SIAM J. IMAGING SCIENCES
Vol. 4, No. 1, pp. 166–199

Large Scale Bayesian Inference and Experimental Design for Sparse Linear Models*

Matthias W. Seeger† and Hannes Nickisch‡

Abstract. Many problems of low-level computer vision and image processing, such as denoising, deconvolution, tomographic reconstruction or superresolution, can be addressed by maximizing the posterior distribution of a sparse linear model (SLM). We show how higher-order Bayesian decision-making problems, such as optimizing image acquisition in magnetic resonance scanners, can be addressed by querying the SLM posterior covariance, unrelated to the density's mode. We propose a scalable algorithmic framework, with which SLM posteriors over full, high-resolution images can be approximated for the first time, solving a variational optimization problem which is convex if and only if posterior mode finding is convex. These methods successfully drive the optimization of sampling trajectories for real-world magnetic resonance imaging through Bayesian experimental design, which has not been attempted before. Our methodology provides new insight into similarities and differences between sparse reconstruction and approximate Bayesian inference, and has important implications for compressive sensing of real-world images. Parts of this work have been presented at conferences [M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf, in *Advances in Neural Information Processing Systems* 21, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds., Curran Associates, Red Hook, NY, 2009, pp. 1441–1448; H. Nickisch and M. Seeger, in *Proceedings of the* 26*th International Conference on Machine Learning*, L. Bottou and M. Littman, eds., Omni Press, Madison, WI, 2009, pp. 761–768].

Key words. sparse linear model, sparsity prior, experimental design, sampling optimization, image acquisition, variational approximate inference, Bayesian statistics, compressive sensing, sparse reconstruction, magnetic resonance imaging

AMS subject classifications. Primary, 68T37; Secondary, 62F15, 49N45, 62C12, 65K10

DOI. 10.1137/090758775

1. Introduction.

Natural images have a sparse low-level statistical signature, represented in the prior distribution of a sparse linear model (SLM). Imaging problems such as reconstruction, denoising, or deconvolution can successfully be solved by maximizing their posterior density (maximum a posteriori (MAP) estimation), a convex program for certain SLMs, for which efficient solvers are available. The success of these techniques in modern imaging practice has somewhat shrouded their limited scope as Bayesian techniques: of all the information in the posterior distribution, they make use of the density's mode only.

Consider the problem of *optimizing image acquisition*, our major motivation in this work.

*Received by the editors May 11, 2009; accepted for publication (in revised form) November 19, 2010; published electronically March 3, 2011. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, and the Excellence Initiative of the German Research Foundation (DFG).

http://www.siam.org/journals/siims/4-1/75877.html

†School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland (matthias.seeger@epfl.ch).

‡Max Planck Institute for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany (hn@tuebingen.mpg.de).

Magnetic resonance images are reconstructed from Fourier samples. With scan time proportional to the number of samples, a central question to ask is which sampling designs of minimum size still lead to MAP reconstructions of diagnostically useful image quality? This is not a direct reconstruction problem; the focus is on improving measurement designs to better support subsequent reconstruction. Goal-directed acquisition optimization cannot sensibly be addressed by MAP estimation, yet we show how to successfully drive it by Bayesian posterior information beyond the mode. Advanced decision-making of this kind needs uncertainty quantification (posterior covariance) rather than point estimation, requiring us to step out of the *sparse reconstruction* scenario and approximate *sparse Bayesian inference* instead.

The Bayesian inference relaxation we focus on is not new [11, 23, 17], yet when it comes to problem characterization or efficient algorithms, previous inference work lags far behind standards established for MAP reconstruction. Our contributions range from theoretical characterizations over novel scalable solvers to applications not previously attempted. The inference relaxation is shown to be a convex optimization problem if and only if this holds for MAP estimation (section 3), a property not previously established for this or any other SLM inference approximation. Moreover, we develop novel *scalable* double loop algorithms to solve the variational problem orders of magnitude faster than previous methods we are aware of (section 4). These algorithms expose an important link between variational Bayesian inference and sparse MAP reconstruction, reducing the former to calling variants of the latter a few times, interleaved by approximations of Gaussian covariance or principal components analysis (PCA). By way of this reduction, the massive recent interest in MAP estimation can play a role for variational Bayesian inference as well. To complement these similarities and clarify confusion in the literature, we discuss computational and statistical differences of sparse estimation and Bayesian inference in detail (section 5).

The ultimate motivation for novel developments presented here is sequential Bayesian experimental design (section 6), applied to acquisition optimization for medical imaging. We present a powerful variant of adaptive compressive sensing, which succeeds on real-world image data where theoretical proposals for nonadaptive compressive sensing [9, 6, 10] fail (section 6.1). Among our experimental results is part of the first successful study for Bayesian sampling optimization of magnetic resonance imaging, learned and evaluated on real-world image data (section 7.4).

An implementation of the algorithms presented here is publicly available, as part of the `glm-ie` toolbox (section 4.6) available at http://mloss.org/software/view/269/.

2. Sparse Bayesian inference. Variational approximations. In a sparse linear model (SLM), the image $\boldsymbol{u} \in \mathbb{R}^n$ of n pixels is unknown, and m linear measurements $\boldsymbol{y} \in \mathbb{R}^m$ are given, where $m \ll n$ in many situations of practical interest. The measurement model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \tag{2.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is the *design matrix* and $\boldsymbol{\varepsilon}$ is Gaussian noise of variance σ^2, implying the Gaussian likelihood $P(\boldsymbol{y}|\boldsymbol{u}) = N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I})$. Natural images are characterized by histograms of simple filter responses (derivatives, wavelet coefficients) exhibiting *super-Gaussian* (or *sparse*) form: most coefficients are close to zero, while a small fraction have significant sizes [35, 28] (a precise definition of super-Gaussianity is given in section 2.1). Accordingly, SLMs have

super-Gaussian prior distributions $P(\boldsymbol{u})$. The MAP estimator $\hat{\boldsymbol{u}}_{\mathrm{MAP}} := \mathrm{argmax}_{\boldsymbol{u}}[\log P(\boldsymbol{y}|\boldsymbol{u}) + \log P(\boldsymbol{u})]$ can outperform maximum likelihood $\hat{\boldsymbol{u}}_{\mathrm{ML}} := \mathrm{argmax}_{\boldsymbol{u}} \log P(\boldsymbol{y}|\boldsymbol{u})$ when $\boldsymbol{u}$ represents an image.

In this paper, we focus on priors of the form $P(\boldsymbol{u}) \propto \prod_{i=1}^{q} t_i(s_i)$, where $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$. The potential functions $t_i(\cdot)$ are positive and bounded. The operator $\boldsymbol{B} \in \mathbb{R}^{q\times n}$ may contain derivative filters or a wavelet transform. Both $\boldsymbol{X}$ and $\boldsymbol{B}$ have to be structured or sparse in order for any SLM algorithm to be scalable. *Laplace* (or double exponential) potentials are sparsity-enforcing:

$$t_i(s_i) = e^{-\tau_i|s_i|}, \quad \tau_i > 0. \tag{2.2}$$

For this particular prior and the Gaussian likelihood (2.1), MAP estimation corresponds to a quadratic program, known as *LASSO* [37], for $\boldsymbol{B} = \boldsymbol{I}$. Note that $\log t_i(s_i)$ is concave. In general, if *log-concavity* holds for all model potentials, MAP estimation is a convex problem. Another example of sparsity potentials are *Student's t:*

$$t_i(s_i) = (1 + (\tau_i/\nu)s_i^2)^{-(\nu+1)/2}, \quad \tau_i, \nu > 0. \tag{2.3}$$

For these, $\log t_i(s_i)$ is not concave, and MAP estimation is not (in general) a convex program. Note that $-\log t_i(s_i)$ is also known as a Lorentzian penalty function.

2.1. Variational lower bounds. Bayesian inference amounts to computing moments of the posterior distribution

$$\begin{aligned} P(\boldsymbol{u}|\boldsymbol{y}) &= Z^{-1} N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) \prod_{i=1}^{q} t_i(s_i), \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}, \\ Z &= \int N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) \prod_{i=1}^{q} t_i(s_i)\, d\boldsymbol{u}. \end{aligned} \tag{2.4}$$

This is not analytically tractable in general for SLMs, due to two reasons coming together: $P(\boldsymbol{u}|\boldsymbol{y})$ is highly coupled ($\boldsymbol{X}$ is not block-diagonal) and non-Gaussian. We focus on *variational* approximations here, rooted in statistical physics. The *log partition function* $\log Z$ (also known as log evidence or log marginal likelihood) is the prime target for variational methods [41]. Formally, the potentials $t_i(s_i)$ are replaced by Gaussian terms of parameterized width, the posterior $P(\boldsymbol{u}|\boldsymbol{y})$ by a Gaussian approximation $Q(\boldsymbol{u}|\boldsymbol{y})$. The width parameters are adjusted by fitting $Q(\boldsymbol{u}|\boldsymbol{y})$ to $P(\boldsymbol{u}|\boldsymbol{y})$, in what amounts to the variational optimization problem.

Our variational relaxation exploits the fact that all potentials $t_i(s_i)$ are *(strongly) super-Gaussian*: there exists a $b_i \in \mathbb{R}$ such that $\tilde{g}_i(s_i) := \log t_i(s_i) - b_i s_i$ is even, and convex and decreasing as a function of $x_i := s_i^2$ [23]. We write $g_i(x_i) := \tilde{g}_i(x_i^{1/2})$, $x_i \geq 0$, in what follows. This implies that

$$t_i(s_i) = \max_{\gamma_i \geq 0} e^{b_i s_i - s_i^2/(2\gamma_i) - h_i(\gamma_i)/2}, \quad h_i(\gamma_i) := \max_{x_i \geq 0}(-x_i/\gamma_i - 2g_i(x_i)). \tag{2.5}$$

A super-Gaussian $t_i(s_i)$ has tight Gaussian-form lower bounds of all widths (see Figure 1). We replace $t_i(s_i)$ by these lower bounds in order to step from $P(\boldsymbol{u}|\boldsymbol{y})$ to the family of approximations $Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) e^{\boldsymbol{b}^T\boldsymbol{s} - \frac{1}{2}\boldsymbol{s}^T(\mathrm{diag}\,\boldsymbol{\gamma})^{-1}\boldsymbol{s}}$, where $\boldsymbol{\gamma} = (\gamma_i)$.

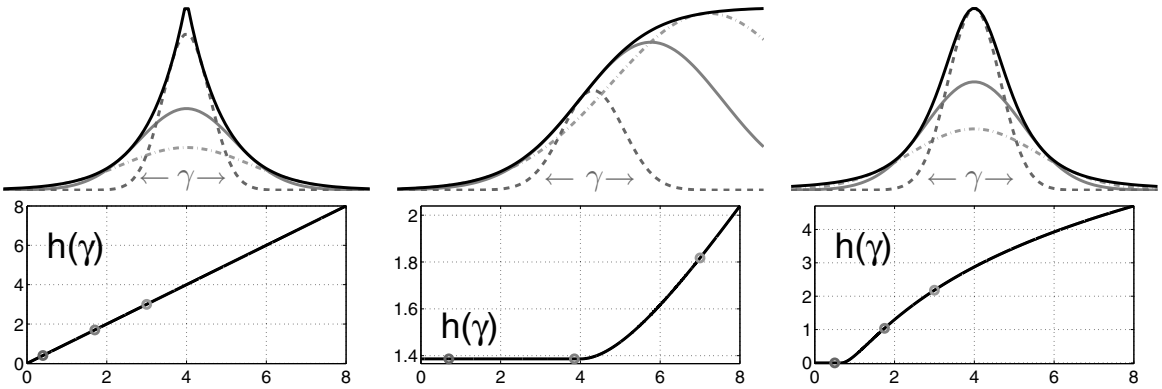

Figure 1. *Super-Gaussian distributions admit tight Gaussian-form lower bounds of any width γ. Left: Laplace (2.2); middle: Bernoulli (2.6); right: Student's t (2.3).*

To establish (2.5), note that the extended-value function $g_i(x_i)$ (assigning $g_i(x_i) = \infty$ for $x_i < 0$) is a closed proper convex function and thus can be represented by Fenchel duality [26, sect. 12]: $g_i(x_i) = \sup_{\lambda_i} x_i\lambda_i - g_i^*(\lambda_i)$, where the conjugate function $g_i^*(\lambda_i) = \sup_{x_i} x_i\lambda_i - g_i(x_i)$ is closed convex as well. For $x_i \geq 0$ and $\lambda_i > 0$, we have that $x_i\lambda_i - g_i(x_i) \geq x_i\lambda_i - g_i(0) \to \infty$ with $x_i \to \infty$, which implies that $g_i^*(\lambda_i) = \infty$ for $\lambda_i > 0$. Also, $g_i^*(0) = -\lim_{x_i\to\infty} g_i(x_i)$, so that $-g_i^*(0) < g_i(x_i)$ for any $x_i < \infty$. Therefore, for any $x_i \in [0, \infty)$, $g_i(x_i) = \sup_{\lambda_i<0} x_i\lambda_i - g_i^*(\lambda_i)$. Reparameterizing $\gamma_i := -1/(2\lambda_i)$, we have that $g_i(x_i) = \max_{\gamma_i\geq 0} -x_i/(2\gamma_i) - g_i^*(-1/(2\gamma_i))$ (note that in order to accommodate $g_i(0)$ in general, we have to allow for $\gamma_i = 0$). Finally, $h_i(\gamma_i) := 2g_i^*(-1/(2\gamma_i))$.

The class of super-Gaussian potentials is large. All scale mixtures (mixtures of zero-mean Gaussians $t_i(s_i) = \mathrm{E}_{\gamma_i}[N(s_i|0, \gamma_i)]$) are super-Gaussian, and $h_i(\gamma_i)$ can be written in terms of the density for γ_i [23]. Both Laplace (2.2) and Student's t potentials (2.3) are super-Gaussian, with $h_i(\gamma_i)$ given in Appendix A.6. Bernoulli potentials, used as binary classification likelihoods,

$$t_i(s_i) = \left(1 + e^{-y_i\tau_i s_i}\right)^{-1}, \quad y_i \in \{\pm 1\},\ \tau_i > 0, \tag{2.6}$$

are super-Gaussian, with $b_i = y_i\tau_i/2$ [17, sect. 3.B]. While the corresponding $h_i(\gamma_i)$ cannot be determined analytically, this is not required in our algorithms, which can be implemented based on $g_i(x_i)$ and its derivatives only. Finally, if the $t_i^{(l)}(s_i)$ are super-Gaussian, so is the positive mixture $\sum_{l=1}^{L} \alpha_l t_i^{(l)}(s_i)$, $\alpha_l > 0$, because the logsumexp function $\boldsymbol{x} \mapsto \log \mathbf{1}^T \exp(\boldsymbol{x})$ [4, sect. 3.1.5] is strictly convex on $\mathbb{R}^L$ and increasing in each argument, and the log-mixture is the concatenation of logsumexp with $(\log t_i^{(l)}(s_i) + \log \alpha_l)_l$, the latter convex and decreasing for $x_i = s_i^2 > 0$ in each component [4, sect. 3.2.4].

A natural criterion for fitting $Q(\boldsymbol{u}|\boldsymbol{y})$ to $P(\boldsymbol{u}|\boldsymbol{y})$ is obtained by plugging these bounds into the partition function Z of (2.4):

$$Z \geq e^{-h(\boldsymbol{\gamma})/2} \int N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) e^{\boldsymbol{b}^T\boldsymbol{s} - \frac{1}{2}\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}}\, d\boldsymbol{u},\ h(\boldsymbol{\gamma}) := \sum_{i=1}^{q} h_i(\gamma_i), \tag{2.7}$$

where $\boldsymbol{\Gamma} := \operatorname{diag}\boldsymbol{\gamma}$ and $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$. The right-hand side is a Gaussian integral and can be evaluated easily. The variational problem is to maximize this lower bound w.r.t. the variational parameters $\boldsymbol{\gamma} \succeq \boldsymbol{0}$ ($\gamma_i \geq 0$ for all $i = 1, \ldots, q$), with the aim of tightening the approximation to $\log Z$. This criterion can be interpreted as a divergence function: if the family of all $Q(\boldsymbol{u}|\boldsymbol{y})$ contained the true posterior $P(\boldsymbol{u}|\boldsymbol{y})$, the latter would maximize the bound.

This relaxation has frequently been used before [11, 23, 17] on inference problems of moderate size. In the following, we provide results that extend its scope to large scale imaging problems of interest here. In the next section, we characterize the convexity of the underlying optimization problem precisely. In section 4, we provide a new class of algorithms for solving this problem orders of magnitude faster than previously used techniques.

3. Convexity properties of variational inference. In this section, we characterize the variational optimization problem of maximizing the right-hand side of (2.7). We show that it is a convex problem if and only if all potentials $t_i(s_i)$ are log-concave, which is equivalent to MAP estimation being convex for the same model.

We start by converting the lower bound (2.7) into a more amenable form. The Gaussian posterior $Q(\boldsymbol{u}|\boldsymbol{y})$ has the covariance matrix

$$\operatorname{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}] = \boldsymbol{A}^{-1}, \quad \boldsymbol{A} := \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{B}, \quad \boldsymbol{\Gamma} = \operatorname{diag}\boldsymbol{\gamma}. \tag{3.1}$$

We have that

$$\int P(\boldsymbol{y}|\boldsymbol{u}) e^{\boldsymbol{b}^T\boldsymbol{s} - \frac{1}{2}\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}}\, d\boldsymbol{u} = |2\pi\boldsymbol{A}^{-1}|^{1/2} \max_{\boldsymbol{u}} P(\boldsymbol{y}|\boldsymbol{u}) e^{\boldsymbol{b}^T\boldsymbol{s} - \frac{1}{2}\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}}, \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u},$$

since $\max_{\boldsymbol{u}} Q(\boldsymbol{u}|\boldsymbol{y}) = |2\pi\boldsymbol{A}^{-1}|^{-1/2} \int Q(\boldsymbol{u}|\boldsymbol{y})\, d\boldsymbol{u}$. We find that $Z \geq C_1 e^{-\phi(\boldsymbol{\gamma})/2}$, where

$$\phi(\boldsymbol{\gamma}) := \log|\boldsymbol{A}| + h(\boldsymbol{\gamma}) + \min_{\boldsymbol{u}} R(\boldsymbol{u}, \boldsymbol{\gamma}), \quad R := \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s} - 2\boldsymbol{b}^T\boldsymbol{s}, \tag{3.2}$$

and $C_1 = (2\pi)^{(n-m)/2}\sigma^{-m}$. The variational problem is $\min_{\boldsymbol{\gamma}\succeq\boldsymbol{0}} \phi(\boldsymbol{\gamma})$, and the Gaussian posterior approximation is $Q(\boldsymbol{u}|\boldsymbol{y})$ with the final parameters $\boldsymbol{\gamma}$ plugged in. We will also write $\phi(\boldsymbol{u}, \boldsymbol{\gamma}) := \log|\boldsymbol{A}| + h(\boldsymbol{\gamma}) + R(\boldsymbol{u}, \boldsymbol{\gamma})$, so that $\phi(\boldsymbol{\gamma}) = \min_{\boldsymbol{u}} \phi(\boldsymbol{u}, \boldsymbol{\gamma})$.

It is instructive to compare this variational inference problem with MAP estimation:

$$\begin{aligned} \min_{\boldsymbol{u}} -2\log P(\boldsymbol{u}|\boldsymbol{y}) &= \min_{\boldsymbol{u}} \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 - 2\sum_i \log t_i(s_i) + C_2 \\ &= \min_{\boldsymbol{u},\boldsymbol{\gamma}\succeq\boldsymbol{0}} h(\boldsymbol{\gamma}) + R(\boldsymbol{u}, \boldsymbol{\gamma}) + C_2, \end{aligned} \tag{3.3}$$

where C_2 is a constant. The difference between these problems rests on the $\log|\boldsymbol{A}|$ term, present in $\phi(\boldsymbol{\gamma})$ yet absent in MAP. Ultimately, this observation is the key to the characterization provided in this section and to the scalable solvers developed in the subsequent section. Its full relevance will be clarified in section 5.

3.1. Convexity results. In this section, we prove that $\phi(\boldsymbol{\gamma})$ is convex if all potentials $t_i(s_i)$ are log-concave, with this condition being necessary in general. We address each term in (3.2) separately.

Theorem 3.1. *Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, $\boldsymbol{B} \in \mathbb{R}^{q\times n}$ be arbitrary matrices, and let*

$$\tilde{\boldsymbol{A}}(\boldsymbol{d}) := \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T(\operatorname{diag}\boldsymbol{d})\boldsymbol{B}, \quad \boldsymbol{d} \succ \boldsymbol{0},$$

so that $\tilde{\boldsymbol{A}}(\boldsymbol{d})$ is positive definite for all $\boldsymbol{d} \succ \boldsymbol{0}$.

(1) *Let $f_i(\gamma_i)$ be twice continuously differentiable functions into $\mathbb{R}_+$ so that $\log f_i(\gamma_i)$ are convex for all i and γ_i. Then, $\boldsymbol{\gamma} \mapsto \log|\tilde{\boldsymbol{A}}(f(\boldsymbol{\gamma}))|$ is convex. Especially, $\boldsymbol{\gamma} \mapsto \log|\boldsymbol{A}|$ is convex.*

(2) *Let $f_i(\pi_i) \geq 0$ be concave functions. Then, $\boldsymbol{\pi} \mapsto \log|\tilde{\boldsymbol{A}}(f(\boldsymbol{\pi}))|$ is concave. Especially, $\boldsymbol{\gamma}^{-1} \mapsto \log|\boldsymbol{A}|$ is concave.*

(3) *Let $f_i(\gamma_i) \geq 0$ be concave functions. Then, $\boldsymbol{\gamma} \mapsto \boldsymbol{1}^T(\log f(\boldsymbol{\gamma})) + \log|\tilde{\boldsymbol{A}}(f(\boldsymbol{\gamma})^{-1})|$ is concave. Especially, $\boldsymbol{\gamma} \mapsto \boldsymbol{1}^T(\log\boldsymbol{\gamma}) + \log|\boldsymbol{A}|$ is concave.*

(4) *Let $Q(\boldsymbol{u}|\boldsymbol{y})$ be the approximate posterior with covariance matrix given by (3.1). Then, for all i,*

$$\operatorname{Var}_Q[s_i|\boldsymbol{y}] = \boldsymbol{\delta}_i^T\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T\boldsymbol{\delta}_i \leq \gamma_i.$$

A proof is provided in Appendix A.1. Instantiating part (1) with $f_i(\gamma_i) = \gamma_i^{-1}$, we see that $\boldsymbol{\gamma} \mapsto \log|\boldsymbol{A}|$ is convex. Other valid examples are $f_i(\gamma_i) = \gamma_i^{-\beta_i}$, $\beta_i > 0$. For $f_i(\gamma_i) = e^{\gamma_i}$, we obtain the convexity of $\boldsymbol{\gamma} \mapsto \log|\tilde{\boldsymbol{A}}(\exp(\boldsymbol{\gamma}))|$, generalizing the logsumexp function to matrix values. Parts (2) and (3) will be required in section 4.2. Finally, part (4) gives a precise characterization of γ_i as sparsity parameter, regulating the variance of s_i.

Theorem 3.2. *The function*

$$\boldsymbol{\gamma} \mapsto \phi(\boldsymbol{\gamma}) - h(\boldsymbol{\gamma}) = \log|\boldsymbol{A}| + \min_{\boldsymbol{u}}\left(\sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s} - 2\boldsymbol{b}^T\boldsymbol{s}\right)$$

is convex for $\boldsymbol{\gamma} \succ \boldsymbol{0}$, where $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$.

Proof. The convexity of $\log|\boldsymbol{A}|$ has been shown in Theorem 3.1(1). $\sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 - 2\boldsymbol{b}^T\boldsymbol{s}$ is convex in $\boldsymbol{u}$, and $(\boldsymbol{u}, \boldsymbol{\gamma}) \mapsto \boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}$ is *jointly* convex, since the quadratic-over-linear function $(s_i, \gamma_i) \mapsto s_i^2/\gamma_i$ is jointly convex for $\gamma_i > 0$ [4, sect. 3.1.5]. Therefore, $\min_{\boldsymbol{u}} R(\boldsymbol{u}, \boldsymbol{\gamma})$ is convex for $\boldsymbol{\gamma} \succ \boldsymbol{0}$ [4, sect. 3.2.5]. ∎

To put this result into context, note that

$$\phi(\boldsymbol{\gamma}) - h(\boldsymbol{\gamma}) = -2\log\int P(\boldsymbol{y}|\boldsymbol{u})e^{\boldsymbol{b}^T\boldsymbol{s} - \frac{1}{2}\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}}\,d\boldsymbol{u}, \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u},$$

is the *negative* log partition function of a Gaussian with natural parameters $\boldsymbol{\gamma}^{-1}$: it is well known that $\boldsymbol{\gamma}^{-1} \mapsto \phi(\boldsymbol{\gamma}) - h(\boldsymbol{\gamma})$ is a *concave* function [41]. However, $\boldsymbol{\gamma}^{-1} \mapsto h(\boldsymbol{\gamma})$ is *convex* for a model with super-Gaussian potentials (recall that $h_i(\gamma_i) = 2g_i^*(-1/(2\gamma_i))$, where $g_i^*(\cdot)$ is convex as a dual function of $g_i(\cdot)$), which means that in general $\boldsymbol{\gamma}^{-1} \mapsto \phi(\gamma)$ need not be convex or concave. The *convexity* of this negative log partition function w.r.t. $\boldsymbol{\gamma}$ seems specific to the Gaussian case.

Given Theorem 3.2, if all $h_i(\gamma_i)$ are convex, the whole variational problem $\min_{\boldsymbol{\gamma}\succeq\boldsymbol{0}}\phi$ is convex. With the following theorem, we characterize this case precisely.

Theorem 3.3. *Consider a model with Gaussian likelihood (2.1) and a prior $P(\boldsymbol{u}) \propto \prod_{i=1}^q t_i(s_i)$, $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$, so that all $t_i(s_i)$ are strongly super-Gaussian, meaning that $\tilde{g}_i(s_i) = \log t_i(s_i) - b_i s_i$ is even, and $g_i(x_i) = \tilde{g}_i(x_i^{1/2})$ is strictly convex and decreasing for $x_i > 0$.*

(1) *If* $\tilde{g}_i(s_i)$ *is concave and twice continuously differentiable for* $s_i > 0$*, then* $h_i(\gamma_i)$ *is convex. On the other hand, if* $\tilde{g}_i''(s_i) > 0$ *for some* $s_i > 0$*, then* $h_i(\gamma_i)$ *is not convex at some* $\gamma_i > 0$*.*

(2) *If all* $\tilde{g}_i(s_i)$ *are concave and twice continuously differentiable for* $s_i > 0$*, then the variational problem* $\min_{\boldsymbol{\gamma} \succeq \mathbf{0}} \phi$ *is a convex optimization problem. On the other hand, if* $\tilde{g}_i''(s_i) > 0$ *for some* i *and* $s_i > 0$*, then* $h(\boldsymbol{\gamma})$ *is not convex, and there exist some* $\boldsymbol{X}$*,* $\boldsymbol{B}$*, and* $\boldsymbol{y}$ *such that* $\phi(\boldsymbol{\gamma})$ *is not a convex function.*

The proof is given in Appendix A.2. Our theorem provides a complete characterization of convexity for the variational inference relaxation of section 2, which is *the same* as for MAP estimation. Log-concavity of all potentials is sufficient, and necessary in general, for the convexity of either. We are not aware of a comparable equivalence having been established for any other nontrivial approximate inference method for continuous variable models.

We close this section with some examples. For Laplacians (2.2), $h_i(\gamma_i) = \tau_i^2 \gamma_i$ (see Appendix A.6). For SLMs with these potentials, MAP estimation is a convex quadratic program. Our result implies that variational inference is a convex problem as well, albeit with a differentiable criterion. Bernoulli potentials (2.6) are log-concave. MAP estimation for generalized linear models with these potentials is known as penalized logistic regression, a convex problem typically solved by the iteratively reweighted least squares (IRLS) algorithm. Variational inference for this model is also a convex problem, and our algorithms introduced in section 4 make use of IRLS as well. Finally, Student's t potentials (2.3) are not log-concave, and $h_i(\gamma_i)$ is neither convex nor concave (see Appendix A.6). Neither MAP estimation nor variational inference is convex in this case.

Convexity of an algorithm is desirable for many reasons. No restarting is needed to avoid local minima. Typically, the result is robust to small perturbations of the data. These stability properties become all the more important in the context of sequential experimental design (see section 6), or when Bayesian model selection[1] is used. However, the convexity of $\phi(\boldsymbol{\gamma})$ does not necessarily imply that the minimum point can be found efficiently. In the next section, we propose a class of algorithms that solve the variational problem for very large instances, by decoupling the criterion (3.2) in a novel way.

4. Scalable inference algorithms. In this section, we propose novel algorithms for solving the variational inference problem $\min_{\boldsymbol{\gamma}} \phi$ in a scalable way. Our algorithms can be used whether $\phi(\boldsymbol{\gamma})$ is convex or not; they are guaranteed to converge to a stationary point. All efforts are reduced to well-known, scalable algorithms of signal reconstruction and numerical mathematics, with little extra technology required, and no additional heuristics or step size parameters to be tuned.

We begin with the special case of log-concave potentials $t_i(s_i)$, such as Laplace (2.2) or Bernoulli (2.6), extending our framework to full generality in section 4.1. The variational inference problem is convex in this case (Theorem 3.3). Previous algorithms for solving $\min_{\boldsymbol{\gamma}} \phi(\boldsymbol{\gamma})$ [11, 23] are of the *coordinate descent* type, minimizing ϕ w.r.t. one γ_i at a time. Unfortunately, such algorithms cannot be scaled up to imaging problems of interest here. An update of γ_i depends on the marginal posterior $Q(s_i|\boldsymbol{y})$, whose computation requires the solution of a

[1]Model selection (or hyperparameter learning) is not discussed in this paper. It can be implemented easily by maximizing the lower bound $-\phi(\boldsymbol{\gamma})/2 + \log C_1 \leq \log Z$ w.r.t. hyperparameters.

linear system with matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. At the projected scale, neither $\boldsymbol{A}$ nor a decomposition thereof can be maintained; systems have to be solved iteratively. Now, each of the q potentials has to be visited at least once, typically several times. With q, n, and m in the hundred thousands, it is certainly infeasible to solve $O(q)$ linear systems. In contrast, the algorithms we develop here often converge after fewer than a hundred systems have been solved. We could also feed $\phi(\boldsymbol{\gamma})$ and its gradient $\nabla_{\boldsymbol{\gamma}}\phi$ into an off-the-shelf gradient-based optimizer. However, as already noted in section 3, $\phi(\boldsymbol{\gamma})$ is the sum of a standard penalized least squares (MAP) part and a highly coupled, computationally difficult term. The algorithms we propose take account of this decomposition, decoupling the troublesome term in inner loop standard form problems which can be solved by any of a large number of specialized algorithms not applicable to $\min_{\boldsymbol{\gamma}} \phi(\boldsymbol{\gamma})$. The expensive part of $\nabla_{\boldsymbol{\gamma}}\phi$ has to be computed only a few times for our algorithms to converge.

We make use of a powerful idea known as *double loop* or *concave-convex* algorithms. Special cases of such algorithms are frequently used in machine learning, computer vision, and statistics: the expectation-maximization (EM) algorithm [8], variational mean field Bayesian inference [1], or concave-convex procedures for discrete approximate inference [48], among many others. The idea is to tangentially upper bound ϕ by decoupled functions $\phi_{\boldsymbol{z}}$ which are much simpler to minimize than ϕ itself: algorithms iterate between refitting $\phi_{\boldsymbol{z}}$ to ϕ and minimizing $\phi_{\boldsymbol{z}}$. For example, in the EM algorithm for maximizing a log marginal likelihood, these stages correspond to an "E step" and an "M step": while the criterion could well be minimized directly (at the expense of one "E step" per criterion evaluation), "M step" minimizations are much easier to do.

As noted in section 3, if the variational criterion (3.2) lacked the $\log|\boldsymbol{A}|$ part, it would correspond to a penalized least squares MAP objective (3.3), and simple efficient algorithms would apply. As discussed in section 4.4, evaluating $\log|\boldsymbol{A}|$ or its gradient is computationally challenging. Crucially, this term satisfies a concavity property. As shown in section 4.2, Fenchel duality implies that $\log|\boldsymbol{A}| \le \boldsymbol{z}_1^T(\boldsymbol{\gamma}^{-1}) - g_1^*(\boldsymbol{z}_1)$. For any fixed $\boldsymbol{z}_1 \succ \boldsymbol{0}$, the upper bound is tangentially tight, convex in $\boldsymbol{\gamma}$, and decouples additively. If $\log|\boldsymbol{A}|$ is replaced by this upper bound, the resulting objective $\phi_{\boldsymbol{z}_1}(\boldsymbol{u},\boldsymbol{\gamma}) := \boldsymbol{z}_1^T(\boldsymbol{\gamma}^{-1}) + h(\boldsymbol{\gamma}) + R(\boldsymbol{u},\boldsymbol{\gamma}) - g_1^*(\boldsymbol{z}_1)$ is of the same decoupled penalized least squares form as a MAP criterion (3.3). This decomposition suggests a double loop algorithm for solving $\min_{\boldsymbol{\gamma}} \phi(\boldsymbol{\gamma})$. In *inner loop minimizations*, we solve $\min_{\boldsymbol{u},\boldsymbol{\gamma}\succeq\boldsymbol{0}} \phi_{\boldsymbol{z}_1}$ for fixed $\boldsymbol{z}_1 \succ \boldsymbol{0}$, and in interjacent *outer loop updates*, we refit $\boldsymbol{z}_1 \leftarrow \operatorname{argmin} \phi_{\boldsymbol{z}_1}(\boldsymbol{u},\boldsymbol{\gamma})$.

The MAP estimation objective (3.3) and $\phi_{\boldsymbol{z}_1}(\boldsymbol{u},\boldsymbol{\gamma})$ have a similar form. Specifically, recall that $-2g_i(x_i) = \min_{\gamma_i\ge 0} x_i/\gamma_i + h_i(\gamma_i)$, where $g_i(x_i) = \tilde{g}_i(x_i^{1/2})$ and $\tilde{g}_i(s_i) = \log t_i(s_i) - b_i s_i$. The inner loop problem is

$$\min_{\boldsymbol{u},\boldsymbol{\gamma}\succeq\boldsymbol{0}} \phi_{\boldsymbol{z}_1}(\boldsymbol{u},\boldsymbol{\gamma}) = \min_{\boldsymbol{u}} \sigma^{-2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{u}\|^2 - 2\sum\nolimits_{i=1}^{q}\left(g_i(z_{1,i}+s_i^2) + b_i s_i\right), \tag{4.1}$$

where $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$. This is a smoothed version of the MAP estimation problem, which would be obtained for $z_{1,i} = 0$. However, $z_{1,i} > 0$ in our approximate inference algorithm at all times (see section 4.2). Upon inner loop convergence to $\boldsymbol{u}_*$, $\gamma_{*,i} = -1/[2(dg_i/dx_i)|_{x_i=z_{1,i}+s_{*,i}^2}]$, where $\boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$. Note that in order to run the algorithm, the analytic form of $h_i(\gamma_i)$ need not be known. For Laplace potentials (2.2), the inner loop penalizer is $2\sum_i \tau_i\sqrt{z_{1,i}+s_i^2}$, and

$\gamma_{*,i} = \sqrt{z_{1,i} + s^2_{*,i}}/\tau_i$.

Importantly, the inner loop problem (4.1) is of the same simple penalized least squares form as MAP estimation, and any of the wide range of recent efficient solvers can be plugged into our method. For example, the *iteratively reweighted least squares* (IRLS) algorithm [14], a variant of the Newton–Raphson method, can be used (details are given in section 4.3). Each Newton step requires the solution of a linear system with a matrix of the same form as $\boldsymbol{A}$ (3.1), and the convergence rate of IRLS is quadratic. It follows from the derivation of (3.2) that once an inner loop has converged to $(\boldsymbol{u}_*, \boldsymbol{\gamma}_*)$, the minimizer $\boldsymbol{u}_*$ is the mean of the approximate posterior $Q(\boldsymbol{u}|\boldsymbol{y})$ for $\boldsymbol{\gamma}_*$.

The rationale behind our algorithms lies in *decoupling* the variational criterion ϕ via a Fenchel duality upper bound, thereby matching algorithmic scheduling to the computational complexity structure of ϕ. To appreciate this point, note that in an off-the-shelf optimizer applied to $\min_{\boldsymbol{\gamma} \succeq \boldsymbol{0}} \phi(\boldsymbol{\gamma})$, both $\phi(\boldsymbol{\gamma})$ and the gradient $\nabla_{\boldsymbol{\gamma}} \phi$ have to be computed frequently. In this respect, the $\log|\boldsymbol{A}|$ coupling term proves by far more computationally challenging than the rest (see section 4.4). This obvious computational difference between parts of $\phi(\boldsymbol{\gamma})$ is not exploited in standard gradient-based algorithms: they require all of $\nabla_{\boldsymbol{\gamma}} \phi$ in each iteration, and all of $\phi(\boldsymbol{\gamma})$ in every single line search step. As discussed in section 4.4, computing $\log|\boldsymbol{A}|$ to high accuracy is not feasible for models of interest here, and most off-the-shelf optimizers with fast convergence rates are very hard to run with such approximately computed criteria. In our algorithm, the critical part is identified and decoupled, and resulting inner loop problems can be solved by robust and efficient standard code, requiring a minimal effort of adaptation. The bound $\phi_{\boldsymbol{z}_1}$ has to be refitted only once per outer loop iteration: $\boldsymbol{z}_1 \leftarrow \nabla_{\boldsymbol{\gamma}^{-1}} \log|\boldsymbol{A}|$ (see section 4.2), the computation of which constitutes the most demanding part of $\nabla_{\boldsymbol{\gamma}} \phi$. Fenchel duality bounding[2] is used to minimize the number of these costly steps (further advantages are noted at the end of section 4.4). Resulting double loop algorithms are simple to implement based on efficient penalized least squares reconstruction code, taking full advantage of the very well researched state of the art for this setup.

4.1. The general case. In this section, we generalize the double loop algorithm along two directions. First, if potentials $\log t_i(s_i)$ are not log-concave, the inner loop problems (4.1) are not convex in general (Theorem 3.3), yet a simple variant can be used to remedy this defect. Second, as detailed in section 4.2, there are different ways of decoupling $\log|\boldsymbol{A}|$, giving rise to different algorithms. In this section, we concentrate on developing these variants; their practical differences and the implications thereof are elaborated on in section 5.

If $t_i(s_i)$ is not log-concave, then $h_i(\gamma_i)$ is not convex in general (Theorem 3.3). In this case, we can write $h_i(\gamma_i) = h_{\cap,i}(\gamma_i) + h_{\cup,i}(\gamma_i)$, where $h_{\cap,i}$ is concave and nondecreasing and $h_{\cup,i}$ is convex. Such a decomposition is not unique and has to be chosen for each h_i at hand. In hindsight, $h_{\cap,i}$ should be chosen as small as possible (for example, $h_{\cap,i} \equiv 0$ if $t_i(s_i)$ is log-concave, as in the case treated above), and if IRLS is to be used for inner loop minimizations (see section 4.3), $h_{\cup,i}$ should be twice continuously differentiable. For Student's t potentials

[2]Note that Fenchel duality bounding is also used in difference-of-convex programming, a general framework to address nonconvex, typically nonsmooth optimization problems in a double loop fashion. In our application, $\phi(\boldsymbol{\gamma})$ is smooth in general and convex in many applications (see section 3): our reasons for applying bound minimization are different.

(2.3), such a decomposition is given in Appendix A.6. We define $h_\cap(\boldsymbol{\gamma}) = \sum_i h_{\cap,i}(\gamma_i)$, $h_\cup(\boldsymbol{\gamma}) = \sum_i h_{\cup,i}(\gamma_i)$ and modify outer loop updates by applying a second Fenchel duality bounding operation, $h_\cap(\boldsymbol{\gamma}) \le \tilde{\boldsymbol{z}}_2^T\boldsymbol{\gamma} - \tilde{g}_2^*(\tilde{\boldsymbol{z}}_2)$, resulting in a variant of the inner loop criterion (4.1). If $h_{\cap,i}$ is differentiable, the outer loop update is $\tilde{z}_2 \leftarrow h'_{\cap,i}(\gamma_i)$; otherwise any element from the subgradient can be chosen (note that $\tilde{z}_2 \ge 0$, as $h_{\cap,i}$ is nondecreasing). Moreover, as shown in section 4.2, Fenchel duality can be employed in order to bound $\log|\boldsymbol{A}|$ in two different ways, one employed above, and the other being $\log|\boldsymbol{A}| \le \boldsymbol{z}_2^T\boldsymbol{\gamma} - \mathbf{1}^T(\log\boldsymbol{\gamma}) - g_2^*(\boldsymbol{z}_2)$, $\boldsymbol{z}_2 \succeq \mathbf{0}$. Combining these bounds (by adding $\tilde{\boldsymbol{z}}_2$ to $\boldsymbol{z}_2$), we obtain

$$\phi(\boldsymbol{\gamma},\boldsymbol{u}) \le \phi_{\boldsymbol{z}}(\boldsymbol{u},\boldsymbol{\gamma}) := \boldsymbol{z}_1^T(\boldsymbol{\gamma}^{-1}) + \boldsymbol{z}_2^T\boldsymbol{\gamma} - \boldsymbol{z}_3^T(\log\boldsymbol{\gamma}) + h_\cup(\boldsymbol{\gamma}) + R(\boldsymbol{u},\boldsymbol{\gamma}) - g^*(\boldsymbol{z}),$$

where $z_{3,i} \in \{0,1\}$ and $g^*(\boldsymbol{z})$ collects the offsets of all Fenchel duality bounds. Note that $\boldsymbol{z}_j \succeq \mathbf{0}$ for $j = 1,2,3$, and for each i, either $z_{1,i} > 0$ and $z_{3,i} = 0$, or $z_{1,i} = 0$ and $z_{2,i} > 0, z_{3,i} = 1$. We have that

$$\begin{aligned} \phi_{\boldsymbol{z}}(\boldsymbol{u}) &:= \min_{\boldsymbol{\gamma}\succeq\mathbf{0}} \phi_{\boldsymbol{z}}(\boldsymbol{u},\boldsymbol{\gamma}) = \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + 2\sum\nolimits_{i=1}^{q} h_i^*(s_i) - 2\boldsymbol{b}^T\boldsymbol{s}, \\ h_i^*(s_i) &:= \frac{1}{2}\min_{\gamma_i\ge 0} \frac{z_{1,i} + s_i^2}{\gamma_i} + z_{2,i}\gamma_i - z_{3,i}\log\gamma_i + h_{\cup,i}(\gamma_i), \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}. \end{aligned} \tag{4.2}$$

Note that $h_i^*(s_i)$ is convex as the minimum (over $\gamma_i \ge 0$) of a jointly convex argument [4, sect. 3.2.5]. The inner loop minimization problem $\min_{\boldsymbol{u}}(\min_{\boldsymbol{\gamma}} \phi_{\boldsymbol{z}})$ is of penalized least squares form and can be solved with the same array of efficient algorithms applicable to the special case (4.1). An application of the second-order IRLS method is detailed in section 4.3. A schema for the full variational inference algorithm is given in Algorithm 1.

Algorithm 1. Double loop variational inference algorithm.

repeat
  if first outer loop iteration **then**
    Initialize bound $\phi_{\boldsymbol{z}}$. $\boldsymbol{u} = \mathbf{0}$.
  else
    Outer loop update: Refit upper bound $\phi_{\boldsymbol{z}}$ to ϕ (tangent at $\boldsymbol{\gamma}$).
    Requires marginal variances $\hat{\boldsymbol{z}} = \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T)$ (section 4.4).
    Initialize $\boldsymbol{u} = \boldsymbol{u}_*$ (previous solution).
  end if
  repeat
    Newton (IRLS) iteration to minimize $\min_{\boldsymbol{\gamma}\succeq\mathbf{0}} \phi_{\boldsymbol{z}}$ (4.2) w.r.t. $\boldsymbol{u}$.
    Entails solving a linear system (by linear conjugate gradients) and line search (section 4.3).
  until $\boldsymbol{u}_* = \mathrm{argmin}_{\boldsymbol{u}}(\min_{\boldsymbol{\gamma}\succeq\mathbf{0}} \phi_{\boldsymbol{z}})$ converged
  Update $\boldsymbol{\gamma} = \mathrm{argmin}_{\boldsymbol{\gamma}\succeq\mathbf{0}} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*,\cdot)$.
until outer loop converged

The algorithms are specialized to the $t_i(s_i)$ through $h_i^*(s_i)$ and its derivatives. The important special case of log-concave $t_i(s_i)$ has been detailed above. For Student's t potentials (2.3),

a decomposition is detailed in Appendix A.6. In this case, the overall problem $\min_{\boldsymbol{\gamma}\succeq\mathbf{0}}\phi(\boldsymbol{\gamma})$ is not convex, yet our double loop algorithm iterates over standard-form convex inner loop problems. Finally, for log-concave $t_i(s_i)$ and $z_{2,i}\neq 0$ (type B bounding, section 4.2), our algorithm can be implemented generically, as detailed in Appendix A.5.

We close this section by establishing some characteristics of these algorithms. First, we found it useful to initialize them with constant $\boldsymbol{z}_1$ and/or $\boldsymbol{z}_2$ of small size, and with $\boldsymbol{u}=\mathbf{0}$. Moreover, each subsequent inner loop minimization is started with $\boldsymbol{u}=\boldsymbol{u}_*$ from the last round. The development of our algorithms is inspired by the sparse estimation method of [43]; relationships to this method are discussed in section 5. Our algorithms are globally convergent; a stationary point of $\phi(\boldsymbol{\gamma})$ is found from any starting point $\boldsymbol{\gamma}\succ\mathbf{0}$ (recall from Theorem 3.3 that for log-concave potentials, this stationary point is a global solution). This is detailed in [43]. Intuitively, at the beginning of each outer loop iteration, $\phi_{\boldsymbol{z}}$ and ϕ have the same tangent plane at $\boldsymbol{\gamma}$, so that each inner loop minimization significantly decreases ϕ unless $\nabla_{\boldsymbol{\gamma}}\phi=\mathbf{0}$. Note that this convergence proof requires that outer loop updates are done exactly; this point is elaborated on at the end of section 4.4.

Our variational *inference* algorithms differ from previous methods[3] in that models that are orders of magnitude larger can successfully be addressed. They apply to the particular variational relaxation introduced in section 3, whose relationship to other inference approximations is detailed in [29]. While most previous relaxations attain scalability through many factorization assumptions concerning the approximate posterior, $Q(\boldsymbol{u}|\boldsymbol{y})$ in our method is fully coupled, sharing its conditional independence graph with the true posterior $P(\boldsymbol{u}|\boldsymbol{y})$.

4.2. Bounding $\log|A|$. We need to upper bound $\log|\boldsymbol{A}|$ by a term which is convex and decoupling in $\boldsymbol{\gamma}$. This can be done in two different ways using Fenchel duality, giving rise to bounds with different characteristics. Details for the development here are given in Appendix A.4.

Recall our assumption that $\boldsymbol{A}\succ\mathbf{0}$ for each $\boldsymbol{\gamma}\succ\mathbf{0}$. If $\boldsymbol{\pi}=\boldsymbol{\gamma}^{-1}$, then $\boldsymbol{\pi}\mapsto\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})|=\log|\boldsymbol{A}|$ is concave for $\boldsymbol{\pi}\succ\mathbf{0}$ (Theorem 3.1(2) with $f\equiv\mathrm{id}$). Moreover, $\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})|$ is increasing and unbounded in each component of $\boldsymbol{\pi}$ (Theorem 5.1). Fenchel duality [26, sect. 12] implies that $\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})|=\min_{\boldsymbol{z}_1\succ\mathbf{0}}\boldsymbol{z}_1^T\boldsymbol{\pi}-g_1^*(\boldsymbol{z}_1)$ for $\boldsymbol{\pi}\succ\mathbf{0}$; thus $\log|\boldsymbol{A}|=\min_{\boldsymbol{z}_1\succ\mathbf{0}}\boldsymbol{z}_1^T(\boldsymbol{\gamma}^{-1})-g_1^*(\boldsymbol{z}_1)$ for $\boldsymbol{\gamma}\succ\mathbf{0}$. Therefore, $\log|\boldsymbol{A}|\leq\boldsymbol{z}_1^T(\boldsymbol{\gamma}^{-1})-g_1^*(\boldsymbol{z}_1)$. For fixed $\boldsymbol{\gamma}\succ\mathbf{0}$, this is an equality for

$$\boldsymbol{z}_{1,*}=\nabla_{\boldsymbol{\gamma}^{-1}}\log|\boldsymbol{A}|=\hat{\boldsymbol{z}}:=(\mathrm{Var}_Q[s_i|\boldsymbol{y}])=\mathrm{diag}^{-1}\left(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T\right)\succ\mathbf{0},$$

and $g_1^*(\boldsymbol{z}_{1,*})=\boldsymbol{z}_{1,*}^T(\boldsymbol{\gamma}^{-1})-\log|\boldsymbol{A}|$. This is called bounding type A in what follows.

On the other hand, $\boldsymbol{\gamma}\mapsto\mathbf{1}^T(\log\boldsymbol{\gamma})+\log|\boldsymbol{A}|$ is concave for $\boldsymbol{\gamma}\succ\mathbf{0}$ (Theorem 3.1(3) with $f\equiv\mathrm{id}$). Employing Fenchel duality once more, we have that $\log|\boldsymbol{A}|\leq\boldsymbol{z}_2^T\boldsymbol{\gamma}-\mathbf{1}^T(\log\boldsymbol{\gamma})-g_2^*(\boldsymbol{z}_2)$, $\boldsymbol{z}_2\succeq\mathbf{0}$. For any fixed $\boldsymbol{\gamma}$, equality is attained at $\boldsymbol{z}_{2,*}=\boldsymbol{\gamma}^{-1}\circ(\mathbf{1}-\boldsymbol{\gamma}^{-1}\circ\hat{\boldsymbol{z}})$, and $g_2^*(\boldsymbol{z}_{2,*})=\boldsymbol{z}_{2,*}^T\boldsymbol{\gamma}-\log|\boldsymbol{A}|-\mathbf{1}^T(\log\boldsymbol{\gamma})$ at this point. This is referred to as bounding type B.

In general, type A bounding is tighter for γ_i away from zero, while type B bounding is tighter for γ_i close to zero (see Figure 2); implications of this point are discussed in section 5. Whatever bounding type we use, refitting the corresponding upper bound to $\log|\boldsymbol{A}|$ requires

[3]This comment holds for approximate *inference* methods. For sparse *estimation*, large scale algorithms are available (see section 5).

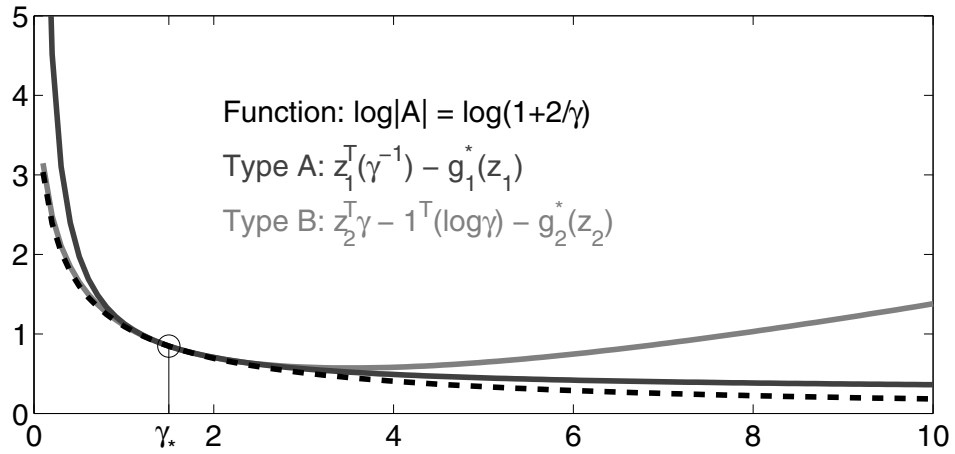

Figure 2. *Comparison of type A and B upper bounds on* $\log(1+2/\gamma)$.

the computation of $\hat{\boldsymbol{z}} = (\mathrm{Var}_Q[s_i|\boldsymbol{y}])$: all marginal *variances* of the Gaussian distribution $Q(\boldsymbol{s}|\boldsymbol{y})$. In general, computing Gaussian marginal variances is a hard numerical problem, which is discussed in more detail in section 4.4.

4.3. The inner loop optimization. The inner loop minimization problem is given by (4.2), and its special case (4.1) for log-concave potentials and $\log|\boldsymbol{A}|$ bounding type A is given by $h_i^*(s_i) = -g_i(z_{1,i} + s_i^2)$. This problem is of standard penalized least squares form, and a large number of recent algorithms [12, 3, 46] can be applied with little customization effort. In this section, we provide details about how to apply the IRLS algorithm [14], a special case of the Newton–Raphson method.

We describe a single IRLS step here, starting from $\boldsymbol{u}$. Let $\boldsymbol{r} := \boldsymbol{X}\boldsymbol{u} - \boldsymbol{y}$ denote the residual vector. If $\theta_i := (h_i^*)'(s_i) - b_i$, $\rho_i := (h_i^*)''(s_i)$, then

$$\nabla_{\boldsymbol{u}}(\phi_{\boldsymbol{z}}/2) = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{r} + \boldsymbol{B}^T\boldsymbol{\theta}, \quad \nabla\nabla_{\boldsymbol{u}}(\phi_{\boldsymbol{z}}/2) = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T(\mathrm{diag}\,\boldsymbol{\rho})\boldsymbol{B}.$$

Note that $\rho_i \geq 0$ by the convexity of h_i^*. The Newton search direction is

$$\boldsymbol{d} := -(\sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T(\mathrm{diag}\,\boldsymbol{\rho})\boldsymbol{B})^{-1}(\sigma^{-2}\boldsymbol{X}^T\boldsymbol{r} + \boldsymbol{B}^T\boldsymbol{\theta}).$$

The computation of $\boldsymbol{d}$ requires us to solve a system with a matrix of the same form as $\boldsymbol{A}$, a reweighted least squares problem otherwise used to compute the means in a Gaussian model of the structure of $Q(\boldsymbol{u}|\boldsymbol{y})$. We solve these systems approximately by the preconditioned *linear conjugate gradients* (LCG) algorithm [13]. The cost per iteration of LCG is dominated by matrix-vector multiplications (MVMs) with $\boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{B}$, and $\boldsymbol{B}^T$. A line search along $\boldsymbol{d}$ can be run in negligible time. If $f(t) := \phi_{\boldsymbol{z}}(\boldsymbol{u} + t\boldsymbol{d})/2$, then $f'(t) = \sigma^{-2}((\boldsymbol{X}\boldsymbol{d})^T\boldsymbol{r} + t\|\boldsymbol{X}\boldsymbol{d}\|^2) + (\boldsymbol{B}\boldsymbol{d})^T\boldsymbol{\theta}^{(t)}$, where $\boldsymbol{\theta}^{(t)}$ is the gradient at $\boldsymbol{s}^{(t)} = \boldsymbol{s} + t\boldsymbol{B}\boldsymbol{d}$. With $(\boldsymbol{X}\boldsymbol{d})^T\boldsymbol{r}$, $\|\boldsymbol{X}\boldsymbol{d}\|^2$, and $\boldsymbol{B}\boldsymbol{d}$ precomputed, $f(t)$ and $f'(t)$ can be evaluated in $O(q)$ without any further MVMs. The line search is started with $t_0 = 1$. Finally, once $\boldsymbol{u}_* = \mathrm{argmin}_{\boldsymbol{u}}\,\phi_{\boldsymbol{z}}(\boldsymbol{u})$ is found, $\boldsymbol{\gamma}$ is explicitly updated as $\mathrm{argmin}\,\phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \cdot)$. Note that at this point, $\boldsymbol{u}_* = \mathrm{E}_Q[\boldsymbol{u}|\boldsymbol{y}]$, which follows from the derivation at the beginning of section 3.

4.4. Estimation of Gaussian variances. Variational inference does require marginal variances $\hat{\boldsymbol{z}} = \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T) = (\mathrm{Var}_Q[s_i|\boldsymbol{y}])$ of the Gaussian $Q(\boldsymbol{s}|\boldsymbol{y})$ (see section 4.2), which are much harder to approximate than means. In this section, we discuss a general method for (co)variance approximation. Empirically, the performance of our double loop algorithms is remarkably robust in light of substantial overall variance approximation errors; some insights into this finding are given below.

Marginal posterior variances have to be computed in any approximate Bayesian inference method, although they are not required in typical sparse point estimation techniques (see section 5). Our double loop algorithms reduce approximate inference to point estimation and Gaussian (co)variance approximation. Not only do they expose the latter as the missing link between sparse estimation and variational inference, but their main rationale is that Gaussian variances have to be computed only a few times, while off-the-shelf variational optimizers query them for every single criterion evaluation.

Marginal variance approximations have been proposed for sparsely connected Gaussian Markov random fields (MRFs), iterating over embedded spanning tree models [42] or exploiting rapid correlation decay in models with homogeneous prior [20]. In applications of interest here, $\boldsymbol{A}$ neither is sparse nor has useful graphical model structure. Committing to a low-rank approximation of the covariance $\boldsymbol{A}^{-1}$ [20, 27], an optimal choice in terms of preserving variances is principal components analysis (PCA), based on the smallest eigenvalues (resp., eigenvectors) vectors of $\boldsymbol{A}$ (resp., the largest of $\boldsymbol{A}^{-1}$). The *Lanczos algorithm* [13] provides a scalable approximation to PCA and was employed for variance estimation in [27]. After k iterations, we have an orthonormal basis $\boldsymbol{Q}_k \in \mathbb{R}^{n\times k}$, within which extremal eigenvectors of $\boldsymbol{A}$ are rapidly well approximated (due to the nearly *linear* spectral decay of typical $\boldsymbol{A}$ matrices (Figure 6, upper panel), both largest and smallest eigenvalues are obtained). As $\boldsymbol{Q}_k^T\boldsymbol{A}\boldsymbol{Q}_k = \boldsymbol{T}_k$ is tridiagonal, the Lanczos variance approximation $\hat{\boldsymbol{z}}_k = \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{Q}_k\boldsymbol{T}_k^{-1}\boldsymbol{Q}_k^T\boldsymbol{B}^T)$ can be computed efficiently. Importantly, $\hat{z}_{k,i} \le \hat{z}_{k+1,i} \le \hat{z}_i$ for all k and i. Namely, if $\boldsymbol{Q}_k = [\boldsymbol{q}_1 \dots \boldsymbol{q}_k]$ and $\boldsymbol{T}_k$ has main diagonal (α_l) and subdiagonal (β_l), let (e_l) and (d_l) be the main diagonal and subdiagonal of the bidiagonal Cholesky factor $\boldsymbol{L}_k$ of $\boldsymbol{T}_k$. Then, $d_{k-1} = \beta_{k-1}/e_{k-1}$, $e_k = (\alpha_k - d_{k-1}^2)^{1/2}$, with $d_0 = 0$. If $\boldsymbol{V}_k := \boldsymbol{B}\boldsymbol{Q}_k\boldsymbol{L}_k^{-T}$, we have $\boldsymbol{V}_k = [\boldsymbol{v}_1 \dots \boldsymbol{v}_k]$, $\boldsymbol{v}_k = (\boldsymbol{B}\boldsymbol{q}_k - d_{k-1}\boldsymbol{v}_{k-1})/e_k$. Finally, $\hat{\boldsymbol{z}}_k = \hat{\boldsymbol{z}}_{k-1} + \boldsymbol{v}_k^2$ (with $\hat{\boldsymbol{z}}_0 = \boldsymbol{0}$).

Unfortunately, the Lanczos algorithm is much harder to run in practice than LCG, and its cost grows superlinearly in k. A promising variant of selectively reorthogonalized Lanczos [24] is given in [2], where contributions from undesired parts of the spectrum ($\boldsymbol{A}$'s largest eigenvalues in our case) are filtered out by replacing $\boldsymbol{A}$ with polynomials of itself. Recently, randomized PCA approaches have become popular [15], although their relevance for variance approximation is unclear. Nevertheless, for large scale problems of interest, standard Lanczos can be run for $k \ll n$ iterations only, at which point most of the $\hat{z}_{k,i}$ are severely underestimated (see section 7.3). Since Gaussian variances are essential for variational Bayesian inference, but scalable, uniformly accurate variance estimators are not known, *robustness* to variance approximations errors is critical for any large scale approximate inference algorithm.

What do the Lanczos variance approximation errors imply for our double loop algorithms? First, the global convergence proof of section 4.1 requires exact variances $\hat{\boldsymbol{z}}$; it may be compromised if $\hat{\boldsymbol{z}}_k$ is used instead. This problem is analyzed in [31]: the convergence proof

remains valid with the PCA approximation, which, however, is different from the Lanczos[4] approximation. Empirically, we have not encountered convergence problems so far.

Surprisingly, while $\hat{\boldsymbol{z}}_k$ is much smaller than $\hat{\boldsymbol{z}}$ in practice, there is little indication of substantial negative impact on performance. This important robustness property is analyzed in section 7.3 for an SLM with Laplace potentials. The underestimation bias has systematic structure (Figure 6, middle and lower panel): moderately small $\hat{z}_i$ are damped most strongly, while large $\hat{z}_i$ are approximated accurately. This happens because the largest coefficients $\hat{z}_i$ depend most strongly on the largest covariance eigenvectors, which are shaped in early Lanczos iterations. This particular error structure has statistical consequences. Recalling the inner loop penalty for Laplacians (2.2) $h_i^*(s_i) = \tau_i(\hat{z}_i + s_i^2)^{1/2}$, the smaller $\hat{z}_i$ is, the stronger it enforces sparsity. If $\hat{z}_i$ is underestimated, the penalty on s_i is stronger than intended, yet this strengthening does not happen uniformly. Coefficients s_i deemed most relevant with exact variance computation (largest $\hat{z}_i$) are least affected (as $\hat{z}_{k,i} \approx \hat{z}_i$ for those), while already subdominant ones (smaller $\hat{z}_i$) are suppressed even more strongly (as $\hat{z}_{k,i} \ll \hat{z}_i$). At least in our experience so far (with sparse linear models), this selective variance underestimation effect seems benign or even somewhat beneficial.

4.5. Extension to group potentials. There is substantial recent interest in methods incorporating sparse *group penalization*, meaning that a number of latent coefficients (such as the column of a matrix or the incoming edge weights for a graph) are penalized jointly [47, 40]. Our algorithms are easily generalized to models with potentials of the form $t_i(\|\boldsymbol{s}_i\|)$, with $\boldsymbol{s}_i$ a subvector of $\boldsymbol{s}$ and $\|\cdot\|$ the Euclidean norm, if $t_i(\cdot)$ is even and super-Gaussian. Such *group potentials* are frequently used in practice. The isotropic total variation penalty is the sum of $\|(dx_i, dy_i)\|$, dx_i, dy_i differences along coordinate axes. It corresponds to group Laplace potentials. In our magnetic resonance imaging (MRI) application (section 7.4), we deal with complex-valued $\boldsymbol{u}$ and $\boldsymbol{s}$. Each entry is treated as element in $\mathbb{R}^2$, and potentials are placed on $|s_i| = \|(\Re s_i, \Im s_i)\|$. Note that with t_i on $\|\boldsymbol{s}_i\|$, the single parameter γ_i is shared by the coefficients $\boldsymbol{s}_i$.

The generalization of our algorithms to group potentials is almost automatic. For example, if all $\boldsymbol{s}_i$ have the same dimensionality, $\boldsymbol{\Gamma}^{-1}$ is replaced by $\boldsymbol{\Gamma}^{-1} \otimes \boldsymbol{I}$ in the definition of $\boldsymbol{A}$, and $\hat{\boldsymbol{z}}$ is replaced by $(\boldsymbol{I} \otimes \boldsymbol{1}^T)\,\mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T)$ in section 4.2. Moreover, $x_i = s_i^2$ is replaced by $x_i = \|\boldsymbol{s}_i\|^2$, whereas the definition of $g_i(x_i)$ remains the same. Apart from these simple replacements, only IRLS inner loop iterations have to be modified (at no extra cost), as detailed in Appendix A.7.

4.6. Publicly available code: The `glm-ie` toolbox. Algorithms and techniques presented in this paper are implemented[5] as part of the *generalized linear model inference and estimation* toolbox (`glm-ie`), maintained as `mloss.org` project at http://mloss.org/software/view/269/. The code runs with both MATLAB 7 and the free `Octave 3.2`. It comprises algorithms for MAP (penalized least squares) estimation and variational inference in generalized linear models (section 4), along with Lanczos code for Gaussian variances (section 4.4).

[4]While Lanczos can be used to compute the PCA approximation (fixed number L of smallest eigenvalues/eigenvectors of $\boldsymbol{A}$), this is rather wasteful.

[5]Our experiments in section 7 use different `C++` and Fortran code, which differs from `glm-ie` mainly by being somewhat faster on large problems.

The generic design of the `glm-ie` toolbox allows for a range of applications, as illustrated by a number of example programs included in the package. Many super-Gaussian potentials $t_i(s_i)$ are included, and others can easily be added by the user. In particular, the toolbox contains a range of solvers for MAP and inner loop problems, from IRLS (or truncated Newton; see section 4.3) over conjugate gradients to quasi-Newton, as well as a range of commonly used operators for constructing $\boldsymbol{X}$ and $\boldsymbol{B}$ matrices.

5. Sparse estimation and sparse Bayesian inference. In this section, we contrast approximate Bayesian inference with point estimation for sparse linear models (SLMs): *sparse Bayesian inference* versus *sparse estimation*. These problem classes serve distinct goals and come with different algorithmic characteristics yet are frequently confused in the literature. Briefly, the goal in sparse estimation is to eliminate variables not needed for the task at hand, while sparse inference aims at quantifying uncertainty in decisions and dependencies between components. While variable elimination is a boon for efficient computation, it cannot be relied upon in sparse inference. Sensible uncertainty estimates like posterior covariance, at the heart of decision-making problems such as Bayesian experimental design, are eliminated alongside.

We restrict ourselves to super-Gaussian SLM problems in terms of variables $\boldsymbol{u}$ and $\boldsymbol{\gamma} \succeq \boldsymbol{0}$, relating the sparse Bayesian inference relaxation $\min_{\boldsymbol{\gamma} \succeq \boldsymbol{0}} \phi(\boldsymbol{\gamma})$ with two sparse estimation principles: maximum a posteriori (MAP) reconstruction (3.3) and automatic relevance determination (ARD) [43], a sparse reconstruction method which inspired our algorithms. We begin by establishing a key difference between these settings. Recall from Theorem 3.1(4) that $\gamma_i = 0$ implies[6] $\mathrm{Var}_Q[s_i|\boldsymbol{y}] = 0$: s_i is *eliminated*, fixed at zero with absolute certainty. Exact sparsity in $\boldsymbol{\gamma}$ does not happen for *inference*, while *estimation* methods are characterized by fixing many γ_i to zero.

Theorem 5.1. *Let $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, $\boldsymbol{B} \in \mathbb{R}^{q \times n}$ be matrices such that $\tilde{\boldsymbol{A}}(\boldsymbol{\pi}) = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T(\mathrm{diag}\,\boldsymbol{\pi})\boldsymbol{B} \succ \boldsymbol{0}$ for each $\boldsymbol{\pi} \succ \boldsymbol{0}$ and no row of $\boldsymbol{B}$ is equal to $\boldsymbol{0}^T$.*

- *The function $\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})|$ is increasing in each component π_i, and is unbounded above. For any sequence $\boldsymbol{\pi}_t$ with $\|\boldsymbol{\pi}_t\| \to \infty$ $(t \to \infty)$ and $\boldsymbol{\pi}_t \succeq \varepsilon\boldsymbol{1}$ for some $\varepsilon > 0$, we have that $\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi}_t)| \to \infty$ $(t \to \infty)$.*
- *Assume that $\log P(\boldsymbol{u}|\boldsymbol{y})$ is bounded above as a function of $\boldsymbol{u}$. Recall the variational criterion $\phi(\boldsymbol{\gamma})$ from (3.2). For any bounded sequence $\boldsymbol{\gamma}_t$ with $(\boldsymbol{\gamma}_t)_i \to 0$ $(t \to \infty)$ for some $i \in \{1, \ldots, q\}$, we have that $\phi(\boldsymbol{\gamma}_t) \to \infty$.*
  In particular, any local minimum point $\boldsymbol{\gamma}_*$ of the variational problem $\min_{\boldsymbol{\gamma} \succeq \boldsymbol{0}} \phi(\boldsymbol{\gamma})$ must have positive components; i.e., $\boldsymbol{\gamma}_* \succ \boldsymbol{0}$.

A proof is given in Appendix A.3. $\log|\boldsymbol{A}|$ acts as a barrier function for $\boldsymbol{\gamma} \succ \boldsymbol{0}$. Any local minimum point $\boldsymbol{\gamma}_*$ of (3.2) is positive throughout, and $\mathrm{Var}_Q[s_i|\boldsymbol{y}] > 0$ for all $i = 1, \ldots, q$. Coefficient elimination does not happen in variational Bayesian inference.

Consider MAP estimation (3.3) with even super-Gaussian potentials $t_i(s_i)$. Following [25], a sufficient condition for sparsity is that $-\log t_i(s_i)$ is concave for $s_i > 0$. In this case, if $\mathrm{rank}\,\boldsymbol{X} = m$ and $\mathrm{rank}\,\boldsymbol{B} = n$, then any local MAP solution $\boldsymbol{u}_*$ is exactly sparse: no more than m coefficients of $\boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$ are nonzero. Examples are $t_i(s_i) = e^{-\tau_i|s_i|^p}$, $p \in (0, 1]$, including Laplace potentials ($p = 1$). Moreover, $\gamma_{*,i} = 0$ whenever $s_{*,i} = 0$ in this case (see Appendix A.3). Local minimum points of SLM MAP estimation are substantially exactly

[6] While the proof of Theorem 3.1(4) holds for $\boldsymbol{\gamma} \succ \boldsymbol{0}$, $\mathrm{Var}_Q[s_i|\boldsymbol{y}]$ is a continuous function of $\boldsymbol{\gamma}$.

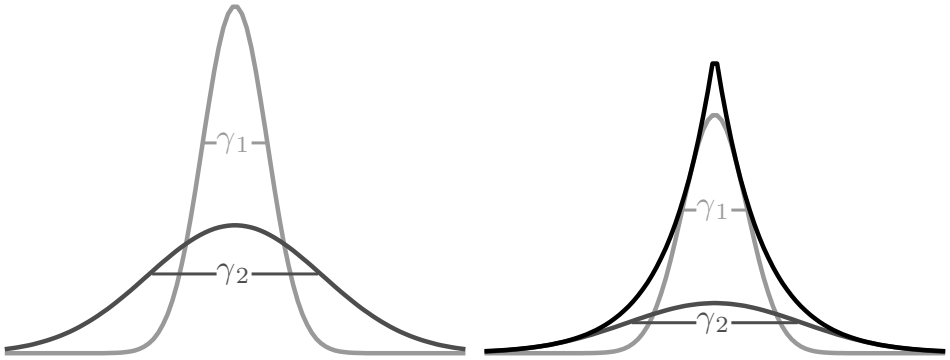

Figure 3. *Different roles of Gaussian functions (width γ_i) in sparse estimation versus sparse inference. Left: sparse estimation (ARD). Gaussian functions are normalized, and there is incentive to drive $\gamma_i \to 0$. Right: variational inference for Laplace potentials (2.2). Gaussian functions are lower bounds of $t_i(s_i)$, and their mass vanishes as $\gamma_i \to 0$. There is no incentive to eliminate γ_i.*

sparse, with matching sparsity patterns of $\boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$ and $\boldsymbol{\gamma}_*$.

A powerful sparse estimation method, automatic relevance determination (ARD) [43], has inspired our approximate inference algorithms developed above. The ARD criterion ϕ_{ARD} is (3.2) with $h(\boldsymbol{\gamma}) = \mathbf{1}^T(\log \boldsymbol{\gamma})$, obtained as the zero-temperature limit ($\nu \to 0$) of variational inference with Student's t potentials (2.3). The function $h_i(\gamma_i)$ is given in Appendix A.6, and $h_i(\gamma_i) \to \log \gamma_i$ ($\nu \to 0$) if additive constants independent of γ_i are dropped.[7] ARD can also be seen as marginal likelihood maximization: $\phi_{\mathrm{ARD}}(\boldsymbol{\gamma}) = -2 \log \int P(\boldsymbol{y}|\boldsymbol{u}) N(\boldsymbol{s}|\mathbf{0}, \boldsymbol{\Gamma})\, d\boldsymbol{u}$ up to an additive constant. Sparsity penalization is implied by the fact that the prior $N(\boldsymbol{s}|\mathbf{0}, \boldsymbol{\Gamma})$ is *normalized* (see Figure 3, left). The ARD problem is not convex. A provably convergent double loop ARD algorithm is obtained by employing bounding type B (section 4.2); along lines similar to those in section 4.1 we obtain

$$\min_{\boldsymbol{\gamma} \succeq \mathbf{0}} \phi_{\mathrm{ARD}}(\boldsymbol{\gamma}) = \min_{\boldsymbol{z}_2 \succeq \mathbf{0}} \left(\min_{\boldsymbol{u}} \sigma^{-2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + 2 \sum_{i=1}^{q} z_{2,i}^{1/2} |s_i| \right) - g_2^*(\boldsymbol{z}_2).$$

The inner problem is ℓ_1 penalized least squares estimation, a reweighted variant of MAP reconstruction for Laplace potentials. Its solutions $\boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$ are exactly sparse, along with corresponding $\boldsymbol{\gamma}_*$ (since $\gamma_{*,i} = z_{2,i}^{-1/2} |s_{*,i}|$). ARD is enforcing sparsity more aggressively than Laplace (ℓ_1) MAP reconstruction [44]. The $\log |\boldsymbol{A}|$ barrier function is counterbalanced by $h(\boldsymbol{\gamma}) = \mathbf{1}^T(\log \boldsymbol{\gamma}) = \log |\boldsymbol{\Gamma}|$. If $\boldsymbol{B} = \boldsymbol{I}$, then

$$\log |\boldsymbol{A}| + \log |\boldsymbol{\Gamma}| = \log |\boldsymbol{I} + \sigma^{-2} \boldsymbol{X} \boldsymbol{\Gamma} \boldsymbol{X}^T| \to 0 \ (\boldsymbol{\gamma} \to \mathbf{0}).$$

The conceptual difference between ARD and our variational inference relaxation is illustrated in Figure 3. In sparse inference, Gaussian functions $e^{-s_i^2/(2\gamma_i) - h_i(\gamma_i)/2}$ lower bound $t_i(s_i)$. Their mass vanishes as $\gamma_i \to 0$, driving $\phi(\boldsymbol{\gamma}) \to \infty$. For ARD, Gaussian functions $N(s_i|0, \gamma_i)$ are normalized, and $\gamma_i \to 0$ is encouraged.

[7] Note that the term dropped (C_i in Appendix A.6) becomes unbounded as $\nu \to 0$. Removing it is essential to obtaining a well-defined problem.

At this point, the roles of different bounding types introduced in section 4.2 become transparent. $\log|\boldsymbol{A}|$ is a barrier function for $\boldsymbol{\gamma} \succ \boldsymbol{0}$ (Theorem 5.1), as is its type A bound $\boldsymbol{z}_1^T(\boldsymbol{\gamma}^{-1}) - g_1^*(\boldsymbol{z}_1)$, $\boldsymbol{z}_1 \succ \boldsymbol{0}$ (see Figure 2). On the other hand, $\log|\boldsymbol{A}| + \boldsymbol{1}^T(\log\boldsymbol{\gamma})$ is bounded below, as is its type B bound $\boldsymbol{z}_2^T\boldsymbol{\gamma} - g_2^*(\boldsymbol{z}_2)$. These facts suggest that type A bounding should be preferred for variational inference, while type B bounding is best suited for sparse estimation. Indeed, experiments in section 7.1 show that for approximate inference with Laplace potentials, type A bounding is by far the better choice, while for ARD, type B bounding leads to the very efficient algorithm just sketched.

Sparse estimation methods eliminate a substantial fraction of $\boldsymbol{\gamma}$'s coefficients, while variational inference methods do not zero any of them. This difference has important computational and statistical implications. First, exact sparsity in $\boldsymbol{\gamma}$ is computationally beneficial. In this regime, even coordinate descent algorithms can be scaled up to large problems [39]. Within the ARD sparse estimation algorithm, variances $\hat{\boldsymbol{z}} \leftarrow \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T)$ have to be computed, but since $\hat{\boldsymbol{z}}$ is as sparse as $\boldsymbol{\gamma}$, this is not a hard problem. Variational inference methods have to cope without exact sparsity. The double loop algorithms of section 4 are scalable nevertheless, reducing to numerical techniques whose performance does not depend on the sparsity of $\boldsymbol{\gamma}$.

While exact sparsity in $\boldsymbol{\gamma}$ implies computational simplifications, it also rules out proper model-based uncertainty quantification.[8] If $\gamma_i = 0$, then $\mathrm{Var}_Q[s_i|\boldsymbol{y}] = 0$. If $Q(\boldsymbol{u}|\boldsymbol{y})$ is understood as representation of uncertainty, it asserts that there is *no posterior variance* in s_i at all: s_i is eliminated with absolute certainty, along with all correlations between s_i and other s_j. Sparsity in $\boldsymbol{\gamma}$ is computationally useful only if most $\gamma_i = 0$. $Q(\boldsymbol{u}|\boldsymbol{y})$, a degenerate distribution with mass only in the subspace corresponding to surviving coefficients, cannot be regarded as approximation to a Bayesian posterior. As zero is just zero, even basic queries such as a confidence ranking over eliminated coefficients cannot be based on a degenerate $Q(\boldsymbol{u}|\boldsymbol{y})$.

In particular, Bayesian experimental design (section 6), based on sparse *inference* methods [36, 28, 32, 34], excels for a range of real-world scenarios (see section 7.4), but it cannot sensibly be driven by sparse *estimation* technology. The latter is attempted in [18], employing the *sparse Bayesian learning* estimator [38] in order to drive inference queries, an approach which fails badly on real-world image data [32]. Started with a few initial measurements, it identifies a very small subspace of noneliminated coefficients (as expected for sparse estimation fed with little data), in which it essentially remains locked ever after. In order to sensibly score a candidate $\boldsymbol{X}_*$, we have to reason about what happens to *all* coefficients, which is not possible based on a "posterior" $Q(\boldsymbol{u}|\boldsymbol{y})$ that rules out most of them with full certainty.

Finally, even if the goal is point reconstruction from given data, the sparse inference posterior mean $\mathrm{E}_Q[\boldsymbol{u}|\boldsymbol{y}]$ (obtained as byproduct $\boldsymbol{u}_*$ in the double loop algorithm of section 4) can be an important alternative to an exactly sparse estimator. For the former, $\mathrm{E}_Q[\boldsymbol{s}|\boldsymbol{y}] = \boldsymbol{B}\boldsymbol{u}_*$ is not sparse in general, and the degree to which coefficients are penalized (but not eliminated) is determined by the choice of $t_i(s_i)$. To illustrate this point, we compare the mean estimators for Laplace and Student's t potentials (different $\nu > 2$) in section 7.2. These results demonstrate that, contrary to some folklore in signal and image processing, *sparser*

[8] Uncertainty quantification may also be obtained by running sparse estimation many times in a bootstrapping fashion [21]. While such procedures cure some robustness issues of MAP estimation, they are probably too costly to run in order to drive experimental design, where dependencies between variables are of interest.

is not necessarily better for point reconstruction of real-world images. Enforcing sparsity too strongly leads to fine details being smoothed out, which is not acceptable in medical imaging (fine features are often diagnostically most relevant) or photography postprocessing (most users strongly dislike unnaturally hard edges and oversmoothed areas).

Sparse *estimation* methodology has seen impressive advancements toward what it is intended to do: solving a given overparameterized reconstruction problem by eliminating nonessential variables. However, it is ill-suited for addressing decision-making scenarios driven by Bayesian *inference*. For the latter, a useful (nondegenerate) posterior approximation has to be obtained without relying on computational benefits of exact sparsity. We show how this can be done by reducing variational inference to numerical techniques (LCG and Lanczos) which can be scaled up to large problems without exact variable sparsity.

6. Bayesian experimental design. In this section, we show how to optimize the image acquisition matrix $\boldsymbol{X}$ by way of Bayesian sequential experimental design (also known as Bayesian active learning), maximizing the expected amount of information gained. Unrelated to the output of point reconstruction methods, information gain scores depend on the posterior covariance matrix $\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}]$ over full images $\boldsymbol{u}$, which within our large scale variational inference framework is approximated by the Lanczos algorithm.

In each round, a part $\boldsymbol{X}_* \in \mathbb{R}^{d\times n}$ is appended to the design $\boldsymbol{X}$, and a new (partial) measurement $\boldsymbol{y}_*$ is appended to $\boldsymbol{y}$. Candidates $\{\boldsymbol{X}_*\}$ are ranked by the *information gain*[9] score $\mathrm{H}[P(\boldsymbol{u}|\boldsymbol{y})] - \mathrm{E}_{P(\boldsymbol{y}_*|\boldsymbol{y})}[\mathrm{H}[P(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{y}_*)]]$, where $P(\boldsymbol{u}|\boldsymbol{y})$ and $P(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{y}_*)$ are posteriors for $(\boldsymbol{X},\boldsymbol{y})$ and $(\boldsymbol{X}\cup\boldsymbol{X}_*,\boldsymbol{y}\cup\boldsymbol{y}_*)$, respectively, and $P(\boldsymbol{y}_*|\boldsymbol{y}) = \int N(\boldsymbol{y}_*|\boldsymbol{X}_*\boldsymbol{u},\sigma^2\boldsymbol{I})P(\boldsymbol{u}|\boldsymbol{y})\,d\boldsymbol{u}$. Replacing $P(\boldsymbol{u}|\boldsymbol{y})$ by its best Gaussian variational approximation $Q(\boldsymbol{u}|\boldsymbol{y}) = N(\boldsymbol{u}_*,\boldsymbol{A}^{-1})$ and $P(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{y}_*)$ by $Q(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{y}_*) \propto N(\boldsymbol{y}_*|\boldsymbol{X}_*\boldsymbol{u},\sigma^2\boldsymbol{I})Q(\boldsymbol{u}|\boldsymbol{y})$, we obtain an approximate information gain score

$$\Delta(\boldsymbol{X}_*) := -\log|\boldsymbol{A}| + \log\left|\boldsymbol{A} + \sigma^{-2}\boldsymbol{X}_*^T\boldsymbol{X}_*\right| = \log\left|\boldsymbol{I} + \sigma^{-2}\boldsymbol{X}_*\boldsymbol{A}^{-1}\boldsymbol{X}_*^T\right|. \tag{6.1}$$

Note that $Q(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{y}_*)$ has the same variational parameters $\boldsymbol{\gamma}$ as $Q(\boldsymbol{u}|\boldsymbol{y})$, which simplifies and robustifies score computations. Refitting of $\boldsymbol{\gamma}$ is done at the end of each round, once the score maximizer $\boldsymbol{X}_*$ is appended along with a new measurement $\boldsymbol{y}_*$.

With N candidates of size d to be scored, a naive computation of (6.1) would require $N\cdot d$ linear systems to be solved, which is not tractable (for example, $N = 240$, $d = 512$ in section 7.4). We can make use of the Lanczos approximation once more (see section 4.4). If $\boldsymbol{Q}_k^T\boldsymbol{A}\boldsymbol{Q}_k = \boldsymbol{T}_k = \boldsymbol{L}_k\boldsymbol{L}_k^T$ ($\boldsymbol{L}_k$ is bidiagonal, computed in $O(k)$), let $\boldsymbol{V}_* := \sigma^{-1}\boldsymbol{X}_*\boldsymbol{Q}_k\boldsymbol{L}_k^{-T} \in \mathbb{R}^{d\times k}$. Then, $\Delta(\boldsymbol{X}_*) \approx \log|\boldsymbol{I} + \boldsymbol{V}_*\boldsymbol{V}_*^T| = \log|\boldsymbol{I} + \boldsymbol{V}_*^T\boldsymbol{V}_*|$ (the latter is preferable if $k < d$), at a total cost of k MVMs with $\boldsymbol{X}_*$ and $O(\max\{k,d\}\cdot\min\{k,d\}^2)$. Just as with marginal variances, Lanczos approximations of $\Delta(\boldsymbol{X}_*)$ are underestimates, nondecreasing in k. The impact of Lanczos approximation errors on design decisions is analyzed in [31]. While absolute score values are much too small, decisions depend only on the *ranking* among the highest-scoring candidates $\boldsymbol{X}_*$, which often is faithfully reproduced even for $k \ll n$. To understand this point, note that $\Delta(\boldsymbol{X}_*)$ measures the alignment of $\boldsymbol{X}_*$ with the directions of largest variance in $Q(\boldsymbol{u}|\boldsymbol{y})$. For example, the single best unit-norm filter $\boldsymbol{x}_* \in \mathbb{R}^n$ is given by the maximal eigenvector of $\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}] = \boldsymbol{A}^{-1}$, which is obtained by a few Lanczos iterations.

[9] $\mathrm{H}[P(\boldsymbol{u})] = \mathrm{E}_P[-\log P(\boldsymbol{u})]$ is the (differential) entropy, measuring the amount of uncertainty in $P(\boldsymbol{u})$. For a Gaussian, $\mathrm{H}[N(\boldsymbol{\mu},\boldsymbol{\Sigma})] = \frac{1}{2}\log|2\pi e\boldsymbol{\Sigma}|$.

In the context of Bayesian experimental design, the convexity of our variational inference relaxation (with log-concave potentials) is an important asset. In contrast to single image reconstruction, which can be tuned by the user until a desired result is obtained, sequential acquisition optimization is an autonomous process consisting of many individual steps (a real-world example is given in section 7.4), each of which requires a variational refitting $Q(\boldsymbol{u}|\boldsymbol{y}) \to Q(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{y}_*)$. Within our framework, each of these has a unique solution which is found by a very efficient algorithm. While we are not aware of Bayesian acquisition optimization being realized at comparable scales with other inference approximations, this would be difficult to do indeed. Different variational approximations are nonconvex problems coming with notorious local minima issues. For Markov chain Monte Carlo methods, there are not even reliable automatic tests of convergence. If approximate inference drives a multistep automated scheme free of human expert interventions, properties like convexity and robustness gain relevance normally overlooked in the literature.

6.1. Compressive sensing of natural images. The main application we address in section 7.4, automatic acquisition optimization for magnetic resonance imaging, is an advanced real-world instance of *compressive sensing* [6, 5]. Given that real-world images come with low entropy super-Gaussian statistics, how can we tractably reconstruct them from a sample below the Nyquist–Shannon limit? What do small successful designs $\boldsymbol{X}$ for natural images look like? Recent celebrated results about recovery properties of convex sparse estimators [10, 6, 5] have been interpreted as suggesting that up from a certain size, successful designs $\boldsymbol{X}$ may simply be drawn blindly at random. Technically speaking, these results are about highly exactly sparse signals (see section 5), yet advancements for *image* reconstruction are typically being implied [6, 5]. In contrast, Bayesian experimental design is an adaptive approach, optimizing $\boldsymbol{X}$ based on real-world training images. Our work is of the latter kind, as are [18, 32, 16] for much smaller scales.

The question of whether a design $\boldsymbol{X}$ is useful for measuring images, can (and should) be resolved empirically. Indeed, it takes no more than some reconstruction code and a range of realistic images (natural photographs, MR images) to convince oneself that MAP estimation from a subset of Fourier coefficients drawn uniformly at random (say, at 1/4 Nyquist) leads to very poor results. This failure of blindly drawn designs is well established by now both for natural images and MR images [32, 34, 19, 7], and it is not hard to motivate. In a nutshell, the assumptions which current compressive sensing theory relies upon do not sensibly describe realistic images. Marginal statistics of the latter are not exactly sparse but exhibit a power law (super-Gaussian) decay. More important, their sparsity is highly structured, a fact which is ignored in assumptions made by current compressive sensing theory, and therefore is not reflected in recovery conditions (such as incoherence) or in designs $\boldsymbol{X}$ drawn uniformly at random. Such designs fail for a number of reasons. First, they do not sample where the image energy is [32, 7]. A more subtle problem is the inherent variability of *independent* sampling in Fourier space: large gaps occur with high probability, which leads to serious MAP reconstruction errors. These points are reinforced in [32, 34]. The former study finds that for good reconstruction quality of real-world images, the choice of $\boldsymbol{X}$ is far more important than the type of reconstruction algorithm used.

In real-world imaging applications, adaptive approaches promise remedies for these prob-

lems (other proposals in this direction are [18] and [16], which, however, have not successfully been applied to real-world images). Instead of relying on simplistic signal assumptions, they learn a design $\boldsymbol{X}$ from realistic image data. Bayesian experimental design provides a general framework for adaptive design optimization, driven not by point reconstruction but by predicting information gains through posterior covariance estimates.

7. Experiments. We begin with a set of experiments designed to explore aspects and variants of our algorithms and to help us understand approximation errors. Our main application concerns the optimization of sampling trajectories in magnetic resonance imaging (MRI) sequences, with the aim of obtaining useful images faster than previously possible.

7.1. Type A versus type B bounding for Laplace potentials. Recall that the critical coupling term $\log|\boldsymbol{A}|$ in the variational criterion $\phi(\boldsymbol{\gamma})$ can be upper bounded in two different ways, referred to as type A and type B in section 4.2. Type A is tight for moderate and large γ_i, and type B is tight for small γ_i (section 5). In this section, we run our inference algorithm with type A and type B bounding, respectively, comparing the speed of convergence. The setup (SLM with Laplace potentials) is as detailed in section 7.4, with a design $\boldsymbol{X}$ of 64 phase encodes (1/4 Nyquist). Results are given in Figure 4, averaged over 7 different slices from `sg88` (256×256 pixels, $n = 131072$).

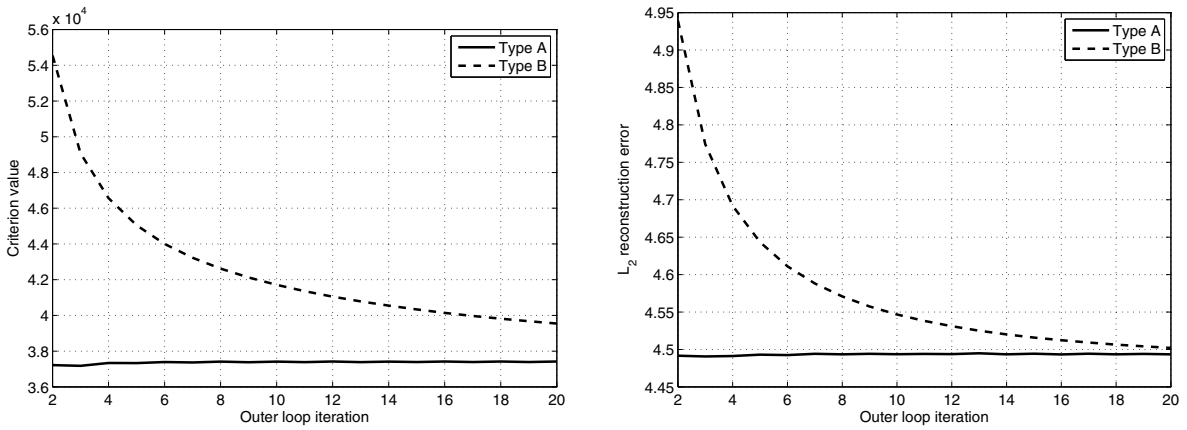

Figure 4. *Comparison of bounding types A, B for SLM with Laplace potentials. Shown are $\phi(\boldsymbol{\gamma})$ criterion values (left) and ℓ_2 errors of posterior mean estimates (not MAP, as in section 7.4), at the end of each outer loop iteration, starting from the second (right).*

In this case, the bounding type strongly influences the algorithm's progress. While two outer loop iterations suffice for convergence with type A, convergence is not attained even after 20 outer loop steps with type B. More inner loop steps are done for type A (30 in first outer loop iteration, 3–4 afterwards) than for type B (5–6 in first outer loop iteration, 3–4 afterwards). The double loop strategy, to make substantial progress with far less expensive inner loop updates, works for type A, but not for type B bounding. These results indicate that bounding type A should be preferred for SLM variational *inference*, certainly with Laplace potentials. Another indication comes from comparing inner loop penalties $h_i^*(s_i)$. For type A, $h_i^*(s_i) = \tau_i(z_{1,i} + s_i^2)^{1/2}$ is sparsity-enforcing for small $z_{1,i}$, retaining an important property

of $\phi(\boldsymbol{\gamma})$, while for type B, $h_i^*(s_i)$ does not enforce sparsity at all (see Appendix A.6).

7.2. Student's t potentials. In this section, we compare SLM variational inference with Student's t (2.3) potentials to the Laplace setup of section 7.1. Student's t potentials are not log-concave, so neither MAP estimation nor variational inference is a convex problem. Student's t potentials enforce sparsity more strongly than Laplacians do, which is often claimed to be more useful for image reconstruction. Their parameters are ν (degrees of freedom, regulating sparsity) and $\alpha = \nu/\tau$ (scale). We compare Laplace and Student's t potentials of same variance (the latter has a variance for $\nu > 2$ only): $\alpha_a = 2(\nu - 2)/\tau_a^2$, where τ_a is the Laplace parameter, α_r or α_i, respectively. The model setup is the same as in section 7.1, using slice 8 of `sg88` only. Result are given in Figure 5.

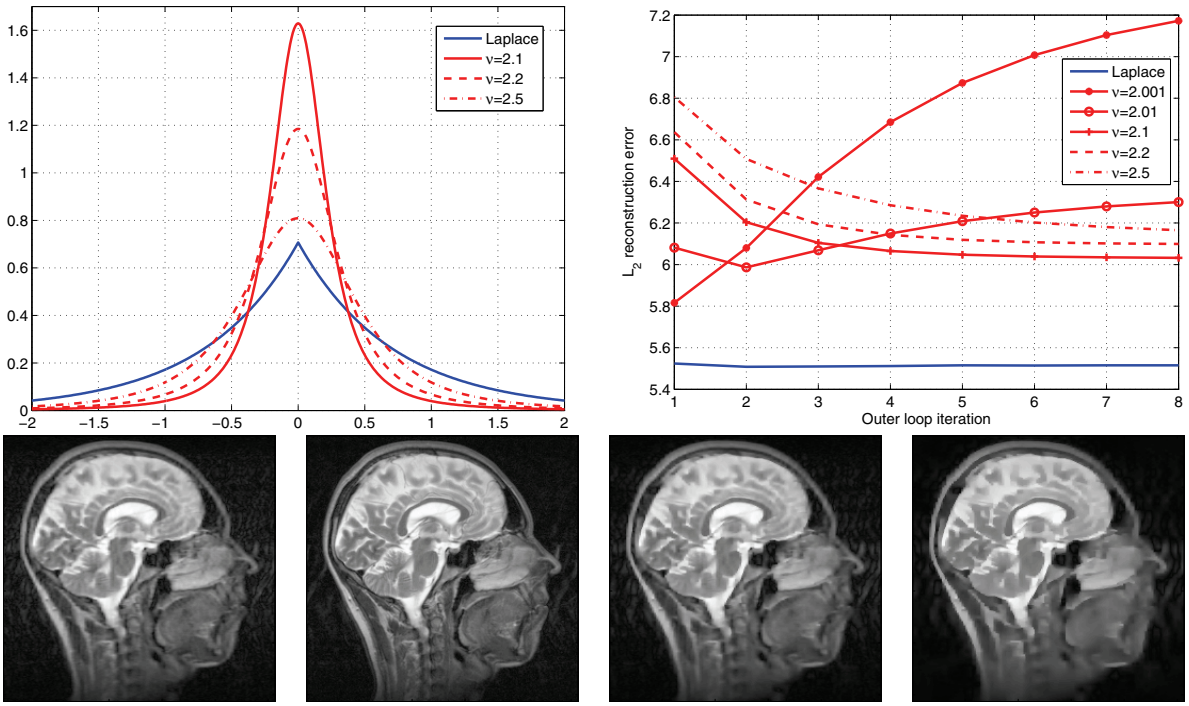

Figure 5. *Comparison of SLM with Student's t and Laplace potentials (type A bounding). Shown are potential density functions (upper left) and ℓ_2 errors of posterior mean estimates (upper right), over 8 outer loop iterations. Lower row: posterior mean reconstructions ($|u_{*,i}|$) after 8 outer loop iterations: $\nu = 2.1$; ground truth; $\nu = 2.01$; $\nu = 2.001$.*

Compared to the Laplace setup, reconstruction errors for Student's t SLMs are worse across all values of ν. While $\nu = 2.1$ outperforms larger values, the reconstruction error *grows* with iterations for $\nu = 2.01$, $\nu = 2.001$. This is not a problem of sluggish convergence: $\phi(\boldsymbol{\gamma})$ decreases rapidly[10] in this case. A glance at the mean reconstructions ($|u_{*,i}|$) (Figure 5, lower row) indicates what happens. For $\nu = 2.01$, 2.001, image sparsity is clearly enforced *too strongly*, leading to fine features being smoothed out. The reconstruction for $\nu = 2.001$

[10]For Student's t potentials (as opposed to Laplacians), type A and type B bounding behave very similarly in these experiments.

is merely a caricature of the real image complexity and is rather useless as the output of a medical imaging procedure. When it comes to *real-world image* reconstruction, more sparsity does not necessarily lead to better results.

7.3. Inaccurate Lanczos variance estimates. The difficulty of large scale Gaussian variance approximation is discussed in section 4.4. In this section, we analyse errors of the Lanczos variance approximation that we employ in our experiments. We downsampled our MRI data to 64×64 to allow for ground truth exact variance computations. The setup is the same as above (Laplacians, type A bounding), with $\boldsymbol{X}$ consisting of 30 phase encodes. Starting with a single common outer loop iteration, we compare different ways of updating $\boldsymbol{z}_1$: exact variance computations versus Lanczos approximations of different size k. Results are given in Figure 6 (upper and middle rows).

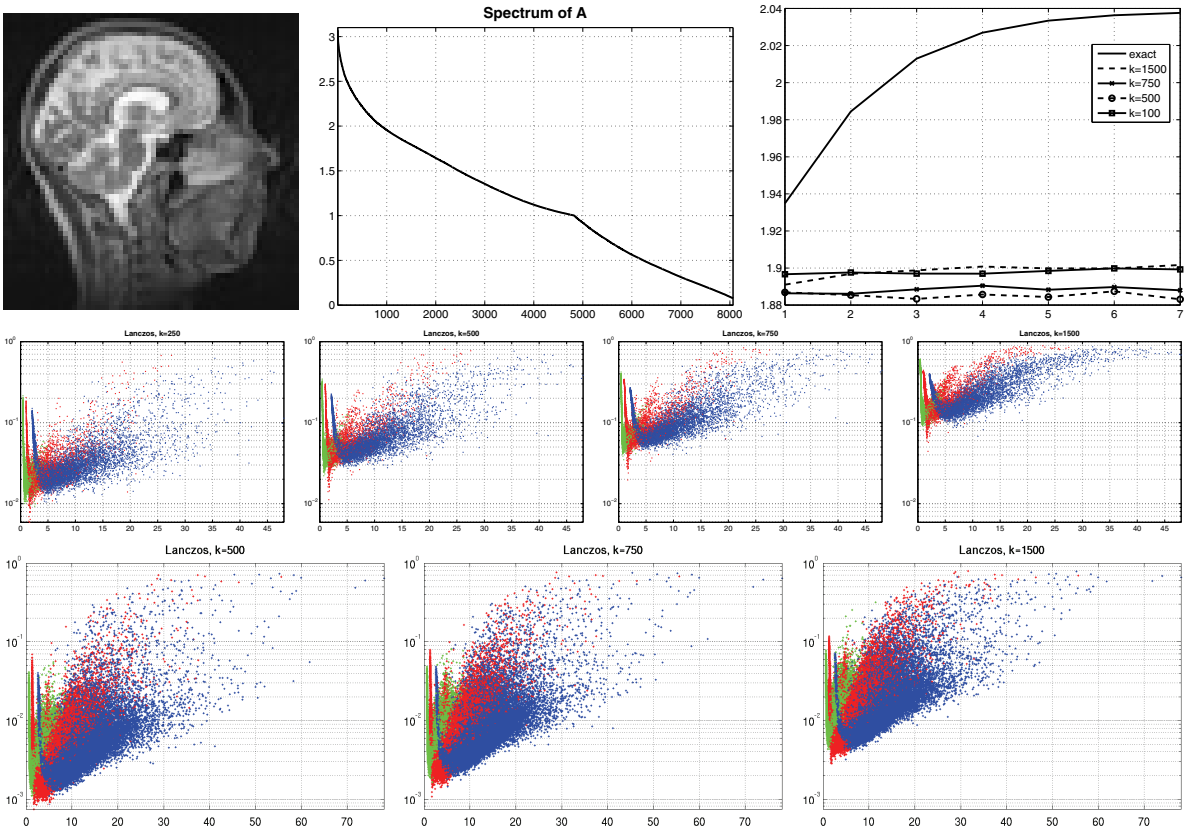

Figure 6. *Lanczos approximations of Gaussian variances, at beginning of second outer loop iteration. For 64×64 data (upper left), spectral decay of inverse covariance matrix $\boldsymbol{A}$ is roughly linear (upper middle). ℓ_2 reconstruction error of posterior mean estimate after subsequent outer loop iterations for exact variance computation versus $k = 250, 500, 750, 1500$ Lanczos steps (upper right). Middle row: relative accuracy $\hat{z}_i \mapsto \hat{z}_{k,i}/\hat{z}_i$ at beginning of second outer loop iteration, separately for "a" potentials (on wavelet coefficients; red), "r" potentials (on derivatives; blue), and "i" potentials (on $\Im(\boldsymbol{u})$; green); see section 7.4. Lower row: relative accuracy $\hat{z}_i \mapsto \hat{z}_{k,i}/\hat{z}_i$ at beginning of second outer loop iteration for full size setup (256×256), $k = 500, 750, 1500$ (ground truth $\hat{z}_i$ determined by separate LCG runs).*

The spectrum of $\boldsymbol{A}$ at the beginning of the second outer loop iteration shows a roughly linear decay. Lanczos approximation errors are rather large (middle row). Interestingly, the algorithm does not work better with exact variance computations (judged by the development of posterior mean reconstruction errors, upper right). We offer a heuristic explanation in section 4.4. A clear structure in the relative errors emerges from the middle row: the largest (and also smallest) true values $\hat{z}_i$ are approximated rather accurately, while smaller true entries are strongly damped. The role of sparsity potentials $t_i(s_i)$, or of γ_i within the variational approximation, is to shrink coefficients selectively. The structure of Lanczos variance errors serves to strengthen this effect. We repeated the relative error estimation for the full-scale setup used in the previous sections and below (256×256); ground truth values $\hat{z}_i$ were obtained by separate conjugate gradients runs. The results (shown in the lower row) exhibit the same structure, although relative errors are larger in general.

Both our experiments and our heuristic explanation are given for SLM inference, we do not expect them to generalize to other models. Within the same model and problem class, the impact of Lanczos approximations on final design outcomes is analyzed in [31]. As noted in section 4.4, understanding the real impact of Lanczos (or PCA) approximations on approximate inference and decision-making is an important topic for future research.

7.4. Sampling optimization for magnetic resonance imaging. Magnetic resonance imaging (MRI) [45] is among the most important medical imaging modalities. Without applying any harmful ionizing radiation, a wide range of parameters, from basic anatomy to blood flow, brain function, or metabolite distribution, can be visualized. Image slices are reconstructed from coefficients sampled along smooth trajectories in Fourier space (*phase encodes*). In *Cartesian MRI*, phase encodes are dense columns or rows in discrete Fourier space. The most serious limiting factor[11] is long scan time, which is proportional to the number of phase encodes acquired. MRI is a prime candidate for compressive sensing (section 6.1) in practice [19, 34]: if images of diagnostic quality can be reconstructed from an undersampled design, time is saved at no additional hardware costs or risks to patients.

In this section, we address the problem of MRI sampling optimization: which smallest subset of phase encodes results in MAP reconstructions of useful quality? To be clear, we do not use approximate Bayesian technology to improve reconstruction from fixed designs (see section 5) but aim to optimize the design $\boldsymbol{X}$ itself, so as to best support subsequent standard MAP reconstruction on real-world images. As discussed in sections 5 and 6.1, the focus for most work on compressive sensing is on the reconstruction algorithm; the question of how to choose $\boldsymbol{X}$ is typically not addressed (exceptions include [18, 16]). We follow the *adaptive* Bayesian experimental design scenario described in section 6, where $\{\boldsymbol{X}_*\}$ are phase encodes (columns in Fourier space) and $\boldsymbol{u}$ the unknown (complex-valued) image. Implementing this proposal requires the approximation of dominating posterior covariance directions for a large scale non-Gaussian SLM ($n = 131072$), which to our knowledge has not been attempted before. Results shown below are part of a larger study [34] on human brain data acquired with a Siemens 3T scanner (TSE, 23 echos/exc, 120° refocusing pulses, $1 \times 1 \times 4\,\mathrm{mm}^3$ voxels,

[11] Patient movement (blood flow, heartbeat, thorax) is strongly detrimental to image quality, which necessitates uncomfortable measures such as breath-hold or fixation. In dynamic MRI, temporal resolution is limited by scan time.

resolution 256×256). Note that for Nyquist dense acquisition, resolution is dictated by the number of phase encodes, 256 in this setting. We employ two datasets, `sg92` and `sg88`, here (sagittal orientation, echo time ≈90ms).

We use an SLM with Laplace potentials (2.2). In MRI, $\boldsymbol{u}$, $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$ are naturally complex-valued, and we make use of the group potential extension discussed in section 4.5 (coding $\mathbb{C}$ as $\mathbb{R}^2$). The vector $\boldsymbol{s}$ is composed of multiscale wavelet coefficients $\boldsymbol{s}_a$, first derivatives (horizontal and vertical) $\boldsymbol{s}_r$, and the imaginary part $\boldsymbol{s}_i = \Im(\boldsymbol{u})$. An MVM with $\boldsymbol{X}$ requires a fast Fourier transform, while an MVM with $\boldsymbol{B}$ costs only $O(n)$. Laplace scale parameters were $\tau_a = 0.07$, $\tau_r = 0.04$, $\tau_i = 0.1$). The algorithms described above were run with $n = 131072$, $q = 261632$, candidate size $d = 512$, and $m = d \cdot N_{\mathrm{col}}$, where N_{col} is the number of phase encodes in $\boldsymbol{X}$. We compare different ways of constructing designs $\boldsymbol{X}$, all of which start with the central 32 columns (lowest horizontal frequencies): Bayesian sequential optimization, with all remaining 224 columns as candidates (`op`); filling the grid from the center outwards (`ct`; such low-pass designs are typically used with *linear* MRI reconstruction); covering the grid with equidistant columns (`eq`); and drawing encodes at random (without replacement), using the variable-density sampling approach of [19] (`rd`). The latter is motivated by compressive sensing theory (see section 6.1), yet is substantially refined compared to naive independent and identically distributed sampling.[12] Results for sparse MAP reconstruction of the most difficult slice in `sg92` are shown in Figure 7 (the error metric is ℓ_2 distance $\| |\boldsymbol{u}_*| - |\boldsymbol{u}_{\mathrm{true}}| \|$, where $\boldsymbol{u}_{\mathrm{true}}$ is the complete data reconstruction).

Obtained with the same standard sparse reconstruction method (convex ℓ_1 MAP estimation), results for fixed N_{col} differ "only" in terms of the composition of $\boldsymbol{X}$ (recall that scan time grows in proportion to N_{col}). Designs chosen by our Bayesian technique substantially outperform all other choices. These results, along with [32, 34], are in stark contrast to claims that independent random sampling is a good way to choose designs for sub-Nyquist reconstruction *of real-world images*. The improvement of Bayesian optimized over randomly drawn designs is larger for smaller N_{col}. In fact, variable-density sampling does worse than conventional low-pass designs below 1/2 Nyquist. Similar findings are obtained in [32] for different natural images. In the regime far below the Nyquist limit, it is all the more important to judiciously optimize the design, using criteria informed by realistic image data in the first place.

A larger range of results is given in [34]. Even at 1/4 Nyquist, designs optimized by our method lead to images where most relevant details are preserved. In Figure 7, testing and design optimization is done on the same dataset. The generalization capability of our optimized designs is tested in this larger study, where they are applied to a range of data from different subjects, different contrasts, and different orientations, achieving improvements on these test sets comparable to what is shown in Figure 7. Finally, we have concentrated on single image slice optimization in our experiments. In realistic MRI experiments, a number of neighboring slices are acquired in an interleaved fashion. Strong statistical dependencies between slices can be exploited, both in reconstruction and joint design optimization, by combining our framework with structured graphical model message passing [30].

[12] Results for drawing *phase encodes* uniformly at random are much worse than the alternatives show, even if started with the same central 32 columns. Reconstructions become even worse when Fourier *coefficients* are drawn uniformly at random.

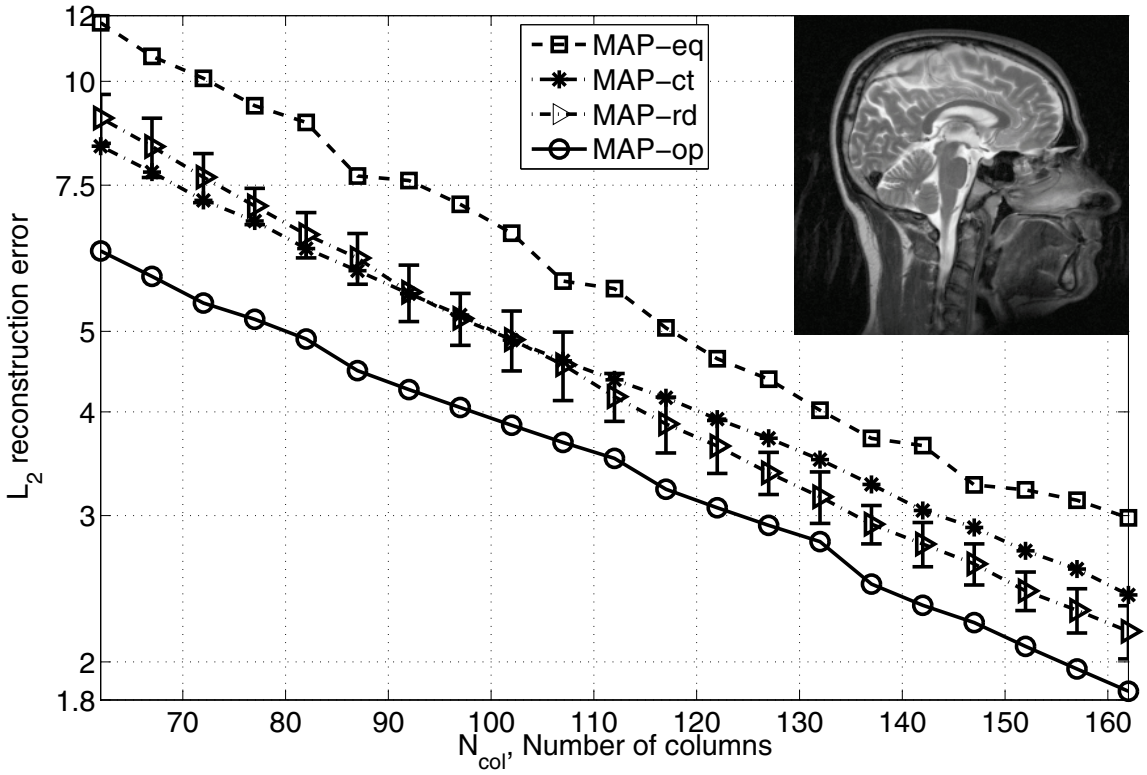

Figure 7. *Results for Cartesian undersampling with different measurement designs, on sagittal slice (TSE, $TE = 92ms$). All designs contain 32 central columns. Equispaced* [eq]; *low-pass* [ct]; *random with variable density* [rd]; *(averaged over 10 repetitions); optimized by our Bayesian technique* [op]. *Shown are ℓ_2 distances to $\boldsymbol{u}_{\text{true}}$ for MAP reconstruction with the Laplace SLM. Designs are optimized on the same data.*

8. Discussion. In this paper, we introduce scalable algorithms for approximate *Bayesian inference* in SLMs, complementing the large body of work on *point estimation* for these models. If the Bayesian posterior is not simply used as a criterion to be optimized, but as a global picture of uncertainty in a reconstruction problem, advanced decision-making problems such as model calibration, feature relevance ranking, or Bayesian experimental design can be addressed. We settle a longstanding question for continuous-variable variational Bayesian inference, proving that the relaxation of interest here [17, 23, 11] has the same convexity profile as MAP estimation. Our double loop algorithms are scalable by reduction to common computational problems: penalized least squares optimization and Gaussian covariance estimation (or PCA). The large and growing body of work for the latter, both in theory and algorithms, is put to novel use in our methods. Moreover, the reductions offer valuable insight into similarities and differences between sparse estimation and approximate Bayesian inference, as does our focus on decision-making problems beyond point reconstruction.

We apply our algorithms to the design optimization problem of improving sampling trajectories for MRI. To the best of our knowledge, this has not been attempted before in the context of sparse nonlinear reconstruction. Ours is the first approximate Bayesian framework

for adaptive compressive sensing that scales up to and succeeds on full high-resolution real-world images. Results here are part of a larger MRI study [34], where designs optimized by our Bayesian technique are found to significantly and robustly improve sparse image reconstruction on a wide range of test datasets for measurements far below the Nyquist limit.

In future work, we will address advanced joint design scenarios, such as MRI sampling optimization for multiple image slices, three-dimensional MRI, and parallel MRI with array coils. Our technique can be sped up along many directions, from algorithmic improvements (advanced algorithms for inner loop optimization, modern Lanczos variants) to parallel computation on graphics hardware. An important future goal, currently out of reach, is supporting real-time MRI applications by automatic on-line sampling optimization.

Appendix. Details and proofs.

A.1. Proof of Theorem 3.1. In this section, we provide a proof of Theorem 3.1, whose statement is reproduced for convenience. *Let* $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, $\boldsymbol{B} \in \mathbb{R}^{q\times n}$ *be arbitrary matrices, and let*

$$\tilde{\boldsymbol{A}}(\boldsymbol{d}) := \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T(\operatorname{diag}\boldsymbol{d})\boldsymbol{B}, \quad \boldsymbol{d} \succ \boldsymbol{0},$$

so that $\tilde{\boldsymbol{A}}(\boldsymbol{d})$ *is positive definite for all* $\boldsymbol{d} \succ \boldsymbol{0}$.

(1) *Let* $f_i(\gamma_i)$ *be twice continuously differentiable functions into* $\mathbb{R}_+$ *so that* $\log f_i(\gamma_i)$ *are convex for all* i *and* γ_i. *Then,* $\boldsymbol{\gamma} \mapsto \log|\tilde{\boldsymbol{A}}(f(\boldsymbol{\gamma}))|$ *is convex. Especially,* $\boldsymbol{\gamma} \mapsto \log|\boldsymbol{A}|$ *is convex.*

(2) *Let* $f_i(\pi_i) \geq 0$ *be concave functions. Then,* $\boldsymbol{\pi} \mapsto \log|\tilde{\boldsymbol{A}}(f(\boldsymbol{\pi}))|$ *is concave. Especially,* $\boldsymbol{\gamma}^{-1} \mapsto \log|\boldsymbol{A}|$ *is concave.*

(3) *Let* $f_i(\gamma_i) \geq 0$ *be concave functions. Then,* $\boldsymbol{\gamma} \mapsto \boldsymbol{1}^T(\log f(\boldsymbol{\gamma})) + \log|\tilde{\boldsymbol{A}}(f(\boldsymbol{\gamma})^{-1})|$ *is concave. Especially,* $\boldsymbol{\gamma} \mapsto \boldsymbol{1}^T(\log\boldsymbol{\gamma}) + \log|\boldsymbol{A}|$ *is concave.*

(4) *Let* $Q(\boldsymbol{u}|\boldsymbol{y})$ *be the approximate posterior with covariance matrix given by* (3.1). *Then, for all* i, $\operatorname{Var}_Q[s_i|\boldsymbol{y}] = \boldsymbol{\delta}_i^T\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T\boldsymbol{\delta}_i \leq \gamma_i$.

For notational convenience, we absorb σ^{-2} into $\boldsymbol{X}^T\boldsymbol{X}$ by replacing $\boldsymbol{X}$ by $\sigma^{-1}\boldsymbol{X}$. We begin with part (2). It is well known that $\boldsymbol{\pi} \mapsto \log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})|$ is concave and nondecreasing for $\boldsymbol{\pi} \succ \boldsymbol{0}$ [4, sect. 3.1.5]. Both properties carry over to the extended-value function.[13] The statement follows from the concatenation rules of [4, sect. 3.2.4].

We continue with part (1). Write $\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{A}}(f(\boldsymbol{\gamma}))$, $\psi_1 := \log|\tilde{\boldsymbol{A}}|$, $\boldsymbol{\Gamma} = \operatorname{diag}\boldsymbol{\gamma}$, and $f(\boldsymbol{\Gamma}) = \operatorname{diag} f(\boldsymbol{\gamma})$. First, $\boldsymbol{\gamma} \mapsto \psi_1$ is the composition of twice continuously differentiable mappings, therefore twice continuously differentiable itself. Now, $d\psi_1 = \operatorname{tr}\boldsymbol{S}f'(\boldsymbol{\Gamma})(d\boldsymbol{\Gamma})$, where $\boldsymbol{S} := \boldsymbol{B}\tilde{\boldsymbol{A}}^{-1}\boldsymbol{B}^T$; moreover, $d^2\psi_1 = -\operatorname{tr}\boldsymbol{S}f'(\boldsymbol{\Gamma})(d\boldsymbol{\Gamma})\boldsymbol{S}f'(\boldsymbol{\Gamma})(d\boldsymbol{\Gamma}) + \operatorname{tr}\boldsymbol{S}f''(\boldsymbol{\Gamma})(d\boldsymbol{\Gamma})^2 = \operatorname{tr}(d\boldsymbol{\Gamma})\boldsymbol{S}(d\boldsymbol{\Gamma})\boldsymbol{E}_1$, where $\boldsymbol{E}_1 := f''(\boldsymbol{\Gamma}) - f'(\boldsymbol{\Gamma})\boldsymbol{S}f'(\boldsymbol{\Gamma})$. Since $\boldsymbol{S} \succeq \boldsymbol{0}$, we have $\boldsymbol{S} = \boldsymbol{V}\boldsymbol{V}^T$ for some matrix $\boldsymbol{V}$, and $d^2\psi_1 = \operatorname{tr}((d\boldsymbol{\Gamma})\boldsymbol{V})^T\boldsymbol{E}_1(d\boldsymbol{\Gamma})\boldsymbol{V}$. Now, if $\boldsymbol{E}_1 \succeq \boldsymbol{0}$, then for $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma} + t(\Delta\boldsymbol{\gamma})$, we have $\psi_1''(0) = \operatorname{tr}\boldsymbol{N}^T\boldsymbol{E}_1\boldsymbol{N} \geq 0$ for any $\Delta\boldsymbol{\gamma}$, where $\boldsymbol{N} := (\operatorname{diag}\Delta\boldsymbol{\gamma})\boldsymbol{V}$, so that ψ_1 is convex.

The log-convexity of $f_i(\gamma_i)$ implies that $f_i(\gamma_i)f_i''(\gamma_i) \geq (f_i'(\gamma_i))^2$ for all γ_i so that

$$\begin{aligned}\boldsymbol{E}_1 &= f(\boldsymbol{\Gamma})^{-1}(f(\boldsymbol{\Gamma})f''(\boldsymbol{\Gamma})) - f'(\boldsymbol{\Gamma})\boldsymbol{S}f'(\boldsymbol{\Gamma}) \succeq f(\boldsymbol{\Gamma})^{-1}(f'(\boldsymbol{\Gamma}))^2 - f'(\boldsymbol{\Gamma})\boldsymbol{S}f'(\boldsymbol{\Gamma}) \\ &= f'(\boldsymbol{\Gamma})\left(f(\boldsymbol{\Gamma})^{-1} - \boldsymbol{S}\right)f'(\boldsymbol{\Gamma}).\end{aligned}$$

[13] In general, we extend convex continuous functions $f(\boldsymbol{\pi})$ on $\boldsymbol{\pi} \succ \boldsymbol{0}$ by $f(\boldsymbol{\pi}) = \lim_{\boldsymbol{d}\searrow\boldsymbol{\pi}} f(\boldsymbol{d})$, $\boldsymbol{\pi} \succeq \boldsymbol{0}$, and by $f(\boldsymbol{\pi}) = \infty$ elsewhere.

Therefore, it remains to show that $f(\boldsymbol{\Gamma})^{-1} - \boldsymbol{S} \succeq \boldsymbol{0}$. We use the identity

$$\boldsymbol{v}^T \boldsymbol{M}^{-1} \boldsymbol{v} = \max_{\boldsymbol{x}} 2\boldsymbol{v}^T \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x}, \tag{A.1}$$

which holds whenever $\boldsymbol{M} \succ \boldsymbol{0}$. For any $\boldsymbol{r} \in \mathbb{R}^q$,

$$\boldsymbol{r}^T \boldsymbol{B} \tilde{\boldsymbol{A}}^{-1} \boldsymbol{B}^T \boldsymbol{r} = \max_{\boldsymbol{x}} 2\boldsymbol{r}^T \boldsymbol{B}\boldsymbol{x} - \boldsymbol{x}^T \left(\boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{B}^T f(\boldsymbol{\Gamma}) \boldsymbol{B} \right) \boldsymbol{x} \le \max_{\boldsymbol{k} = \boldsymbol{B}\boldsymbol{x}} 2\boldsymbol{r}^T \boldsymbol{k} - \boldsymbol{k}^T f(\boldsymbol{\Gamma}) \boldsymbol{k},$$

using (A.1) and $\boldsymbol{x}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{x} = \|\boldsymbol{X}\boldsymbol{x}\|^2 \ge 0$. Therefore, $\boldsymbol{r}^T \boldsymbol{S} \boldsymbol{r} \le \max_{\boldsymbol{k}} 2\boldsymbol{r}^T \boldsymbol{k} - \boldsymbol{k}^T f(\boldsymbol{\Gamma}) \boldsymbol{k} = \boldsymbol{r}^T f(\boldsymbol{\Gamma})^{-1} \boldsymbol{r}$, using (A.1) once more, which implies $f(\boldsymbol{\Gamma})^{-1} - \boldsymbol{S} \succeq \boldsymbol{0}$. This completes the proof of part (1). Since $\boldsymbol{A} = \tilde{\boldsymbol{A}}(\boldsymbol{\gamma}^{-1})$, we can employ this argument with $f_i(\gamma_i) = \gamma_i^{-1}$ and $\boldsymbol{r} = \boldsymbol{\delta}_i$ in order to establish part (4).

We continue with part (3). Write $\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{A}}(f(\boldsymbol{\gamma})^{-1})$ and $\psi_2 := \boldsymbol{1}^T (\log f(\boldsymbol{\gamma})) + \log |\tilde{\boldsymbol{A}}|$. Assume for now that $\boldsymbol{X}^T \boldsymbol{X} \succ \boldsymbol{0}$. Let $\boldsymbol{B} = (\boldsymbol{B}_{<q}^T \tilde{\boldsymbol{b}})^T$ (so that $\tilde{\boldsymbol{b}}^T$ is the last row of $\boldsymbol{B}$), and define $\tilde{\boldsymbol{A}}_{<q} = \boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{B}_{<q}^T f(\boldsymbol{\Gamma}_{<q})^{-1} \boldsymbol{B}_{<q}$, where $f(\boldsymbol{\gamma}_{<q}) = (f_i(\gamma_i))_{i<q} \in \mathbb{R}_+^{q-1}$. We make use of the well-known determinant identity $|\boldsymbol{I} + \boldsymbol{v}\boldsymbol{v}^T| = 1 + \boldsymbol{v}^T \boldsymbol{v}$. Namely,

$$\begin{aligned}
&\log f_q(\gamma_q) + \log \left| \tilde{\boldsymbol{A}}_{<q} + f_q(\gamma_q)^{-1} \tilde{\boldsymbol{b}} \tilde{\boldsymbol{b}}^T \right| \\
&= \log f_q(\gamma_q) + \log \left| \tilde{\boldsymbol{A}}_{<q} \right| + \log \left| \boldsymbol{I} + f_q(\gamma_q)^{-1} (\tilde{\boldsymbol{A}}_{<q})^{-1} \tilde{\boldsymbol{b}} \tilde{\boldsymbol{b}}^T \right| \\
&= \log \left| \tilde{\boldsymbol{A}}_{<q} \right| + \log f_q(\gamma_q) + \log \left(1 + f_q(\gamma_q)^{-1} \tilde{\boldsymbol{b}}^T (\tilde{\boldsymbol{A}}_{<q})^{-1} \tilde{\boldsymbol{b}} \right) \\
&= \log \left| \tilde{\boldsymbol{A}}_{<q} \right| + \log \left(f_q(\gamma_q) + \tilde{\boldsymbol{b}}^T (\tilde{\boldsymbol{A}}_{<q})^{-1} \tilde{\boldsymbol{b}} \right).
\end{aligned} \tag{A.2}$$

Since the extended-value function $\log(\cdot)$ (assigning $-\infty$ to arguments ≤ 0) is concave and nondecreasing, the concatenation rules of [4, sect. 3.2.4] imply the concavity of the final term in (A.2) whenever $f_q(\gamma_q) + \tilde{\boldsymbol{b}}^T (\tilde{\boldsymbol{A}}_{<q})^{-1} \tilde{\boldsymbol{b}}$ is concave. We will use induction on q, the dimensionality of $\boldsymbol{\gamma}$. For $q = 1$, ψ_2 is given by (A.2) with $\tilde{\boldsymbol{A}}_{<1} = \boldsymbol{X}^T \boldsymbol{X}$, and its concavity follows from the concavity of $f_1(\gamma_1)$. For $q > 1$, (A.2) implies

$$\psi_2 = \boldsymbol{1}^T (\log f(\boldsymbol{\gamma}_{<q})) + \log |\tilde{\boldsymbol{A}}_{<q}| + \log \left(f_q(\gamma_q) + \tilde{\boldsymbol{b}}^T (\tilde{\boldsymbol{A}}_{<q})^{-1} \tilde{\boldsymbol{b}} \right).$$

Both the sum of the first two terms and $f_q(\gamma_q)$ are concave by assumption, so that the concavity of ψ_2 is implied by the concavity of $\boldsymbol{\gamma} \mapsto \tilde{\boldsymbol{b}}^T (\tilde{\boldsymbol{A}}_{<q})^{-1} \tilde{\boldsymbol{b}}$. Using (A.1), we have

$$\tilde{\boldsymbol{b}}^T (\tilde{\boldsymbol{A}}_{<q})^{-1} \tilde{\boldsymbol{b}} = \max_{\boldsymbol{x}} 2\tilde{\boldsymbol{b}}^T \boldsymbol{x} - \boldsymbol{x}^T \tilde{\boldsymbol{A}}_{<q} \boldsymbol{x} = \max_{\boldsymbol{x}} 2\tilde{\boldsymbol{b}}^T \boldsymbol{x} - \|\boldsymbol{X}\boldsymbol{x}\|^2 - \boldsymbol{v}^T f(\boldsymbol{\Gamma}_{<q})^{-1} \boldsymbol{v}$$

with $\boldsymbol{v} := \boldsymbol{B}_{<q} \boldsymbol{x}$. Now, $(\boldsymbol{x}, \boldsymbol{f}) \mapsto 2\tilde{\boldsymbol{b}}^T \boldsymbol{x} - \|\boldsymbol{X}\boldsymbol{x}\|^2 - \boldsymbol{v}^T (\mathrm{diag}\, \boldsymbol{f})^{-1} \boldsymbol{v}$ is jointly concave for $\boldsymbol{f} \succ \boldsymbol{0}$ (see proof of Theorem 3.2), so that $\kappa(\boldsymbol{f}) := \tilde{\boldsymbol{b}}^T \tilde{\boldsymbol{A}}_{<q} (\boldsymbol{f}^{-1})^{-1} \tilde{\boldsymbol{b}}$ is concave for $\boldsymbol{f} \succ \boldsymbol{0}$ [4, sect. 3.2.5] (recall that $\tilde{\boldsymbol{A}}_{<q}(\boldsymbol{f}^{-1}) = \boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{B}_{<q}^T (\mathrm{diag}\, \boldsymbol{f}^{-1}) \boldsymbol{B}_{<q}$). To finish the argument, we plug in $\boldsymbol{f} := f(\boldsymbol{\gamma}_{<q})$ and use the concatenation theorems of [4, sect. 3.2.4]. What remains to be shown in this context is that $\kappa(\boldsymbol{f})$ is nondecreasing in each argument. Pick any $i \in \{1, \ldots, q\}$, $\boldsymbol{f} \succ \boldsymbol{0}$, and any $\Delta > 0$. Then,

$$\kappa(\boldsymbol{f} + \Delta \boldsymbol{\delta}_i) = \tilde{\boldsymbol{b}}^T \left(\tilde{\boldsymbol{A}}_{<q}(\boldsymbol{f}^{-1}) - \frac{\Delta}{f_i (f_i + \Delta)} \boldsymbol{b}_i \boldsymbol{b}_i^T \right)^{-1} \tilde{\boldsymbol{b}} \ge \tilde{\boldsymbol{b}}^T \tilde{\boldsymbol{A}}_{<q} (\boldsymbol{f}^{-1})^{-1} \tilde{\boldsymbol{b}} = \kappa(\boldsymbol{f}),$$

where $\boldsymbol{b}_i = \boldsymbol{B}^T\boldsymbol{\delta}_i$. This concludes the proof of part (3), under the assumption that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible. If $\boldsymbol{X}^T\boldsymbol{X}$ is singular, define ψ_2^ε as above, but with $\boldsymbol{X}^T\boldsymbol{X} \to \boldsymbol{X}^T\boldsymbol{X} + \varepsilon\boldsymbol{I}$. We saw that ψ_2^ε is concave for any $\varepsilon > 0$. For any $\boldsymbol{\gamma} \succ \boldsymbol{0}$ such that $\psi_2(\boldsymbol{\gamma}) > -\infty$, ψ_2^ε converges uniformly to ψ_2 on a closed environment of $\boldsymbol{\gamma}$ (ψ_2 and all ψ_2^ε are continuous), so that ψ_2 is concave at $\boldsymbol{\gamma}$. This completes the proof of part (3).

A.2. Proof of Theorem 3.3. In this section, we provide the proof of Theorem 3.3, whose statement is reproduced for convenience. *Consider a model with Gaussian likelihood* (2.1) *and a prior* $P(\boldsymbol{u}) \propto \prod_{i=1}^q t_i(s_i)$, $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$, *so that all* $t_i(s_i)$ *are strongly super-Gaussian, meaning that* $\tilde{g}_i(s_i) = \log t_i(s_i) - b_i s_i$ *is even, and* $g_i(x_i) = \tilde{g}_i(x_i^{1/2})$ *is strictly convex and decreasing for* $x_i > 0$.

(1) *If* $\tilde{g}_i(s_i)$ *is concave and twice continuously differentiable for* $s_i > 0$, *then* $h_i(\gamma_i)$ *is convex. On the other hand, if* $\tilde{g}_i''(s_i) > 0$ *for some* $s_i > 0$, *then* $h_i(\gamma_i)$ *is not convex at some* $\gamma_i > 0$.

(2) *If all* $\tilde{g}_i(s_i)$ *are concave and twice continuously differentiable for* $s_i > 0$, *then the variational problem* $\min_{\boldsymbol{\gamma}\geq\boldsymbol{0}} \phi$ *is a convex optimization problem. On the other hand, if* $\tilde{g}_i''(s_i) > 0$ *for some* i *and* $s_i > 0$, *then* $h(\boldsymbol{\gamma})$ *is not convex, and there exist some* $\boldsymbol{X}$, $\boldsymbol{B}$, *and* $\boldsymbol{y}$ *such that* $\phi(\boldsymbol{\gamma})$ *is not a convex function.*

We begin with part (1), focusing on a single potential i and dropping its index. Since $b \neq 0$ is dealt with separately, assume that $t(s)$ is even: $\log t(s) = \tilde{g}(s) = g(s^2) = g(x)$, where $x = s^2$. If $f(x,\gamma) := -x/\gamma - 2g(x)$ and $\tilde{f}(s,\gamma) := -s^2/\gamma - 2\tilde{g}(s)$, then $h(\gamma) = \max_{x\geq 0} f(x,\gamma)$, and $f(x,\gamma) = \tilde{f}(x^{1/2},\gamma)$. It suffices to consider $s \geq 0$. Denote $x_* = x_*(\gamma) := \operatorname{argmax}_{x\geq 0} f(x,\gamma)$ (which is unique, since $g(x)$ is strictly convex). If $\gamma_0 := \sup\{\gamma \,|\, f(x,\gamma) \leq -2g(0) \text{ for all } x\}$ ($\gamma_0 = 0$ for an empty set), then

- $x_* = 0$, $h(\gamma) = -2g(0)$ for $\gamma \in (0,\gamma_0]$;
- $x_* > 0$, $h(\gamma)$ is strictly increasing for $\gamma > \gamma_0$.

Namely, if $\gamma_0 < \gamma_1 < \gamma_2$, then $x_*(\gamma_1) > 0$ by definition of γ_0, and $h(\gamma_1) = -x_*(\gamma_1)/\gamma_1 - 2g(x_*(\gamma_1)) < -x_*(\gamma_1)/\gamma_2 - 2g(x_*(\gamma_1)) \leq -x_*(\gamma_2)/\gamma_2 - 2g(x_*(\gamma_2)) = h(\gamma_2)$. Note that $\gamma_0 > 0$ if and only if $\lim_{\varepsilon\searrow 0} g'(\varepsilon)$ is finite. It suffices to show that h is convex at all $\gamma > \gamma_0$, where $x_* = s_*^2 > 0$.

We use the notation $\tilde{f}_s = \partial\tilde{f}/(\partial s)$, and functions are evaluated at $(x_* = s_*^2, \gamma)$ unless otherwise noted. Now, $\tilde{f}_s = -2s_*/\gamma - 2\tilde{g}_s(s_*) = 0$ so that $\tilde{g}_s(s_*) = -s_*/\gamma$. Next, $g(x)$ is twice continuously differentiable, and $x_* = s_*^2$ at γ. Therefore, $f_x = \partial f/(\partial x)$ is continuously differentiable. Moreover, $g_{x,x}(x) > 0$ by the strict convexity of $g(x)$. By the implicit function theorem, $x_*(\gamma)$ is continuously differentiable at γ, and since $h(\gamma) = f(x_*(\gamma),\gamma)$, $h'(\gamma)$ exists. Moreover, $0 = (d/d\gamma)f_x(x_*(\gamma),\gamma) = f_{x,\gamma} + f_{x,x}\cdot(dx_*)/(d\gamma)$ so that $(dx_*)/(d\gamma) = \gamma^{-2}/(2g_{x,x}(x_*)) > 0$: $x_*(\gamma)$ is increasing. From $f_x = 0$, we have that $h'(\gamma) = f_\gamma = s_*^2/\gamma^2 = (\tilde{g}_s(s_*))^2$, since $\tilde{g}_s(s_*) = -s_*/\gamma$. Now, $\tilde{g}_s(s)$ is nonincreasing by the concavity of $\tilde{g}(s)$, and $\tilde{g}_s(s_*) < 0$ so that $s_* \mapsto h'(\gamma)$ is nondecreasing. Since $s_*^2 = x_*$ is increasing in γ, so is s_*. Therefore, $\gamma \mapsto h'(\gamma)$ is nondecreasing, which means that $h(\gamma)$ is convex for $\gamma > \gamma_0$.

The concavity of $\tilde{g}(s)$ is necessary. Suppose that $\tilde{g}_{s,s}(\tilde{s}) > 0$ for some $\tilde{s} > 0$. If $\tilde{x} = \tilde{s}^{1/2}$, $g(x)$ is differentiable at $\tilde{x}$, and if $\tilde{\gamma} = -1/(2g'(\tilde{x}))$, then $s_*(\tilde{\gamma}) = \tilde{s}$. But if $\tilde{g}_{s,s}(s_*) > 0$ at $\tilde{\gamma}$, then $s_* \mapsto h'(\gamma)$ is decreasing at $s_* = \tilde{s}$, and, just as above, $\gamma \mapsto h'(\gamma)$ is decreasing at $\tilde{\gamma}$, so that h is not convex at $\tilde{\gamma}$. This concludes the proof of part (1).

Part (2) is a direct consequence of part (1) and Theorem 3.2. For the final statement, suppose that $h_i'(\gamma_i)$ is decreasing at $\gamma_i = \tilde{\gamma}_i$. Pick the other coefficients in $\tilde{\boldsymbol{\gamma}} \succ \mathbf{0}$ arbitrary, and choose $m = n = 1$, $\boldsymbol{y} = \mathbf{0}$, $\boldsymbol{X} = X$, $\boldsymbol{B} = \boldsymbol{\delta}_i$ so that $\phi(\boldsymbol{\gamma}) - h(\boldsymbol{\gamma}) = r(\gamma_i) := \log(1+X^2\gamma_i) - \log\gamma_i$, ignoring additive constants. Consider $\tilde{\phi}(t) = \phi(\tilde{\boldsymbol{\gamma}} + t\boldsymbol{\delta}_i)$. Since $r'(\tilde{\gamma}_i) = X^2/(1+X^2\tilde{\gamma}_i) - 1/\tilde{\gamma}_i \to 0$ for $X \to \infty$, $\tilde{\phi}'(t)$ is decreasing at $t = 0$ for large enough X, and ϕ is not convex at $\tilde{\boldsymbol{\gamma}}$.

A.3. Proof of Theorem 5.1. In this section, we provides proofs related to section 5. We begin with Theorem 5.1, whose statement is reproduced for convenience. *Let* $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, $\boldsymbol{B} \in \mathbb{R}^{q\times n}$ *be matrices such that* $\tilde{\boldsymbol{A}}(\boldsymbol{\pi}) = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T(\operatorname{diag}\boldsymbol{\pi})\boldsymbol{B} \succ \mathbf{0}$ *for each* $\boldsymbol{\pi} \succ \mathbf{0}$ *and no row of* $\boldsymbol{B}$ *is equal to* $\mathbf{0}^T$.

- *The function* $\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})|$ *is increasing in each component* π_i*, and is unbounded above. For any sequence* $\boldsymbol{\pi}_t$ *with* $\|\boldsymbol{\pi}_t\| \to \infty$ *(*$t \to \infty$*) and* $\boldsymbol{\pi}_t \succeq \varepsilon\mathbf{1}$ *for some* $\varepsilon > 0$*, we have that* $\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi}_t)| \to \infty$ *(*$t \to \infty$*).*
- *Assume that* $\log P(\boldsymbol{u}|\boldsymbol{y})$ *is bounded above as a function of* $\boldsymbol{u}$*. Recall the variational criterion* $\phi(\boldsymbol{\gamma})$ *from* (3.2)*. For any bounded sequence* $\boldsymbol{\gamma}_t$ *with* $(\boldsymbol{\gamma}_t)_i \to 0$ *(*$t \to \infty$*) for some* $i \in \{1,\dots,q\}$*, we have that* $\phi(\boldsymbol{\gamma}_t) \to \infty$.
  In particular, any local minimum point $\boldsymbol{\gamma}_*$ *of the variational problem* $\min_{\boldsymbol{\gamma}\succeq\mathbf{0}}\phi(\boldsymbol{\gamma})$ *must have positive components; i.e.,* $\boldsymbol{\gamma}_* \succ \mathbf{0}$.

For the first part, fix any $\boldsymbol{\pi} \succ \mathbf{0}$, any $i \in \{1,\dots,q\}$, and any $\Delta > 0$. If $\boldsymbol{b}_i = \boldsymbol{B}^T\boldsymbol{\delta}_i \neq \mathbf{0}$, then, using the determinant identity previously employed in Appendix A.1, we have

$$\begin{aligned}\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi}+\Delta\boldsymbol{\delta}_i)| &= \log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})| + \log\left|\boldsymbol{I} + \Delta\tilde{\boldsymbol{A}}(\boldsymbol{\pi})^{-1}\boldsymbol{b}_i\boldsymbol{b}_i^T\right| \\ &= \log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})| + \log(1+\Delta\boldsymbol{b}_i^T\tilde{\boldsymbol{A}}(\boldsymbol{\pi})^{-1}\boldsymbol{b}_i) > \log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})|,\end{aligned}$$

since $\boldsymbol{b}_i^T\tilde{\boldsymbol{A}}(\boldsymbol{\pi})^{-1}\boldsymbol{b}_i > 0$ and $\log(1+x) > 0$ for $x > 0$. Therefore, $\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})|$ is increasing in each component. Moreover, we have that $\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi}+\Delta\boldsymbol{\delta}_i)| \to \infty$ ($\Delta \to \infty$), since $\log(1+x)$ is unbounded above for $x \to \infty$. If $\boldsymbol{\pi}_t$ is a sequence with $\|\boldsymbol{\pi}_t\| \to \infty$ and $\boldsymbol{\pi}_t \succeq \varepsilon\mathbf{1}$, there must be some $i \in \{1,\dots,q\}$ such that $(\boldsymbol{\pi}_t)_i \to \infty$. If $\tilde{\boldsymbol{\pi}}_t := \varepsilon\mathbf{1} + ((\boldsymbol{\pi}_t)_i - \varepsilon)\boldsymbol{\delta}_i \succeq \boldsymbol{\pi}_t$, then $\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi}_t)| \geq \log|\tilde{\boldsymbol{A}}(\tilde{\boldsymbol{\pi}}_t)| \to \infty$ ($t \to \infty$).

For the second part, recall that

$$\begin{aligned}\phi(\boldsymbol{\gamma}) &= \log|\boldsymbol{A}| + h(\boldsymbol{\gamma}) + \min_{\boldsymbol{u}}\left\{R(\boldsymbol{u},\boldsymbol{\gamma}) = \sigma^{-2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{u}\|^2 + \boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s} - 2\boldsymbol{b}^T\boldsymbol{s}\right\}, \\ -2\log P(\boldsymbol{u}|\boldsymbol{y}) &= \min_{\boldsymbol{\gamma}\succeq\mathbf{0}} h(\boldsymbol{\gamma}) + R(\boldsymbol{u},\boldsymbol{\gamma}) + C_2, \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u},\end{aligned}$$

for some constant C_2. If $\boldsymbol{\gamma}_t$ is a bounded sequence such that $(\boldsymbol{\gamma}_t)_i \to 0$ ($t \to \infty$) for some $i \in \{1,\dots,q\}$, then $\log|\boldsymbol{A}(\boldsymbol{\gamma}_t)| = \log|\tilde{\boldsymbol{A}}(\boldsymbol{\gamma}_t^{-1})| \to \infty$. Suppose that $\phi(\boldsymbol{\gamma}_t)$ remains bounded above. Let $\boldsymbol{u}_t = \operatorname{argmin}_{\boldsymbol{u}} R(\boldsymbol{u},\boldsymbol{\gamma}_t)$. Then, $\phi(\boldsymbol{\gamma}_t) - \log|\boldsymbol{A}(\boldsymbol{\gamma}_t)| = h(\boldsymbol{\gamma}_t) + R(\boldsymbol{u}_t,\boldsymbol{\gamma}_t) \to -\infty$, so that $-2\log P(\boldsymbol{u}_t|\boldsymbol{y}) - C_2 \leq h(\boldsymbol{\gamma}_t) + R(\boldsymbol{u}_t,\boldsymbol{\gamma}_t) \to -\infty$, in contradiction to the boundedness of the log posterior. This concludes the proof.

Next, assume we run MAP estimation (3.3) with even super-Gaussian potentials $t_i(s_i)$, so that $|s_i| \mapsto -\log t_i(s_i) = -\tilde{g}_i(s_i)$ is concave. As argued in section 5, any local minimum point $\boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$ is exactly sparse. We show that the corresponding $\boldsymbol{\gamma}_*$ has the same sparsity pattern: $\gamma_{*,i} = 0$ whenever $s_{*,i} = 0$. Dropping the index, since $\gamma_* \in \operatorname{argmin}_{\gamma\geq 0} s_*^2/\gamma + h(\gamma)$, we have to show that $h(\gamma) > h(0)$ for all $\gamma > 0$ (or, in terms of Appendix A.2, that $\gamma_0 = 0$).

Fix $\gamma > 0$, and recall that $h(\gamma) = \max_{s\geq 0}\{\tilde{f}(s,\gamma) = -s^2/\gamma - 2\tilde{g}(s)\} \geq -2\tilde{g}(0) = h(0)$. Now, $\tilde{f}(s,\gamma) = -2\tilde{g}(s) + O(s^2)$, $s \searrow 0$, where $-2\tilde{g}(s)$ is concave, nondecreasing, and not constant. Therefore, $\lim_{s\searrow 0} \partial\tilde{f}/(\partial s) \in (0,\infty]$, and $\tilde{f}(\tilde{s},\gamma) > \tilde{f}(0,\gamma)$ for some $\tilde{s} > 0$, so that $h(\gamma) \geq \tilde{f}(\tilde{s},\gamma) > h(0)$.

A.4. Details for bounding $\log|A|$. In this section, we provide details concerning the $\log|\boldsymbol{A}|$ bounds discussed in section 4.2. Recall that $\tilde{\boldsymbol{A}}(\boldsymbol{\pi}) = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T(\operatorname{diag}\boldsymbol{\pi})\boldsymbol{B}$ for $\boldsymbol{\pi} \succ \boldsymbol{0}$. Define the extended-value extension $g_1(\boldsymbol{\pi}) = \lim_{\boldsymbol{d}\searrow\boldsymbol{\pi}} \log|\tilde{\boldsymbol{A}}(\boldsymbol{d})|$, $\boldsymbol{\pi} \succeq \boldsymbol{0}$, and $g_1(\boldsymbol{\pi}) = -\infty$ elsewhere (note that $\log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})|$ is continuous). Since g_1 is lower semicontinuous, and concave for $\boldsymbol{\pi} \succ \boldsymbol{0}$ (Theorem 3.1(2)), it is a closed proper concave function. Fenchel duality [26, sect. 12] implies that $g_1(\boldsymbol{\pi}) = \inf_{\boldsymbol{z}_1} \boldsymbol{z}_1^T\boldsymbol{\pi} - g_1^*(\boldsymbol{z}_1)$, where $g_1^*(\boldsymbol{z}_1) = \inf_{\boldsymbol{\pi}} \boldsymbol{z}_1^T\boldsymbol{\pi} - g_1(\boldsymbol{\pi})$ is closed concave as well. As $g_1(\boldsymbol{\pi})$ is unbounded above as $\|\boldsymbol{\pi}\| \to \infty$ (Theorem 5.1), $\boldsymbol{z}_1^T\boldsymbol{\pi} - g_1(\boldsymbol{\pi})$ is unbounded below whenever $z_{1,i} \leq 0$ for any i, and $g_1^*(\boldsymbol{z}_1) = -\infty$ in this case. Moreover, for any $\boldsymbol{\pi} \succ \boldsymbol{0}$, the corresponding minimizer $\boldsymbol{z}_{1,*}$ is given in section 4.2, so that $g_1(\boldsymbol{\pi}) = \min_{\boldsymbol{z}_1\succ\boldsymbol{0}} \boldsymbol{z}_1^T\boldsymbol{\pi} - g_1^*(\boldsymbol{z}_1)$.

Second, define the extended-value extension $g_2(\boldsymbol{\gamma}) = \lim_{\boldsymbol{d}\searrow\boldsymbol{\gamma}} \boldsymbol{1}^T(\log\boldsymbol{d}) + \log|\tilde{\boldsymbol{A}}(\boldsymbol{d})|$, $\boldsymbol{\gamma} \succeq \boldsymbol{0}$, and $g_2(\boldsymbol{\gamma}) = -\infty$ elsewhere (note that $\boldsymbol{1}^T(\log\boldsymbol{\pi}) + \log|\tilde{\boldsymbol{A}}(\boldsymbol{\pi})|$ is continuous). Since g_2 is lower semicontinuous, and concave for $\boldsymbol{\gamma} \succ \boldsymbol{0}$ (Theorem 3.1(3)), it is a closed proper concave function. Fenchel duality [26, sect. 12] implies that $g_2(\boldsymbol{\gamma}) = \inf_{\boldsymbol{z}_2} \boldsymbol{z}_2^T\boldsymbol{\gamma} - g_2^*(\boldsymbol{z}_2)$, where $g_2^*(\boldsymbol{z}_2) = \inf_{\boldsymbol{\gamma}} \boldsymbol{z}_2^T\boldsymbol{\gamma} - g_2(\boldsymbol{\gamma})$ is closed concave as well. Since $\boldsymbol{z}_2^T\boldsymbol{\gamma} - g_2(\boldsymbol{\gamma})$ is unbounded below whenever $z_{2,i} < 0$ for any i, we see that $g_2^*(\boldsymbol{z}_2) = -\infty$ in this case. For any $\boldsymbol{\gamma} \succ \boldsymbol{0}$, the corresponding minimizer $\boldsymbol{z}_{2,*}$ is given in section 4.2, so that $g_2(\boldsymbol{\gamma}) = \min_{\boldsymbol{z}_2\succeq\boldsymbol{0}} \boldsymbol{z}_2^T\boldsymbol{\gamma} - g_2^*(\boldsymbol{z}_2)$.

A.5. Implicit computation of h_i and h_i^*. Recall from sections 4 and 4.3 that our algorithms can be run whenever $h_i^*(s_i)$ and its derivatives can be evaluated. For log-concave potentials, these evaluations can be done generically, even if no closed form for $h_i(\gamma_i)$ is available. We focus on a single potential i and drop its index. As noted in section 4, if $z_2 = z_3 = 0$, then $h^*(s) = -g(z_1 + s^2)$. With $p := z_1 + s^2$, we have that $\theta = -2g'(p)s - b$, $\rho = -4g''(p)s^2 - 2g'(p)$. With a view to Appendix A.7, $\tilde{\theta} = -2g'(p)$, $p = z_1 + \|\boldsymbol{s}\|^2$, and $\kappa = 2[g''(p)]^{1/2}$.

If $z_2 \neq 0$ (and $t(s)$ is log-concave), we have to employ scalar convex minimization. We require $h^*(s) = \frac{1}{2}\min_\gamma k(x,\gamma)$, $k := (z_1 + x)/\gamma + z_2\gamma - z_3\log\gamma + h(\gamma)$, $x = s^2$, as well as $\theta = (h^*)'(s)$ and $\rho = (h^*)''(s)$. Let $\gamma_* = \operatorname{argmin} k(x,\gamma)$. Assuming for now that h and its derivatives are available, γ_* is found by univariate Newton minimization, where $\gamma^2 k_\gamma = -(z_1 + x) - z_3\gamma + \gamma^2(z_2 + h'(\gamma))$, $\gamma^3 k_{\gamma,\gamma} = 2(z_1 + x) + \gamma z_3 + \gamma^3 h''(\gamma)$. Now, $k_\gamma = 0$ (always evaluated at (x,γ_*)), so that $\theta = (h^*)'(s) = s/\gamma_*$. Moreover, $0 = (d/ds)k_\gamma = k_{s,\gamma} + k_{\gamma,\gamma}\cdot(d\gamma_*)/(ds)$, so that $\rho = (h^*)''(s) = \gamma_*^{-1}(1 - s\gamma_*^{-1}(d\gamma_*)/(ds)) = \gamma_*^{-1}(1 - 2x/(\gamma_*^3 k_{\gamma,\gamma}(x,\gamma_*)))$. With a view to Appendix A.7, $\tilde{\theta} = 1/\gamma_*$ and $\kappa = [2/(\gamma_*^4 k_{\gamma,\gamma}(x,\gamma_*))]^{1/2}$ (note that $\tilde{\theta} \geq \rho$).

By Fenchel duality, $h(\gamma) = -\min_x l(x,\gamma)$, $l := x/\gamma + 2g(x)$, where $g(x)$ is strictly convex and decreasing. We need methods to evaluate $g(x)$ and its first and second derivatives (note that $g''(x) > 0$). The minimizer $x_* = x_*(\gamma)$ is found by convex minimization once more, started from the last recently found x_* for this potential. Note that $x_* = 0$ if and only if $\gamma \leq \gamma_0 := -1/(2g'(0))$ (where $\gamma_0 = 0$ if $g'(x) \to -\infty$ as $x \to 0$), which has to be checked up front. Given x_*, we have that $\gamma h(\gamma) = -x_* - 2\gamma g(x_*)$. Since $l_x = 0$ for $\gamma > \gamma_0$ (always evaluated at (x_*,γ)), we have that $\gamma^2 h'(\gamma) = -\gamma^2 l_\gamma = x_*$ (this holds even if $l_x > 0$ and $x_* = 0$). Moreover, if $x_* > 0$ (for $\gamma > \gamma_0$), then $(d/d\gamma)l_x(x_*,\gamma) = 0$, so that $(dx_*)/(d\gamma) = \gamma^{-2}/(2g''(x_*))$, and

$\gamma^3 h''(\gamma) = (2\gamma g''(x_*))^{-1} - 2x_*$. If $x_* = 0$ and $l_x > 0$, then $x_*(\tilde{\gamma}) = 0$ for $\tilde{\gamma}$ close to γ, so that $h''(\gamma) = 0$. A critical case is $x_* = 0$ and $l_x = 0$, which happens for $\gamma = \gamma_0 > 0$: $h''(\gamma)$ does not exist at this point in general. This is not a problem for our code, since we employ a robust Newton/bisection search for γ_*. If $\gamma > \gamma_*$, but is very close, then $(dx_*)/(d\gamma) \approx \xi_0/\gamma$ with $\xi_0 := -g'(0)/g''(0)$, and therefore $x_*(\gamma) \approx \int_{\gamma_0}^{\gamma} \xi_0/t\, dt = \xi_0(\log\gamma - \log\gamma_0)$. We use $\gamma^2 h'(\gamma) = x_* \approx \xi_0(\log\gamma - \log\gamma_0)$ and $\gamma^3 h''(\gamma) \approx \xi_0 - 2x_*$ in this case.

A.6. Details for specific potentials. Our algorithms are configured by the dual functions $h_i(\gamma_i)$ for each non-Gaussian $t_i(s_i)$, and the inner loops require $h_i^*(s_i)$ and its derivatives (see (4.2), and recall that for each i, either $z_{1,i} > 0$ and $z_{3,i} = 0$, or $z_{1,i} = 0$ and $z_{2,i} > 0, z_{3,i} = 1$). In this section, we show how these are computed for the potentials used in this paper. We use the notation of Appendix A.5, focus on a single potential i, and drop its index.

Laplace potentials. These are $t(s) = \exp(-\tau|s|)$, $\tau > 0$, so that $g(x) = \tau x^{1/2}$. We have that $h(\gamma) = h_\cup(\gamma) = \tau^2\gamma$, so that $k(x,\gamma) = (z_1 + x)/\gamma + (z_2 + \tau^2)\gamma - z_3 \log\gamma$. The stationary equation for γ_* is $(z_2 + \tau^2)\gamma^2 - z_3\gamma - (z_1 + x) = 0$. If $z_3 = 0$ (bounding type A), this is just a special case of Appendix A.5. With $p := z_1 + x$, $q := (z_2 + \tau^2)^{1/2}$, we have that $\gamma_* = p^{1/2}/q$, $h^*(s) = qp^{1/2}$, and $\theta = (h^*)'(s) = qp^{-1/2}s$, $\rho = (h^*)''(s) = qz_1p^{-3/2}$. With a view to Appendix A.7, $\tilde{\theta} = qp^{-1/2}$ and $\kappa = [qp^{-3/2}]^{1/2}$.

If $z_3 = 1$ (bounding type B), note that $z_1 = 0$, $z_2 > 0$. Let $q := 2(z_2 + \tau^2)$, $p := (1 + 2qx)^{1/2}$. Then, $\gamma_* = (p+1)/q$, and $k(x,\gamma_*) = p - \log(p+1) + \log q$ after some algebra, so that $h^*(s) = \frac{1}{2}(p - \log(p+1) + \log q)$. With $dp/(ds) = 2qp^{-1}s$, we have $\theta = qs/(p+1)$, $\rho = q/(p(p+1))$. With a view to Appendix A.7, $\tilde{\theta} = q/(p+1)$. Using $p^2 - 1 = 2xq$, some algebra gives $\kappa = (2/p)^{1/2}q/(p+1) = (2/p)^{1/2}\tilde{\theta}$.

Student's t potentials. These are $t(s) = (1 + (\tau/\nu)x)^{-(\nu+1)/2}$, $\nu > 0$, $\tau > 0$. If $\alpha := \nu/\tau$, the critical point of Appendix A.2 is $\gamma_0 := \alpha/(\nu+1)$, and $h(\gamma) = [\alpha/\gamma + (\nu+1)\log\gamma + C]\mathrm{I}_{\{\gamma \ge \gamma_0\}}$ with $C := -(\nu+1)(\log\gamma_0 + 1)$. $h(\gamma)$ is not convex. We choose a decomposition such that $h_\cup(\gamma)$ is convex and twice continuously differentiable, ensuring that $h^*(s)$ is continuously differentiable and the inner loop optimization runs smoothly. Since $h(\gamma)$ does not have a second derivative at γ_0, neither has $h_\cap(\gamma)$:

$$h_\cap(\gamma) = \begin{cases} (\nu + 1 - z_3)\log\gamma & | \ \gamma \ge \gamma_0, \\ (2(\nu+1) - z_3)\log\gamma - a(\gamma - \gamma_0) - b & | \ \gamma < \gamma_0, \end{cases}$$

$$h_\cup(\gamma) = \begin{cases} \alpha/\gamma + C & | \ \gamma \ge \gamma_0, \\ -2(\nu+1)\log\gamma + a(\gamma - \gamma_0) + b & | \ \gamma < \gamma_0, \end{cases}$$

where $b := (\nu+1)\log\gamma_0$, $a := (\nu+1)/\gamma_0$. Here, the $-z_3\log\gamma$ term of $k(x,\gamma)$ is folded into $h_\cap(\gamma)$.

We follow Appendix A.5 in determining $h^*(s)$ and its derivatives, but solve for γ_* directly. Note that $z_2 > 0$ even if $z_3 = 0$ (bounding type A), due to the Fenchel bound on $h_\cap(\gamma)$. We minimize $k(x,\gamma)$ for $\gamma \ge \gamma_0$, $\gamma < \gamma_0$, respectively, and pick the minimum. For $\gamma \ge \gamma_0$, $k(x,\gamma) = (z_1 + \alpha + x)/\gamma + z_2\gamma + C$, whose minimum point $\gamma_{*,1} := [(z_1 + \alpha + x)/z_2]^{1/2}$ is a candidate if $\gamma_{*,1} \ge \gamma_0$, with $k(x,\gamma_{*,1}) = 2[z_2(z_1 + \alpha + x)]^{1/2} + C$. For $\gamma < \gamma_0$, $k(x,\gamma) = (z_1 + x)/\gamma + (z_2 + a)\gamma - 2(\nu+1)\log\gamma + b - a\gamma_0$, with minimum point $\gamma_{*,2} := [\nu + 1 + ((\nu+1)^2 + (z_2 + a)(z_1 + x))^{1/2}]/(z_2 + a) < \gamma_0$. If $z_2 \le a$, then $\gamma_{*,2} \ge \gamma_0$ (not a candidate).

This can be tested up front. If $c := (z_2 + a)(z_1 + x)$, $d := ((\nu+1)^2 + c)^{1/2} \geq \nu + 1$, then $k(x, \gamma_{*,2}) = c/(\nu + 1 + d) + d + (\nu + 1)[2\log(z_2 + a) - 2\log(\nu + 1 + d) + \log\gamma_0] = 2d + (\nu + 1)[2\log(z_2 + a) - 2\log(\nu + 1 + d) + \log\gamma_0 - 1]$. Now, θ and ρ are computed as in Appendix A.5 ($h(\gamma)$ there is $h_\cup(\gamma)$ here, and $z_3 = 0$, since this is folded into $h_\cap$ here), where $\gamma_*^3 h''_\cup(\gamma_*) = 2\alpha$ for $\gamma_* \geq \gamma_0$, and $\gamma_*^3 h''_\cup(\gamma_*) = 2(\nu+1)\gamma_*$ for $\gamma_* < \gamma_0$.

Bernoulli potentials. These are $t(s) = (1+e^{-y\tau s})^{-1} = e^{y\tau s/2}(2\cosh v)^{-1}$, $v := (y\tau/2)x^{1/2} = (y\tau/2)|s|$. They are not even; $b = y\tau/2$. While $h(\gamma)$ is not known analytically, we can plug these expressions into the generic setup of Appendix A.5. Namely, $g(x) = -\log(\cosh v) - \log 2$, so that $g'(x) = -C(\tanh v)/v$, $g''(x) = (C/2)x^{-1}((\tanh v)/v + \tanh^2 v - 1)$, $C := (y\tau/2)^2/2$. For x close to zero, we use $\tanh v = v - v^3/3 + 2v^5/15 + O(v^7)$ for these computations. Moreover, $\gamma_0 = 1/(2C)$ and $\xi_0 = 3/(2C)$.

A.7. Group potentials. An extension of our framework to group potentials $t_i(\|\boldsymbol{s}_i\|)$ is described in section 4.5. Recall the details about the IRLS algorithm from section 4.3. For group potentials, the inner Hessian is not diagonal anymore. $h_i^*(s_i)$ becomes $h_i^*(\|\boldsymbol{s}_i\|)$, and $x_i = \|\boldsymbol{s}_i\|^2$. If θ_i, ρ_i are as in section 4.3, $ds_i \to d\|\boldsymbol{s}_i\|$, and $\tilde{\theta}_i := \theta_i/\|\boldsymbol{s}_i\|$, we have that $\nabla_{\boldsymbol{s}_i} h_i^* = \tilde{\theta}_i \boldsymbol{s}_i$, since $\nabla_{\boldsymbol{s}_i}\|\boldsymbol{s}_i\| = \boldsymbol{s}_i/\|\boldsymbol{s}_i\|$. Therefore, the gradient $\boldsymbol{\theta}$ is given by $\boldsymbol{\theta}_i = \tilde{\theta}_i \boldsymbol{s}_i$. Moreover,

$$\nabla\nabla_{\boldsymbol{s}_i} h_i^* = \tilde{\theta}_i \boldsymbol{I} - (\tilde{\theta}_i - \rho_i)\|\boldsymbol{s}_i\|^{-2}\boldsymbol{s}_i\boldsymbol{s}_i^T.$$

For simplicity of notation, assume that all $\boldsymbol{s}_i$ have the same dimensionality. From Appendix A.5, we see that $\tilde{\theta}_i \geq \rho_i$. Let $\kappa_i := (\tilde{\theta}_i - \rho_i)^{1/2}/\|\boldsymbol{s}_i\|$, and let $\hat{\boldsymbol{s}} := ((\operatorname{diag}\boldsymbol{\kappa}) \otimes \boldsymbol{I})\boldsymbol{s}$. The Hessian w.r.t. $\boldsymbol{s}$ is

$$\boldsymbol{H}^{(s)} = (\operatorname{diag}\tilde{\boldsymbol{\theta}}) \otimes \boldsymbol{I} - \sum_i \boldsymbol{w}_i \boldsymbol{w}_i^T, \quad \boldsymbol{w}_i = (\boldsymbol{\delta}_i \boldsymbol{\delta}_i^T \otimes \boldsymbol{I})\hat{\boldsymbol{s}}.$$

If $\boldsymbol{w}$ is given by $w_i = \hat{\boldsymbol{s}}_i^T \boldsymbol{v}_i$, then $\boldsymbol{H}^{(s)}\boldsymbol{v} = ((\operatorname{diag}\tilde{\boldsymbol{\theta}}) \otimes \boldsymbol{I})\boldsymbol{v} - ((\operatorname{diag}\boldsymbol{w}) \otimes \boldsymbol{I})\hat{\boldsymbol{s}}$. The system matrix for the Newton direction is $\sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T\boldsymbol{H}^{(s)}\boldsymbol{B}$. For numerical reasons, $\tilde{\theta}_i$ and κ_i should be computed directly rather than via θ_i, ρ_i.

If $\boldsymbol{s}_i \in \mathbb{R}^2$, we can avoid the subtraction in computing $\boldsymbol{H}^{(s)}\boldsymbol{v}$ and gain numerical stability. Namely, $\nabla\nabla_{\boldsymbol{s}_i} h_i^* = \rho_i \boldsymbol{I} + \kappa_i^2 \left(\|\boldsymbol{s}_i\|^2 \boldsymbol{I} - \boldsymbol{s}_i\boldsymbol{s}_i^T\right)$. Since $\|\boldsymbol{s}_i\|^2\boldsymbol{I} - \boldsymbol{s}_i\boldsymbol{s}_i^T = \boldsymbol{M}\boldsymbol{s}_i(\boldsymbol{M}\boldsymbol{s}_i)^T$, $\boldsymbol{M} = \boldsymbol{\delta}_2\boldsymbol{\delta}_1^T - \boldsymbol{\delta}_1\boldsymbol{\delta}_2^T$, if we redefine $\hat{\boldsymbol{s}} := ((\operatorname{diag}\boldsymbol{\kappa}) \otimes (\boldsymbol{\delta}_2\boldsymbol{\delta}_1^T - \boldsymbol{\delta}_1\boldsymbol{\delta}_2^T))\boldsymbol{s}$, then

$$\boldsymbol{H}^{(s)}\boldsymbol{v} = ((\operatorname{diag}\boldsymbol{\rho}) \otimes \boldsymbol{I})\boldsymbol{v} + ((\operatorname{diag}\boldsymbol{w}) \otimes \boldsymbol{I})\hat{\boldsymbol{s}}, \quad \boldsymbol{w} = \left(\hat{\boldsymbol{s}}_i^T \boldsymbol{v}_i\right).$$

Acknowledgments. The MRI application is joint work with Rolf Pohmann and Bernhard Schölkopf, MPI for Biological Cybernetics, Tübingen. We thank Florian Steinke and David Wipf for helpful discussions.

REFERENCES

[1] H. Attias, *A variational Bayesian framework for graphical models*, in Advances in Neural Information Processing Systems 12, S. Solla, T. Leen, and K.-R. Müller, eds., MIT Press, Cambridge, MA, 2000, pp. 209–215.

[2] C. Bekas, E. Kokiopoulou, and Y. Saad, *Computation of large invariant subspaces using polynomial filtered Lanczos iterations with applications in density functional theory*, SIAM J. Matrix. Anal. Appl., 30 (2008), pp. 397–418.

[3] J. Bioucas-Dias and M. Figueiredo, *Two-step iterative shrinkage/thresholding algorithms for image restoration*, IEEE Trans. Image Process., 16 (2007), pp. 2992–3004.
[4] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.
[5] A. M. Bruckstein, D. L. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, SIAM Rev., 51 (2009), pp. 34–81.
[6] E. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.
[7] H. Chang, Y. Weiss, and W. Freeman, *Informative sensing*, IEEE Trans. Inform. Theory, submitted; available online from http://arxiv.org/abs/0901.4275.
[8] A. Dempster, N. Laird, and D. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Statist. Soc. Ser. B, 39 (1977), pp. 1–38.
[9] D. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
[10] D. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 2197–2202.
[11] M. Girolami, *A variational method for learning sparse and overcomplete representations*, Neural Comput., 13 (2001), pp. 2517–2532.
[12] T. Goldstein and S. Osher, *The split Bregman method for L1-regularized problems*, SIAM J. Imaging Sci., 2 (2009), pp. 323–343.
[13] G. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[14] P. Green, *Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives*, J. Roy. Statist. Soc. Ser. B, 46 (1984), pp. 149–192.
[15] N. Halko, P. Martinsson, and J. Tropp, *Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions*, Technical report, 2009; available online from http://www.arxiv.org/abs/0909.4061.
[16] J. Haupt, R. Castro, and R. Nowak, *Distilled sensing: Selective sampling for sparse signal recovery*, in Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, D. van Dyk and M. Welling, eds., 2009, pp. 216–223.
[17] T. Jaakkola, *Variational Methods for Inference and Estimation in Graphical Models*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1997.
[18] S. Ji and L. Carin, *Bayesian compressive sensing and projection optimization*, in Proceedings of the 24th International Conference on Machine Learning, Z. Ghahramani, ed., Omni Press, Madison, WI, 2007, pp. 377–384.
[19] M. Lustig, D. Donoho, and J. Pauly, *Sparse MRI: The application of compressed sensing for rapid MR imaging*, Magn. Reson. Med., 58 (2007), pp. 1182–1195.
[20] D. Malioutov, J. Johnson, and A. Willsky, *Low-rank variance estimation in large-scale GMRF models*, in Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2006.
[21] N. Meinshausen and P. Bühlmann, *Stability selection*, J. R. Stat. Soc. Ser. B Stat. Methodol., 72 (2010), pp. 417–473.
[22] H. Nickisch and M. Seeger, *Convex variational Bayesian inference for large scale generalized linear models*, in Proceedings of the 26th International Conference on Machine Learning, L. Bottou and M. Littman, eds., Omni Press, Madison, WI, 2009, pp. 761–768.
[23] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao, *Variational EM algorithms for non-Gaussian latent variable models*, in Advances in Neural Information Processing Systems 18, Y. Weiss, B. Schölkopf, and J. Platt, eds., MIT Press, Cambridge, MA, 2006, pp. 1059–1066.
[24] B. Parlett and D. Scott, *The Lanczos algorithm with selective orthogonalization*, Math. Comp., 33 (1979), pp. 217–238.
[25] B. Rao, K. Engan, S. Cotter, J. Palmer, and K. Kreutz-Delgado, *Subset selection in noise based on diversity measure minimization*, IEEE Trans. Signal Process., 51 (2003), pp. 760–770.
[26] R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[27] M. K. Schneider and A. S. Willsky, *Krylov subspace estimation*, SIAM J. Sci. Comput., 22 (2001), pp. 1840–1864.

[28] M. Seeger, *Bayesian inference and optimal design for the sparse linear model*, J. Mach. Learn. Res., 9 (2008), pp. 759–813.
[29] M. Seeger, *Sparse linear models: Variational approximate inference and Bayesian experimental design*, J. Phys.: Conf. Ser., 197 (2009), 012001.
[30] M. Seeger, *Speeding up magnetic resonance image acquisition by Bayesian multi-slice adaptive compressed sensing*, in Advances in Neural Information Processing Systems 22, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, eds., Curran Associates, Red Hook, NY, 2009, pp. 1633–1641.
[31] M. Seeger, *Gaussian covariance and scalable variational inference*, in Proceedings of the 27th International Conference on Machine Learning, J. Fürnkranz and T. Joachims, eds., Omni Press, Madison, WI, 2010.
[32] M. Seeger and H. Nickisch, *Compressed sensing and Bayesian experimental design*, in Proceedings of the 25th International Conference on Machine Learning, A. McCallum, S. Roweis, and R. Silva, eds., Omni Press, Madison, WI, 2008, pp. 912–919.
[33] M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf, *Bayesian experimental design of magnetic resonance imaging sequences*, in Advances in Neural Information Processing Systems 21, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds., Curran Associates, Red Hook, NY, 2009, pp. 1441–1448.
[34] M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf, *Optimization of k-space trajectories for compressed sensing by Bayesian experimental design*, Magn. Reson. Med., 63 (2010), pp. 116–126.
[35] E. Simoncelli, *Modeling the joint statistics of images in the wavelet domain*, in Wavelet Applications in Signal and Image Processing VII, Proceedings of the SPIE, Vol. 3813, 1999, pp. 188–195.
[36] F. Steinke, M. Seeger, and K. Tsuda, *Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models*, BMC Syst. Biol., 1 (2007).
[37] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[38] M. Tipping, *Sparse Bayesian learning and the relevance vector machine*, J. Mach. Learn. Res., 1 (2001), pp. 211–244.
[39] M. Tipping and A. Faul, *Fast marginal likelihood maximisation for sparse Bayesian models*, in Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, C. Bishop and B. Frey, eds., 2003.
[40] J. Tropp, *Algorithms for simultaneous sparse approximation.* II: *Convex relaxation*, Signal Process., 86 (2006), pp. 589–602.
[41] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*, Found. Trends Mach. Learn., 1 (2008), pp. 1–305.
[42] M. Wainwright, E. Sudderth, and A. Willsky, *Tree-based modeling and estimation of Gaussian processes on graphs with cycles*, in Advances in Neural Information Processing Systems 13, T. Leen, T. Dietterich, and V. Tresp, eds., MIT Press, Cambridge, MA, 2001, pp. 661–667.
[43] D. Wipf and S. Nagarajan, *A new view of automatic relevance determination*, in Advances in Neural Information Processing Systems 20, J. Platt, D. Koller, Y. Singer, and S. Roweis, eds., Curran Associates, Red Hook, NY, 2008, pp. 1625–1632.
[44] D. Wipf and S. Nagarajan, *Latent variable Bayesian models for promoting sparsity*, IEEE Trans. Inform. Theory, to appear.
[45] G. Wright, *Magnetic resonance imaging: From basic physics to imaging principles*, IEEE Signal Processing Magazine, 14 (1997), pp. 56–66.
[46] S. Wright, R. Nowak, and M. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Process., 57 (2009), pp. 2479–2493.
[47] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. Ser. B Methodol., 68 (2006), pp. 49–67.
[48] A. Yuille and A. Rangarajan, *The concave-convex procedure*, Neural Comput., 15 (2003), pp. 915–936.