expand this out and write:

$$
\begin{aligned}
\tilde{\mu}_N &= \tau(\Sigma_\beta^{-1} + U_N U_N^\top X^\top \tau X U_N U_N^\top)^{-1} U_N U_N^\top X^\top (X\beta + \frac{1}{\tau}\epsilon) \\
&= \tau \left\{ \Sigma_\beta^{-1} + U_N \left[\tau \text{diag}(\lambda^{(N)} \odot \lambda^{(N)}) \right] U_N^\top \right\}^{-1} U_N \text{diag}(\lambda^{(N)}) V_N^\top \left[V_N \text{diag}(\lambda^{(N)}) U_N^\top \beta + \frac{1}{\tau}\epsilon \right] \\
&= \left\{ \Sigma_\beta^{-1} + U_N \left[\tau \text{diag}(\lambda^{(N)} \odot \lambda^{(N)}) \right] U_N^\top \right\}^{-1} U_N \left[\tau \text{diag}(\lambda^{(N)} \odot \lambda^{(N)}) \right] \left[U_N^\top \beta + \text{diag}(\lambda^{(N)})^{-1} V_N^\top \frac{1}{\tau}\epsilon \right] \\
&= \Sigma_\beta U_N \left[U_N^\top \Sigma_\beta U_N + \tau^{-1} \text{diag}(\lambda^{(N)})^{-2} \right]^{-1} \left[U_N^\top \beta + \text{diag}(\lambda^{(N)})^{-1} V_N^\top \frac{1}{\tau}\epsilon \right] \\
&\xrightarrow{P} \Sigma_\beta U_* (U_*^\top \Sigma_\beta U_*)^{-1} U_*^\top \beta,
\end{aligned}
$$

where in the fourth line we use the matrix identity, $(R^{-1} + W^\top Q W)^{-1} W^\top Q = R W^\top (W R W^\top + Q^{-1})^{-1}$ (Petersen & Pedersen, 2008). Convergence in probability in the last line follows since $\text{diag}(\lambda^{(N)^{-2}}) \xrightarrow{P} 0$ (Vershynin, 2012) and $U_N \xrightarrow{P} U$.

E.4. Proof of Lemma E.2

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

Define β_* as

$$
\beta_* = \underset{v \in \mathbb{R}^D}{\arg\min} \, \|v\|_{\Sigma_\beta^{-1}} \;\; \text{subject to} \;\; U^\top v = U^\top \beta \tag{E.14}
$$

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

$$
\langle \Sigma_\beta U[:, i], \beta_* \rangle = U[:, i]^\top \beta \tag{E.15}
$$

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 $\left[A \right]$ (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 the vectors $\left\{ \Sigma_\beta U[:, i] \right\}_{i=1}^M$; that is, for some c in $\mathbb{R}^M$

$$
\beta_* = \Sigma_\beta U c. \tag{E.16}
$$

Our constraints in Eq. (E.15) then demand that $\langle \Sigma_\beta U[:, i], \Sigma_\beta U c \rangle = U[:, i]^\top \beta$ for each i, or equivalently that $U^\top \Sigma_\beta \Sigma_\beta^{-1} \Sigma_\beta U c = U^\top \beta$. This implies that $c = (U^\top \Sigma_\beta U)^{-1} U^\top \beta$. Plugging this into Eq. (E.16) yields $\beta_* = \Sigma_\beta U (U^\top \Sigma_\beta U)^{-1} U^\top \beta$, as desired.

E.5. Proof of Corollary 4.3

Recall that we wish to show that, for conjugate Bayesian regression, under $\tilde{p}$ the uncertainty (i.e., posterior variance) for any linear combination of parameters, $\text{Var}_{\tilde{p}}[v^\top \beta]$, is no smaller than the exact posterior variance. First, we note that this statement is formally equivalent to stating that $v^\top \tilde{\Sigma}_N v \geq v^\top \Sigma_N v$, or that $E := \tilde{\Sigma}_N - \Sigma_N \succeq 0$ (where $\succeq$ denotes positive definiteness). By Theorem 4.1, $\Sigma_N^{-1} - \tilde{\Sigma}_N^{-1} = \bar{U} \text{diag}(\bar{\lambda}^2) \bar{U}^\top \succeq 0$. 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.

E.6. Information loss due the LR-GLM approximation

We see similar behavior to that demonstrated in Corollary 4.3 in the following corollary, which shows that our approximate posterior never has lower entropy than the exact posterior. Concretely, we look at 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 E.1. *The entropy $H\big[\tilde{p}(\beta|X,Y)\big]$ is no less than $H\big[p(\beta|X,Y)\big]$. Furthermore, 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 $H\big[\tilde{p}(\beta|X,Y)\big] - H\big[p(\beta|X,Y)\big] \leq \frac{\tau\sigma_\beta^2}{2}\sum_{i=1}^{D-M}\bar{\lambda}_i^2$.*

This result formalizes the intuition that the LR-GLM approximation reduces the information about the parameter that we are able to extract from the data. Additionally, the upper bound tells us that when U is obtained via an M truncated SVD, at most $\tau\sigma_\beta^2\bar{\lambda}_1^2/2$ 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:

$$H(p) = -\frac{1}{2}\log|2\pi e\Sigma_N^{-1}| = -\frac{1}{2}\big[D\log 2\pi e + \sum_{i=1}^{D}\log(\sigma_\beta^{-2} + \tau\lambda_i^2)\big]$$

and

$$H(\tilde{p}) = -\frac{1}{2}\log|2\pi e\tilde{\Sigma}_N^{-1}| = -\frac{1}{2}\big[D\log 2\pi e + \sum_{i=1}^{M}\log(\sigma_\beta^{-2} + \tau\lambda_i^2) - \sum_{i=M+1}^{D}\log\sigma_\beta^{-2}\big].$$

Therefore, we conclude that

$$\begin{aligned}
H\big[\tilde{p}(\beta|X)\big] - H\big[p(\beta|X)\big] &= -\frac{1}{2}\sum_{i=1}^{D-M}\log\sigma_\beta^{-2} + \frac{1}{2}\sum_{i=1}^{D-M}\log(\sigma_\beta^{-2} + \tau\bar{\lambda}_i^2) \\
&= \frac{1}{2}\sum_{i=1}^{D-M}\log\frac{\sigma_\beta^{-2} + \tau\bar{\lambda}_i^2}{\sigma_\beta^{-2}} \\
&= \frac{1}{2}\sum_{i=1}^{D-M}\log(1 + \frac{\tau}{\sigma_\beta^{-2}}\bar{\lambda}_i^2) \\
&\leq \frac{1}{2}\sum_{i=1}^{D-M}\frac{\tau}{\sigma_\beta^{-2}}\bar{\lambda}_i^2 = \frac{\tau\sigma_\beta^2}{2}\sum_{i=1}^{D-M}\bar{\lambda}_i^2.
\end{aligned}$$

That $H\big[\tilde{p}(\beta|X)\big] - H\big[p(\beta|X)\big] > 0$ follows from the monotonicity of $\log$, that $\log(1) = 0$, and that $\tau\sigma_\beta^2\bar{\lambda}_i^2 > 0$ for $i = 1,\ldots,D-M$. □

F. Proofs and further results for LR-Laplace in non-conjugate models

In the main text we introduced LR-Laplace as a method which takes advantage of low-rank approximations to provide computational gains when computing a Laplace approximation to the Bayesian posterior. In what follows we verify the theoretical justifications for this approach. Appendix F.1 provides a derivation of Algorithm 1 and demonstrates the time complexities of each step, serving as a proof of Theorem 5.1. The remainder of the section is devoted to the proofs and discussion of the theoretical properties of LR-Laplace.

F.1. Proof of Theorem 5.1

Proof of Theorem 5.1. The LR-Laplace approximation is defined by mean and covariance parameters, $\hat{\mu}$ and $\hat{\Sigma}$. We prove Theorem 5.1 in two parts. First, we show that $\hat{\mu}$ and $\hat{\Sigma}$ do in fact define the Laplace approximation of $\tilde{p}(\beta|X,Y)$, i.e. the

construction of $\hat{\mu}$ in Line 9 satisfies $\hat{\mu} = \arg\max_\beta \tilde{p}(\beta|X,Y)$ and that $\hat{\Sigma} = \big(-\nabla^2_\beta \log \tilde{p}(\beta|X,Y)|_{\beta=\hat{\mu}}\big)^{-1}$. Second, we show that each step of Algorithm 1 may be computed in $O(NDM)$ time with $O(DM+NM)$ storage.

Correctness of $\hat{\mu}$ and $\hat{\Sigma}$

In Line 8, the definition of γ_* implies that $\gamma_* = \arg\max_{\gamma\in\mathbb{R}^M} \tilde{p}_{U^\top\beta|X,Y}(\gamma|X,Y)$ since

$$\begin{aligned}
\log \tilde{p}_{U^\top\beta|X,Y}(\gamma|X,Y) &= \log \tilde{p}_{U^\top\beta}(\gamma) + \log \tilde{p}_{Y|X,U^\top\beta}(Y|X,\gamma) + C \\
&= \log p_{U^\top\beta}(\gamma) + \log p_{Y|X,\beta}(Y|X,U\gamma) + C \\
&= \log \mathcal{N}(\gamma|U^\top\mu_\beta, U^\top\Sigma_\beta U) + \sum_{i=1}^N \log p_{y|x,\beta}(y_i|x_i,U\gamma) + C \\
&= -\frac{1}{2}\gamma^\top U^\top\Sigma_\beta U\gamma + \sum_{i=1}^N \phi(y_i, x_i^\top U\gamma) + C',
\end{aligned}$$

where line 1 uses Bayes' rule, line 2 uses the definition of $\tilde{p}$ in Eq. (1), line 3 uses the normality the prior, and the assumed conditional independence of the responses given β, and line 4 follows from the definition of $\phi(\cdot,\cdot)$ and the assumption that $\mu_\beta = 0$. C and C' are constants which do not depend on γ. This together with the following result (proved in Appendix F.2) implies that as defined in Line 9 of Algorithm 1, $\hat{\mu} = \arg\max_\beta \tilde{p}(\beta|X,Y)$.

Lemma F.1. *Suppose a Gaussian prior $p(\beta) = \mathcal{N}(\mu_\beta, \Sigma_\beta)$, and let $\gamma_* := \arg\max_{\gamma\in\mathbb{R}^M} \log \tilde{p}_{U^\top\beta|X,Y}(\gamma|X,Y)$. Then $\hat{\mu} := \arg\max_{\beta\in\mathbb{R}^D} \log\tilde{p}(\beta|X,Y)$ may be written as $\hat{\mu} = U\gamma_* + \bar{U}\bar{U}^\top\Sigma_\beta U(U^\top\Sigma_\beta U)^{-1}\gamma_*$.*

We now show that as defined in Line 12 of Algorithm 1, $\hat{\Sigma}$ is inverse of the Hessian of the negative log posterior, H. We see this by writing

$$\begin{aligned}
H :&= \nabla^2_\beta - \log\tilde{p}(\beta|X,Y)|_{\beta=\hat{\mu}} \\
&= \nabla^2_\beta - \log\mathcal{N}(\beta|\mu_\beta,\Sigma_\beta)|_{\beta=\hat{\mu}} + \nabla^2_\beta \sum_{i=1}^N -\phi(y_i, x_i^\top UU^\top\beta)|_{\beta=\hat{\mu}} \\
&= \Sigma_\beta^{-1} + \sum_{i=1}^N -\phi''(y_i, x_i^\top UU^\top\hat{\mu}) x_i UU^\top x_i^\top \\
&= \Sigma_\beta^{-1} + UU^\top X^\top \mathrm{diag}\big(-\vec{\phi}''(Y, XUU^\top\hat{\mu})\big) XUU^\top,
\end{aligned}$$

where $\vec{\phi}''$ is the second derivative of ϕ. The Woodbury matrix lemma then provides that we may compute $\hat{\Sigma}_N := H^{-1}$ as

$$\hat{\Sigma}_N = \Sigma_\beta - \Sigma_\beta U\left(U^\top\Sigma_\beta U - \left\{U^\top X^\top \mathrm{diag}\left[\vec{\phi}''(Y, XUU^\top\hat{\mu})\right] XU\right\}^{-1}\right)^{-1} U^\top\Sigma_\beta,$$

which we have written as $\hat{\Sigma} := \Sigma_\beta - \Sigma_\beta UWU^\top\Sigma_\beta$ in Line 12 with $W^{-1} = U^\top\Sigma_\beta U - \left\{U^\top X^\top \mathrm{diag}\left[\vec{\phi}''(Y, XUU^\top\hat{\mu})\right] XU\right\}^{-1}$.

Time complexity of Algorithm 1

We now prove the asserted time and memory complexities for each line of Algorithm 1.

Algorithm 1 begins with the computation of the M-truncated SVD of $X^\top \approx U\mathrm{diag}(\lambda)V$. As discussed in Section 4.1, U may be found in $O(ND\log M)$ time. At the end of this step we must store the projected data $XU \in \mathbb{R}^{N,M}$ and the left singular vectors, $U \in \mathbb{R}^{D,M}$. Which demands $O(NM+DM)$ memory, and the matrix multiply for XU requires $O(NDM)$ time and is the bottleneck step of the algorithm. The matrix V need not be explicitly computed or stored.

The next stage of the algorithm is solving for $\hat{\mu} = \arg\max_\beta \log\tilde{p}(\beta|X,Y)$. This is done in two stages: in Line 8 find $\gamma_* = \arg\max_{\gamma\in\mathbb{R}^M} \log\tilde{p}_{U^\top\beta|X,Y}(\gamma|X,Y)$ as the solution to a convex optimization problem, and in Line 9 find $\hat{\mu}$ as $\hat{\mu} = U\gamma_* + \bar{U}\bar{U}^\top\Sigma_\beta U(U^\top\Sigma_\beta U)^{-1}\gamma_*$. Beginning with Line 8, we note that the function $\log\tilde{p}(U^\top\beta|X,Y) =$

$\log p(\beta) + \log \tilde{p}(Y|X,\beta) + c \stackrel{c}{=} \log \mathcal{N}(U^\top\beta|U^\top\mu_\beta, U^\top\Sigma_\beta U) + \sum_{i=1}^N \log p(y_i|x_i^\top UU^\top\beta)$ is a finite sum of functions concave in β and therefore also in $U^\top\beta$. γ_* may therefore be solved to a fixed precision in $O(NM)$ time under the assumptions of our theorem using stochastic optimization algorithms such as stochastic average gradient (Schmidt et al., 2017). In our experiments we use more standard batch convex optimization algorithm (L-BFGS-B (Zhu et al., 1997)) which takes at most $O(N^2M)$ time. This latter upper bound on complexity may be seen from observing each gradient evaluation takes $O(NM)$ time (the cost for the likelihood evaluation, since computing the log prior and its gradient is $O(M^2)$ after computing $U^\top\Sigma_\beta U$ once, which takes $O(DM^2)$ time by assumption) and the number of iterations required can grow up to linearly in the maximum eigenvalue of Hessian, which in turn grows linearly in N (Boyd & Vandenberghe, 2004).

The second step is computing $\hat{\mu} = U\gamma_* + \bar{U}\bar{U}^\top\Sigma_\beta U(U^\top\Sigma_\beta U)^{-1}\gamma_*$. Given γ_*, this may be computed in $O(DM)$ time, which one may see by noting that $\bar{U}\bar{U}^\top$ (which we never explicitly compute) may be written as $\bar{U}\bar{U}^\top = (I - UU^\top)$, and finding $\hat{\mu}$ as $\hat{\mu} = U\gamma_* + \Sigma_\beta U(U^\top\Sigma_\beta U)^{-1}\gamma_* - UU^\top\Sigma_\beta U(U^\top\Sigma_\beta U)^{-1}\gamma_*$. By assumption, the structure of Σ_β allows us to compute $U^\top\Sigma_\beta U$ in $O(DM^2)$ time and matrix vector products with Σ_β in $O(D)$ time.

We now turn to the third stage of the algorithm, solving for the posterior covariance $\hat{\Sigma}$, which is represented as an expression of U, Σ_β and W, defined in Line 11. Computing W requires $O(DM)$ and $O(NM^2)$ matrix multiplications (since we have precomputed XU), and two $O(M^3)$ matrix inversions which comes to $O(NM^2 + DM)$ time. The memory complexity of this step is $O(NM)$ since it involves handling XU. Once W has been computed we may use the representation $\hat{\Sigma} = \Sigma_\beta - \Sigma_\beta UWU^\top\Sigma_\beta$ as presented in Line 12. This representation does not entail performing any additional computation (which is why we have written $O(0)$), but as this expression includes U, storing $\hat{\Sigma}$ requires $O(DM)$ memory.

Lastly, we may immediately see that computing posterior variances and covariances takes only $O(M^2)$ time as it involves only indexing into Σ_β and U and $O(M^2)$ matrix-vector multiplies. □

F.2. Proof of Lemma F.1

We prove the lemma by constructing a rotation of the parameter space by the matrix of singular vectors $[U, \bar{U}]$, in which we have the prior

$$p\Big(\begin{bmatrix}U^\top\beta\\ \bar{U}^\top\beta\end{bmatrix}\Big) = \mathcal{N}\Big(\begin{bmatrix}U^\top\beta\\ \bar{U}^\top\beta\end{bmatrix}\Big|\begin{bmatrix}U^\top\mu_\beta\\ \bar{U}^\top\mu_\beta\end{bmatrix}, \begin{bmatrix}U^\top\Sigma_\beta U, & U^\top\Sigma_\beta\bar{U}\\ \bar{U}^\top\Sigma_\beta U, & \bar{U}^\top\Sigma_\beta\bar{U}\end{bmatrix}\Big).$$

We have that

$$\begin{aligned}
\hat{\mu} :&= \arg\max_{\beta\in\mathbb{R}^D} \log\tilde{p}(\beta|X,Y)\\
&= [U\,\bar{U}] \arg\max_{U^\top\beta\in\mathbb{R}^M, \bar{U}^\top\beta\in\mathbb{R}^{D-M}} \log\tilde{p}\Big(\begin{bmatrix}U^\top\beta\\ \bar{U}^\top\beta\end{bmatrix}|X,Y\Big)\\
&= U \arg\max_{U^\top\beta\in\mathbb{R}^M} \Big(\log\tilde{p}(U^\top\beta|X,Y) + \bar{U} \arg\max_{\bar{U}^\top\beta\in\mathbb{R}^{D-M}} \log\tilde{p}(\bar{U}^\top\beta|U^\top\beta,X,Y)\Big)\\
&= U \arg\max_{U^\top\beta\in\mathbb{R}^M} \log\tilde{p}(U^\top\beta|X,Y) +\\
&\quad \bar{U} \arg\max_{\bar{U}^\top\beta\in\mathbb{R}^{D-M}} \log\mathcal{N}\big(\bar{U}^\top\beta|\bar{U}^\top\Sigma_\beta U(U^\top\Sigma_\beta U)^{-1}U^\top\beta, \bar{U}\Sigma_\beta\bar{U} - \bar{U}\Sigma_\beta U(U^\top\Sigma_\beta U)U^\top\Sigma_\beta\bar{U}\big)\\
&= U\gamma_* + \bar{U}\bar{U}^\top\Sigma_\beta U(U^\top\Sigma_\beta U)^{-1}\gamma_*.
\end{aligned}$$

In the second line we simply move to the rotated parameter space. In the third line, we use the chain rule of probability to separate out two terms. To produce the fourth line, we note that since $\tilde{p}(Y|X,\beta) = p(Y|XUU^\top\beta) = \tilde{p}(Y|X,U^\top\beta)$, that Y and $\bar{U}^\top\beta$ are conditionally independent given $U^\top\beta$. We next note that though $\arg\max_{\bar{U}^\top\beta\in\mathbb{R}^{D-M}} \log p(\bar{U}^\top\beta|U^\top\beta)$ depends on $U^\top\beta$, $\max_{\bar{U}^\top\beta\in\mathbb{R}^{D-M}} \log p(\bar{U}^\top\beta|U^\top\beta)$ does not depend $U^\top\beta$. This allows us to use the definition of γ_* to arrive at the fifth line, as desired.

In the special case that Σ_β is diagonal, this expression reduces to $U\gamma_*$. This can be seen by recognizing that $\bar{U}^\top\Sigma_\beta U$ is then $\mathrm{diag}(\mathbf{0})$.

F.3. Proof of Theorem 5.2

Our approach to proving Theorem 5.2 follows a similar approach to that taken to prove Theorem 4.1. In particular, we begin by upper bounding the norm of the error of the gradients at the approximate MAP. Noting that the strong log concavity of the exact posterior, which having been assumed to hold globally, must then also hold on $\{t\hat{\mu} + (1-t)\bar{\mu} | t \in [0,1]\}$, we obtain an upper-bound on $\|\hat{\mu} - \bar{\mu}\|_2$ by again applying Lemma E.1.

To begin, we first recall that the exact and LR-GLM posteriors may be written as

$$\log p(\beta|X,Y) = \log p(\beta) + \sum_{n=1}^{N} \phi(y_n | x_n^\top \beta) - \log Z$$

and

$$\log \tilde{p}(\beta|X,Y) = \log p(\beta) + \sum_{n=1}^{N} \phi(y_n, x_n^\top U U^\top \beta) - \log \tilde{Z}$$

where $\phi(\cdot,\cdot)$ is such that $\phi(y,a) = \log p(y|x^\top\beta = a)$, and Z and $\tilde{Z}$ are the normalizing constants of the exact and approximate posteriors. As a result, the gradients of these log densities are given as

$$\nabla_\beta \log p(\beta|X,Y) = \nabla_\beta \log p(\beta) + X^\top \vec{\phi}'(Y, X\beta)$$

and

$$\nabla_\beta \log \tilde{p}(\beta|X,Y) = \nabla_\beta \log p(\beta) + UU^\top X^\top \vec{\phi}'(Y, XUU^\top \beta),$$

where $\vec{\phi}'(Y, X\beta) \in \mathbb{R}^N$ is such that for each $n \in [N]$, $\vec{\phi}'(Y, X\beta)_n = \frac{d}{da}\phi(y_n, a)|_{a = x_n^\top \beta}$.

And the difference in the gradients is

$$\nabla_\beta \log p(\beta|X,Y) - \nabla_\beta \log \tilde{p}(\beta|X,Y) = X^\top \vec{\phi}'(Y, X\beta) - UU^\top X^\top \vec{\phi}'(Y, XUU^\top \beta). \tag{F.1}$$

Appealing to Taylor's theorem, we may write for any β that

$$\phi'(y_n, x_n^\top UU^\top \beta) = \phi'(y_n, x_n^\top \beta) + (x_n^\top UU^\top \beta - x_n^\top \beta)\phi''(y_n, a_n)$$

for some $a_n \in [x_n^\top UU^\top \beta, x_n^\top \beta]$, where $\phi''(y,a) := \frac{d^2}{da^2}\phi(y,a)$.

Using this and introducing vectorized notation for ϕ'' to match that used for $\vec{\phi}'$, we may rewrite the difference in the gradients as

$$\begin{aligned}
&\nabla_\beta \log p(\beta|X,Y) - \nabla_\beta \log \tilde{p}(\beta|X,Y) \\
&= X^\top \vec{\phi}'(Y, X\beta) - UU^\top X^\top \vec{\phi}'(Y, X\beta) - UU^\top X^\top \big[(XUU^\top \beta - X^\top \beta) \circ \vec{\phi}''(Y, A)\big] \\
&= \bar{U}\bar{U}^\top X^\top \vec{\phi}'(Y, X\beta) + UU^\top X^\top \big[(X\bar{U}\bar{U}^\top \beta) \circ \vec{\phi}''(Y, A)\big],
\end{aligned}$$

where $A \in \mathbb{R}^N$ is such that for each $n \in [N]$, $A_n \in [x_n^\top UU^\top \beta, x_n^\top \beta]$, and $\circ$ denotes element-wise scalar multiplication.

We can use this to derive an upper bound on the norm of the difference of the gradients as

$$\begin{aligned}
\|\nabla_\beta \log p(\beta|X,Y) - \nabla_\beta \log \tilde{p}(\beta|X,Y)\|_2 &= \|\bar{U}\bar{U}^\top X^\top \vec{\phi}' + UU^\top X^\top \big[(X\bar{U}\bar{U}^\top \beta) \circ \vec{\phi}''\big]\|_2 \\
&\leq \|\bar{U}\bar{U}^\top X^\top \vec{\phi}'\|_2 + \|UU^\top X^\top \big[(X\bar{U}\bar{U}^\top \beta) \circ \vec{\phi}''\big]\|_2 \\
&\leq \bar{\lambda}_1 \|\vec{\phi}'\|_2 + \lambda_1 \|(X\bar{U}\bar{U}^\top \beta) \circ \vec{\phi}''\|_2 \\
&\leq \bar{\lambda}_1 \|\vec{\phi}'\|_2 + \lambda_1 \bar{\lambda}_1 \|\bar{U}^\top \beta\|_2 \|\vec{\phi}''\|_\infty \\
&= \bar{\lambda}_1 \big(\|\vec{\phi}'\|_2 + \lambda_1 \|\bar{U}^\top \beta\|_2 \|\vec{\phi}''\|_\infty\big),
\end{aligned}$$

where we have written $\vec{\phi}'$ and $\vec{\phi}''$ in place of $\vec{\phi}'(Y, X\beta)$ and $\vec{\phi}''(Y, A)$, respectively, for brevity despite their dependence on β.

Next, let α be the strong log-concavity parameter of $p(\beta|X,Y)$. Lemma E.1 then implies that

$$\|\hat{\mu} - \bar{\mu}\|_2 \leq \frac{\bar{\lambda}_1\big(\|\vec{\phi}'(Y, X\hat{\mu})\|_2 + \lambda_1\|\bar{U}^\top\hat{\mu}\|_2\|\vec{\phi}''(Y,A)\|_\infty\big)}{\alpha}$$

as desired, where for each $n \in [N]$, $A_n \in [x_n^\top UU^\top\hat{\mu}, x_n^\top\hat{\mu}]$.

F.4. Bounds on derivatives of higher order for the log-likelihood in logistic regression and other GLMs

We here provide some additional support for the claim that in Remark 5.3 that the higher order derivatives of the log-likelihood function, ϕ, are well-behaved. For logistic regression (which we explore in detail below), for any y in $\{-1,1\}$ and a in $\mathbb{R}$, it holds that $|\frac{\partial}{\partial a}\phi(y,a)| \leq 1$ and $|\frac{\partial^2}{\partial^2 a}\phi(y,a)| \leq \frac{1}{4}$. For Poisson regression with $\phi(y,a) = \log \text{Pois}\big(y|\lambda = \log(1+\exp\{a\})\big)$, both $|\frac{\partial}{\partial a}(y,a)|$ and $|\frac{\partial^2}{\partial^2 a}\phi(y,a)|$ are bounded by a small constant factor of y. Additionally, in these cases $|\frac{\partial^3}{\partial a^3}\phi(y,a)|$ is also well behaved, a fact relevant to Corollary 5.6. However, for alternative mapping functions for Poisson regression, e.g. defining $\mathbb{E}[y_i|x_i,\beta] = \exp\{x_i^\top\beta\}$, these derivatives will grow exponentially quickly with $x_i^\top\beta$, which illustrates that our provided bounds are sensitive to the particular form chosen for the GLM likelihood.

We now move to compute explicit upper bounds on the derivatives of the log likelihood in logistic regression. This produces the constants mentioned above, and permits easy computation of upper bounds on the bounds on the approximation error of LR-Laplace provided in Theorem 5.2 and Corollary 5.6. In particular the logistic regression mapping function (Huggins et al., 2017) is given as

$$\phi(y_n, x_n^\top\beta) = -\log\big(1 + \exp\{-y_n x_n^\top\beta\}\big), \tag{F.2}$$

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

The first three derivatives of this mapping function and bounds on their absolute values are as follows:

$$\phi'(y_n, x_n^\top\beta) := \frac{d}{da}\phi(y_n,a)\big|_{a=x_n^\top\beta} = y_n\frac{\exp\{-yx_n^\top\beta\}}{1+\exp\{-y_n x_n^\top\beta\}} \tag{F.3}$$

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

$$\phi''(y_n, x_n^\top\beta) := \frac{d^2}{da^2}\phi(y,a)\big|_{a=x_n^\top\beta} = -(1+\exp\{x_n^\top\beta\})^{-1}(1+\exp\{-x_n^\top\beta\})^{-1}. \tag{F.4}$$

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

$$\frac{d^3}{da^3}\phi(y,a) = \phi'''(a) = \frac{\big(\exp\{a\}(\exp(-a)-1)\big)}{(1+\exp\{a\})^3} \tag{F.5}$$

which for any a in $\mathbb{R}$ satisfies, $-\frac{1}{6\sqrt{3}} \leq \phi'''(a) \leq \frac{1}{6\sqrt{3}}$.

F.5. Asymptotic inconsistency of the approximate posterior mean within the span of the projections

Consider a Bayesian logistic regression, in which

$$x_i \sim \mathcal{N}\Big(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\ 0 & 0.99\end{bmatrix}\Big), \qquad \beta = \begin{bmatrix}10\\1000\end{bmatrix}, \qquad y_i \sim \text{Bern}\big((1+\exp\{x_i^\top\beta\})^{-1}\big).$$

In this setting, a rank 1 approximation of the design will capture only the first dimension of data (i.e. $U_N \to U_* = [1,0]$). However the second dimension explains almost all of the variance in the responses. As such $y_i|U_*^\top x_i, \beta \overset{d}{\approx} \text{Bern}(1/2)$ and we will get $U_*^\top\beta|X,Y = \beta_1|X,Y \approx 0.0$ under $\tilde{p}$.

F.6. Proof of Corollary 5.6

Our proof proceeds via an upper bound on the $(2, \hat{p})$-Fisher distance between $\hat{p}$ and $\bar{p}$ (Huggins et al., 2018). Specifically, the $(2, \hat{p})$-Fisher distance given by

$$d_{2,\hat{p}}(\hat{p}, \bar{p}) = \left(\int \|\nabla_\beta \log \hat{p}(\beta) - \nabla_\beta \log \bar{p}(\beta)\|_2^2 dp(\beta) \right)^{\frac{1}{2}}. \tag{F.6}$$

Given the strong log-concavity of $\bar{p}$, our upper bound on this Fisher distance immediately provides an upper-bound on the 2-Wasserstein distance (Huggins et al., 2018).

We first recall that $\hat{p}$ and $\bar{p}$ are defined by Laplace approximations of $\tilde{p}(\beta|X, Y)$ and $p(\beta|X, Y)$ respectively. As such we have that

$$\log \hat{p}(\beta) \stackrel{c}{=} -\frac{1}{2}(\beta - \hat{\mu})^\top \big(\Sigma_\beta^{-1} - UU^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, XUU^\top \hat{\mu}))XUU^\top\big)(\beta - \hat{\mu})$$

where $\vec{\phi}''(Y, XUU^\top \hat{\mu})$ is defined as in Algorithm 1 such that $\vec{\phi}''(Y, X\beta)_i = \frac{d^2}{da^2} \log p(y_i|x^\top \beta = a)|_{a=x_i^\top \beta}$, and

$$\log \bar{p}(\beta) \stackrel{c}{=} -\frac{1}{2}(\beta - \bar{\mu})^\top \big(\Sigma_\beta^{-1} - X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu}))X\big)(\beta - \bar{\mu}).$$

Accordingly,

$$\nabla_\beta \log \hat{p}(\beta) = -(\beta - \hat{\mu})^\top \big(\Sigma_\beta^{-1} - UU^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, XUU^\top \hat{\mu}))XUU^\top\big)$$

and

$$\nabla_\beta \log \bar{p}(\beta) = -(\beta - \bar{\mu})^\top \big[\Sigma_\beta^{-1} - X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu}))X\big]$$

To define an upper bound on $d_{2,\hat{p}}(\hat{p}, p)$, we must consider the difference between the gradients,

$$\begin{aligned}
\nabla_\beta \log \hat{p}(\beta) - \nabla_\beta \log \bar{p}(\beta) =& -(\beta - \hat{\mu})^\top \big\{\Sigma_\beta^{-1} - UU^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, XUU^\top \hat{\mu})]XUU^\top\big\} \\
&+ (\beta - \bar{\mu})^\top \big\{\Sigma_\beta^{-1} - X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]X\big\} \\
&= (\hat{\mu} - \bar{\mu})\Sigma_\beta^{-1} + (\beta - \hat{\mu})^\top UU^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, XUU^\top \hat{\mu})]XUU^\top \\
&- (\beta - \bar{\mu})^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]X.
\end{aligned}$$

Appealing to Taylor's theorem, we can rewrite $\vec{\phi}''(Y, XUU^\top \hat{\mu})$ as

$$\begin{aligned}
\vec{\phi}''(Y, XUU^\top \hat{\mu}) &= \vec{\phi}''(Y, X\bar{\mu}) + (XUU^\top \hat{\mu} - X\bar{\mu}) \circ \vec{\phi}'''(Y, A) \\
&= \vec{\phi}''(Y, X\bar{\mu}) + (XUU^\top \hat{\mu} - X\hat{\mu} + X(\hat{\mu} - \bar{\mu})) \circ \vec{\phi}'''(Y, A) \\
&= \vec{\phi}''(Y, X\bar{\mu}) - X\bar{U}\bar{U}^\top \circ \vec{\phi}'''(Y, A) + X(\hat{\mu} - \bar{\mu}) \circ \vec{\phi}'''(Y, A) \\
&= \vec{\phi}''(Y, X\bar{\mu}) + R,
\end{aligned}$$

where the first line follows from Taylor's theorem by appropriately choosing each $A_i \in [x_i^\top UU^\top \hat{\mu}, x_i^\top \bar{\mu}]$, and in the fourth line we substitute in $R := -X\bar{U}\bar{U}^\top \circ \vec{\phi}'''(Y, A) + X(\hat{\mu} - \bar{\mu}) \circ \vec{\phi}'''(Y, A)$.

We now can rewrite the difference in the gradients as

$$\begin{aligned}
\nabla_\beta \log \hat{p}(\beta) - \nabla_\beta \log \bar{p}(\beta) &= (\hat{\mu} - \bar{\mu})\Sigma_\beta^{-1} \\
&\quad + (\beta - \hat{\mu})^\top U U^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})] X U U^\top \\
&\quad + (\beta - \hat{\mu})^\top U U^\top X^\top \mathrm{diag}(R) X U U^\top \\
&\quad - (\beta - \bar{\mu})^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu})) X \\
&= (\hat{\mu} - \bar{\mu})^\top (\Sigma_\beta^{-1} - U U^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})] X U U^\top) \\
&\quad + (\beta - \hat{\mu})^\top U U^\top X^\top \mathrm{diag}(R) X U U^\top \\
&\quad - (\beta - \bar{\mu})^\top U U^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu})) X \bar{U}\bar{U}^\top \\
&\quad - (\beta - \bar{\mu})^\top \bar{U}\bar{U}^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu})) X U U^\top \\
&\quad - (\beta - \bar{\mu})^\top \bar{U}\bar{U}^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu})) X \bar{U}\bar{U}^\top.
\end{aligned}$$

Which is obtained by first writing $X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})] X$ in the fourth line as $(U U^\top X^\top + \bar{U}\bar{U}^\top X^\top)\mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})](X U U^\top + X\bar{U}\bar{U}^\top)$, multiplying through and rearranging the resulting terms.

Given this form of the difference in the gradients, we may upper bound its norm as

$$\begin{aligned}
\|\nabla_\beta \log \hat{p}(\beta) - \nabla_\beta \log \bar{p}(\beta)\|_2 &\leq \|\hat{\mu} - \bar{\mu}\|_2 \|\Sigma_\beta^{-1} - U U^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})] X U U^\top\|_2 \\
&\quad + \|\beta - \hat{\mu}\|_2 \|U U^\top X^\top \mathrm{diag}(R) X U U^\top\|_2 \\
&\quad + \|\beta - \bar{\mu}\|_2 \|U U^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})] X \bar{U}\bar{U}^\top + \\
&\qquad \bar{U}\bar{U}^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})] X U U^\top + \bar{U}\bar{U}^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})] X \bar{U}\bar{U}^\top\|_2 \\
&\leq \|\hat{\mu} - \bar{\mu}\|_2 \|\Sigma_\beta^{-1} - U U^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})] X U U^\top\|_2 \\
&\quad + \|\beta - \hat{\mu}\|_2 \|U U^\top X^\top \mathrm{diag}(R) X U U^\top\|_2 \\
&\quad + \|\beta - \bar{\mu}\|_2 \big\{ \|\bar{U}\bar{U}^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})] X \bar{U}\bar{U}^\top\|_2 + \\
&\qquad 2\|\bar{U}\bar{U}^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})] X U U^\top\|_2 \big\}
\end{aligned}$$

by the triangle inequality.

$$\begin{aligned}
&\leq \|\hat{\mu} - \bar{\mu}\|_2 \Big\{ \|\Sigma_\beta^{-1}\|_2 + \|U\mathrm{diag}(\lambda)V^\top\|_2 \|\mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]\|_2 \|V\mathrm{diag}(\lambda)U^\top\|_2 \Big\} \\
&\quad + \|\beta - \hat{\mu}\|_2 \|U\mathrm{diag}(\lambda)V^\top\|_2 \|\mathrm{diag}(R)\|_2 \|V\mathrm{diag}(\lambda)U^\top\|_2 \\
&\quad + \|\beta - \bar{\mu}\|_2 \big\{ \|\bar{U}\mathrm{diag}(\bar{\lambda})\bar{V}^\top\|_2 \|\mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]\|_2 \|\bar{V}\mathrm{diag}(\bar{\lambda})\bar{U}^\top\|_2 + \\
&\qquad 2\|\bar{U}^\top \mathrm{diag}(\bar{\lambda})\bar{V}^\top\|_2 \|\mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]\|_2 \|V\mathrm{diag}(\lambda)U^\top\|_2 \big\}
\end{aligned}$$

by again using the triangle inequality, and decomposing $X^\top$ into $U\mathrm{diag}(\lambda)V^\top + \bar{U}\mathrm{diag}(\bar{\lambda})\bar{V}^\top$.

$$\leq \|\hat{\mu} - \bar{\mu}\|_2 \big(\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2 \|\vec{\phi}''\|_\infty\big) + \lambda_1^2 \|\beta - \hat{\mu}\|_2 \|R\|_\infty + (\bar{\lambda}_1^2 + 2\lambda_1\bar{\lambda}_1)\|\beta - \bar{\mu}\|_2 \|\vec{\phi}''\|_2,$$

where in the last line we have shortened $\vec{\phi}''(Y, X\bar{\mu})$ to $\vec{\phi}''$ for convenience.

Next noting that $\|\bar{\mu} - \hat{\mu}\|_2 \leq \bar{\lambda}_1 c$ for $c := \frac{\|\vec{\phi}'(Y, X\hat{\mu})\|_2 + \lambda_1 \|\bar{U}^\top \hat{\mu}\|_2 \|\vec{\phi}''(Y, A)\|_\infty}{\alpha}$, where α is the strong log concavity parameter of $p(\beta|X, Y)$ (which follows from Theorem 5.2), we can see that $\|R\|_\infty \leq \bar{\lambda}_1 r$ where $r := (\|U^\top \hat{\mu}\|_\infty \|\vec{\phi}'''(Y, A)\|_\infty + \lambda_1 c \|\vec{\phi}'''(Y, A)\|_\infty)$. That r is bounded follows from the assumption that $\log p(y|x, \beta)$ has bounded third derivatives, an equivalent to a Lipschitz condition on ϕ''. We can next simplify this upper bound to

$$\begin{aligned}
\|\nabla_\beta \log \hat{p}(\beta) - \nabla_\beta \log \bar{p}(\beta)\|_2 &\leq \bar{\lambda}_1 c [\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2 \|\vec{\phi}''\|_\infty] + \lambda_1^2 \bar{\lambda}_1 r \|\beta - \hat{\mu}\|_2 + \bar{\lambda}_1(\bar{\lambda}_1 + 2\lambda_1)\|\beta - \bar{\mu}\|_2 \|\vec{\phi}''\|_\infty \\
&= \bar{\lambda}_1 \big[c(\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2 \|\vec{\phi}''\|_\infty) + \lambda_1^2 r \|\beta - \hat{\mu}\|_2 + (\bar{\lambda}_1 + 2\lambda_1)\|\beta - \bar{\mu}\|_2 \|\vec{\phi}''\|_\infty\big] \\
&\leq \bar{\lambda}_1 \big[c(\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2 \|\vec{\phi}''\|_\infty) + \lambda_1^2 r \|\beta - \hat{\mu}\|_2 + (\bar{\lambda}_1 + 2\lambda_1)(\|\hat{\mu} - \bar{\mu}\|_2 + \|\beta - \hat{\mu}\|_2)\|\vec{\phi}''\|_\infty\big]
\end{aligned}$$

by the triangle inequality.

$$
\begin{aligned}
&\leq \bar{\lambda}_1\left[c(\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2\|\vec{\phi}''\|_\infty) + \lambda_1^2 r\|\beta - \hat{\mu}\|_2 + (\bar{\lambda}_1 + 2\lambda_1)(\bar{\lambda}_1 c + \|\beta - \hat{\mu}\|_2)\|\vec{\phi}''\|_\infty\right] \\
&= \bar{\lambda}_1\left[c(\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2\|\vec{\phi}''\|_\infty) + c(\bar{\lambda}_1^2 + 2\lambda_1\bar{\lambda}_1)\|\vec{\phi}''\|_\infty + (\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty)\|\beta - \hat{\mu}\|_2\right] \\
&= \bar{\lambda}_1\left[c(\|\Sigma_\beta^{-1}\|_2 + (\lambda_1 + \bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty) + (\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty)\|\beta - \hat{\mu}\|_2\right].
\end{aligned}
$$

Thus, taking the expectation of this upper bound on the norm squared over β with respect to $\hat{p}$ we get

$$
\begin{aligned}
d_{2,\hat{p}}^2(\hat{p}, p) &\leq \mathbb{E}_{\hat{p}(\beta)}\left(\bar{\lambda}_1^2\left\{c\left[\|\Sigma_\beta^{-1}\|_2 + (\lambda_1 + \bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right] + \left[\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty\right]\|\beta - \hat{\mu}\|_2\right\}^2\right) \\
&\leq 2\bar{\lambda}_1^2\mathbb{E}_{\hat{p}(\beta)}\left\{c^2\left[\|\Sigma_\beta^{-1}\|_2 + (\lambda_1 + \bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right]^2 + \left[\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty\right]^2\|\beta - \hat{\mu}\|_2^2\right\} \\
&\text{since } \forall a, b \in \mathbb{R}, (a+b)^2 \leq 2(a^2 + b^2) \\
&= 2\bar{\lambda}_1^2\left\{c^2\left[\|\Sigma_\beta^{-1}\|_2 + (\lambda_1 + \bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right]^2 + \left[\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty\right]^2\mathbb{E}_{\hat{p}(\beta)}[\|\beta - \hat{\mu}\|_2^2]\right\} \\
&= 2\bar{\lambda}_1^2\left\{c^2\left[\|\Sigma_\beta^{-1}\|_2 + (\lambda_1 + \bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right]^2 + \left[\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty\right]^2\operatorname{tr}(\hat{\Sigma})\right\}.
\end{aligned}
$$

Next noting that $\bar{p}$ is strongly $\|\bar{\Sigma}\|_2^{-1}$ log-concave, we may apply Theorem F.1, stated below, to obtain that

$$
\begin{aligned}
W_2(\hat{p}, \bar{p}) &\leq \|\bar{\Sigma}\|_2\sqrt{2\bar{\lambda}_1^2\left\{c^2\left[\|\Sigma_\beta^{-1}\|_2 + (\lambda_1 + \bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right]^2 + \left[\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty\right]^2\operatorname{tr}(\hat{\Sigma})\right\}} \\
&\leq \sqrt{2}\bar{\lambda}_1\|\bar{\Sigma}\|_2\left\{c\left[\|\Sigma_\beta^{-1}\|_2 + (\lambda_1 + \bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right] + \left[\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty\right]\sqrt{\operatorname{tr}(\hat{\Sigma})}\right\},
\end{aligned}
$$

which is our desired upper bound.

Theorem F.1. *Suppose that $p(\beta)$ and $q(\beta)$ are twice continuously differentiable and that q is α-strongly log concave. Then*

$$
W_2(p, q) \leq \alpha^{-1} d_{2,p}(p, q),
$$

where W_2 denotes the 2-Wasserstein distance between p and q.

Proof. This follows from Huggins et al. (2018) Theorem 5.2, or similarly from Bolley et al. (2012) Lemma 3.3 and Proposition 3.10. □

F.7. Proof of bounded asymptotic error

We here provide a formal statement and proof of Theorem 5.7, detailing the required regularity conditions.

Theorem F.2 (Asymptotic)**.** *Assume $x_i \overset{\text{i.i.d.}}{\sim} p_*$ for some distribution p_* such that $\mathbb{E}_{p_*}[x_i x_i^\top]$ exists and is non-singular with diagonalization $\mathbb{E}_{p_*}[x_i x_i^\top] = U_*^\top \operatorname{diag}(\lambda) U_* + \bar{U}_*^\top \operatorname{diag}(\bar{\lambda})\bar{U}_*$ such that $\min(\lambda) > \max(\bar{\lambda})$. Additionally, for a strictly concave (in its second argument), twice differentiable log-likelihood function ϕ with bounded second derivatives (in both arguments) and some $\beta \in \mathbb{R}^D$, let $y_i|x_i \sim \exp\{\phi(y_i, x_i^T\beta)\}$. Also, suppose that $\mathbb{E}\|y_i\|_2^2 < \infty$. Then if $p(\beta)$ is log-concave and positive on $\mathbb{R}^D$, the asymptotic error (in N) of the exact relative to approximate maximum a posteriori parameters, $\hat{\mu} = \lim_{N\to\infty}\hat{\mu}_N$ and $\bar{\mu} = \lim_{N\to\infty}\bar{\mu}_N$ is finite (where $\hat{\mu}_N$ and $\bar{\mu}_N$ are the approximate and exact MAP estimates, respectively, after N data-points), i.e., $\lim_{n\to\infty}\|\hat{\mu}_N - \bar{\mu}_N\|$ exists and is finite.*

Proof. Before beginning, let $\mathbb{P}$ denote a Borel probability measure on the sample space on which our random variables, $\{x_i\}$ and $\{y_i\}$, are defined such that these random variables are distributed as assumed according to $\mathbb{P}$. In what follows we demonstrate the asymptotic error is finite $\mathbb{P}$-almost surely. To this end, it suffices to show that $\hat{\mu}_N \overset{a.s.}{\to} \hat{\mu}$ and $\bar{\mu}_N \overset{a.s.}{\to} \bar{\mu}$ for some $\hat{\mu}, \bar{\mu}$ in $\mathbb{R}^D$.

Strong convergence of the exact MAP ($\bar{\mu}_N \overset{a.s.}{\to} \bar{\mu}$)

This follows from Doob's consistency theorem (Van der Vaart, 2000, Theorem 10.10). The only nuance required in the application of this theorem here is that we must accommodate the regression setting. However by constructing a single measure $\mathbb{P}$ governing both the covariates and responses, this simply becomes a special case of the usual theorem for unconditional models.

Strong convergence of the approximate MAP ($\hat{\mu}_N \overset{a.s.}{\to} \hat{\mu}$)

In contrast to the strong consistency of $\bar{\mu}_N$, showing convergence of $\hat{\mu}_N$ requires more work. This is because we cannot rely on standard results such as Bernstein–Von Mises or Doob's consistency theorem, which require correct model specification. Since we have introduced the likelihood approximation $\tilde{p}(y|x,\beta) \neq p(y|x,\beta)$, the vector $\hat{\mu}_N$ is the MAP estimate under a misspecified model.

We demonstrate almost sure convergence in two steps; first we show that $U_*^\top \hat{\mu}_N$ converges almost surely to some $\gamma^* \in \mathbb{R}^M$; then we show that $\hat{\mu}_N = U_* U_*^\top \hat{\mu}_N + \bar{U}_N \bar{U}_N^\top \hat{\mu}_N$ must converge as a result. Since $U_N U_N^\top \overset{a.s.}{\to} U_* U_*^\top$ (as follows from entry-wise almost sure convergence of $\frac{1}{N} X^\top X \to \mathbb{E}_{p_*}[x_i x_i^\top]$ and the Davis–Kahan Theorem (Davis & Kahan, 1970)), this guarantees strong convergence of $\hat{\mu}_N = U_N U_N^\top \hat{\mu}_N + \bar{U}_N \bar{U}_N^\top \hat{\mu}_N$.

Part I: strong convergence of the projected approximate MAP, $U_* \hat{\mu}_N \overset{a.s.}{\to} \gamma^*$

Let $U_* \in \mathbb{R}^{D,M}$ be the top M eigenvectors of $\mathbb{E}_{p_*}[x_i x_i^\top]$, and recall that by assumption for any $y, \phi(y, x_i^T \beta)$ is a strictly concave function of $x_i^\top \beta$, in the sense that for any y and any b, b' in $\mathbb{R}$ and t in $(0,1)$ with $b \neq b'$, $\phi(y, tb + (1-t)b') > t\phi(y,b) + (1-t)\phi(y,b')$. Then by Lemma F.2 we have that there is a unique maximizer $\gamma^* = \arg\max_{\gamma \in R^M} \mathbb{E}[\phi(y, x^\top U_* \gamma)]$

We next note that the Hessian of the expected approximate negative log likelihood with respect to γ is positive definite everywhere,

$$\nabla_\gamma^2 - \mathbb{E}_{y \sim p(y|x,\beta), x \sim p_*}[\phi(y, x^\top U_* \gamma)] = -\mathbb{E}[\big(\nabla_\gamma \phi'(y, x^\top U_* \gamma)\big) x^\top U_*] = -U_*^\top \mathbb{E}[x \phi''(y, x^\top U_* \gamma) x^\top] U_* \succ 0$$

since the strict log concavity and twice differentiability of ϕ ensure that $-\mathbb{E}[x \phi''(y, x^\top U_* \gamma) x^\top] \succ 0$.

Now consider any compact neighborhood $K \subset \mathbb{R}^M$ containing γ^* as an interior point. Then, by Lemma F.3 the set $\mathcal{F} = \{f_\gamma : X \times Y \to \mathbb{R}, (x,y) \mapsto \phi(y, x^\top U_* \gamma) | \gamma \in K\}$ is $\mathbb{P}$-Glivenko–Cantelli. As such $\sup_{f_\gamma \in \mathcal{F}} |\frac{1}{N} \sum_{i=1}^N f_\gamma(x_i, y_i) - \mathbb{E}[f_\gamma(x_i, y_i)]| \overset{a.s.}{\to} 0$, that is to say, the empirical average log-likelihood converges uniformly to its expectation across all $\gamma \in K$. As a result, we have that for $\gamma_N := \arg\max_{\gamma \in K} \log \tilde{p}(U_* \beta = \gamma | X, Y) = \arg\max_{\gamma \in K} \frac{1}{N} \big[\log p(U_*^T \beta = \gamma) + \sum_{i=1}^N \phi(y_i, x_i^\top U_* \gamma)\big]$, $\gamma_N \overset{a.s.}{\to} \gamma^*$.

It remains in this part only to show that convergence of the approximate MAP parameter within this subset K implies convergence of $U_*^\top \hat{\mu}$, the approximate MAP parameter (across all of $\mathbb{R}^M$). However, this follows immediately from the strict log concavity of the posterior; because $\gamma^* \in K^\circ$, for N large enough each $\gamma_N \in K^\circ$ and we may construct a sub-level set such that $\gamma_N \in C_N \subset K$ such that $\forall \gamma \notin C_N, \log p(U_*^\top \beta = \gamma) + \sum_{i=1}^N \phi(y_i, x_i^\top U_* \gamma) < \log p(U_*^\top \beta = \gamma_N) + \sum_{i=1}^N \phi(y_i, x_i^\top U_* \gamma_N)$.

Part II: convergence of $\bar{U}_* \bar{U}_*^\top \hat{\mu}_N + U_* \gamma$

Using the result of Part I, we can write that $\hat{\mu}_N = U_* U_*^\top \hat{\mu}_N + \bar{U}_* \bar{U}_*^\top \hat{\mu}_N \to U_* \gamma^* + \bar{U}_* \bar{U}_*^\top \hat{\mu}_N$. However, since $\bar{U}_* \bar{U}_*^\top \beta \perp X, Y | U_*^\top \beta$ under $\mathbb{P}$, convergence of $U_*^\top \hat{\mu}_N \to \gamma^*$ implies convergence of $\arg\max_{\bar{U}_*^\top \beta} \tilde{p}(\bar{U}_*^\top \beta | U_*^\top \beta = U_*^\top \hat{\mu}_N, X, Y) = \arg\max_{\bar{U}_*^\top \beta} \tilde{p}(\bar{U}_*^\top \beta | U_*^\top \beta = U_*^\top)$ to some $\bar{U}_*^\top \hat{\mu}_N$ since continuity of $p(\beta)$ and $\tilde{p}(Y|X,\beta)$ imply continuity of the arg-max. Thus both $\hat{\mu}_N$ and $\bar{\mu}_N$ converge, guaranteeing convergence of the asymptotic error. □

Lemma F.2. *For any $\phi(\cdot,\cdot)$ which is strictly concave in its second argument, if there is a global maximizer $\beta^* = \arg\max_{\beta \in \mathbb{R}^D} V(\beta) = \mathbb{E}_{x \sim p_*, y \sim p(y|x,\beta)}[\phi(y, x^\top \beta)]$, then there is a unique global maximizer,*

$$\gamma^* = \arg\max_{\gamma \in \mathbb{R}^M} V(U_* \gamma)$$

Proof. We first note that $V(\cdot)$ must have bounded sub-level sets. Thus $W(\cdot) := V(U_* \cdot)$ must also have bounded sub-level sets since $V^{-1}([a,\infty]) = \{\beta | V(\beta) \geq a\} \supset \{\beta | \exists \gamma \in \mathbb{R}^M \ s.t. \ \beta = U_* \gamma \ and \ V(U_* \gamma) \geq a\} = U_* W^{-1}([a,\infty])$. Thus, since W is strictly concave and has bounded sub-level sets, it has a unique maximizer. □

Lemma F.3. *Let $K \subset \mathbb{R}^M$ be compact and denote by X and Y the domains of the covariates and responses, respectively. Then under the assumptions of Theorem F.2, the set $\mathcal{F} = \{f_\gamma : X \times Y \to \mathbb{R}, (x, y) \mapsto \phi(y, x^\top U\gamma) | \gamma \in K\}$ is $\mathbb{P}$-Glivenko–Cantelli.*

Proof. This result follows from Theorem 19.4 in (Van der Vaart, 2000), and builds from example 19.7 of the same reference; in particular, the condition of bounded second derivatives of ϕ implies that for any $f_\gamma, f_{\gamma'}$ in $\mathcal{F}$ and x in X, y in Y, we have $|f_\gamma(x, y) - f_{\gamma'}(x, y)| \leq C\|x\|_2^2$. The previous condition is sufficient to ensure finite bracketing numbers, and the result follows. Notably, in keeping with example 19.7 we have that for all x, y and for all γ and γ' in K,

$$
\begin{aligned}
|f_\gamma(x, y) - f_{\gamma'}(x, y)| &= |\int_{x^\top U\gamma'}^{x^\top U\gamma} \phi'(y, a)da| \\
&= |\int_{x^\top U\gamma'}^{x^\top U\gamma} \phi'(y, x^\top U\gamma') + \int_{x^\top U\gamma'}^{a} \phi''(y, b)db\, da| \\
&\leq \|x^\top U(\gamma - \gamma')\phi'(y, x^\top U\gamma')\|_2 + \frac{1}{2}\|x^\top U(\gamma - \gamma')\|_2^2 \sup_{a\in\mathbb{R}} \phi''(y, a) \\
&\leq \|x^\top U\|_2(\|y\|_2 + \|x^\top U\|_2\|\gamma'\|_2)\phi''_{\max}\|\gamma - \gamma'\|_2 + \frac{1}{2}\|x^\top U\|_2^2\|\gamma - \gamma'\|_2^2\phi''_{\max} \\
&\leq [\frac{3}{2}\|x^\top U\|_2^2 \mathrm{diam}(K)\phi''_{\max} + \|x^\top U\|_2\|y\|_2\phi''_{\max}]\|\gamma - \gamma'\|_2 \\
&\leq C(\|x^\top U\|_2^2 + \|x^\top U\|_2\|y\|_2)\|\gamma - \gamma'\|_2,
\end{aligned}
\tag{F.7}
$$

where in the first and second lines we use the fundamental theorem of calculus, and in the fourth and fifth lines we rely on the boundedness of the second derivatives of ϕ and that the compactness subsets of $\mathbb{R}^M$ implies boundedness. In the final line C is an absolute constant.

Finally, we note that $\mathbb{E}_\mathbb{P}\|x^\top U\|_2^2 < \infty$ since $\mathbb{E}_\mathbb{P}\|x^\top U\|_2^2 = \mathbb{E}_\mathbb{P}x^\top UU^\top x < \mathbb{E}_\mathbb{P}x^\top x = \mathrm{Tr}(\mathbb{E}_{p_*}xx^\top) < \infty$, and by Cauchy Schwartz, $\mathbb{E}_\mathbb{P}\|x^\top U\|_2\|y\|_2 \leq \sqrt{\mathbb{E}_\mathbb{P}\|x^\top U\|_2^2\mathbb{E}_\mathbb{P}\|y\|_2^2} \leq \infty$. This confirms (as in example 19.7 (Van der Vaart, 2000)) that for all $\epsilon > 0$, the ϵ-bracketing number of $\mathcal{F}$ is finite. By Theorem 19.4 of (Van der Vaart, 2000), this proves that $\mathcal{F}$ is $\mathbb{P}$-Glivenko-Cantelli. □

F.8. Factorized Laplace approximations underestimate marginal variances

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

$$\Sigma = \begin{bmatrix} a & b \\ b & c \end{bmatrix},$$

for which the Hessian evaluated anywhere is

$$\Sigma^{-1} = \frac{1}{ac - b^2}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}.$$

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

$$\tilde{\Sigma} = \begin{bmatrix} a - \frac{b^2}{c} & 0 \\ 0 & c - \frac{b^2}{a} \end{bmatrix}.$$

This approximation reports marginal variances which are lower than the exact marginal variances.

That this approximation underestimates marginal variances in the more general $D > 2$ dimensional case may be easily seen from considering the block matrix inversion of Σ, with blocks of dimension 1×1, $(D-1) \times 1$, $1 \times (D-1)$ and $(D-1) \times (D-1)$, and noting that the Schur complement of a positive definite covariance matrix will always be positive definite.

G. LR-MCMC

We provide the LR-MCMC algorithm for performing fast MCMC in generalized linear models with low-rank data approximations.

Algorithm 2 LR-MCMC for Bayesian inference in GLMs with low-rank data approximations.

	Pseudo-Code	**Time Complexity**	**Memory Complexity**
1:	***Input:*** *prior $p(\beta)$, data $X \in \mathbb{R}^{N,D}$, rank $M \ll D$, GLM mapping ϕ, MCMC transition kernel $q(\cdot,\cdot)$, number of MCMC iterations T. Time and memory complexities that are not included depend on the specific choice of MCMC transition kernel.*		
5:	*Data preprocessing — M-Truncated SVD*		
6:	$U, \text{diag}(\lambda), V := \text{truncated-SVD}(X^\top, M)$	$O(NDM)$	$O(NM + DM)$
7:	$X_U = XU$	$O(NM)$	$O(NDM)$
8:	*Propose $\beta^{(t)} \in \mathbb{R}^D$, compute likelihood*		
9:	$\beta^{(t)} \sim q(\beta^{(t)}, \beta^{(t-1)})$	—	—
10:	$\mathcal{L}_t := \sum_{i=1}^N \phi(y_i, x_i^\top U U^\top \beta^{(t)}) + \log p(\beta^{(t)})$	$O(1)$	$O(NM + MD)$
11:	*Accept or Reject*		
12:	Acceptance probability $p_A := \min\left(1, \frac{\mathcal{L}_t}{\mathcal{L}_{t-1}}\right)$	$O(1)$	$O(1)$
13:	Accept $\beta^{(t)}$ with probability p_A	$O(1)$	$O(1)$
14:	*Repeat steps 3-6 for T iterations*		

The transition in Line 9 may additionally benefit from the LR-GLM approximation. In particular, widely used algorithms such as Hamiltonian Monte Carlo and the No-U-Turn Sampler rely on many $O(ND)$-time likelihood and gradient evaluations, the cost of which can be reduced to $O(NM + DM)$ with LR-GLM. An implementation of this approximation is given in the `Stan` model in Appendix A.3 with performance results in Figures A.3 and A.6.

H. LR-Laplace with non-Gaussian priors

As discussed in the main text, we can maintain computational advantages of LR-GLM even when we have non-Gaussian priors. This admits the procedure provided in Algorithm 3.

In order for this more general LR-Laplace algorithm to be computationally efficient, we still require that the prior have some properties which can accommodate efficiency. In particular Line 11 demands that the Hessian of the prior is computed and inverted, as will true even in the high-dimensional setting when, for example, the prior factorizes across dimensions. Additionally, properties of the prior such as log concavity will facilitate efficient optimisation in Line 8.

[7] To keep notation concise we use $\vec{\phi}''_{\hat{\mu}}$ to denote $\vec{\phi}''(Y, XUU^\top \hat{\mu})$

Algorithm 3 LR-Laplace for Bayesian inference in GLMs with low-rank data approximations and twice differentiable prior. Time and memory complexities which are not included depend on the choice of prior and optimisation method, which can be problem specific.

	Pseudo-Code	**Time Complexity**	**Memory Complexity**
1:	***Input:*** *twice differentiable prior $p(\beta)$, data $X \in \mathbb{R}^{N,D}$, rank $M \ll D$, GLM mapping ϕ with ϕ'' (see Eq. (5) and Section 5.1)*		
5:	*Data preprocessing — M-Truncated SVD*		
6:	$U, \mathrm{diag}(\lambda), V := \text{truncated-SVD}(X^T, M)$	$O(NDM)$	$O(NM + DM)$
7:	*Optimize to find approximate MAP estimate (in D-dimensional space)*		
8:	$\hat{\mu} := \arg\max_{\mu \in \mathbb{R}^D} \sum_{i=1}^{N} \phi(y_i, x_i U U^\top \mu) + \log p(\beta = \mu)$	—	—
9:	*Compute approximate posterior covariance*[7]		
10:	$\hat{\Sigma}^{-1} := -\nabla^2_\beta \log p_\beta(\hat{\mu}) - UU^\top X^\top \mathrm{diag}(\vec{\phi}''_{\hat{\mu}}) XUU^\top$	—	—
11:	$K := [\nabla^2_\beta \log p(\beta)|_{\beta=\hat{\mu}}]^{-1}$	—	—
12:	$\hat{\Sigma} := -K + KU\left([U^\top X^\top \mathrm{diag}(\vec{\phi}''_{\hat{\mu}}) XU]^{-1} + U^\top KU\right)^{-1} U^\top K$	—	—
13:	*Compute variances and covariances of parameters*		
14:	$\mathrm{Var}_{\hat{p}}(\beta_i) = e_i^\top \hat{\Sigma} e_i$	—	—
15:	$\mathrm{Cov}_{\hat{p}}(\beta_i, \beta_j) = e_i^\top \hat{\Sigma} e_j$	—	—