LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

Brian L. Trippe [1] **Jonathan H. Huggins** [2] **Raj Agrawal** [1] **Tamara Broderick** [1]

Abstract

Due to the ease of modern data collection, applied statisticians often have access to a large set of covariates that they wish to relate to some observed outcome. Generalized linear models (GLMs) offer a particularly interpretable framework for such an analysis. In these high-dimensional problems, the number of covariates is often large relative to the number of observations, so we face non-trivial inferential uncertainty; a Bayesian approach allows coherent quantification of this uncertainty. Unfortunately, existing methods for Bayesian inference in GLMs require running times roughly cubic in parameter dimension, and so are limited to settings with at most tens of thousand parameters. We propose to reduce time and memory costs with a low-rank approximation of the data in an approach we call LR-GLM. When used with the Laplace approximation or Markov chain Monte Carlo, LR-GLM provides a full Bayesian posterior approximation and admits running times reduced by a full factor of the parameter dimension. We rigorously establish the quality of our approximation and show how the choice of rank allows a tunable computational–statistical trade-off. Experiments support our theory and demonstrate the efficacy of LR-GLM on real large-scale datasets.

1. Introduction

Scientists, engineers, and social scientists are often interested in characterizing the relationship between an outcome and a set of covariates, rather than purely optimizing predictive accuracy. For example, a biologist may wish to understand the effect of natural genetic variation 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 yields particularly large sets of covariates for data analysts to study. While these rich data should ultimately aid understanding, they pose a number of practical challenges for data analysis. One challenge is how to discover interpretable relationships between the covariates and the outcome. Generalized linear models (GLMs) are widely used in part because they provide such interpretability – as well as the flexibility to accommodate a variety of different outcome types (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 covariates, there is likely to be non-trivial uncertainty about these relationships.

A Bayesian approach to GLM inference provides the desired coherent uncertainty quantification as well as favorable calibration properties (Dawid, 1982, Theorem 1). Bayesian methods additionally provide the ability to improve inference by incorporating expert information and sharing power across experiments. Using Bayesian GLMs leads to computational challenges, however. Even when the Bayesian posterior can be computed exactly, conjugate inference costs $O(N^2D)$ in the case of $D \gg N$. And most models are sufficiently complex as to require expensive approximations.

In this work, we propose to reduce the effective dimensionality of the feature set as a pre-processing step to speed up Bayesian inference, while still performing inference in the original parameter space; in particular, we show that low-rank descriptions of the data permit fast Markov chain Monte Carlo (MCMC) samplers 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 are exactly low-rank. When the data are merely approximately low-rank, our proposal is an approximation. Through both theory and experiments, we demonstrate that low-rank data approximations provide a number of properties that are desirable in an efficient posterior approximation method: (1) *soundness:* our approximations admit error bounds directly on the quantities that practitioners report as well as practical interpretations

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA [2]Department of Biostatistics, Harvard, Cambridge, MA. Correspondence to: Brian L. Trippe <btrippe@mit.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Table 1. Time complexities of naive inference and LR-GLM with a rank M approximation when $D \geq N$.

METHOD	NAIVE	LR-GLM	SPEEDUP
LAPLACE	$O(N^2D)$	$O(NDM)$	N/M
MCMC (ITER.)	$O(ND)$	$O(NM + DM)$	N/M

of those bounds; (2) *tunability:* the choice of the rank of the approximation defines a tunable trade-off between the computational demands of inference and statistical precision; and (3) *conservativeness:* our approximation never reports less uncertainty than the exact posterior, where uncertainty is quantified via either posterior variance or information entropy. Together, these properties allow a practitioner to choose how much information to extract from the data on the basis of computational resources while being able to confidently trust the conclusions of their analysis.

2. Bayesian inference in GLMs

Suppose we have N data points $\{(x_n, y_n)\}_{n=1}^N$. We collect our covariates, where x_n has dimension D, 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 GLM likelihood $p(Y \mid X, \beta) = p(Y \mid X\beta)$. That is, β_d describes the effect size of the dth covariate (e.g., the influence of a non-reference allele on an individual's height in a genomic association study). Completing our Bayesian model specification, we assume a prior $p(\beta)$, which describes our knowledge of β before seeing data. Bayes' theorem gives the Bayesian posterior $p(\beta \mid Y, X) = p(\beta)p(Y \mid X\beta)/\int p(\beta')p(Y \mid X\beta')d\beta'$, which captures the updated state of our knowledge after observing data. We often summarize the β posterior via its mean and covariance. In all but the simplest settings, though, computing these posterior summaries is analytically intractable, and these quantities must be approximated.

Related work. In the setting of large D and large N, existing Bayesian inference methods for GLMs may lead to unfavorable trade-offs between accuracy and computation; see Appendix B for further discussion. While Markov chain Monte Carlo (MCMC) can approximate Bayesian GLM posteriors arbitrarily well given enough time, standard methods can be slow, with $O(DN)$ time per likelihood evaluation. Moreover, in practice, mixing time may scale poorly with dimension and sample size; algorithms thus require many iterations and hence many likelihood evaluations. Subsampling MCMC methods can speed up inference, but they are effective only with tall data ($D \ll N$; Bardenet et al., 2017).

An alternative to MCMC is to use a deterministic approximation such as the Laplace approximation (Bishop, 2006, Chap. 4.4), integrated nested Laplace approximation (Rue et al., 2009), variational Bayes (VB; Blei et al., 2017), or an alternative likelihood approximation (Huggins et al., 2017; Campbell & Broderick, 2019; Huggins et al., 2016). However these methods are computationally efficient only when $D \ll N$ (and in some cases also when $N \ll D$). For example, the Laplace approximation requires inverting the Hessian, which uses $O(\min(N, D)ND)$ time (Appendix C). Improving computational tractability by, for example, using a mean field approximation with VB or a factorized Laplace approximation can produce substantial bias and uncertainty underestimation (MacKay, 2003; Turner & Sahani, 2011).

A number of papers have explored using random projections and low-rank approximations in both Bayesian (Lee & Oh, 2013; Spantini et al., 2015; Guhaniyogi & Dunson, 2015; Geppert et al., 2017) and non-Bayesian (Zhang et al., 2014; Wang et al., 2017) settings. The Bayesian approaches have a variety of limitations. E.g., Lee & Oh (2013); Geppert et al. (2017); Spantini et al. (2015) give results only for certain conjugate Gaussian models. And Guhaniyogi & Dunson (2015) provide asymptotic guarantees for prediction but do not address parameter estimation.

See Section 6 for a demonstration of the empirical disadvantages of mean field VB, factored Laplace, and random projections in posterior inference.

3. LR-GLM

The intuition for our low-rank GLM (LR-GLM) approach is as follows. Supervised learning problems in high-dimensional settings often exhibit strongly correlated covariates (Udell & Townsend, 2019). In these cases, the data may provide little information about the parameter along certain directions of parameter space. This observation suggests the following procedure: first identify a relatively lower-dimensional subspace within which the data most directly inform the posterior, and then perform the data-dependent computations of posterior inference (only) within this subspace, at lower computational expense. In the context of GLMs with Gaussian priors, the singular value decomposition (SVD) of the design matrix X provides a natural and effective mechanism for identifying a subspace. We will see that this perspective gives rise to simple, efficient, and accurate approximate inference procedures. In models with non-Gaussian priors the approximation enables more efficient inference by facilitating faster likelihood evaluations.

Formally, the first step of LR-GLM is to choose an integer M such that $0 < M < D$. For any real design matrix X, its SVD exists and may be written as

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

where $U \in \mathbb{R}^{D\times M}, \bar{U} \in \mathbb{R}^{D\times(D-M)}, V \in \mathbb{R}^{N\times M}$, and $\bar{V} \in \mathbb{R}^{N\times(D-M)}$ are matrices of orthonormal rows, and

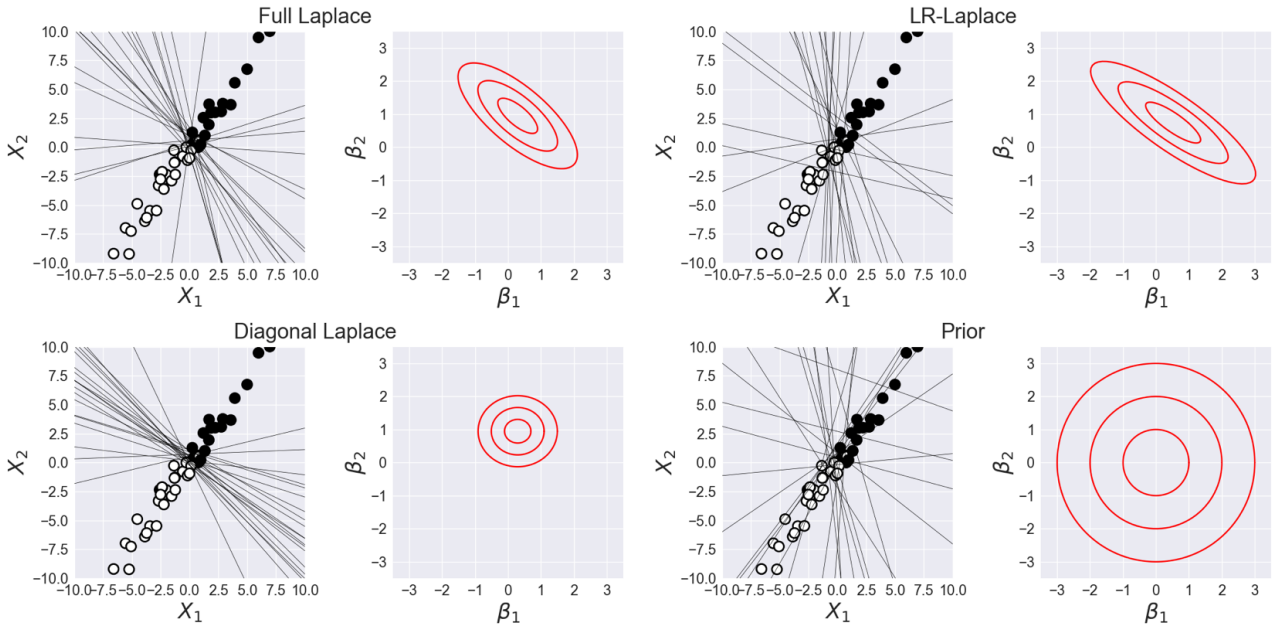

Figure 1. LR-Laplace with a rank-1 data approximation closely matches the Bayesian posterior of a toy logistic regression model. In each pair of plots, the left panel depicts the same 2-dimensional dataset with points in two classes (black and white dots) and decision boundaries (black lines) separating the two classes, which are sampled from the given posterior approximation (see title for each pair). In the right panel, the red contours represent the marginal posterior approximation of the parameter β (a bias parameter is integrated out).

$\lambda \in \mathbb{R}^M$ and $\bar{\lambda} \in \mathbb{R}^{D-M}$ are vectors of non-increasing singular values $\lambda_1 \geq \cdots \geq \lambda_M \geq \bar{\lambda}_1 \geq \cdots \geq \bar{\lambda}_{D-M} \geq 0$. We replace X with the low-rank approximation $XUU^\top$. Note that the resulting posterior approximation $\tilde{p}(\beta \mid X, Y)$ is still a distribution over the full D-dimensional β vector:

$$\tilde{p}(\beta \mid X, Y) := \frac{p(\beta)p(Y \mid XUU^\top\beta)}{\int p(\beta')p(Y \mid XUU^\top\beta')d\beta'} \qquad (1)$$

In this way, we cast low-rank data approximations for approximate Bayesian inference as a likelihood approximation. This perspective facilitates our analysis of posterior approximation quality and provides the flexibility either to use the likelihood approximation in an otherwise exact MCMC algorithm or to make additional fast approximations such as the Laplace approximation.

We let *LR-Laplace* denote the combination of LR-GLM and the Laplace approximation. Figure 1 illustrates LR-Laplace on a toy problem and compares it to full Laplace, the prior, and diagonal Laplace. Diagonal Laplace refers to a factorized Laplace approximation in which the Hessian of the log posterior is approximated with only its diagonal. While this example captures some of the essence of our proposed approach, we emphasize that our focus in this paper is on problems that are high-dimensional.

4. Low-rank data approximations for conjugate Gaussian regression

We now consider the quality of approximate Bayesian inference using our LR-GLM approach in the case of conjugate Gaussian regression. We start by assuming that the data is exactly low rank since it most cleanly illustrates the computational gains from LR-GLM. We then move on to the case of conjugate regression with approximately low-rank data and rigorously characterize the quality of our approximation via interpretable error bounds. We consider non-conjugate GLMs in Section 5. We defer all proofs to the Appendix.

4.1. Conjugate regression with exactly low-rank data

Classic linear regression fits into our GLM likelihood framework with $p(Y|X,\beta) = \mathcal{N}(Y|X\beta, (\tau I_N)^{-1})$, where $\tau > 0$ is the precision and I_N is the identity matrix of size N. For the conjugate prior $p(\beta) = \mathcal{N}(\beta|0, \Sigma_\beta)$, we can write the posterior in closed form: $p(\beta|Y,X) = \mathcal{N}(\beta|\mu_N, \Sigma_N)$, where $\Sigma_N := (\Sigma_\beta^{-1} + \tau X^T X)^{-1}$ and $\mu_N := \tau\Sigma_N X^\top Y$.

While conjugacy avoids the cost of approximating Bayesian inference, it does not avoid the often prohibitive $O(ND^2 + D^3)$ cost of calculating Σ_N (which requires computing and then inverting Σ_N^{-1}) and the $O(D^2)$ memory demands of storing it. In the $N \ll D$ setting, these costs can be mitigated by using the Woodbury formula to obtain μ_N and Σ_N in $O(N^2D)$ time with $O(ND)$ memory (Appendix C). But this alternative becomes computationally prohibitive as well

when both N and D are large (e.g., $D \approx N > 20,000$).

Now suppose that X is rank $M \ll \min(D, N)$ and can therefore be written as $X = XUU^T$ exactly, where $U \in \mathbb{R}^{D\times M}$ denotes the top M right singular vectors of X. Then, if $\Sigma_\beta = \sigma_\beta^2 I_D$ and 1_M is the ones vector of length M, we can write (see Appendix D.1 for details)

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

where multiplication ($\odot$) and division in the diag input are component-wise across the vector λ. Eq. (2) provides a more computationally efficient route to inference. The singular vectors in U may be obtained in $O(ND \log M)$ time via a randomized SVD (Halko et al., 2011) or in $O(NDM)$ time using more standard deterministic methods (Press et al., 2007). The bottleneck step is finding λ via $\mathrm{diag}(\lambda \odot \lambda) = U^\top X^\top X U$, which can be computed in $O(NDM)$ time. As for storage, this approach requires keeping only U, λ, and $V^\top Y$, which takes just $O(MD)$ space. In sum, utilizing low-rank structure via Eq. (2) provides an order $\min(N, D)/M$-fold improvement in both time and memory over more naive inference.

4.2. Conjugate regression with low-rank approximations

While the case with exactly low-rank data is illustrative, real data are rarely exactly low rank. So, more generally, LR-GLM 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 Section 4.1. We next provide upper bounds on the error from our approximation. Since practitioners typically report posterior means and covariances, we focus on how well LR-GLM approximates these functionals.

Theorem 4.1. *For conjugate Bayesian linear regression, the LR-GLM approximation Eq. (1) satisfies*

$$\|\tilde{\mu}_N - \mu_N\|_2 \leq \frac{\bar{\lambda}_1 \left(\bar{\lambda}_1 \|\bar{U}^\top \tilde{\mu}_N\|_2 + \|\bar{V}^\top Y\|_2 \right)}{\|\tau\Sigma_\beta\|_2^{-1} + \bar{\lambda}_{D-M}^2} \quad (3)$$
$$\text{and} \quad \Sigma_N^{-1} - \tilde{\Sigma}_N^{-1} = \tau(X^\top X - UU^\top X^\top X UU^\top). \quad (4)$$

In particular, $\|\Sigma_N^{-1} - \tilde{\Sigma}_N^{-1}\|_2 = \tau\bar{\lambda}_1^2$.

The major driver of the approximation error of the posterior mean and covariance is $\bar{\lambda}_1 = \|X - XUU^\top\|_2$, the largest truncated singular value of X. This result accords with the intuition that if the data are "approximately low-rank" then LR-GLM should perform well.

The following corollary shows that the posterior mean estimate is not, in general, consistent for the true parameter. But it does exhibit reasonable asymptotic behavior. In particular, $\tilde{\mu}_N$ is consistent within the span of U and converges to the *a priori* most probable vector with this characteristic (see the toy example in Figure E.1).

Corollary 4.2. *Suppose* $x_n \overset{i.i.d.}{\sim} p_*$, *for some distribution* p^*, *and* $y_n \mid x_n \overset{indep}{\sim} \mathcal{N}(x_n^\top \mu_*, \tau^{-1})$, *for some* $\mu_* \in \mathbb{R}^D$. *Assume* $\mathbb{E}_{p_*}[x_n x_n^\top]$ *is nonsingular. Let the columns of* $U_* \in \mathbb{R}^{D\times M}$ *be the top eigenvectors of* $\mathbb{E}_{p_*}[x_n x_n^T]$. *Then* $\tilde{\mu}_N$ *converges weakly to the maximum a priori vector* $\tilde{\mu}$ *satisfying* $U_*^\top \tilde{\mu} = U_*^\top \mu_*$.

In the special case that Σ_β is diagonal this result implies that $\tilde{\mu}_N \xrightarrow{p} U_* U_*^\top \mu_*$ (Appendices E.3 and F.2). Thus Corollary 4.2 reflects the intuition that we are not learning anything about the relation between response and covariates in the data directions that we truncate away with our approach. If the response has little dependence on these directions, $\bar{U}_* \bar{U}_*^\top \mu_* = \lim_{N\to\infty} \tilde{\mu}_N - \mu_*$ will be small and the error in our approximation will be low (Appendix E.3). If the response depends heavily on these directions, our error will be higher. This challenge is ubiquitous in dealing with projections of high-dimensional data. Indeed, 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).

Our next corollary captures that LR-GLM never underestimates posterior uncertainty (the *conservativeness* property).

Corollary 4.3. *LR-GLM approximate posterior uncertainty in any linear combination of parameters is no less than the exact posterior uncertainty. Equivalently,* $\tilde{\Sigma}_N - \Sigma_N$ *is positive semi-definite.*

See Figure 1 for an illustration of this result. 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. This behavior stands in sharp contrast to alternative fast approximate inference methods, such as diagonal Laplace approximations (Appendix F.8) and variational Bayes (MacKay, 2003), which can dramatically underestimate uncertainty. We further characterize the conservativeness of LR-GLM in Corollary E.1, which shows that the LR-GLM posterior never has lower entropy than the exact posterior and quantifies the bits of information lost due to approximation.

[1]This manipulation is purely symbolic. See Appendix F.1 for details.

Algorithm 1 LR-Laplace for Bayesian inference in GLMs with low-rank data approximations and zero-mean prior – with computation costs. See Appendix H for the general algorithm.

Input: *prior* $p(\beta) = \mathcal{N}(\mathbf{0}, \Sigma_\beta)$*, data* $X \in \mathbb{R}^{N,D}$*, rank* $M \ll D$*, GLM mapping* ϕ *with* $\vec{\phi}''$ *(see Eq. (5) and Section 5.1)*

Pseudo-Code	**Time Complexity**	**Memory Complexity**
Data preprocessing — M-Truncated SVD		
$U, \mathrm{diag}(\lambda), V := \text{truncated-SVD}(X^T, M)$	$O(NDM)$	$O(NM + DM)$
Optimize in projected space and find approximate MAP estimate		
$\gamma_* := \arg\max_{\gamma \in \mathbb{R}^M} \sum_{i=1}^N \phi(y_i, x_i U\gamma) - \frac{1}{2}\gamma^\top U^\top \Sigma_\beta U\gamma$	$O(NM + DM^2)$	$O(N + M^2)$
$\hat{\mu} = U\gamma_* + \bar{U}\bar{U}^\top \Sigma_\beta U(U^\top \Sigma_\beta U)^{-1}\gamma_*$	$O(DM)$	$O(DM)$
Compute approximate posterior covariance		
$W^{-1} := U^\top \Sigma_\beta U - (U^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, XUU^\top \hat{\mu}))XU)^{-1}$	$O(NM^2 + DM)$	$O(NM)$
$\hat{\Sigma} := \Sigma_\beta - \Sigma_\beta UWU^\top \Sigma_\beta$	0 (see footnote[1])	$O(DM)$
Compute variances and covariances of parameters		
$\mathrm{Var}_{\hat{p}}(\beta_i) = e_i^\top \hat{\Sigma} e_i$	$O(M^2)$	$O(DM)$
$\mathrm{Cov}_{\hat{p}}(\beta_i, \beta_j) = e_i^\top \hat{\Sigma} e_j$	$O(M^2)$	$O(DM)$

5. Non-conjugate GLMs with approximately low-rank data

While the conjugate linear setting facilitates intuition and theory, GLMs are a larger and more broadly useful class of models for which efficient and reliable Bayesian inference is of significant practical concern. Assuming conditional independence of the observations given the covariates and parameter, the posterior for a GLM likelihood can be written

$$\log p(\beta \mid X, Y) = \log p(\beta) + \sum_{n=1}^{N} \phi(y_n, x_n^\top \beta) + Z \quad (5)$$

for some real-valued mapping function ϕ and log normalizing constant Z. For priors and mapping functions that do not form a conjugate pair, accessing posterior functionals of interest is analytically intractable and requires posterior approximation. One possibility is to use a Monte Carlo method such as MCMC, which has theoretical guarantees asymptotic in running time but is relatively slow in practice. The usual alternative is a deterministic approximation such as VB or Laplace. These approximations are typically faster but do not become arbitrarily accurate in the limit of infinite computation. We next show how LR-GLM can be applied to facilitate faster MCMC samplers and Laplace approximations for Bayesian GLMs. We also characterize the additional error introduced to Laplace approximations by low-rank data approximations.

5.1. LR-GLM for fast Laplace approximations

The Laplace approximation refers to a Gaussian approximation obtained via a second-order Taylor approximation of the log density. In the Bayesian setting, the Laplace approximation $\bar{p}(\beta \mid X, Y)$ is typically formed at the maximum a posteriori (MAP) parameter: $\bar{p}(\beta \mid X, Y) := \mathcal{N}(\beta \mid \bar{\mu}, \bar{\Sigma})$, where $\bar{\mu} := \arg\max_\beta \log p(\beta \mid X, Y)$ and $\bar{\Sigma}^{-1} := -\nabla^2_\beta \log p(\beta \mid X, Y)|_{\beta=\bar{\mu}}$. When computing and analyzing Laplace approximations for GLMs, we will often refer to vectorized first, second, and third derivatives $\vec{\phi}', \vec{\phi}'', \vec{\phi}''' \in \mathbb{R}^N$ of the mapping function ϕ. For $Y, A \in \mathbb{R}^N$, we define $\vec{\phi}'(Y, A)_n := \frac{\partial}{\partial a}\phi(Y_n, a)|_{a=A_n}$. The higher-order derivative definitions are analogous, with the derivative order of $\frac{\partial}{\partial a}$ increased commensurately.

Laplace approximations are typically much faster than MCMC for moderate/large N and small D, but they become expensive or intractable for large D. In particular, they require inverting a $D \times D$ Hessian matrix, which is in general an $O(D^3)$ time operation, and storing the resulting covariance matrix, which requires $O(D^2)$ memory.[2]

As in the conjugate case, LR-GLM permits a faster and more memory-efficient route to inference. Here, we say that the *LR-Laplace approximation*, $\hat{p}(\beta \mid X, Y) = \mathcal{N}(\beta \mid \hat{\mu}, \hat{\Sigma})$, denotes the Laplace approximation to the LR-GLM approximate posterior. The special case of LR-Laplace with zero-mean prior is given in Algorithm 1 as it allows us to easily analyze time and memory complexity. For the more general LR-Laplace algorithm, see Appendix H.

Theorem 5.1. *In a GLM with a zero-mean, structured-Gaussian prior*[3] *and a log-concave likelihood,*[4] *the rank-*

[2] Notably, as in the conjugate setting, an alternative matrix inversion using the Woodbury identity reduces this cost when $N < D$ to $O(N^2 D)$ time and $O(ND)$ memory (Appendix C).

[3] For example (banded) diagonal or diagonal plus low-rank, such that matrix vector multiplies may be computed in $O(D)$ time.

[4] This property is standard for common GLMs such as logistic

M LR-Laplace approximation may be computed via Algorithm 1 in $O(NDM)$ time with $O(DM + NM)$ memory. Furthermore, any posterior covariance entry can be computed in $O(M^2)$ time.

Algorithm 1 consists of three phases: (1) computation of the M-truncated SVD of $X^\top$; (2) MAP optimization to find $\hat{\mu}$; and (3) estimation of $\hat{\Sigma}$. In the second phase we are able to efficiently compute $\hat{\mu}$ by first solving a lower-dimensional optimization for the quantity $\gamma_* \in \mathbb{R}^M$ (Line 8), from which $\hat{\mu}$ is available analytically. Notably, in the common case that $p(\beta)$ is isotropic Gaussian, the expression for $\hat{\mu}$ reduces to $U\gamma_*$ and the full time complexity of MAP estimation is $O(NM + DM)$. Though computing the covariance for each pair of parameters and storing $\hat{\Sigma}$ explicitly would of course require a potentially unacceptable $O(D^2)$ storage, the output of Algorithm 1 is smaller and enables arbitrary parameter variances and covariances to be computed in $O(M^2)$ time. See Appendix F.1 for additional details.

5.2. Accuracy of the LR-Laplace approximation

We now consider the quality of the LR-Laplace approximate posterior relative to the usual Laplace approximation. Our first result concerns the difference of the posterior means

Theorem 5.2 (Non-asymptotic)**.** *In a generalized linear model with an α–strongly log concave posterior, the exact and approximate MAP values, $\hat{\mu} = \arg\max_\beta \tilde{p}(\beta \mid X, Y)$ and $\bar{\mu} = \arg\max_\beta p(\beta \mid X, Y)$, satisfy*

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

for some vector $A \in \mathbb{R}^N$ such that $A_n \in [x_n^\top U U^\top \hat{\mu}, x_n^\top \hat{\mu}]$.

This bound reveals several characteristics of the regimes in which LR-Laplace performs well. As in conjugate regression, we see that the bound tightens to 0 as the rank of the approximation increases to capture all of the variance in the covariates and $\bar{\lambda}_1 \to 0$.

Remark 5.3. For many common GLMs, $\|\vec{\phi}'\|_2$, $\|\vec{\phi}''\|_\infty$, and $\|\vec{\phi}'''\|_\infty$ are well controlled; see Appendix F.4. $\|\vec{\phi}'''\|_\infty$ appears in an upcoming corollary.

Remark 5.4. The α–strong log concavity of the posterior is satisfied for any strongly log concave prior (e.g., a Gaussian, in which case we have $\alpha \geq \|\Sigma_\beta\|_2^{-1}$) and $\phi(y, \cdot)$ is concave for all y. In this common case, Theorem 5.2 provides a computable upper bound on the posterior mean error.

Remark 5.5. In contrast to the conjugate case (Corollary 4.2), general LR-GLM parameter estimates are not necessarily consistent within the span of the projection. That is, $U^\top \hat{\mu}_N$ may not converge to $U^\top \beta$ (see Appendix F.5).

and Poisson regression.

We next consider the distance between our approximation and target posterior under a Wasserstein metric (Villani, 2008). Let $\Gamma(\hat{p}, \bar{p})$ be the set of all couplings of distributions $\hat{p}$ and $\bar{p}$, i.e. joint distributions $\gamma(\cdot, \cdot)$ satisfying $\hat{p}(\beta) = \int \gamma(\beta, \beta')d\beta'$ and $\bar{p}(\beta) = \int \gamma(\beta', \beta)d\beta'$ for all β. Then the 2-Wasserstein distance between $\hat{p}$ and $\bar{p}$ is defined

$$W_2(\hat{p}, \bar{p}) = \inf_{\gamma \in \Gamma(\hat{p}, \bar{p})} \mathbb{E}_\gamma[\|\hat{\beta} - \bar{\beta}\|_2^2]^{\frac{1}{2}}. \tag{6}$$

Wasserstein bounds provide tight control of many functionals of interest, such as means, variances, and standard deviations (Huggins et al., 2018). For example, if $\xi_i \sim q_i$ for any distribution q_i $(i = 1, 2)$, then $|\mathbb{E}[\xi_1] - \mathbb{E}[\xi_2]| \leq W_2(q_1, q_2)$ and $|\mathrm{Var}[\xi_1]^{\frac{1}{2}} - \mathrm{Var}[\xi_2]^{\frac{1}{2}}| \leq 2W_2(q_1, q_2)$.

We provide a finite-sample upper bound on the 2-Wasserstein distance between the Laplace and LR-Laplace approximations. In particular, the 2-Wasserstein will decrease to 0 as the rank of the LR-Laplace approximation increases since the largest truncated singular value $\bar{\lambda}_1$ will approach zero.

Corollary 5.6. *Assume the prior $p(\beta)$ is Gaussian with covariance Σ_β and the mapping function $\phi(y, a)$ has bounded 2nd and 3rd derivatives with respect to a. Take A and α as in Theorem 5.2. Then $\bar{p}(\beta)$ and $\hat{p}(\beta)$ satisfy*

$$W_2(\hat{p}, \bar{p}) \leq \sqrt{2}\bar{\lambda}_1 \|\bar{\Sigma}\|_2 \Big\{ c\big[\|\Sigma_\beta^{-1}\|_2 + (\lambda_1 + \bar{\lambda}_1)^2 \|\vec{\phi}''\|_\infty\big] + \big[\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty\big]\sqrt{\mathrm{tr}(\hat{\Sigma})}\Big\}, \tag{7}$$

where $c := \big(\|\vec{\phi}'(Y, X\hat{\mu})\|_2 + \lambda_1 \|\bar{U}^\top \hat{\mu}\|_2 \|\vec{\phi}''(Y, A)\|_\infty\big)/\alpha$ and $r := \|U^\top \hat{\mu}\|_\infty \|\vec{\phi}'''\|_\infty + \lambda_1 c \|\vec{\phi}'''\|_\infty$.

When combined with Huggins et al. (2018, Prop. 6.1), this result guarantees closeness in 2-Wasserstein of LR-Laplace to the exact posterior.

We conclude with a result showing that the error due to the LR-GLM approximation cannot grow without bound as the sample size increases.

Theorem 5.7 (Asymptotic)**.** *Under mild regularity conditions, the error in the posterior means, $\|\hat{\mu}_n - \bar{\mu}_n\|_2$, converges as $n \to \infty$, and the limit is finite almost surely.*

For the formal statement see Theorem F.2 in Appendix F.7.

5.3. LR-MCMC for faster MCMC in GLMs

LR-Laplace is inappropriate when the posterior is poorly approximated by a Gaussian. This may be the case, for example, when the posterior is multi-modal, a common characteristic of GLMs with sparse priors. To remedy this limitation of LR-Laplace, we introduce *LR-MCMC*, a wrapper around the Metropolis–Hastings algorithm using the LR-GLM approximation. For a GLM, each full likelihood and gradient

computation takes $O(ND)$ time but only $O(NM + DM)$ time with the LR-GLM approximation, resulting in the same $\min(N, D)/M$-fold speedup obtained by LR-Laplace. See Appendix G for further details on LR-MCMC.

6. Experiments

We empirically evaluated LR-GLM on real and synthetic datasets. For synthetic data experiments, we considered logistic regression with covariates of dimension $D = 250$ and $D = 500$. In each replicate, we generated the latent parameter from an isotropic Gaussian prior, $\beta \sim \mathcal{N}(0, I_D)$, correlated covariates from a multivariate Gaussian, and responses from the logistic regression likelihood (see Appendix A.1 for details). We compared to the standard Laplace approximation, the diagonal Laplace approximation, the Laplace approximation with a low-rank data approximation obtained via random projections rather than the SVD ("Random-Laplace"), and mean-field automatic differentiation variational inference in `Stan` (ADVI-MF).[5]

Computational–statistical trade-offs. Figure 2A shows empirically the tunable computational–statistical trade-off offered by varying M in our low-rank data approximation. This plot depicts the error in posterior mean and variance estimates relative to results from the No-U-Turn Sampler (NUTS) in `Stan` (Hoffman & Gelman, 2014; Carpenter et al., 2017), which we treat as ground truth. As expected, LR-Laplace with larger M takes longer to run but yields lower errors. Random-Laplace was usually faster but provided a poor posterior approximation. Interestingly, the error of the Random-Laplace approximate posterior mean actually increased with the dimension of the projection. We conjecture this behavior may be due to Random-Laplace prioritizing covariate directions that are correlated with directions where the parameter, β, is large.

We also consider predictive performance via the classification error rate and the average negative log likelihood. In particular, we generated a *test* dataset with covariates drawn from the same distribution as the observed dataset and an *out-of-sample* dataset with covariates drawn from a different distribution (see Appendix A.1). The computation time vs. performance trade-offs, presented in Figure A.1 on the test and out-of-sample datasets, mirror the results for approximating the posterior mean and variances. In this evaluation, correctly accounting for posterior uncertainty appears less important for in-sample prediction. But in the out-of-sample case, we see a dramatic difference in negative log likelihood. Notably, ADVI-MF and diagonal Laplace exhibit much worse performance. These results support the utility of correctly estimating Bayesian uncertainty when making out-of-sample predictions.

[5] We also tested ADVI using a full rank Gaussian approximation but found it to provide near uniformly worse performance compared to ADVI-MF. So we exclude full-rank ADVI from the presented results.

Conservativeness. A benefit of LR-GLM is that the posterior approximation never underestimates the posterior uncertainty (see Corollary 4.3). Figure 2C illustrates this property for LR-Laplace applied to logistic regression. When LR-Laplace misestimates posterior variances, it always overestimates. Also, when LR-Laplace misestimates means (Figure A.2), the estimates shrink closer to the prior mean, zero in this case. These results suggest that LR-GLM interpolates between the exact posterior and the prior. Notably, this property is not true of all methods. The diagonal Laplace approximation, by contrast, dramatically underestimates posterior marginal variances (see Appendix F.8).

Reliability and calibration. Bayesian methods enjoy desirable calibration properties under correct model specification. But since LR-Laplace serves as a likelihood approximation, it does not retain this theoretical guarantee. Therefore, we assessed its calibration properties empirically by examining the credible sets of both parameters and predictions. We found that the parameter credible sets of LR-Laplace are extremely well calibrated for all values of M between 20 and 400 (Figure 2B and Figure A.4). The prediction credible sets were well calibrated for all but the smallest value of M tested ($M = 20$); in the $M = 20$ case, LR-Laplace yielded under-confident predictions (Figure A.5). The good calibration of LR-Laplace stood in sharp contrast to the diagonal Laplace approximation and ADVI-MF. Random-Laplace also provided inferior calibration (Figures A.4 and A.5).

LR-GLM with MCMC and non-Gaussian priors. In Section 5.3 we argued that LR-GLM speeds up MCMC for GLMs by decreasing the cost of likelihood and gradient evaluations in black-box MCMC routines. We first examined LR-MCMC with NUTS using `Stan` on the same synthetic datasets as we did for LR-Laplace. In Figures A.3 and A.6, we see a similar conservativeness and computational–statistical trade-off as for LR-Laplace, and superior performance relative to alternative methods.

We expect MCMC to yield high-quality posterior approximations across a wider range of models than Laplace approximations. For example, for multimodal posteriors and other posteriors that deviate substantially from Gaussianity. We next demonstrate that LR-MCMC is useful in these more general cases. In high-dimensional settings, practitioners are often interested in identifying a sparse subset of parameters that significantly influence responses. This belief may be incorporated in a Bayesian setting through a sparsity-inducing prior such as the spike and slab prior or the horseshoe (George & McCulloch, 1993; Carvalho et al., 2009). However, posteriors in these cases may be multimodal, and scalable Bayesian inference with such priors is a challenging, active area of research (Guan & Stephens,

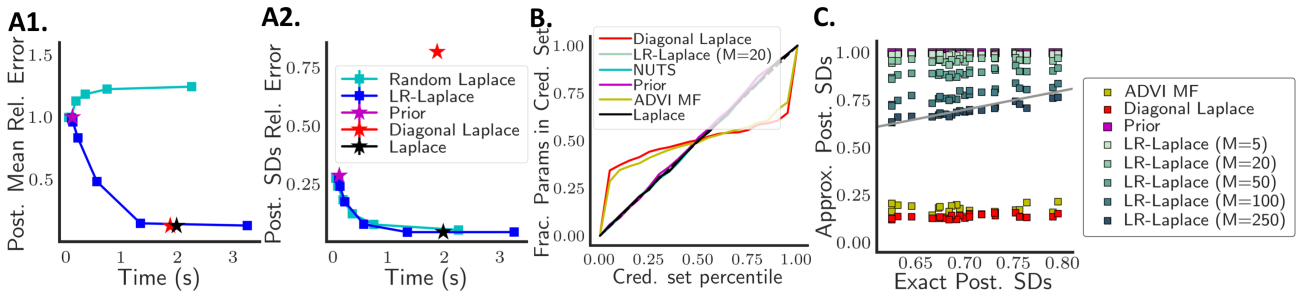

Figure 2. Left: Error of the approximate posterior (A1.) mean and (A2.) variances relative to ground truth (running NUTS with `Stan`). Lower and further left is better. *Right* (B.): Credible set calibration across all parameters and repeated experiments. (C.): Approximate posterior standard deviations for a subset of parameters. The grey line reflects zero error.

2011; Yang et al., 2016; Johndrow et al., 2017). To demonstrate the applicability of low-rank data approximations to this setting, we ran NUTS using `Stan` on a logistic regression model with a regularized horseshoe prior (Carvalho et al., 2009; Piironen & Vehtari, 2017). In Figure A.7, we see an attractive trade-off between computational investment and approximation error. For example, we obtained relative mean and standard deviation errors of only about 10^{-2} while reducing computation time by a factor of three.

We also applied LR-MCMC to linear regression with the regularized horseshoe prior on a dataset with very correlated covariates and $D = 6{,}238$. However, this sampler exhibited severe mixing problems, both with and without the approximation, as diagnosed by large $\hat{R}$ values in `pyStan`. These issues reflect the innate challenges of high-dimensional Bayesian inference with the horseshoe prior and correlated covariates.

Scalability to large-scale real datasets. Finally, we explored the applicability of LR-Laplace to two real, large-scale logistic regression tasks (Figure 3). The first is the UCI Farm-Ads dataset, which consists of $N = 4{,}143$ online advertisements for animal-related topics together with binary labels indicating whether the content provider approved of the ad; there are $D = 54{,}877$ bag-of-words features per ad (Dheeru & Karra Taniskidou, 2017). As with the synthetic datasets, we evaluated the error in the approximations of posterior means and variances. As a baseline to evaluate this error, we use the usual Laplace approximation because the computational demands of MCMC preclude the possibility of using it as a baseline.

As a second real dataset we evaluated our approach on the Reuters RCV1 text categorization test collection (Amini et al., 2009; Chang & Lin, 2011). RCV1 consists of $D = 47{,}236$ bag-of-words features for $N = 20{,}241$ English documents grouped into two different categories. We were unable to compare to the full Laplace approximation due to the high-dimensionality, so we used LR-Laplace with $M = 20{,}000$ as a baseline. For both datasets, we find that as we increase the rank of the data approximation, we incur longer running times but reduced errors in posterior means and variances. Laplace and Diagonal Laplace do not provide the same computation–accuracy trade-off.

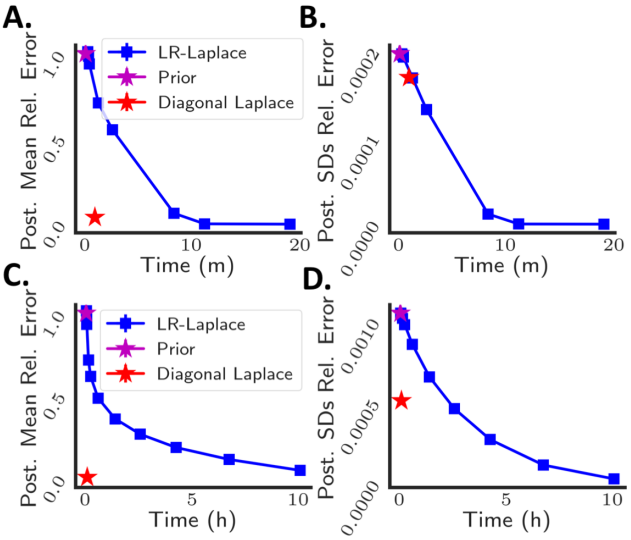

Figure 3. LR-Laplace approximation quality on Farm-Ads (top) and RCV-1 (bottom) datasets with varying M. (A.) Farm-Ads error in the posterior mean and (B.) Farm-Ads error in posterior variances (C.) RCV-1 error in posterior mean and (D.) RCV-1 error in posterior variances.

Choosing M. Applying LR-GLM requires choosing the rank M of the low rank approximation. As we have shown, this choice characterizes a computational–statistical trade-off whereby larger M leads to linearly larger computational demands, but increases the precision of the approximation. As a practical rule of thumb, we recommend setting M to be as large as is allowable for the given application without the resulting inference becoming too slow. For our experiments with LR-Laplace, this limit was $M \approx 20{,}000$. For LR-MCMC, the largest manageable choice of M will be problem dependent but will typically be much smaller than 20,000.

Acknowledgments

This research is supported in part by an NSF CAREER Award, an ARO YIP Award, a Google Faculty Research Award, a Sloan Research Fellowship, and ONR. BLT is supported by NSF GRFP.

References

Amini, M., Usunier, N., and Goutte, C. Learning from Multiple Partially Observed Views – an Application to Multilingual Text Categorization. In *Advances in Neural Information Processing Systems*, pp. 28–36, 2009.

Bardenet, R., Doucet, A., and Holmes, C. On Markov Chain Monte Carlo Methods for Tall Data. *The Journal of Machine Learning Research*, 18(1):1515–1557, 2017.

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

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

Bolley, F., Gentil, I., and Guillin, A. Convergence to Equilibrium in Wasserstein Distance for Fokker–Planck Equations. *Journal of Functional Analysis*, 263(8):2430–2457, 2012.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. Optimal Rates of Convergence for Covariance Matrix Estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

Campbell, T. and Broderick, T. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 698–706, 2018.

Campbell, T. and Broderick, T. Automated Scalable Bayesian Inference via Hilbert Coresets. *Journal of Machine Learning Research*, 20(1):551–588, 2019.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017.

Carvalho, C. M., Polson, N. G., and Scott, J. G. Handling Sparsity via the Horseshoe. In *Artificial Intelligence and Statistics*, pp. 73–80, 2009.

Chang, C.-C. and Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

Davis, C. and Kahan, W. M. The Rotation of Eigenvectors by a Perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Dawid, A. P. The Well-Calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Dheeru, D. and Karra Taniskidou, E. UCI Machine Learning Repository, 2017.

George, E. I. and McCulloch, R. E. Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Geppert, L. N., Ickstadt, K., Munteanu, A., Quedenfeld, J., and Sohler, C. Random Projections for Bayesian Regression. *Statistics and Computing*, 27(1):79–101, 2017.

Guan, Y. and Stephens, M. Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems. *The Annals of Applied Statistics*, pp. 1780–1815, 2011.

Guhaniyogi, R. and Dunson, D. B. Bayesian Compressed Regression. *Journal of the American Statistical Association*, 110(512):1500–1514, 2015.

Halko, N., Martinsson, P.-G., and Tropp, J. A. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2):217–288, 2011.

Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika*, 57 (1), 1970.

Hoffman, M. D. and Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Huggins, J., Campbell, T., and Broderick, T. Coresets for Scalable Bayesian Logistic Regression. In *Advances in Neural Information Processing Systems*, pp. 4080–4088, 2016.

Huggins, J., Adams, R. P., and Broderick, T. PASS-GLM: Polynomial Approximate Sufficient Statistics for Scalable Bayesian GLM Inference. In *Advances in Neural Information Processing Systems*, pp. 3611–3621, 2017.

Huggins, J. H., Campbell, T., Kasprzak, M., and Broderick, T. Practical Bounds on the Error of Bayesian Posterior Approximations: A Nonasymptotic Approach. *arXiv preprint arXiv:1809.09505*, 2018.

Johndrow, J. E., Orenstein, P., and Bhattacharya, A. Scalable MCMC for Bayes Shrinkage Priors. *arXiv preprint arXiv:1705.00841*, 2017.

Lee, J. and Oh, H.-S. Bayesian Regression Based on Principal Components for High-Dimensional Data. *Journal of Multivariate Analysis*, 117:175–192, 2013.

Luenberger, D. G. *Optimization by Vector Space Methods*. John Wiley & Sons, 1969.

MacKay, D. J. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning*, 12:2825–2830, 2011.

Petersen, K. B. and Pedersen, M. S. The Matrix Cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Piironen, J. and Vehtari, A. Sparsity Information and Regularization in the Horseshoe and Other Shrinkage Priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 2007.

Rue, H., Martino, S., and Chopin, N. Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392, 2009.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing Finite Sums with the Stochastic Average Gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Spantini, A., Solonen, A., Cui, T., Martin, J., Tenorio, L., and Marzouk, Y. Optimal Low-Rank Approximations of Bayesian Linear Inverse Problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487, 2015.

Turner, R. E. and Sahani, M. Two Problems with Variational Expectation Maximisation for Time-Series Models. *Bayesian Time Series Models*, 1(3.1):3–1, 2011.

Udell, M. and Townsend, A. Why Are Big Data Matrices Approximately Low Rank? *SIAM Journal of Mathematical Data Science*, 2019.

Van der Vaart, A. W. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Vershynin, R. How Close is the Sample Covariance Matrix to the Actual Covariance Matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

Wang, J., Lee, J. D., Mahdavi, M., Kolar, M., and Srebro, N. 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.

Yang, Y., Wainwright, M. J., and Jordan, M. I. On the Computational Complexity of High-Dimensional Bayesian Variable Selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

Zhang, L., Mahdavi, M., Jin, R., Yang, T., and Zhu, S. Random Projections for Classification: A Recovery Approach. *IEEE Transactions on Information Theory*, 60 (11):7300–7316, 2014.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.