Variational Inference of Penalized Regression with Submodular Functions

Koh Takeuchi
NTT
koh.t@acm.org

Yuichi Yoshida
National Institute of Informatics
yyoshida@nii.ac.jp

Yoshinobu Kawahara
Institute of Mathematics for Industry
Kyushu University
Center for Advanced Intelligence Project
RIKEN
kawahara@imi.kyushu-u.ac.jp

Abstract

Various regularizers inducing structured-sparsity are constructed as Lovász extensions of submodular functions. In this paper, we consider a hierarchical probabilistic model of linear regression and its kernel extension with this type of regularization, and develop a variational inference scheme for the posterior estimate on this model. We derive an upper bound on the partition function with an approximation guarantee, and then show that minimizing this bound is equivalent to the minimization of a quadratic function over the polyhedron determined by the corresponding submodular function, which can be solved efficiently by the proximal gradient algorithm. Our scheme gives a natural extension of the Bayesian Lasso model for the maximum a posteriori (MAP) estimation to a variety of regularizers inducing structured sparsity, and thus this work provides a principled way to transfer the advantages of the Bayesian formulation into those models. Finally, we investigate the empirical performance of our scheme with several Bayesian variants of widely known models such as Lasso, generalized fused Lasso, and non-overlapping group Lasso.

1 INTRODUCTION

The development of penalized regression methods for simultaneous variable selection and coefficient estimation is one of the most important problems in the fields of machine learning and statistics. In particular, Lasso and its generalizations have shown excellent performance in many situations with extensive theoretical appraisals [28, 29]. Furthermore, extensions to structured penalties have attracted attention in these fields, with applications in a variety of engineering and scientific scenarios. In this context, the recent pioneering works by Bach [3] revealed relationships between structured sparsity and submodular functions; many known regularizers inducing structured sparsity coincide with the Lovász extensions of submodular functions. Some novel structured regularizers have been developed based on this relationship with specific submodular functions [5, 27].

From a Bayesian perspective, it is well known that Lasso is the Bayesian posterior mode under independent Laplace priors [22]. In addition, grouped gamma priors yield the Bayesian group Lasso, whose maximum a posteriori (MAP) estimate coincides with group Lasso [24]. The spike-and-slab prior has been considered as another candidate for variable selection [17, 32]. Although Bayesian models could bring us various fruitful benefits in regression analyses, posterior inference in the models in general requires the application of a Markov chain Monte Carlo (MCMC) method such as Gibbs sampling, which can be computationally expensive and limits the range of applications. Therefore, approximate inference methods, such as variational inference and MAP estimate, have been developed for respective models. However, to the best of our knowledge, neither a Bayesian nor related probabilistic methods for regression with structured sparsity from submodular functions have been studied despite their wide coverage of the existing structured penalties.

In this paper, we consider a hierarchical probabilistic formulation of linear regression and its kernel extension regularized with the Lovász extensions of submodular functions, and develop a variational inference method for the posterior inference in this model.

The proposed method gives a natural extension of the Bayesian Lasso model for the MAP estimate to a variety of regularizers inducing structured sparsity, and hence this work provides a principled way to transfer the advantages of the Bayesian formulation into those models. As is the case with hierarchical probabilistic models including the Bayesian Lasso, the main difficulty in the inference of the posterior distribution of our model lies in computing the partition function. To address this difficulty, we derive a variational upper bound on the partition function and show that minimizing this bound is equivalent to the minimization of a quadratic function over the polyhedron determined by the corresponding submodular function, which can be solved efficiently by the proximal gradient algorithm. Then, we give another interpretation of our variational bound, demonstrating that it is proportional to the MAP value of our model. We also give a theoretical approximation guarantee for the bound. Finally, we investigate the empirical performance of our method with several Bayesian variants of widely known models such as Lasso, generalized fused Lasso, and non-overlapping group Lasso.

Variational Bayesian inference in probabilistic models involving submodular functions has received attention recently in machine learning although those works have focused on models for discrete random variables [6, 7, 8]. Variational bridge regression [2] provides a variational lower bound on a partition function, however, this method does not coincide with Lasso. Meanwhile, it is known that a variational bound on the partition function of a Gibbs distribution can be obtained by averaging MAP estimates of randomly perturbed models [13, 14]. Note that, by contrast, our variational bound is obtained from the MAP estimate of a single model, meaning that no averaging scheme is required.

The remainder of this paper is organized as follows. First, in the paragraph that follows, we describe notation used in this paper and give preliminaries on submodular functions. In Section 2, we develop a hierarchical probabilistic model whose MAP estimate coincides with penalized regression with submodular functions. Next, in Section 3, we develop a variational upper bound on the partition function. In Section 4, we discuss a connection between this upper bound and the MAP value of our model. Then, in Section 5, we will see that this connection can be used to produce a general method to compute our variational upper bound, and we discuss more efficient methods for specific cases. In Section 6, we derive an approximation guarantee on our upper bound. In Section 7, we extend our model to kernel regression model. Finally, we empirically demonstrate the effectiveness of our scheme in Section 8, and we conclude the paper in Section 9.

Notation and Preliminaries We first describe the notation used in this paper. For an integer p, let $[p]$ denote the set $\{1, 2, \dots, p\}$. For a positive-semidefinite matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ and a vector $\boldsymbol{b} \in \mathbb{R}^p$, we define $\|\boldsymbol{b}\|_{\boldsymbol{A}}$ as $\sqrt{\boldsymbol{b}^\top \boldsymbol{A} \boldsymbol{b}}$. Let $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be a vector and a matrix. We define a vector $|\boldsymbol{x}| \in \mathbb{R}^p$ as $|\boldsymbol{x}|(i) = |\boldsymbol{x}(i)|$ for each $i \in [p]$. For a set $S \subseteq [p]$, we define $\boldsymbol{x}(S)$ as the sum $\sum_{i \in S} \boldsymbol{x}(i)$.

Next, we review some background on submodular functions necessary for the remaining parts of this paper. A set function $F : 2^{[p]} \to \mathbb{R}$ is said to be *submodular* if

$$F(S) + F(T) \geq F(S \cup T) + F(S \cap T)$$

for every $S, T \subseteq [p]$. For $S \subseteq [p]$ and $i \in [p] \setminus S$, we define $F(i \mid S) = F(S \cup \{i\}) - F(S)$ as the marginal gain of adding i when having had S. It is well known that $F : 2^{[p]} \to \mathbb{R}$ is submodular if and only if $F(i \mid S) \geq F(i \mid T)$ for every $S \subseteq T \subseteq [p]$ and $i \in [p] \setminus S$; this is called the *diminishing return property*. All submodular functions we consider in this work are supposed to be *normalized*, that is, $F(\emptyset) = 0$.

For a submodular function $F : 2^{[p]} \to \mathbb{R}$, the *submodular polyhedron* and the *symmetric submodular polyhedron* of F are defined as

$$P(F) = \{\boldsymbol{x} \in \mathbb{R}^p \mid \boldsymbol{x}(S) \leq F(S)\ \forall S \subseteq [p]\} \quad \text{and}$$
$$|P|(F) = \{\boldsymbol{x} \in \mathbb{R}^p \mid |\boldsymbol{x}| \in P(F)\},$$

respectively. The *base polyhedron* of F is defined as $B(F) = \{\boldsymbol{x} \in P(F) \mid \boldsymbol{x}([p]) = F([p])\}$. The *Lovász extension* $f : \mathbb{R}^p \to \mathbb{R}$ of a submodular function $F : 2^{[p]} \to \mathbb{R}$ is defined as

$$f(\boldsymbol{x}) = \max_{\boldsymbol{z} \in B(F)} \boldsymbol{z}^\top \boldsymbol{x}.$$

By Edmonds' algorithm, we can find the maximizer $\boldsymbol{z}^*$ of the above maximization problem as follows [10]: Let $\pi_1, \dots, \pi_p$ be an ordering of $[p]$ such that $\boldsymbol{x}(\pi_1) \geq \boldsymbol{x}(\pi_2) \geq \cdots \geq \boldsymbol{x}(\pi_p)$. Then, we set $\boldsymbol{z}^*(\pi_i) = f(\pi_i \mid \pi_1, \dots, \pi_{i-1})$. Moreover, any extreme point of $B(F)$ can be constructed in this manner.

We also consider a function $|f| : \mathbb{R}^p \to \mathbb{R}$ defined as $|f|(\boldsymbol{x}) = f(|\boldsymbol{x}|)$. We can verify that

$$|f|(\boldsymbol{x}) = \max_{\boldsymbol{z} \in B(F)} \boldsymbol{z}^\top |\boldsymbol{x}| = \max_{\boldsymbol{z} \in P(F)} \boldsymbol{z}^\top |\boldsymbol{x}| = \max_{\boldsymbol{z} \in |P|(F)} \boldsymbol{z}^\top \boldsymbol{x}.$$

2 INFERENCE WITH PRIORS FROM SUBMODULAR FUNCTIONS

Consider a linear regression model with parameters $\boldsymbol{w} \in \mathbb{R}^p$, i.e., $y = \boldsymbol{x}^\top \boldsymbol{w} + \epsilon$, where $y \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^p$, and $\epsilon \in \mathbb{R}$ is an independent and identically distributed (i.i.d.) Gaussian noise term. Here, we treat the parameters as random quantities, along with a prior distribution whose structure is characterized by a submodular function $F : 2^{[p]} \to \mathbb{R}$. We consider the following hierarchical model:

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{w}, F &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I}) \quad \text{and} \\ \boldsymbol{w} \mid F &\sim \frac{1}{\mathcal{Z}_0} \exp\left(\frac{\eta}{2}\|\boldsymbol{w}\|_2^2 + \lambda r(\boldsymbol{w})\right), \end{aligned} \tag{1}$$

where $\eta > 0$ is a regularization factor for the ℓ_2 term; $r \colon \mathbb{R}^p \to \mathbb{R}$ is determined via F, i.e., $r = f$ or $r = |f|$; $\lambda \in \mathbb{R}_+$ is another regularization factor for r; and $\mathcal{Z}_0 := \int_{\mathbb{R}^p} \exp\left(\frac{\eta}{2}\|\boldsymbol{w}\|_2^2 + \lambda r(\boldsymbol{w})\right) d\boldsymbol{w}$ is the partition function. This model is a natural extension of the Bayesian Lasso of Park and Casella [22], which corresponds to the case that F is the cardinality function and $\eta = 0$. Under this model, it is easy to see that the negative log posterior density for $\boldsymbol{w} \mid \boldsymbol{y}, F$ is given by

$$\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \frac{\eta}{2}\|\boldsymbol{w}\|_2^2 + \lambda r(\boldsymbol{w}),$$

where we drop an additive constant independent of $\boldsymbol{w}$. Consequently, for any fixed F, the posterior mode gives an estimate of a linear regression model with regularization by r and the ℓ_2 term. As described below, this covers several existing regularized regression models. In addition, the posterior distribution provides more than point estimates, i.e., an entire joint distribution, for these models.

As in other Bayesian models, one of the main interests for this model is the posterior inference of the parameters $\boldsymbol{w}$ given data $(\boldsymbol{y}, \boldsymbol{X})$. The posterior distribution is given as

$$p(\boldsymbol{w} \mid \boldsymbol{y}, F) = \frac{1}{\mathcal{Z}} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 - \frac{\eta}{2}\|\boldsymbol{w}\|_2^2 - \lambda r(\boldsymbol{w})}, \tag{2}$$

where $\mathcal{Z}$ is the partition function. For posterior inference, we need the partition function $\mathcal{Z}$. However, we cannot afford to compute it in general because the integral in the partition function cannot be efficiently computed even in the simplest case of Lasso. Our challenge in this paper is to develop a variational inference scheme for this model by constructing a tight upper bound that can be calculated efficiently in most practical cases.

Example Models There are various regularizers that are known to be representable as the Lovász extensions of submodular functions (refer to [3], for examples). Model (1) above is a natural extension of the Bayesian Lasso, in which F is the cardinality function, to more general cases. One popular example is the generalized fused Lasso (GFL), also known as the total variation regularization [29, 15, 31], in which the regularizer is given as the Lovász extension of a cut function associated with an undirected graph with positive weights. Recently, this has been extended to regularizers from hypergraphs, which are also given as the Lovász extension of a submodular function [27]. Another important class of regularizers is the one that induces group sparsity. For example, given a partition $(G_1, \ldots, G_k)$ of $[p]$, the ℓ_1/ℓ_∞-norm $\sum_{j\in[k]} \max_{i\in G_j} \boldsymbol{x}(i)$ is given as the Lovász extension of a coverage function corresponding to this partition. Group sparsity with overlapping groups similar to the ℓ_1/ℓ_∞-norm is also known to be attained by using the Lovász extension of a coverage function for the general case [21]. Other examples include regularization by spectral functions of submatrices of design matrices [30], and regularization with the scale-free property of a network [5].

3 VARIATIONAL INFERENCE

In this section, we develop a variational inference method that approximates $\mathcal{Z}$ in (2). For clarity, we give the explicit form of $\mathcal{Z}$ here:

$$\mathcal{Z} = \int_{\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 - \frac{\eta}{2}\|\boldsymbol{w}\|_2^2 - \lambda r(\boldsymbol{w})} \mathrm{d}\boldsymbol{w}.$$

Let $P(r)$ be the polyhedron associated with r, that is, $P(r) = B(F)$ if $r = f$ and $P(r) = |P|(F)$ if $r = |f|$. Note that $r(\boldsymbol{w}) \geq \boldsymbol{g}^\top \boldsymbol{w}$ for any $\boldsymbol{g} \in P(r)$ from the definition. Then, for any $\boldsymbol{g} \in P(r)$, we can upper-bound $\mathcal{Z}$ by

$$\begin{aligned} \overline{\mathcal{Z}}_{\boldsymbol{g}} &:= \int_{\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 - \frac{\eta}{2}\|\boldsymbol{w}\|_2^2 - \lambda \boldsymbol{g}^\top \boldsymbol{w}} \mathrm{d}\boldsymbol{w} \\ &= \int_{\mathbb{R}^p} e^{-\frac{1}{2}\left(\frac{\|\boldsymbol{y}\|_2^2}{\sigma^2} - \frac{2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w}}{\sigma^2} + \frac{\|\boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} + \eta\|\boldsymbol{w}\|_2^2\right) - \lambda \boldsymbol{g}^\top \boldsymbol{w}} \mathrm{d}\boldsymbol{w} \\ &= e^{-\frac{1}{2\sigma^2}\|\boldsymbol{y}\|_2^2} \times \int_{\mathbb{R}^p} e^{-\frac{1}{2}\left(\frac{\|\boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} + \eta\|\boldsymbol{w}\|_2^2\right) + \left(\frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2} - \lambda \boldsymbol{g}\right)^\top \boldsymbol{w}} \mathrm{d}\boldsymbol{w} \end{aligned} \tag{3}$$

To further simplify (3), we use the following fact on Gaussian integrals:

Lemma 3.1 (See, e.g., Chapter 2 of [26])**.** *For a positive-definite matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ and a vector $\boldsymbol{b} \in$*

$\mathbb{R}^p$, *we have*

$$\int_{\mathbb{R}^p} e^{-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}^\top \boldsymbol{x}} d\boldsymbol{x} = \sqrt{\frac{(2\pi)^p}{\det \boldsymbol{A}}} e^{\frac{1}{2}\|\boldsymbol{b}\|^2_{\boldsymbol{A}^{-1}}}.$$

Note that matrix $\boldsymbol{M}_{\boldsymbol{X},\sigma,\eta} := \boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \eta I$ is positive-definite as $\eta > 0$. Hence, by (3) and Lemma 3.1, we have

$$\overline{\mathcal{Z}}_{\boldsymbol{g}} = e^{-\frac{1}{2\sigma^2}\|\boldsymbol{y}\|_2^2} \sqrt{\frac{(2\pi)^p}{\det \boldsymbol{M}_{\boldsymbol{X},\sigma,\eta}}} e^{\Psi(\boldsymbol{g})}, \qquad (4)$$

where

$$\Psi(\boldsymbol{g}) := \frac{1}{2}\Big\| \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2} - \lambda \boldsymbol{g} \Big\|^2_{\boldsymbol{M}^{-1}_{\boldsymbol{X},\sigma,\eta}}.$$

Finally, we can use the following as an upper bound on $\mathcal{Z}$.

$$\overline{\mathcal{Z}} = \min_{\boldsymbol{g}\in P(r)} \overline{\mathcal{Z}}_{\boldsymbol{g}}.$$

4 RELATION TO MAP INFERENCE

We show that our variational upper bound $\overline{\mathcal{Z}}$ is proportional to the MAP value of the posterior distribution (2). More specifically, we show the following.

Theorem 4.1. *We have*

$$\overline{\mathcal{Z}} = \sqrt{\frac{(2\pi)^p}{\det \boldsymbol{M}_{\boldsymbol{X},\sigma,\eta}}} \cdot \max_{\boldsymbol{w}\in\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 - \frac{\eta}{2}\|\boldsymbol{w}\|_2^2 - \lambda r(\boldsymbol{w})}.$$

We will see in Section 5 that this form is useful for the efficient computation of $\overline{\mathcal{Z}}$.

To prove Theorem 4.1, we need the following auxiliary lemma.

Lemma 4.2. *Let $h : \mathbb{R}^p \to \mathbb{R}$ be a function defined as $h(\boldsymbol{g}) = \frac{1}{2}\|\boldsymbol{b}+\boldsymbol{g}\|^2_{\boldsymbol{A}^{-1}}$, where $\boldsymbol{g} \in \mathbb{R}^p$ is a vector and $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ is a positive-definite matrix. Then, the Fenchel conjugate $h^* : \mathbb{R}^p \to \mathbb{R}$ is of the form*

$$h^*(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2_{\boldsymbol{A}} - \boldsymbol{b}^\top \boldsymbol{w}.$$

Proof. By definition, $h^*(\boldsymbol{w}) = \sup_{\boldsymbol{g}\in\mathbb{R}^p}\{\boldsymbol{g}^\top \boldsymbol{w} - h(\boldsymbol{g})\}$. As h is strongly convex, the supremum is attained at the point $\boldsymbol{g}_{\boldsymbol{w}} \in \mathbb{R}^p$ such that

$$\nabla h(\boldsymbol{g}_{\boldsymbol{w}}) = \boldsymbol{A}^{-1}(\boldsymbol{b}+\boldsymbol{g}_{\boldsymbol{w}}) = \boldsymbol{w},$$

which implies $\boldsymbol{g}_{\boldsymbol{w}} = \boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}$. Therefore, we have

$$\begin{aligned} h^*(\boldsymbol{w}) &= (\boldsymbol{A}\boldsymbol{w}-\boldsymbol{b})^\top \boldsymbol{w} - h(\boldsymbol{A}\boldsymbol{w}-\boldsymbol{b}) \\ &= (\boldsymbol{A}\boldsymbol{w}-\boldsymbol{b})^\top \boldsymbol{w} - \frac{1}{2}\|\boldsymbol{A}\boldsymbol{w}\|^2_{\boldsymbol{A}^{-1}} = \frac{1}{2}\|\boldsymbol{w}\|^2_{\boldsymbol{A}} - \boldsymbol{b}^\top \boldsymbol{w}. \end{aligned}$$

□

Proof of Theorem 4.1. We start with the following observation:

$$\begin{aligned} \min_{\boldsymbol{g}\in P(r)} \Psi(\boldsymbol{g}) &= \min_{\boldsymbol{g}\in P(\lambda r)} \Phi(-\boldsymbol{g}) \\ &= -\min_{\boldsymbol{w}\in\mathbb{R}^p} \Big\{ \Phi^*(\boldsymbol{w}) + \lambda r(\boldsymbol{w}) \Big\}, \end{aligned} \qquad (5)$$

where

$$\Phi(\boldsymbol{g}) := \frac{1}{2}\Big\| \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2} + \boldsymbol{g} \Big\|^2_{\boldsymbol{M}^{-1}_{\boldsymbol{X},\sigma,\eta}},$$

the function $\Phi^* : \mathbb{R}^p \to \mathbb{R}$ is the Fenchel conjugate of Φ, and the second equality is due to the Fenchel duality. By Lemma 4.2, we have

$$\Phi^*(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2_{\boldsymbol{M}_{\boldsymbol{X},\sigma,\eta}} - \frac{1}{\sigma^2}\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w}.$$

We note that

$$\begin{aligned} &\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 + \frac{\eta}{2}\|\boldsymbol{w}\|_2^2 \\ &= \frac{1}{2}\|\boldsymbol{w}\|^2_{\boldsymbol{M}_{\boldsymbol{X},\sigma,\eta}} - \frac{1}{\sigma^2}\boldsymbol{y}^T \boldsymbol{X}\boldsymbol{w} + \frac{1}{2\sigma^2}\|\boldsymbol{y}\|_2^2. \end{aligned}$$

Combining these two equalities, we have

$$\Phi^*(\boldsymbol{w}) = \frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 + \frac{\eta}{2}\|\boldsymbol{w}\|_2^2 - \frac{1}{2\sigma^2}\|\boldsymbol{y}\|_2^2. \qquad (6)$$

By (4), (5), and (6), we have

$$\begin{aligned} \overline{\mathcal{Z}} &= \min_{\boldsymbol{g}\in P(r)} \overline{\mathcal{Z}}_{\boldsymbol{g}} = e^{-\frac{1}{2\sigma^2}\|\boldsymbol{y}\|_2^2} \sqrt{\frac{(2\pi)^p}{\det \boldsymbol{M}_{\boldsymbol{X},\sigma,\eta}}} e^{\min_{\boldsymbol{g}\in P(r)} \Psi(\boldsymbol{g})} \\ &= e^{-\frac{1}{2\sigma^2}\|\boldsymbol{y}\|_2^2} \sqrt{\frac{(2\pi)^p}{\det \boldsymbol{M}_{\boldsymbol{X},\sigma,\eta}}} \\ &\quad \times e^{-\min_{\boldsymbol{w}\in\mathbb{R}^p}\left\{\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 + \frac{\eta}{2}\|\boldsymbol{w}\|_2^2 - \frac{1}{2\sigma^2}\|\boldsymbol{y}\|_2^2 + \lambda r(\boldsymbol{w})\right\}} \\ &= \sqrt{\frac{(2\pi)^p}{\det \boldsymbol{M}_{\boldsymbol{X},\sigma,\eta}}} \cdot \max_{\boldsymbol{w}\in\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 - \frac{\eta}{2}\|\boldsymbol{w}\|_2^2 - \lambda r(\boldsymbol{w})}. \end{aligned}$$

□

5 VARIATIONAL UPPER BOUNDS

5.1 GENERAL CASE

From Theorem 4.1, it suffices to solve the following problem for the computation of $\overline{\mathcal{Z}}$:

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 + \frac{\eta}{2}\|\boldsymbol{w}\|_2^2 + \lambda r(\boldsymbol{w}). \qquad (7)$$

The objective in Eq. (7) is the sum of a quadratic function and the Lovász extension of a submodular function, which is non-smooth convex.

Therefore, it can be solved by the proximal gradient method, and, as shown by Bach [3], its proximal operator can be reduced to the minimization of the ℓ_2-norm over polyhedron $P(r)$. Moreover, if $r = f$, this problem is the so-called minimum-norm-point (MNP) problem and is solvable efficiently by the MNP algorithm. Alternatively, in many practical cases, it can be solved much more efficiently with parametric maximum flow algorithms such as those given by Gallo *et al.* [12] (see, for example, [31]). Meanwhile, when $r = |f|$, the proximal operator for problem (7) is obtained by, first, solving the minimization of the ℓ_2-norm over polyhedron $P(r) = B(F)$ and then by converting the solution to the one for $P(r) = |P|(F)$ as in Proposition 8.9 of [3].

In the next section, We discuss some special cases in which the polyhedron $P(r)$ can be simplified, allowing us to directly compute $\min_{\boldsymbol{g}\in P(r)} \Psi(\boldsymbol{g})$ efficiently.

5.2 SPECIAL CASES

Recall that $\overline{\mathcal{Z}}$ is computed by solving $\min_{\boldsymbol{g}\in P(r)} \Psi(\boldsymbol{g})$ by Eq. (4). We here discuss some special cases in which $P(r)$ can be simplified, allowing us to directly compute $\min_{\boldsymbol{g}\in P(r)} \Psi(\boldsymbol{g})$ efficiently.

Lasso: In this case, we have $r(\boldsymbol{x}) = \sum_{i\in[p]} |\boldsymbol{x}(i)|$. The function r can be identified with $|f| : \mathbb{R}^p \to \mathbb{R}$ associated with a submodular function $F : 2^{[p]} \to \mathbb{R}$ defined as $F(S) = |S|$. Then, a vector $\boldsymbol{g} \in \mathbb{R}^p$ belongs to $|P|(F)$ if and only if $|\boldsymbol{g}|(S) \leq |S|$ for every $S \subseteq [p]$, which is equivalent to $|\boldsymbol{g}|(i) \leq 1$. Hence, we have $P(r) = [-1, 1]^p$.

As is well known, problem (7) for $r(\boldsymbol{x}) = \sum_{i\in[p]} |\boldsymbol{x}(i)|$ can be solved efficiently by the proximal gradient method with soft-thresholding. Thus, the posterior inference of the distribution (2) can be performed using the method described in Section 5. Meanwhile, using the above structure of $P(r)$, we can minimize $\Psi(\boldsymbol{g})$ directly by the proximal gradient method as follows. For a set $\mathcal{C} \subseteq \mathbb{R}$, we define the indicator function $\iota_{\mathcal{C}}$ as $\iota_{\mathcal{C}}(\boldsymbol{x})(i) = 0$ if $\boldsymbol{x}(i) \in \mathcal{C}$ and $\iota_{\mathcal{C}}(\boldsymbol{x})(i) = \infty$ otherwise for any vector $\boldsymbol{x}$ (of arbitrary dimension). Now, computing $\min_{\boldsymbol{g}\in P(r)} \Psi(\boldsymbol{g})$ for Lasso is equivalent to solving the following constrained problem:

$$\min_{\boldsymbol{g}\in\mathbb{R}^p} \frac{1}{2}\Big\|\frac{\boldsymbol{X}^\top\boldsymbol{y}}{\sigma^2} - \lambda\boldsymbol{g}\Big\|^2_{\boldsymbol{M}^{-1}_{\boldsymbol{X},\sigma,\eta}} + \sum_{i\in[p]} \iota_{[-1,1]}(\boldsymbol{g})(i). \quad (8)$$

The first term is a differentiable convex function and the second is a non-differentiable convex function. Hence, we can minimize (8) by the following iterative updating procedure: $\boldsymbol{g}^{t+1} = \mathrm{prox}_{\iota_{[-1,1]}}(\boldsymbol{g}^t - \eta^t\nabla(\boldsymbol{g}))$ for $t = 1, 2, \ldots$, where $\eta^t > 0$ is the step size for the t-th iteration. Moreover, $\mathrm{prox}_{\iota_{[-1,1]}} a$ is -1 if $a < -1$, a if $-1 \leq a \leq 1$, and 1 if $a > 1$.

Generalized fused Lasso: In this case, we have $r(\boldsymbol{x}) = \sum_{\{i,j\}\in E} r_{ij}(\boldsymbol{x})$, where E is a set of (unordered) pairs and $r_{ij} : \mathbb{R}^p \to \mathbb{R}$ is a function defined as $r_{ij}(\boldsymbol{x}) = |\boldsymbol{x}(i) - \boldsymbol{x}(j)|$. Then, the function r is the Lovász extension of a submodular function $F = \sum_{\{i,j\}\in E} F_{ij}$, where $F_{ij} : 2^{[p]} \to \mathbb{R}$ is a submodular function defined as

$$F_{ij}(S) = [(i \in S \wedge j \notin S) \vee (i \notin S \wedge j \in S)],$$

where $[X]$ is one if the predicate X evaluates to true and is zero otherwise.

Now, a vector $\boldsymbol{g} \in \mathbb{R}^p$ belongs to $B(F_{ij})$ if and only if $\boldsymbol{g}(S) = 0$ for $S \subseteq [p]$ with either $i, j \in S$ or $i, j \notin S$ and $\boldsymbol{g}(S) \leq 1$ for other S's. It follows that $\boldsymbol{g}(i) \leq 1$, $\boldsymbol{g}(j) \leq 1$, $\boldsymbol{g}(i) + \boldsymbol{g}(j) = 0$, and $\boldsymbol{g}(k) = 0$ for every $k \in [p] \setminus \{i, j\}$. This means that $\boldsymbol{g}$ is of the form $p_{ij}(\boldsymbol{\chi}_i - \boldsymbol{\chi}_j)$ for $-1 \leq p_{ij} \leq 1$, where $\boldsymbol{\chi}_i \in \mathbb{R}^p$ and $\boldsymbol{\chi}_j \in \mathbb{R}^p$ are the i-th and j-th characteristic vectors, respectively. It follows that a vector $\boldsymbol{g} \in \mathbb{R}^p$ belongs to $B(F)$ if and only if it is of the form $\sum_{\{i,j\}\in E} p_{ij}(\boldsymbol{\chi}_i - \boldsymbol{\chi}_j)$ and $p_{ij} \in [-1, 1]$ for every $\{i, j\} \in E$. Hence, we have

$$P(r) = \Big\{ \sum_{\{i,j\}\in E} p_{ij}(\boldsymbol{\chi}_i - \boldsymbol{\chi}_j) \mid p_{ij} \in [-1, 1]\ \forall\{i, j\} \in E \Big\}.$$

Computing $\min_{\boldsymbol{g}\in P(r)} \Psi(\boldsymbol{g})$ for GFL is equivalent to solving the following problem:

$$\min_{p_{ij}\in\mathbb{R}\ \forall\{i,j\}\in E} \frac{1}{2}\|\boldsymbol{X}^\top\boldsymbol{y} - \lambda \sum_{\{i,j\}\in E} p_{ij}(\boldsymbol{\chi}_i - \boldsymbol{\chi}_j)\|^2_{\boldsymbol{M}^{-1}_{\boldsymbol{X},\sigma,\eta}} + \sum_{\{i,j\}\in E} \iota_{[-1,1]}(p_{ij}). \quad (9)$$

As with the case of Lasso, we can minimize Eq. (9) by the following simple iterative updating procedure: $p^{t+1}_{ij} = \mathrm{prox}_{\iota_{[-1,1]}}(p^t_{ij} - \eta^t_{ij}\nabla(p^t_{ij}))$, where $\eta^t_{ij} > 0$ is the step size for the edge $\{i, j\}$ in the t-th iteration.

Generalized isotonic regression: In the soft-constrained variant of generalized isotonic regression, we have $r(\boldsymbol{x}) = \sum_{(i,j)\in E} r_{ij}(\boldsymbol{x})$, where E is a set of (ordered) pairs and $r_{ij} \colon \mathbb{R}^p \to \mathbb{R}$ is defined as $r_{ij}(\boldsymbol{x}) = \max\{\boldsymbol{x}(i) - \boldsymbol{x}(j), 0\}$. Then, the function r is the Lovász extension of a submodular function $F = \sum_{(i,j)\in E} F_{ij}$, where $F_{ij} : 2^{[p]} \to \mathbb{R}$ is a submodular function defined as $F_{ij}(S) = [i \in S \wedge j \notin S]$.

Now, a vector $\boldsymbol{g} \in \mathbb{R}^p$ belongs to $B(F_{ij})$ if and only if $\boldsymbol{g}(S) \leq 1$ for $S \subseteq [p]$ with $i \in S$ and $j \notin S$ and

$\boldsymbol{g}(S) = 0$ for other S's. It follows that $\boldsymbol{g}(i) \leq 1$, $\boldsymbol{g}(j) \leq 0$, $\boldsymbol{g}(i) + \boldsymbol{g}(j) = 0$, and $\boldsymbol{g}(k) = 0$ for every $k \in [p] \setminus \{i, j\}$. This means that $\boldsymbol{g}$ is of the form $p_{ij}(\boldsymbol{\chi}_i - \boldsymbol{\chi}_j)$ for $0 \leq p_{ij} \leq 1$. It follows that a vector $\boldsymbol{g} \in \mathbb{R}^p$ belongs to $B(F)$ if and only if it is of the form $\sum_{(i,j)\in E} p_{ij}(\boldsymbol{\chi}_i - \boldsymbol{\chi}_j)$ and $p_{ij} \in [0, 1]$ for every $(i, j) \in E$. Hence, we have

$$P(r) = \Big\{ \sum_{\{i,j\}\in E} p_{ij}(\boldsymbol{\chi}_i - \boldsymbol{\chi}_j) \mid p_{ij} \in [0, 1] \; \forall (i, j) \in E \Big\}.$$

The rest of the argument is very similar to the case of the generalized fused Lasso. The only difference is the replacement of $\iota_{[-1,1]}$ with $\iota_{[0,1]}$ in Eq. (9).

Group Lasso: In group Lasso, $r(\boldsymbol{x}) = \sum_{G\in\mathcal{G}} r_G(\boldsymbol{x})$, where $\mathcal{G}$ is a family of disjoint subsets of $[p]$ and $r_G(\boldsymbol{x}) = \max_{g\in G} |\boldsymbol{x}(i)|$. Then, the function r can be identified with $|f| : \mathbb{R}^p \to \mathbb{R}$ associated with a submodular function $F = \sum_{G\in\mathcal{G}} F_G$, where $F_G : 2^{[p]} \to \mathbb{R}$ is a submodular function defined as $F(S) = [S \cap G \neq \emptyset]$.

A vector $\boldsymbol{g} \in \mathbb{R}^p$ belongs to $|P|(F_G)$ if and only if $|\boldsymbol{g}|(S) \leq [S \cap G \neq \emptyset]$ for every $S \subseteq [p]$ which means that $\boldsymbol{g}(i) = 0$ for every $i \in [p] \setminus G$ and $\boldsymbol{g}(G) \leq 1$. It follows that a vector $\boldsymbol{g} \in \mathbb{R}^p$ belongs to $|P|(F)$ if and only if $|\boldsymbol{g}|(G) \leq 1$ for every $G \in \mathcal{G}$. Hence, we have

$$P(r) = \{\boldsymbol{g} \in \mathbb{R}^p \mid |\boldsymbol{g}|(G) \leq 1 \; \forall G \in \mathcal{G}\}.$$

Then, the computation of $\min_{\boldsymbol{g}\in P(r)} \Psi(\boldsymbol{g})$ for group Lasso is equivalent to solving the following problem:

$$\min_{\boldsymbol{g}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{g}\|^2_{\boldsymbol{M}^{-1}_{\boldsymbol{X},\sigma,\eta}} + \sum_{G\in\mathcal{G}} \iota_{[0,1]}(|\boldsymbol{g}|(G)).$$

We regard $|\boldsymbol{g}|(G)$ as a one-dimensional vector. The proximity operator for group Lasso is a projection onto the ℓ_1-ball [9]. Because the second term is a sum of indicator functions, we obtain an upper bound for group Lasso by simply employing a generalization of the proximal gradient method [23].

6 APPROXIMATION GUARANTEE

In Section 3, we saw that $\overline{\mathcal{Z}}$ is an upper bound on $\mathcal{Z}$. In this section, we derive an approximation ratio for $\overline{\mathcal{Z}}$. More specifically, we show the following.

Theorem 6.1. *We have*

$$\mathcal{Z} \leq \overline{\mathcal{Z}} \leq \exp\Big(\frac{B}{2}\Big(1 + \frac{2p}{\eta} + \|X^+\boldsymbol{y}\|_2^2\Big)\Big)\mathcal{Z},$$

where B is the maximum ℓ_2-norm of a point in $P(r)$ and X^+ is the pseudoinverse of X.

As we have mentioned, any extreme point of the base polytope of a submodular function can be constructed by Edmonds' algorithm [10]. Then, we can easily derive that $B = \sqrt{p}$ for Lasso, $B = \sqrt{|\mathcal{G}|}$ for group Lasso, and B is upper-bounded by the square root of the maximum size of a cut for generalized fused Lasso and that of a directed cut for generalized isotonic regression.

To derive a lower bound on $\mathcal{Z}$, we give an upper bound on $r(\boldsymbol{w})$:

Lemma 6.2. *For any $\boldsymbol{w} \in \mathbb{R}^p$, we have*

$$r(\boldsymbol{w}) \leq \frac{B}{2}(1 + \|\boldsymbol{w}\|_2^2),$$

where B is the maximum ℓ_2-norm of a point in $P(r)$.

Proof. We have

$$r(\boldsymbol{w}) = \max_{\boldsymbol{z}\in P(r)} \boldsymbol{z}^\top \boldsymbol{w} \leq \max_{\boldsymbol{z}\in P(r)} \|\boldsymbol{z}\|_2 \|\boldsymbol{w}\|_2$$
$$\leq B\|\boldsymbol{w}\|_2 \leq \frac{B}{2}(1 + \|\boldsymbol{w}\|_2^2). \qquad \square$$

Proof of Theorem 6.1. By Lemma 6.2, we can lower-bound $\mathcal{Z}$ by

$$\underline{\mathcal{Z}}_{\boldsymbol{g}} := \int_{\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 - \frac{(\eta+B)}{2}\|\boldsymbol{w}\|_2^2 - \frac{B}{2}} \mathrm{d}\boldsymbol{w}.$$

Following the calculations in Sections 3 and 4, we obtain

$$\underline{\mathcal{Z}} := \min_{\boldsymbol{g}\in B(F)} \underline{\mathcal{Z}}_{\boldsymbol{g}}$$

$$= \sqrt{\frac{(2\pi)^p}{\det \boldsymbol{M}_{\boldsymbol{X},\sigma,\eta+B}}} \times \max_{\boldsymbol{w}\in\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 - \frac{\eta+B}{2}\|\boldsymbol{w}\|_2^2 - \frac{B}{2}},$$

which is also a lower bound on $\mathcal{Z}$.[1]

Let $\boldsymbol{w}^*$ be the maximizer in the maximization problem in the statement of Theorem 4.1. Then, we have

$$\overline{\mathcal{Z}} = \sqrt{\frac{(2\pi)^p}{\det \boldsymbol{M}_{\boldsymbol{X},\sigma,\eta}}} \times e^{-\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}^*-\boldsymbol{y}\|_2^2 - \frac{\eta}{2}\|\boldsymbol{w}^*\|_2^2 - \lambda r(\boldsymbol{w}^*)}.$$

$$\underline{\mathcal{Z}} \geq \sqrt{\frac{(2\pi)^p}{\det \boldsymbol{M}_{\boldsymbol{X},\sigma,\eta+B}}} \times e^{-\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w}^*-\boldsymbol{y}\|_2^2 - \frac{\eta+B}{2}\|\boldsymbol{w}^*\|_2^2 - \frac{B}{2}}.$$

[1] Although $\max_{\boldsymbol{g}\in B(F)} \underline{\mathcal{Z}}_{\boldsymbol{g}}$ is a tighter lower bound on $\mathcal{Z}$, the present lower bound is more convenient for our analysis.

Further, we note that $\|\boldsymbol{w}^*\|_2 \leq \|X^+\boldsymbol{y}\|_2$. It follows that

$$\frac{\overline{\mathcal{Z}}}{\mathcal{Z}} \leq \frac{\overline{\mathcal{Z}}}{\underline{\mathcal{Z}}} \leq \sqrt{\frac{\det \boldsymbol{M}_{\boldsymbol{X},\sigma,\eta+B}}{\det \boldsymbol{M}_{\boldsymbol{X},\sigma,\eta}}} e^{\frac{B}{2}\|\boldsymbol{w}^*\|_2^2 + \frac{B}{2} - \lambda r(\boldsymbol{w}^*)}$$
$$\leq \left(1+\frac{B}{\eta}\right)^p e^{\frac{B}{2}(1+\|X^+\boldsymbol{y}\|_2^2)} \leq e^{\frac{B}{2}\left(1+\frac{2p}{\eta}+\|X^+\boldsymbol{y}\|_2^2\right)}$$

and the claim holds. □

7 EXTENSION TO KERNEL REGRESSION

In this section, we consider the extension to kernel regression model [25]. In this model, the negative log posterior density for $\boldsymbol{w} \mid \boldsymbol{y}, F$ is given by

$$\frac{1}{2\sigma^2}\|\boldsymbol{K}\boldsymbol{w}-\boldsymbol{y}\|_2^2 + \frac{\eta}{2}\|\boldsymbol{w}\|_{\boldsymbol{K}}^2 + \lambda r(\boldsymbol{w}),$$

where $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ is the kernel matrix. Then, the partition function is given by

$$\mathcal{Z} = \int_{\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{K}\boldsymbol{w}-\boldsymbol{y}\|_2^2 - \frac{\eta}{2}\|\boldsymbol{w}\|_{\boldsymbol{K}}^2 - \lambda r(\boldsymbol{w})} \mathrm{d}\boldsymbol{w}.$$

and we define an upper bound on $\mathcal{Z}$ by following the argument in Section 3. Then, we obtain the following theorem following the argument in Section 4.

Theorem 7.1. *If $\boldsymbol{K}$ is positive definite, then we have*

$$\overline{\mathcal{Z}} = \sqrt{\frac{(2\pi)^p}{\det \boldsymbol{M}'_{\boldsymbol{K},\sigma,\eta}}} \cdot \max_{\boldsymbol{w}\in\mathbb{R}^p} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{K}\boldsymbol{w}-\boldsymbol{y}\|_2^2 - \frac{\eta}{2}\|\boldsymbol{w}\|_{\boldsymbol{K}}^2 - \lambda r(\boldsymbol{w})},$$

where

$$\boldsymbol{M}'_{\boldsymbol{K},\sigma,\eta} := \boldsymbol{K}^\top\boldsymbol{K}/\sigma^2 + \eta\boldsymbol{K}.$$

The modification to the methods for computing $\overline{\mathcal{Z}}$ described in Section 5 is trivial.

8 EXPERIMENTS

In this section, we investigate the empirical performance of our variational upper bound (VUB) on Bayesian Lasso [22], Bayesian group Lasso [18], Bayesian fused Lasso, Spike-and-Slab model [17], and Spike-and-Slab model for group Lasso [32].

As publicly available data sets, we used diabetes data ($p = 10$, $n = 442$) [11], eye data ($p = 200$, $n = 130$) [4], and colon data ($p = 2000$, $n = 62$) [1] for Lasso, used birthwt data ($p = 16$, $n = 189$) [16], splice data ($p = 28$, $n = 400$) [34], and bardet data ($p = 100$, $n = 120$) [33] for group Lasso, and used two leukemia data ($p = 3564$, $n = 36$ and $p = 7128$, $n = 72$) [29] for fused Lasso. We employed the same group and graph structures as in previous works [29, 19].

Tuning hyperparameters Following the Bayesian Lasso, which puts a gamma prior on λ, our upper bound on the log marginal likelihood in Eq. (4) with a gamma prior for a minimizer $\boldsymbol{g}^* \in \mathbb{R}^p$ of $\min_{\boldsymbol{g}\in P(r)} \Psi(\boldsymbol{g})$ can be rephrased as:

$$\frac{1}{2}\left\|\frac{\boldsymbol{X}^\top\boldsymbol{y}}{\sigma^2} - \lambda\boldsymbol{g}^*\right\|_{\boldsymbol{M}_{\boldsymbol{X},\sigma,\eta}^{-1}}^2 - \frac{1}{2}\log\mathcal{Z}_0$$
$$+ (\alpha-1)\log\lambda - \beta\lambda + C,$$

where $\alpha > 0$ and $\beta > 0$ are the shape and scale parameters for the prior, and C is a constant. We set α to 0.1, and set β to p/n, $|\mathcal{G}|/2$, and $|E|/2$ for Lasso, group Lasso, and fused Lasso, and set η to a small value such as 0.1, respectively, which is default hyperparameters of prior experiments.

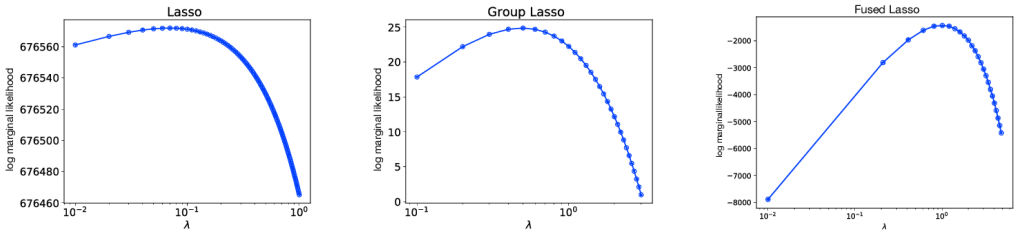
In each of the experiments below, we selected λ that obtained the highest (estimated) marginal likelihood for each dataset. Fig. 1 shows examples of the estimated upper bounds of the marginal likelihood without a constant for some of the datasets.

Prediction performance As we have seen in Section 6, our variational upper bound can approximate the partition function with a guaranteed ratio. To evaluate the effectiveness of our upper bound, we conducted predictive experiments.

For each dataset, we randomly sampled 90% observations as the training data and set the rests as the test data for each experiment. For VUB, we used the MAP solution as an estimator. As baseline methods, we used a Gibbs sampler, which estimates the posterior mean by sampling λ, the MAP solution from five-fold cross validation (CV) with the same λ grid as our method, and the Spike-and-Slab model (Spike-and-Slab)[2]. We ran experiments ten times and measured the root mean squared error (RMSE) for the test data.

Table 1 summarizes the averages and standard deviations of RMSE obtained by our method, the Gibbs sampler, the CV, and the Spike-and-Slab, where boldface denotes the smaller average. We can confirm that our MAP solutions with the selected λ are comparable or even superior to the posterior mean computed by the Gibbs sampler, the CV and the Spike-and-Slab. On colon and leukemia data for which $p \gg n$, the performance of the Gibbs sampler was unstable because its computation contains matrix inversion calculus.

[2]We used spikeslab and MBSGS codes obtained from CRAN



(a) diabetes data (b) bardet data (c) leukemia data

Figure 1: Variational upper bound on log marginal likelihood on real data sets.

Table 1: Average and standard deviation of RMSE

Data set	Model	VUB	Gibbs	CV	Spike-and-Slab
diabetes	Lasso	$\mathbf{0.71 \pm 0.07}$	$\mathbf{0.71 \pm 0.07}$	0.71 ± 0.07	0.93 ± 0.08
eye	Lasso	$\mathbf{0.09 \pm 0.03}$	0.10 ± 0.04	$\mathbf{0.09 \pm 0.03}$	0.68 ± 0.15
colon	Lasso	$\mathbf{0.27 \pm 0.01}$	0.46 ± 0.13	0.30 ± 0.09	0.49 ± 0.12
birthwt	Group Lasso	0.66 ± 0.15	$\mathbf{0.65 \pm 0.16}$	0.66 ± 0.16	0.72 ± 0.23
splice	Group Lasso	$\mathbf{0.28 \pm 0.02}$	$\mathbf{0.28 \pm 0.02}$	$\mathbf{0.28 \pm 0.02}$	1.10 ± 0.10
bardet	Group Lasso	$\mathbf{0.78 \pm 0.46}$	0.88 ± 0.46	0.97 ± 0.97	0.98 ± 0.70
leukemia	Fused Lasso	$\mathbf{0.26 \pm 0.11}$	5.16 ± 1.87	0.47 ± 0.06	-
leukemia (big))	Fused Lasso	$\mathbf{0.24 \pm 0.06}$	12.12 ± 2.79	0.48 ± 0.04	-

Table 2: Robustness against the change of the number of observed samples

	eye data				bardet data		
n	VUB	Gibbs	CV	n	VUB	Gibbs	CV
120	$\mathbf{0.09 \pm 0.04}$	0.10 ± 0.05	$\mathbf{0.09 \pm 0.04}$	120	$\mathbf{0.78 \pm 0.46}$	0.88 ± 0.46	0.97 ± 0.97
100	$\mathbf{0.08 \pm 0.02}$	0.09 ± 0.02	$\mathbf{0.08 \pm 0.02}$	100	$\mathbf{0.78 \pm 0.52}$	0.79 ± 0.52	0.98 ± 1.05
80	$\mathbf{0.08 \pm 0.03}$	0.10 ± 0.03	$\mathbf{0.08 \pm 0.02}$	80	0.80 ± 0.54	$\mathbf{0.77 \pm 0.55}$	0.99 ± 1.13
60	$\mathbf{0.09 \pm 0.03}$	0.10 ± 0.03	$\mathbf{0.09 \pm 0.03}$	60	$\mathbf{0.75 \pm 0.37}$	$\mathbf{0.75 \pm 0.33}$	0.77 ± 0.36
40	0.10 ± 0.04	0.11 ± 0.04	$\mathbf{0.08 \pm 0.04}$	40	$\mathbf{0.75 \pm 0.43}$	0.79 ± 0.38	0.92 ± 0.43

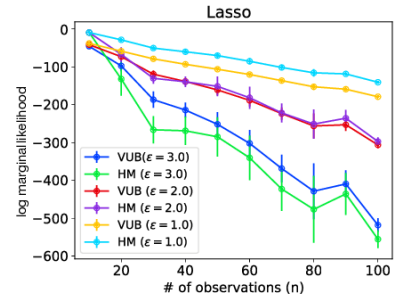
Table 3: Log marginal likelihood on the diabetes data

n	VUB	Gibbs
442	-488.52 ± 1.67	-480.63 ± 48.45
200	-229.85 ± 1.44	-244.47 ± 40.28
100	-118.96 ± 1.00	-141.20 ± 37.75
50	-68.19 ± 0.77	-75.77 ± 18.42
30	-43.18 ± 0.45	-53.18 ± 14.46
20	-30.61 ± 0.35	-37.71 ± 9.85
10	-16.56 ± 0.19	-34.87 ± 6.39

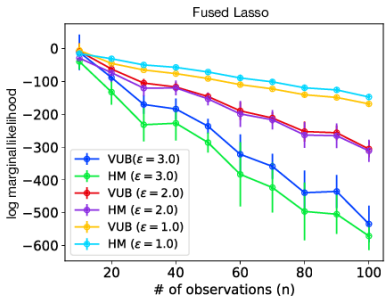
Robustness of prediction Next, we examined the robustness of our method, the Gibbs sampler and the CV on several data sets against the change of the number of observed samples. Table 2 summarizes the averages and standard deviations of RMSE on data sets obtained by subsampling the eye and bardet data. Our method seems to show robust performance comparable or superior to that of baselines the Gibbs sampler and the CV. The computational time for selecting λ on our estimation method was shorter than baseline methods because ours did not need any heavy operation, such as matrix inversions or parameter estimations on validation sets.

Tightness of approximation Finally, we empirically examined the tightness of our variational upper bound on synthetic and diabetes data sets. Although it has been known as a loose estimator of the marginal likelihood, we compared our method with the harmonic mean estimator [20] of the Gibbs sampler because no other method has been proposed to evaluate the marginal likelihood of a hierarchical probabilistic model such as we considered in this work.

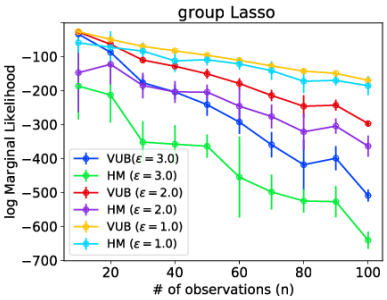
We now describe how we generated synthetic data sets. In the Lasso experiment, all the elements of $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and 20% of elements of $\boldsymbol{w} \in \mathbb{R}^p$ were sampled from the standard normal distribution independently from others. Other elements of $\boldsymbol{w}$ were set to 0. Then, we sampled $\boldsymbol{y}(i)$ for each $i \in [p]$ from the normal distribution with mean $(\boldsymbol{X}\boldsymbol{w})(i)$ and variance $\epsilon > 0$ independently from others. In the fused Lasso experiment, we set 50% of elements of $\boldsymbol{w} \in \mathbb{R}^p$ to -1



(a) Lasso



(b) Fused Lasso



(c) Group Lasso

Figure 2: Log marginal likelihood on synthetic data sets

and the remaining elements to 1. As the underlying graph, we used a path of length $p-1$. Other variables, that is, $\boldsymbol{X}$ and $\boldsymbol{y}$ were generated as in the Lasso experiment.

We run five experiments and calculated the averages and standard deviations of estimators. We also examined the approximated marginal log-likelihood on diabetes data, whose p is the smallest among the datasets considered, with different numbers of observations.

The graphs in Fig. 2 show estimated marginal likelihoods for varying choices of n and ϵ when $p = 10$, $\lambda = 1, \eta = 0.01$. The estimated upper bounds obtained by our method were close to those of the harmonic mean method, and our method seems to show smaller variances than the harmonic mean method. And, Table 3 shows the averages and standard deviations on the diabetes data for the same setting. This also seems to show that our method provides average values similar to the harmonic mean with more stable calculation.

9 CONCLUSIONS

In this paper, we have developed a hierarchical probabilistic formulation of linear regression and its kernel extension regularized by the Lovász extensions of submodular functions, and have proposed a variational inference scheme for posterior inference on this model. We first developed an upper bound on the partition function, and showed that calculating our variational upper bound can be seen as MAP inference on the posterior distribution with a special condition. Then, we showed that minimizing this bound is equivalent to the minimization of a quadratic function over the polyhedron determined by the corresponding submodular function. For some special cases, the minimization problems are equal to constrained convex minimization problems, which can be solved efficiently by the proximal gradient algorithm and its variants. Furthermore, we showed an approximation guarantee of our variational upper bound. Our scheme gives a natural extension of the Bayesian Lasso for MAP estimation to a variety of regularizers inducing structured sparsity, and thus this work provides a principled way to transfer the advantages of the Bayesian formulation into those models. Finally, we empirically confirmed the effectiveness of our scheme with several datasets.

This work was supported by JSPS KAKENHI Grant Numbers 18H03287 and 18H05291 and JST ACT-I Grant Number JPMJPR18UG.

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

[2] A. Armagan. Variational bridge regression. In *Proceedings of AISTATS*, pages 17–24, 2009.

[3] F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6:145–373, 2013.

[4] A. P. Chiang, J. S. Beck, H.-J. Yen, M. K. Tayeh, T. E. Scheetz, R. E. Swiderski, D. Y. Nishimura,

T. A. Braun, K.-Y. A. Kim, J. Huang, et al. Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a bardet–biedl syndrome gene (BBS11). *Proceedings of the National Academy of Sciences*, 103(16):6287–6292, 2006.

[5] A. Defazio and T. S. Caetano. A Convex Formulation for Learning Scale-Free Networks via Submodular Relaxation. In *Proceedings of NIPS*, pages 1250–1258, 2012.

[6] J. Djolonga and A. Krause. From MAP to marginals: Variational inference in Bayesian submodular models. In *Proceedings of NIPS*, pages 244–252, 2014.

[7] J. Djolonga and A. Krause. Scalable variational inference in log-supermodular models. In *Proceedings of ICML*, pages 1804–1813, 2015.

[8] J. Djolonga, S. Tschiatschek, and A. Krause. Variational inference in mixed probabilistic submodular models. In *Proceedings of NIPS*, pages 1759–1767, 2016.

[9] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of ICML*, pages 272–279, 2008.

[10] J. Edmonds. Matroids and the greedy algorithm. *Mathematical Programming*, 1(1):127–136, 1971.

[11] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[12] G. Gallo, M. D. Grigoriadis, and R. E. Tarjan. A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, 18(1):30–55, 1989.

[13] T. Hazan and T. S. Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *Proceedings of ICML*, pages 991–998, 2012.

[14] T. Hazan, S. Maji, and T. Jaakkola. On sampling from the Gibbs distribution with random maximum a-posteriori perturbations. In *Proceedings of NIPS*, pages 1268–1276, 2013.

[15] H. Hoefling. A path algorithm for the fused Lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

[16] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[17] H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

[18] M. Kyung, J. Gill, M. Ghosh, G. Casella, et al. Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

[19] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[20] M. A. Newton and A. E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48, 1994.

[21] G. Obozinski and F. Bach. Convex relaxation for combinatorial penalties. Report, 2012.

[22] T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[23] H. Raguet, J. Fadili, and G. Peyré. A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.

[24] S. Raman, T. Fuchs, P. Wild, E. Dahl, and V. Roth. The Bayesian group-Lasso for analyzing contingency tables. In *Proceedings of ICML*, pages 881–888, 2009.

[25] J. Shawe-Taylor, N. Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[26] H. T. C. Stoof, K. B. Gubbels, and D. B. M. Dickerscheid. *Ultracold Quantum Fields*. Theoretical and Mathematical Physics. Springer Netherlands, 2008.

[27] K. Takeuchi, Y. Kawahara, and T. Iwata. Higher-order fused regularization for supervised learning with grouped parameters. In *Proceedings of ECML-PKDD*, pages 577–593, 2015.

[28] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[29] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 67(1):91–108, 2005.

[30] D. Wipfand and S. Nagarajan. Sparse estimation using general likelihoods and non-factorial priors. In *NIPS*, pages 2071–2079. 2009.

[31] B. Xin, Y. Kawahara, Y. Wang, and W. Gao. Efficient generalized fused Lasso and its application to the diagnosis of Alzheimer's disease. In *Proceedings of AAAI*, pages 2163–2169, 2014.

[32] X. Xu, M. Ghosh, et al. Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936, 2015.

[33] Y. Yang and H. Zou. A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141, 2015.

[34] G. Yeo and C. B. Burge. Maximum entropy modeling of short sequence motifs with applications to rna splicing signals. *Journal of Computational Biology*, 11(2-3):377–394, 2004.