Article

Sequential Sampling and Estimation of Approximately Bandlimited Graph Signals

Sijie Lin ¹, Ke Xu ² , Hui Feng ^{1,*} and Bo Hu ¹

¹ Research Center of Smart Networks and Systems, School of Information Science and Technology, Fudan University, Shanghai 200433, China; sjlin18@fudan.edu.cn (S.L.); bohu@fudan.edu.cn (B.H.)

² Center for Intelligent Medical Electronics, School of Information Science and Technology, Fudan University, Shanghai 200433, China; kexu18@fudan.edu.cn

* Correspondence: hfeng@fudan.edu.cn

Abstract: Graph signal sampling has been widely studied in recent years, but the accurate signal models required by most of the existing sampling methods are usually unavailable prior to any observations made in a practical environment. In this paper, a sequential sampling and estimation algorithm is proposed for approximately bandlimited graph signals, in the absence of prior knowledge concerning signal properties. We approach the problem from a Bayesian perspective in which we formulate the signal prior by a multivariate Gaussian distribution with unknown hyperparameters. To overcome the interconnected problems associated with the parameter estimation, in the proposed algorithm, hyperparameter estimation and sample selection are performed in an alternating way. At each step, the unknown hyperparameters are updated by an expectation maximization procedure based on historical observations, and then the next node in the sampling operation is chosen by uncertainty sampling with the latest hyperparameters. We prove that under some specific conditions, signal estimation in the proposed algorithm is consistent. Subsequent validation of the approach through simulations shows that the proposed procedure yields performances which are significantly better than existing state-of-the-art approaches notwithstanding the additional attribute of robustness in the presence of a broad range of signal attributes.

Keywords: graph signal; sequential sampling; consistent estimation

Citation: Lin, S.; Xu, K.; Feng, H.; Hu, B. Sequential Sampling and Estimation of Approximately Bandlimited Graph Signals. *Sensors* **2021**, *21*, 1460. <https://doi.org/10.3390/s21041460>

Academic Editor:
Vladimir Stankovic

Received: 31 December 2020
Accepted: 15 February 2021
Published: 19 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Graph signal is a powerful tool to represent and analyze data with irregular interconnections by defining signal values on graph vertices and assigning edge weights according to correlation or similarity [1]. In the past decade, graph signal processing theory has developed rapidly, extending classical signal processing techniques such as Fourier analysis [1,2], filtering [3–5], sampling and interpolation [6–8] to graph signal setting. Related concepts and methods have found wide application in sensor networks [9,10], brain analysis [11], image processing [12], three-dimensional (3D) point cloud processing [13], and machine learning [14–16].

Sampling for (lossless) reconstruction or (minimum error) estimation is a fundamental problem in graph signal processing. For example, a sample survey in a social network needs to carefully select interviewees so as to better predict the attitudes of all users. In a sensor network, it is important to optimize sensor placement so that more information can be collected by a restricted number of sensors due to economic constraint [9]. In addition, any graph-based active semisupervised learning task may also be interpreted as a graph signal sampling and estimation problem [14,15].

There have been multiple works on graph signal sampling. The authors of [6,7] try to extend the Nyquist sampling theory to bandlimited graph signals. More works are seeking optimal sampling sets for graph signal estimation, where the graph signals are usually assumed to be bandlimited [17–19] or approximately bandlimited [17,20] in the

frequency domain, or smooth in the vertex domain [17,21]. Sampling strategies mainly include topology-based methods that compute a score for each vertex [18,21], and design-of-experiment (DOE) approaches that optimize certain scalarization form of some target matrix [19,22]. Most of the current sampling methods rely on accurate bandwidth [6,18,19] or prior distribution [22,23] of the graph signal.

It is natural in many applications to assume that the graph signal is somehow bandlimited or smooth. For instance, people with strong social connection are likely to hold similar opinions and temperature sensors located close to each other usually get similar measurements. However, in practice it is hard to know the exact bandwidth or prior distribution of the graph signal before any observation. In such cases, those sampling strategies based on accurate signal model become not applicable.

A possible solution is to sample and estimate in a sequential way. Unknown signal properties can be estimated from previous observations, and the subsequent sampling node can be selected under the latest signal model. Such a sequential framework can overcome the difficulty of incomplete system model and take advantage of model-based sample selection and signal estimation [20].

In this work, we investigate the sequential sampling and estimation of approximately bandlimited (ABL) graph signals, whose bandwidth and energy level are both not known in advance. With the aid of a Butterworth low-pass filter, the ABL graph signal is assumed to follow a multivariate Gaussian distribution with unknown hyperparameters. The variance of observation noise is also considered unknown.

In the proposed algorithm, hyperparameter estimation and sample selection are performed in an alternating way. At each step, expectation maximization (EM) [24] is first used to update the maximum marginal likelihood (MML) estimation of unknown hyperparameters based on historical observations. Substituting these latest estimated values into the signal prior and noise distribution, a Bayesian estimation and prediction procedure can then be performed. According to the uncertainty sampling (US) criterion [25], the node with the largest predictive variance is selected as the next node to sample.

In fact, estimating prior distribution from data, namely, empirical Bayes (EB) method, has long been studied [26]. It has been proved that under certain assumptions, the EB posterior distribution of parameter is consistent at its true value [27]. Meanwhile, an alternating approach analogous to the proposed one has been used in sequential DOE for nonlinear least squares (LS) regression. There, parameter estimation is proved to be consistent, and the design is asymptotically optimal [28].

In this paper, we prove that under specific conditions, signal estimation in the proposed algorithm is consistent at the true value. That is, as the number of samples goes to infinity, the estimated graph signal can get arbitrarily close to the true value. The finite-time performance of the proposed algorithm is validated by simulation results. Our algorithm is able to adjust to ABL graph signals with different properties, and it performs better in estimation error than existing methods.

The rest of the paper is organized as follows. We detail the signal prior and observation model in Section 2, and develop a sequential sampling and estimation algorithm in Section 3. The consistency of the proposed algorithm is proved under specific conditions in Section 4, and its finite-time performance is validated by experiments in Section 5. We conclude the paper in Section 6.

Throughout the paper, we use normal-weight italic lowercase letters (e.g., x) to denote scalar variables, bold italic lowercase letters (e.g., $\mathbf{x}$) to denote vectors, bold italic uppercase letters (e.g., $\mathbf{X}$) to denote matrices, and calligraphic uppercase letters (e.g., $\mathcal{X}$) to denote sets.

2. System Model

2.1. Preliminaries

We consider a simple connected weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, where $\mathcal{V}$ is the set of vertices indexed by $\{1, \dots, N\}$; $\mathcal{E}$ is the set of edges between vertex pairs, e.g., $(i, j) \in \mathcal{E}$ denotes an edge between vertex i and vertex j ; and $\mathbf{W}$ is the weighted adjacency

matrix, with w_{ij} the weight of edge between vertex i and vertex j if they are connected, or 0 otherwise. A graph signal that takes a real value on each vertex of the graph can be represented as $\mathbf{f} \in \mathbb{R}^N$, where f_i is the signal value on vertex i .

The Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D} = \text{diag}(d_1, \dots, d_N)$ with $d_i = \sum_{j:(i,j) \in \mathcal{E}} w_{ij}$ is the degree matrix. Being symmetric and positive semi-definite, $\mathbf{L}$ has a spectral decomposition $\mathbf{L} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, where $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues $0 = \lambda_1 < \lambda_2 \leq \dots \leq \lambda_N$, and $\mathbf{V}$ contains the corresponding orthonormal eigenvectors $\{\mathbf{v}_k\}_{k=1}^N$ as columns. These eigenvectors act as graph Fourier basis, and the associated eigenvalues are regarded as graph frequencies [1]. The frequency-domain coefficients $\hat{\mathbf{f}}$ of a graph signal $\mathbf{f}$ can be calculated via the graph Fourier transform (GFT) as $\hat{\mathbf{f}} = \mathbf{V}^T \mathbf{f}$. The graph signal $\mathbf{f}$ can be expressed as $\mathbf{f} = \mathbf{V}\hat{\mathbf{f}}$, which is known as the inverse GFT (IGFT).

2.2. Signal Prior

In this work, the graph signal $\mathbf{f}$ is assumed to be drawn from the following distribution:

$$p(\mathbf{f}; \alpha, \gamma) = \mathcal{N}(\mathbf{f} | \mathbf{0}, (\alpha \mathbf{I}_N + \gamma \mathbf{L}^{2n})^{-1}), \quad (1)$$

where $\mathbf{I}_N$ denotes the $N \times N$ identity matrix, and $n, \alpha, \gamma > 0$ are hyperparameters, among which α, γ are assumed unknown in prior. The meanings of the signal model and hyperparameters are explained in the frequency domain as follows.

By GFT, the spectral coefficients $\{\hat{f}_k\}_{k=1}^N$ follow independent zero-mean Gaussian distributions with variances conforming to a Butterworth low-pass filter ([29], Section 7.3):

$$p(\hat{f}_k) = \mathcal{N}\left(\hat{f}_k \mid 0, \frac{g^2}{1 + (\lambda_k/\lambda_c)^{2n}}\right), \quad (2)$$

where n is the order, $\lambda_c = \sqrt[n]{\alpha/\gamma}$ is the cut-off frequency, and $g = 1/\sqrt{\alpha}$ is the amplitude gain of the filter. Under such a prior, we say the graph signal is *approximately bandlimited* in a probabilistic sense, where λ_c plays the role of bandwidth, n controls the strictness of bandlimitedness, and g determines the energy level of the graph signal.

Specially, when $n = 1/2$ and $\delta = \alpha/\gamma$ is small, the signal prior Equation (1) becomes

$$p(\mathbf{f}) \propto \exp\left(-\frac{\gamma}{2} \mathbf{f}^T (\mathbf{L} + \delta \mathbf{I}_N) \mathbf{f}\right), \quad (3)$$

which is widely used in graph signal sampling and estimation [9,23]. The restriction on total variation $\mathbf{f}^T \mathbf{L} \mathbf{f} = \sum_{(i,j) \in \mathcal{E}} w_{ij} (f_i - f_j)^2$ can lead to smooth signals over the graph.

If $n \rightarrow \infty$, the out-of-band coefficients become deterministic zeros. The graph signal becomes strictly bandlimited, represented by

$$\mathbf{f} = \mathbf{V}_K \hat{\mathbf{f}}_K, \quad (4)$$

where $K = \max\{k \mid \lambda_k \leq \lambda_c\}$, $\hat{\mathbf{f}}_K$ collects the first K GFT coefficients that may be non-zero, and $\mathbf{V}_K$ contains the first K GFT basis.

In conclusion, our signal prior Equation (1) is flexible to describe approximately bandlimited graph signals with different bandwidth, strictness of bandlimitedness, and energy level. It can also cover the most commonly used smoothness prior [9,23] and bandlimitedness prior [6,7,18,19]. Additionally, this model is simple with three hyperparameters, and can be represented in the vertex domain without eigendecomposition.

Here, hyperparameters α, γ or equivalently λ_c, g in the signal prior are assumed unknown, that is, we do not need to have prior knowledge about the bandwidth or energy level of the graph signal. Only a value for n is required to control the strictness of bandlimitedness of the estimated signal. Compared to those methods for only strictly bandlimited graph signals with known bandwidth or smooth graph signals with known prior distribution, our signal prior is more general.

2.3. Observation Model

To overcome the difficulty caused by unknown hyperparameters, a sequential sampling process is considered, where one node is sampled at each step, so that unknown hyperparameters can be estimated from previous observations, and the next node to sample can be selected using the latest hyperparameters.

Suppose that vertex s_t is sampled at step t , and the observed signal value is y_t . The observation model can be expressed as

$$y_t = \boldsymbol{\psi}_t^T \mathbf{f} + w_t, \quad (5)$$

where $\boldsymbol{\psi}_t^T = \mathbf{e}_{s_t}^T$ is the sampling vector with its s_t -th element equal to 1 and others equal to 0, and w_t is an additive zero-mean Gaussian observation noise with unknown precision β ,

$$p(w_t; \beta) = \mathcal{N}(w_t | 0, \beta^{-1}). \quad (6)$$

In hyperparameter, signal estimation, and sample selection at step t , the historical observations from step 1 to step t are considered together. Let

$$\mathbf{y}_{1:t} = \begin{bmatrix} y_1 \\ \vdots \\ y_t \end{bmatrix}, \quad \boldsymbol{\Psi}_{1:t} = \begin{bmatrix} \boldsymbol{\psi}_1^T \\ \vdots \\ \boldsymbol{\psi}_t^T \end{bmatrix} = \begin{bmatrix} \mathbf{e}_{s_1}^T \\ \vdots \\ \mathbf{e}_{s_t}^T \end{bmatrix}, \quad \mathbf{w}_{1:t} = \begin{bmatrix} w_1 \\ \vdots \\ w_t \end{bmatrix}. \quad (7)$$

The noise values $w_1, \dots, w_t$ involved in different samples (on either different vertices or the same vertex) are assumed to be independent and identically distributed (i.i.d.). According to Equations (5) and (6),

$$\mathbf{y}_{1:t} = \boldsymbol{\Psi}_{1:t} \mathbf{f} + \mathbf{w}_{1:t}, \quad (8)$$

$$p(\mathbf{w}_{1:t}; \beta) = \mathcal{N}(\mathbf{w}_{1:t} | \mathbf{0}, \beta^{-1} \mathbf{I}_t). \quad (9)$$

2.4. Problem Formulation

Under the signal and observation model described in Sections 2.2 and 2.3, at each step t in the sequential sampling and estimation process, the two core problems are (1) how to estimate the unknown hyperparameters α, γ, β as well as the graph signal $\mathbf{f}$ based on the historical observations $\boldsymbol{\Psi}_{1:t}, \mathbf{y}_{1:t}$

$$\alpha'_t, \gamma'_t, \beta'_t, \mathbf{f}'_t \leftarrow \boldsymbol{\Psi}_{1:t}, \mathbf{y}_{1:t}, \quad (10)$$

where $\alpha'_t, \gamma'_t, \beta'_t$ are the estimated values of hyperparameters at step t , and $\mathbf{f}'_t$ denotes the estimated graph signal at step t , and (2) how to select the next node s_{t+1} to be sampled at step $t + 1$ with the latest estimated values of hyperparameters

$$s_{t+1} \leftarrow \boldsymbol{\Psi}_{1:t}, \mathbf{y}_{1:t}, \alpha'_t, \gamma'_t, \beta'_t. \quad (11)$$

Our ultimate goal is to minimize signal estimation error within given budget, or from another angle, reach certain estimation accuracy with least samples.

3. Algorithm

3.1. Hyperparameter and Signal Estimation

We first focus on unknown hyperparameter estimation from observation data. Here, maximum marginal likelihood (MML) estimation of unknown hyperparameters is adopted:

$$\begin{aligned} \alpha'_t, \gamma'_t, \beta'_t &= \arg \max_{\alpha, \gamma, \beta} p(\mathbf{y}_{1:t} | \boldsymbol{\Psi}_{1:t}; \alpha, \gamma, \beta) \\ &= \arg \max_{\alpha, \gamma, \beta} \int_{\mathbb{R}^N} p(\mathbf{y}_{1:t} | \boldsymbol{\Psi}_{1:t}, \mathbf{f}; \beta) p(\mathbf{f}; \alpha, \gamma) d\mathbf{f}. \end{aligned} \quad (12)$$

Then, the signal posterior given estimated hyperparameters is Gaussian:

$$p(f|\mathbf{\Psi}_{1:t}, \mathbf{y}_{1:t}; \alpha'_t, \gamma'_t, \beta'_t) = \mathcal{N}(f | \boldsymbol{\mu}_t, \mathbf{C}_t), \quad (13)$$

with mean vector and covariance matrix ([30], Section 10.6)

$$\boldsymbol{\mu}_t = \beta'_t \mathbf{C}_t \mathbf{\Psi}_{1:t}^T \mathbf{y}_{1:t}, \quad (14)$$

$$\mathbf{C}_t = (\alpha'_t \mathbf{I}_N + \gamma'_t \mathbf{L}^{2n} + \beta'_t \mathbf{\Psi}_{1:t}^T \mathbf{\Psi}_{1:t})^{-1}. \quad (15)$$

The minimum mean square error (MMSE) or maximum a posteriori (MAP) estimation of the graph signal is $f'_t = \boldsymbol{\mu}_t$.

However, direct optimization of Equation (12) is intractable. We view f as a hidden variable, and introduce expectation maximization (EM) ([24], Section 9.4) in our algorithm to estimate α, γ, β , and thereby obtain the posterior distribution of f . To be concise, the subscripts t and $1:t$ are omitted in the introduction of our EM algorithm.

Consider the l -th EM iteration. In E step, hyperparameters α, γ, β are fixed at $\alpha_{l-1}, \gamma_{l-1}, \beta_{l-1}$, and the signal posterior $p(f|\mathbf{\Psi}, \mathbf{y}; \alpha_{l-1}, \gamma_{l-1}, \beta_{l-1})$ with mean $\boldsymbol{\mu}_{l-1}$ and covariance $\mathbf{C}_{l-1}$ is computed as in Equations (13)–(15).

In M step, α, γ, β are updated according to

$$\begin{aligned} \alpha_l, \gamma_l, \beta_l &= \arg \max_{\alpha, \gamma, \beta} \mathbb{E}_{f|\mathbf{\Psi}, \mathbf{y}; \alpha_{l-1}, \gamma_{l-1}, \beta_{l-1}} [\ln p(\mathbf{y}, f|\mathbf{\Psi}; \alpha, \gamma, \beta)] \\ &= \arg \max_{\alpha, \gamma, \beta} \mathbb{E}_{f|\mathbf{\Psi}, \mathbf{y}; \alpha_{l-1}, \gamma_{l-1}, \beta_{l-1}} [\ln p(\mathbf{y}|\mathbf{\Psi}, f; \beta) + \ln p(f; \alpha, \gamma)], \end{aligned} \quad (16)$$

where $\mathbb{E}_{f|\mathbf{\Psi}, \mathbf{y}; \alpha_{l-1}, \gamma_{l-1}, \beta_{l-1}}[\cdot]$ means the expectation is taken with respect to $p(f|\mathbf{\Psi}, \mathbf{y}; \alpha_{l-1}, \gamma_{l-1}, \beta_{l-1})$ computed at E step, the signal prior $p(f; \alpha, \gamma)$ is given in Equation (1), and the signal likelihood $p(\mathbf{y}|\mathbf{\Psi}, f; \beta) = \mathcal{N}(\mathbf{y}|\mathbf{\Psi}f, \beta^{-1}\mathbf{I})$ according to Equations (8) and (9).

For brevity, denote the expectation operator in Equation (16) as $\mathbb{E}_{l-1}[\cdot]$, and the expectation value as E_{l-1} , which is also the objective value of the optimization problem. By direct calculations, we have

$$\begin{aligned} E_{l-1} &= \frac{1}{2} \ln |\alpha \mathbf{I}_N + \gamma \mathbf{L}^{2n}| - \frac{\alpha}{2} \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{f}] - \frac{\gamma}{2} \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{L}^{2n} \mathbf{f}] \\ &\quad + \frac{M}{2} \ln \beta - \frac{\beta}{2} \mathbb{E}_{l-1}[(\mathbf{y} - \mathbf{\Psi} \mathbf{f})^T (\mathbf{y} - \mathbf{\Psi} \mathbf{f})] + \text{const}, \end{aligned} \quad (17)$$

where M is the sample size; “const” denotes terms that are independent of α, γ, β ; and

$$\begin{aligned} \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{f}] &= \mathbb{E}_{l-1}[(\mathbf{f} - \boldsymbol{\mu}_{l-1})^T (\mathbf{f} - \boldsymbol{\mu}_{l-1}) + 2\boldsymbol{\mu}_{l-1}^T \mathbf{f} - \boldsymbol{\mu}_{l-1}^T \boldsymbol{\mu}_{l-1}] \\ &= \text{tr}(\mathbf{C}_{l-1}) + \boldsymbol{\mu}_{l-1}^T \boldsymbol{\mu}_{l-1}, \end{aligned} \quad (18)$$

$$\begin{aligned} \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{L}^{2n} \mathbf{f}] &= \mathbb{E}_{l-1}[(\mathbf{f} - \boldsymbol{\mu}_{l-1})^T \mathbf{L}^{2n} (\mathbf{f} - \boldsymbol{\mu}_{l-1}) + 2\boldsymbol{\mu}_{l-1}^T \mathbf{L}^{2n} \mathbf{f} - \boldsymbol{\mu}_{l-1}^T \mathbf{L}^{2n} \boldsymbol{\mu}_{l-1}] \\ &= \text{tr}(\mathbf{L}^{2n} \mathbf{C}_{l-1}) + \boldsymbol{\mu}_{l-1}^T \mathbf{L}^{2n} \boldsymbol{\mu}_{l-1}, \end{aligned} \quad (19)$$

$$\begin{aligned} \mathbb{E}_{l-1}[(\mathbf{y} - \mathbf{\Psi} \mathbf{f})^T (\mathbf{y} - \mathbf{\Psi} \mathbf{f})] &= \mathbb{E}_{l-1}[(\mathbf{y} - \mathbf{\Psi} \boldsymbol{\mu}_{l-1})^T (\mathbf{y} - \mathbf{\Psi} \boldsymbol{\mu}_{l-1}) \\ &\quad - 2(\mathbf{y} - \mathbf{\Psi} \boldsymbol{\mu}_{l-1})^T \mathbf{\Psi} (\mathbf{f} - \boldsymbol{\mu}_{l-1}) \\ &\quad + (\mathbf{f} - \boldsymbol{\mu}_{l-1})^T \mathbf{\Psi}^T \mathbf{\Psi} (\mathbf{f} - \boldsymbol{\mu}_{l-1})] \\ &= (\mathbf{y} - \mathbf{\Psi} \boldsymbol{\mu}_{l-1})^T (\mathbf{y} - \mathbf{\Psi} \boldsymbol{\mu}_{l-1}) + \text{tr}(\mathbf{\Psi}^T \mathbf{\Psi} \mathbf{C}_{l-1}). \end{aligned} \quad (20)$$

Note that E_{l-1} is concave with respect to α, γ, β , see Appendix A. The maximization problem in Equation (16) can be solved by any standard tool for convex optimization. Here, the first order conditions of the optimization problem are analyzed to give some insights into the hyperparameter estimation, and then a possible efficient search method based on these conditions is provided in Appendix B.

Take the partial derivatives of E_{l-1} with respect to α, γ, β , and set them to zero:

$$\frac{\partial E_{l-1}}{\partial \alpha} = \frac{1}{2} \text{tr}((\alpha \mathbf{I}_N + \gamma \mathbf{L}^{2n})^{-1}) - \frac{1}{2} \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{f}] = 0, \quad (21)$$

$$\frac{\partial E_{l-1}}{\partial \gamma} = \frac{1}{2} \text{tr}((\alpha \mathbf{I}_N + \gamma \mathbf{L}^{2n})^{-1} \mathbf{L}^{2n}) - \frac{1}{2} \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{L}^{2n} \mathbf{f}] = 0, \quad (22)$$

$$\frac{\partial E_{l-1}}{\partial \beta} = \frac{M}{2\beta} - \frac{1}{2} \mathbb{E}_{l-1}[(\mathbf{y} - \mathbf{\Psi} \mathbf{f})^T (\mathbf{y} - \mathbf{\Psi} \mathbf{f})] = 0. \quad (23)$$

Note that $\text{tr}((\alpha \mathbf{I}_N + \gamma \mathbf{L}^{2n})^{-1})$, $\text{tr}((\alpha \mathbf{I}_N + \gamma \mathbf{L}^{2n})^{-1} \mathbf{L}^{2n})$ and $\frac{M}{\beta}$ can be seen as the prior expectations of $\mathbf{f}^T \mathbf{f}$, $\mathbf{f}^T \mathbf{L}^{2n} \mathbf{f}$ and $\mathbf{w}^T \mathbf{w}$ with respect to $p(\mathbf{f}; \alpha, \gamma)$ and $p(\mathbf{w}; \beta)$. The M step actually looks for a group of hyperparameters that makes these prior expectations equal to their posterior ones. In this sense, the estimated hyperparameters agree with the observations.

Repeat the E and M steps as above until convergence, and we will obtain an MML estimation of unknown hyperparameters α, γ, β , together with a posterior of the graph signal $\mathbf{f}$. This process is summarized in Algorithm 1. Although the estimated hyperparameters are not guaranteed to be globally optimal by MML, we find the performance good enough in experiments.

Algorithm 1 Hyperparameter and signal estimation by EM.

- 1: Initialize $\alpha_0, \gamma_0, \beta_0$.
 - 2: **for** $l = 1, 2, \dots$ **do**
 - 3: Compute $\boldsymbol{\mu}_{l-1}, \mathbf{C}_{l-1}$ as in Equations (14) and (15).
 - 4: Update $\alpha_l, \gamma_l, \beta_l$ according to Equation (16).
 - 5: **end for**
-

3.2. Sample Selection

Having been able to estimate the unknown hyperparameters from previous observations by EM, at each decision step, the subsequent sampling node can be selected using the latest estimated values of hyperparameters.

According to the uncertainty sampling (US) criterion in active learning ([25], Chapter 2), we should scan through all the nodes, and pick the one whose observed signal value we are most uncertain about as the next node to sample. Here, predictive variance is regarded as a measurement of uncertainty.

The predictive distribution of an observation y on vertex s with sampling vector $\boldsymbol{\psi}^T = \mathbf{e}_s^T$ given historical observations $\mathbf{\Psi}_{1:t}, \mathbf{y}_{1:t}$ is

$$p(y | \mathbf{\Psi}_{1:t}, \mathbf{y}_{1:t}, \boldsymbol{\psi}^T; \alpha'_t, \gamma'_t, \beta'_t) = \mathcal{N}(y | \mathbf{e}_s^T \boldsymbol{\mu}_t, \mathbf{e}_s^T \mathbf{C}_t \mathbf{e}_s + (\beta'_t)^{-1}), \quad (24)$$

where $\boldsymbol{\mu}_t$ and $\mathbf{C}_t$ are the posterior mean and covariance of the graph signal $\mathbf{f}$, respectively, given $\mathbf{\Psi}_{1:t}$ and $\mathbf{y}_{1:t}$. The predictive variance consists of two parts: estimative variance $\mathbf{e}_s^T \mathbf{C}_t \mathbf{e}_s$ and noise variance $(\beta'_t)^{-1}$, among which the latter is equal for all vertices. The next sampling node s_{t+1} can be decided by

$$s_{t+1} = \arg \max_{s \in \mathcal{V}} \mathbf{e}_s^T \mathbf{C}_t \mathbf{e}_s. \quad (25)$$

3.3. Sequential Sampling and Estimation

Finally, the complete procedure of the proposed sequential sampling and estimation algorithm for approximately bandlimited graph signals is given in Algorithm 2. At each step, first unknown hyperparameters are re-estimated by EM as in Section 3.1, and then the next node to sample is selected by US as in Section 3.2. The total number of sampling steps T is equal to the sampling budget.

Algorithm 2 Sequential sampling and estimation of approximately bandlimited graph signals.

- 1: Choose the first sampling node s_1 arbitrarily.
 - 2: **for** $t = 1, 2, \dots$ **do**
 - 3: Sample the vertex s_t and obtain an observation y_t .
 - 4: Update hyperparameters $\alpha'_t, \gamma'_t, \beta'_t$ and signal posterior μ_t, C_t by EM as in Algorithm 1.
 - 5: Select the next node to sample s_{t+1} by US as in Equation (25).
 - 6: **end for**
-

Our sequential sampling strategy takes into account not only graph topology, but also previous observations, by both estimating hyperparameters based on historical data and selecting sampling nodes based on signal posterior. Making full use of historical observations and deciding which nodes to sample online, the proposed algorithm is able to cope with the situation where signal and noise distributions are not completely available in advance, and efficiently select samples to estimate the underlying graph signal.

The computational complexity of a decision step in the proposed algorithm is $O(n_t^{\text{EM}} \overline{n_t^{\text{bs}}} N^3)$, where n_t^{EM} is the number of EM iterations, $\overline{n_t^{\text{bs}}}$ is the average iterations of binary search in M step, and N is the graph size. The computational cost is acceptable in view of the performance improvement, especially in time-insensitive scenarios with large observation cost. The proposed method is able to efficiently select samples to give accurate estimation of signals thanks to the sequential framework with EM hyperparameter update, and the signal estimation is consistent.

4. Asymptotic Analysis

In the previous section, we have developed a sequential sampling and estimation algorithm for approximately bandlimited graph signals with an incomplete system model, where hyperparameter estimation and sample selection are performed in an alternating way. The performance of the proposed algorithm will be validated via simulation in the next section, where we will see its efficient sample selection and accurate signal estimation with limited sampling budget and little prior knowledge. In this section, we are to emphasize that the proposed algorithm improves limited-budget performance without sacrifice of consistency, which is not trivial for a sequential decision process.

Intuitively, as the number of observations grows, hyperparameters better fitting the true model will be picked out, really important nodes will be selected to sample, and finally an excellent estimation of the graph signal can be obtained. Here, we try to provide some theoretical support for this intuition. The asymptotic performance of the proposed algorithm is analyzed, and we state and prove the consistency of signal estimation in our method.

Before concentrating on the asymptotic performance of signal estimation, we first investigate the asymptotic behavior of sample selection in the proposed algorithm.

Lemma 1. Let $m_{i,t}$ denote the number of times vertex i is sampled up to step t . If $\lim_{t \rightarrow \infty} \frac{1}{t} \frac{\text{tr}(\alpha'_t \mathbf{I}_N + \gamma'_t \mathbf{L}^{2n})}{N\beta'_t} = 0$,

$$\exists \delta > 0, \liminf_{t \rightarrow \infty} \min_{i \in \mathcal{V}} \frac{m_{i,t}}{t} > \delta.$$

That is, as $t \rightarrow \infty$, all the nodes of the graph will be sampled again and again, each maintaining a sampling ratio greater than δ , including the lowest one. A proof of Lemma 1 is given in Appendix C. Then, two direct corollaries follow.

Corollary 1. $\forall i \in \mathcal{V}, \exists \delta > 0, \liminf_{t \rightarrow \infty} \frac{m_{i,t}}{t} > \delta.$

Corollary 2. $\forall i \in \mathcal{V}, \lim_{t \rightarrow \infty} m_{i,t} = \infty.$

Based on them, we are now to state our main theorem on the consistency of signal estimation in the proposed algorithm.

Theorem 2. *The signal posterior $p(f|\mathbf{Y}_{1:t}, \mathbf{y}_{1:t}; \alpha'_t, \gamma'_t, \beta'_t)$ is consistent at the true value f^* , if EM converges at MML hyperparameters satisfying the condition in Lemma 1.*

The consistency of signal posterior here means, as $t \rightarrow \infty$, the posterior distribution will become a point mass at the true value of the graph signal. A proof of Theorem 2 is provided in Appendix D. Then, we have the following corollary.

Corollary 3. *The MMSE/MAP estimation $f'_t = \mu_t$ is consistent at f^* under the conditions in Theorem 2.*

It ensures that under specific conditions, as the sample size grows infinitely large, the MMSE/MAP estimation can get arbitrarily close to the true signal value, i.e., the graph signal can be estimated with arbitrary accuracy. This guarantees the asymptotic performance of the proposed algorithm in signal estimation.

The condition $\lim_{t \rightarrow \infty} \frac{1}{t} \frac{\text{tr}(\alpha'_t I_N + \gamma'_t L^{2n})}{N\beta'_t} = 0$ is satisfied when $\frac{\alpha'_t}{\beta'_t} = o(t)$, $\frac{\gamma'_t}{\beta'_t} = o(t)$, or more sufficiently, if $\limsup_{t \rightarrow \infty} \alpha'_t < \infty$, $\limsup_{t \rightarrow \infty} \gamma'_t < \infty$, $\liminf_{t \rightarrow \infty} \beta'_t > 0$, which ensures that the estimated signal energy and bandwidth do not decrease to zero, and the estimated noise variance keeps finite. We did not observe any violation of these conditions in plenty of experiments, where the true graph signals do not have all same values, and all observations are finite.

5. Experiments

In this section, the finite-time performance of the proposed algorithm is validated via a series of simulation experiments, on both synthetic and real-world data.

The performance of the proposed algorithm is compared to the following methods:

- RndEM: Select sampling nodes randomly and uniformly, and estimate hyperparameters and the signal using the proposed EM algorithm.
- OrcUS: Suppose that the hyperparameters are known in advance (the “oracle”). Select sampling nodes by US, and estimate the signal by MMSE/MAP.
- Anis2016: A heuristic sampling algorithm proposed in Anis et al. [7] to maximize the cut-off frequency such that the sampling set is a uniqueness set for the bandlimited subspace. The graph signal is recovered in the bandlimited subspace by least squares. The estimation order k of cut-off frequency is set to $k = 4$.
- Perraudin2018: A nonuniform random sampling method proposed in Perraudin et al. [21] with probability relevant to local uncertainty. Inpainting is done by minimizing total variation. The kernel in local uncertainty is $\hat{g}(\lambda) = \exp(-\tau\lambda)$ where $\tau = 8$.
- Sakiyama2019: In this method by Sakiyama et al. [31], the graph signal is recovered by a linear combination of localization operators at the sampled nodes. The sampling set is designed such that the energy of operator at each node is large and meanwhile the overlapping areas are small. The kernel in localization operator is $\hat{g}(\lambda) = \exp(-\tau\lambda)$ where $\tau = 0.5$. The Chebyshev polynomial approximation order is $P = 12$, and the signal recovery order is $k = 12$.

RndEM and OrcUS are references to verify the effectiveness of the proposed algorithm. Anis2016, Perraudin2018, Sakiyama2019 are three state-of-the-art methods that do not directly rely on signal model. For Anis2016, the benefit of further increasing k is not noticeable in our experiments, but the computations will become less numerically stable. For Perraudin2018 and Sakiyama2019, we use the same kernel type as in the original papers, and the parameters are chosen by experiments to adjust to our settings.

The normalized l_2 -norm estimation error

$$e = \frac{\|f' - f^*\|_2}{\|f^*\|_2} \quad (26)$$

is taken as a performance index, where f' is the estimated signal, and f^* is the true value.

5.1. Simulations on Synthetic Data

Three graphs of representative types that have widespread applications in the real world are used for testing in our experiments:

- $\mathcal{G}_1$: A random geometric graph with $N = 100$ vertices randomly placed in a 1-by-1 square, and edge weights assigned via a thresholded Gaussian kernel

$$w_{ij} = \begin{cases} \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right), & \text{if } d_{ij} < r \\ 0, & \text{otherwise} \end{cases} \quad (27)$$

where d_{ij} is the Euclid distance between vertex i and vertex j , $r = 0.2$ and $\sigma = 0.1$.

- $\mathcal{G}_2$: A small world graph with $N = 100$ nodes, generated by the Watts–Strogatz model [32] with mean node degree 6 and rewiring probability 0.2.
- $\mathcal{G}_3$: A scale-free graph with $N = 100$ nodes, generated by the Barabási–Albert model [33,34] with 1 initial node, each newly added node connected to 1 existing node, and connecting probabilities proportional to the degrees of existing nodes.

The graph topologies as well as their eigenvalue distributions are visualized in Figure 1. Due to limited space, we will mainly display and analyze our experimental results on $\mathcal{G}_1$ in detail, whereas example results on $\mathcal{G}_2$ and $\mathcal{G}_3$ are shown at the end to demonstrate the topology adaptability.

Figure 1. The three graphs used in our simulations, and the histograms of their eigenvalue distributions. (a) $\mathcal{G}_1$, a random geometric graph and (d) its eigenvalue distribution. (b) $\mathcal{G}_2$, a Watts–Strogatz small world graph and (e) its eigenvalue distribution. (c) $\mathcal{G}_3$, a Barabási–Albert scale-free graph and (f) its eigenvalue distribution.

Graph signals are generated from the ABL prior distribution Equation (1). Moreover, observation noise follows i.i.d. zero-mean Gaussian distribution Equation (6). A variety of hyperparameters are considered, in order to cover ABL graph signals with different bandwidth, strictness of bandlimitedness and energy level, and scenarios with different signal-noise ratio (SNR). SNR here considering random ABL graph signal is defined as

$$SNR_r = 10 \lg \frac{\sum_{k=1}^N 1/(\alpha + \gamma \lambda_k^{2n})}{N/\beta}. \quad (28)$$

For each setting, 100 graph signals are generated to evaluate the average performance of each method. The average estimation errors under different sample sizes are computed and plotted.

First of all, we compare the performance of the proposed algorithm with RndEM and OrcUS on $\mathcal{G}_1$. Hyperparameters are set to $n = 2$, $\lambda_c = 2$, $g = 1$, and $SNR_r = 15$ dB. As shown in Figure 2, although in the initial stage, the estimation error of the proposed algorithm is similar to RndEM, as the sample size grows, our performance gradually approaches that of OrcUS. This result is as expected, as hyperparameter estimation gets better with growing sample size, which can lead to more efficient sample selection and more accurate signal estimation. Our superiority to RndEM (EM estimation without US sampling) demonstrates the effectiveness of choosing sampling nodes by US. The performance gap between our method and OrcUS (US sampling with known hyperparameters) tends to disappear when samples are abundant, also supporting the validity of hyperparameter estimation by EM.

Figure 2. Estimation error with growing sample size of the proposed algorithm versus RndEM and OrcUS on $\mathcal{G}_1$. Hyperparameters are $n = 2$, $\lambda_c = 2$, $g = 1$, and $SNR_r = 15$ dB.

Subsequently, the proposed algorithm is compared to three existing sampling and estimation methods—Anis2016, Perraudin2018, and Sakiyama2019—in scenarios without much prior knowledge about signal and noise model. A series of experiments with different hyperparameters are done on $\mathcal{G}_1$ to test and compare their performance in all kinds of settings. In the first experiment, we fix $g = 1$, $SNR_r = 15$ dB, and repeat the simulation with $n = 1, 2$, $\lambda_c = 2, 3$, corresponding to ABL graph signals with different bandwidth and strictness of bandlimitedness. Simulation results are depicted in Figure 3. From the graphics we can see, the proposed algorithm can obtain better or competitive estimation performance compared to state-of-the-art methods on ABL graph signals with various properties. For Perraudin2018 and Sakiyama2019, the proposed method significantly outperforms them after a small number of initial samples in all the four settings. As for Anis2016, when n is small or λ_c is large, which means the signal is far from strictly bandlimited or occupies large bandwidth, we need far fewer samples than Anis2016 to recover the signal with tolerable error. This is because our method estimates the graph signal at all frequencies from the beginning thanks to the Bayesian framework, and our

signal prior with hyperparameters estimated from data can better fit ABL graph signals with different properties. Therefore, the proposed algorithm is more robust against different signal properties and is able to obtain better estimation results in most cases.

Figure 3. Estimation errors of the proposed algorithm versus Anis2016, Perraudin2018, and Sakiyama2019 on $\mathcal{G}_1$ with different signal bandwidth and strictness of bandlimitedness, i.e., different n and λ_c . We fix $g = 1$ and $SNR_r = 15$ dB. (a) $n = 1, \lambda_c = 2$. (b) $n = 1, \lambda_c = 3$. (c) $n = 2, \lambda_c = 2$. (d) $n = 2, \lambda_c = 3$.

We continue to investigate our performance for ABL graph signals with different energy level as well as in scenarios with different signal–noise ratio. We fix $n = 2, \lambda_c = 2$. When changing g from 0.01 to 100, SNR_r is fixed at 15 dB. When changing SNR_r from 5 dB to 25 dB, g is fixed at 1. Sampling size is set to $M = 60$. According to Figure 4a, the proposed algorithm gets better performance for graph signals with lower energy. For one thing, even though the estimated variance in the signal prior also becomes larger when the signal energy is larger, the zero-mean Gaussian prior still prefers signals with lower energy. For another, as the variance gets larger, the signal prior becomes flatter, and thus weaker, in the sense that it has less preference among different signals. Therefore, more observations are needed to reach the same estimation accuracy. Fortunately, this problem can be overcome by adaptive rescaling, dividing the observed signal values by a sufficiently large number, as long as we know the rough order of magnitudes of the signal energy. As for SNR, from Figure 4b, although lower SNR may result in worse performance of the proposed algorithm, our method keeps outperforming the others when samples are adequate.

Figure 4. Estimation errors of the proposed algorithm versus Anis2016, Perraudin2018, and Sakiyama2019 on $\mathcal{G}_1$ with different signal energy level and signal–noise ratio, i.e., different g and SNR_r . We fix $n = 2$, $\lambda_c = 2$. The sampling size is set to $M = 60$. (a) g is changed from 0.01 to 100, while SNR_r is fixed at 15 dB. (b) SNR_r is changed from 5 dB to 25 dB, while g is fixed at 1.

Then, example simulation results on $\mathcal{G}_2$ and $\mathcal{G}_3$ are given in Figure 5. $n = 2$, $g = 1$, $SNR_r = 15$ dB. $\lambda_c = 3$ for $\mathcal{G}_2$, and $\lambda_c = 1$ for $\mathcal{G}_3$. The proposed algorithm also reaches higher estimation accuracy than the other methods with enough sampling budget. This demonstrates the robustness of the proposed algorithm against different graph topologies.

Figure 5. Estimation errors of the proposed algorithm versus Anis2016, Perraudin2018, and Sakiyama2019 on $\mathcal{G}_2$ and $\mathcal{G}_3$. (a) Results on $\mathcal{G}_2$ with $n = 2$, $\lambda_c = 3$, $g = 1$, $SNR_r = 15$ dB. (b) Results on $\mathcal{G}_3$ with $n = 2$, $\lambda_c = 1$, $g = 1$, $SNR_r = 15$ dB.

From the simulation results above, without much prior knowledge, the proposed algorithm can adjust to ABL graph signals with different bandwidth, strictness of bandlimitedness, energy level, and SNR. In most cases, exceeding a small number of initial samples, the proposed algorithm is able to estimate the graph signals with higher accuracy than the existing methods.

Besides the validation of the limited-budget performance, we also verify the consistency of the proposed algorithm by a simulation experiment. We generate an ABL graph signal on $\mathcal{G}_1$ with $n = 2$, $\lambda_c = 2$, and $g = 1$, and the SNR is set to 15 dB. To observe the asymptotic performance, the maximum sample size is set to 10^5 , which is 10^3 times the graph size. We repeat the proposed algorithm on the graph signal for 50 times to see the mean as well as the variance of the estimation errors.

As shown in Figure 6, the mean normalized l_2 -norm error tends to decrease continuously as the sample size grows, and has decreased to about 5×10^{-3} when the sample size reaches 10^5 . The standard deviation also tends to decline with growing sample size, which is around 3×10^{-4} with 10^5 samples. Other ABL graph signals also have similar results. This agrees with the consistency of the proposed algorithm, which has been stated and theoretically proved in Section 4.

Figure 6. Result of the consistency verification experiment. The graph signal is on $\mathcal{G}_1$ generated with $n = 2$, $\lambda_c = 2$, $g = 1$, and the observation SNR is 15 dB. The curve denotes the mean of the normalized l_2 -norm error, and the region corresponds to ± 1 standard deviation around the mean. In the overview figure, the y-axis is in logarithmic scale to better display the decrease of the mean estimation error, whereas in the partial enlargement, it is linear to present the decline of the standard deviation.

5.2. A Real-World Example of Temperature Sensor Network

As an application example, we apply the proposed algorithm to the air temperatures measured by weather stations in China [35]. This example shows a potential application of the proposed algorithm in sensor selection.

We construct a K -nearest neighbor (K -NN) graph from the $N = 381$ weather stations with $K = 5$, where each node corresponds to a weather station, and the great-circle distances between stations are considered in K -NN search. Edge weights are assigned via a Gaussian kernel:

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right) \text{ for } (i, j) \in \mathcal{E}, \quad (29)$$

where d_{ij} is the great-circle distance between station i and station j in degree, and $\sigma = 2$.

Then, the air temperatures measured by these weather stations can be seen as a graph signal. We consider two such signals in our experiment: f_1 at 0:00 on 1 January 2020, which is a winter night, and f_2 at 12:00 on 1 July 2020, which is a summer noon. The two graph signals and their spectra are visualized in Figure 7. As expected, the signals are smooth in the vertex domain and approximately bandlimited in the frequency domain.

Figure 7. Two graph signals of air temperatures measured by 381 weather stations in China, and their spectra. (a) f_1 at 0:00 on 1 January 2020, and (b) its spectrum. (c) f_2 at 12:00 on 1 July 2020, and (d) its spectrum.

We apply the proposed sequential sampling and estimation algorithm to these graph signals. Additive i.i.d. zero-mean Gaussian observation noise following Equation (6) is added, where $SNR_d = 10 \lg \frac{f^T f}{N/\beta}$ is 15 dB. The known hyperparameter n in the signal prior Equation (1) is set to 1/2, which means a smoothness prior Equation (3) is considered. The observed signal values are divided by 100 as preprocessing, and accordingly the estimated graph signal is multiplied by 100 at output. The performance of the proposed algorithm is compared to Anis2016, Perraudin2018 and Sakiyama2019. As our first sampling node is chosen at random, Perraudin2018 is a random sampling method, and the observation noises also introduce randomness, we test each method on the same signal for 20 times, and compare their average estimation errors.

As shown in Figure 8, although the proposed algorithm needs some initial samples to “warm up”, after that the estimation error decreases rapidly and surpass Anis2016, Perraudin2018, and Sakiyama2019 in very short time, and we keep leading as the sample size further grows. This is because our sequential method with unknown hyperparameters requires a certain number of samples to have a relatively proper estimation of the signal and noise distributions. Then, based on the latest signal bandwidth, energy level, and noise precision estimated from data, we are able to choose the subsequent sampling nodes more efficiently and estimate the graph signal more accurately than other methods that take no account of such signal and noise properties.

Figure 8. Estimation errors of the two temperature graph signals by the proposed algorithm versus Anis2016, Perraudin2018, and Sakiyama2019. (a) Results on f_1 . (b) Results on f_2 .

6. Conclusions

In this paper, a sequential sampling and estimation algorithm is proposed for approximately bandlimited graph signals without complete prior knowledge about signal and noise properties. In the proposed algorithm, hyperparameter estimation by EM based on historical observations and sample selection by US with latest estimated hyperparameters are done in an alternating way. We prove in theory the consistency of signal estimation in the proposed algorithm under specific conditions. Finite-time performance of the proposed algorithm is validated by a series of simulation results. The proposed method provides a novel and practical sequential framework for graph signal sampling with unknown hyperparameters in the system model, and the experiment on temperature data shows our potential application in sensor selection on sensor networks with smooth measurements over the topologies. Other potential applications include opinion prediction in social networks assuming that people with strong social connection are likely to hold similar opinions, and active semisupervised learning tasks on similarity graphs.

Author Contributions: Conceptualization, S.L., H.F., and B.H.; methodology, S.L. and H.F.; software, S.L.; validation, S.L., K.X., and H.F.; formal analysis, S.L. and H.F.; investigation, S.L. and H.F.; resources, H.F. and B.H.; data curation, S.L.; writing—original draft preparation, S.L.; writing—review and editing, K.X. and H.F.; visualization, S.L.; supervision, H.F. and B.H.; project administration, H.F., B.H.; funding acquisition, H.F. and B.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Shanghai Municipal Natural Science Foundation (No. 19ZR1404700), Fudan University-CIOMP Joint Fund (FC2019-003), and 2020 Okawa Foundation Research Grant.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Concavity of Objective Function in M Step

The objective function in Equation (16) is concave with respect to α, γ, β , as the Hessian matrix

$$H = \begin{bmatrix} \frac{\partial^2 E_{l-1}}{\partial \alpha^2} & \frac{\partial^2 E_{l-1}}{\partial \alpha \partial \gamma} & \frac{\partial^2 E_{l-1}}{\partial \alpha \partial \beta} \\ \frac{\partial^2 E_{l-1}}{\partial \gamma \partial \alpha} & \frac{\partial^2 E_{l-1}}{\partial \gamma^2} & \frac{\partial^2 E_{l-1}}{\partial \gamma \partial \beta} \\ \frac{\partial^2 E_{l-1}}{\partial \beta \partial \alpha} & \frac{\partial^2 E_{l-1}}{\partial \beta \partial \gamma} & \frac{\partial^2 E_{l-1}}{\partial \beta^2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \sum_{k=1}^N \frac{1}{(\alpha + \gamma \lambda_k^{2n})^2} & -\frac{1}{2} \sum_{k=1}^N \frac{\lambda_k^{2n}}{(\alpha + \gamma \lambda_k^{2n})^2} & 0 \\ -\frac{1}{2} \sum_{k=1}^N \frac{\lambda_k^{2n}}{(\alpha + \gamma \lambda_k^{2n})^2} & -\frac{1}{2} \sum_{k=1}^N \frac{\lambda_k^{4n}}{(\alpha + \gamma \lambda_k^{2n})^2} & 0 \\ 0 & 0 & -\frac{N}{2\beta^2} \end{bmatrix} \quad (A1)$$

$$= -\frac{1}{2} \sum_{k=1}^N \frac{1}{(\alpha + \gamma \lambda_k^{2n})^2} \begin{bmatrix} 1 \\ \lambda_k^{2n} \\ 0 \end{bmatrix} \begin{bmatrix} 1 & \lambda_k^{2n} & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\frac{N}{2\beta^2} \end{bmatrix}$$

is negative semi-definite for all α, γ, β .

Appendix B. Search Method for Optimization Problem in M Step

To maximize E_{l-1} to solve Equation (16), take the partial derivatives with respect to α, γ, β and set them to zero as in Equations (21)–(23).

From Equation (23), the M-step update of β has closed form

$$\beta_l = \frac{M}{\mathbb{E}_{l-1}[(\mathbf{y} - \mathbf{\Psi}\mathbf{f})^T(\mathbf{y} - \mathbf{\Psi}\mathbf{f})]}, \quad (A2)$$

which is independent of the update of α, γ .

As for α, γ , as the approximately bandlimited prior Equation (1) requires $\alpha, \gamma > 0$, if Equations (21) and (22) hold for some $\alpha, \gamma > 0$, this (α, γ) is the solution for (α_l, γ_l) ; otherwise, (α_l, γ_l) should be determined using gradient information based on concavity.

Figure A1. Two cases of M-step update of α, γ . (a) Case I when $\Delta > 0$: The solution point (α_l, γ_l) locates on PQ between P and Q . (b) Case II when $\Delta \leq 0$: The endpoint Q is the solution for (α_l, γ_l) .

Combining Equations (21) and (22) to have $\alpha \frac{\partial E_{l-1}}{\partial \alpha} + \gamma \frac{\partial E_{l-1}}{\partial \gamma} = 0$, we get

$$\alpha \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{f}] + \gamma \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{L}^{2n} \mathbf{f}] = N, \quad (A3)$$

which constrains (α, γ) on a straight line in the α - γ plane. Only the segment PQ in the first quadrant is considered as shown in Figure A1 to ensure $\alpha, \gamma > 0$. The endpoint P has coordinates $(0, \frac{N}{\mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{L}^{2n} \mathbf{f}]})$, and the endpoint Q is at $(\frac{N}{\mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{f}]}, 0)$.

Define the gradient of E_{l-1} with respect to α, γ as $\mathbf{g} = [\frac{\partial E_{l-1}}{\partial \alpha}, \frac{\partial E_{l-1}}{\partial \gamma}]^T$. Any point A on PQ satisfies $\mathbf{g}|_A \perp OA$. The gradient approaching P

$$\mathbf{g}|_{(\alpha, \gamma) \rightarrow P} = [+ \infty, 0]^T \quad (A4)$$

is always rightward as illustrated in Figure A1. Moreover, the gradient at Q

$$\mathbf{g}|_Q = \left[0, \frac{1}{2}\Delta\right]^T, \text{ where } \Delta = \frac{\text{tr}(\mathbf{L}^{2n})}{N} \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{f}] - \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{L}^{2n} \mathbf{f}], \quad (\text{A5})$$

is either upward or downward, depending on the sign of Δ .

If $\Delta > 0$, the gradient at Q is upward as in Figure A1a. Let $\mathbf{d}$ be the unit vector of direction PQ . Consider the directional derivative

$$\frac{\partial E_{l-1}}{\partial \mathbf{d}} = \langle \mathbf{g}, \mathbf{d} \rangle \propto \frac{\partial E_{l-1}}{\partial \alpha} \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{L}^{2n} \mathbf{f}] - \frac{\partial E_{l-1}}{\partial \gamma} \mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{f}]. \quad (\text{A6})$$

$\frac{\partial E_{l-1}}{\partial \mathbf{d}}|_{(\alpha, \gamma) \rightarrow P} > 0$ and $\frac{\partial E_{l-1}}{\partial \mathbf{d}}|_Q < 0$ are different in sign. By continuity of $\frac{\partial E_{l-1}}{\partial \mathbf{d}}$, there exists a zero-crossing Z of $\frac{\partial E_{l-1}}{\partial \mathbf{d}}$ on PQ between P and Q . Moreover, by concavity of E_{l-1} , $\frac{\partial E_{l-1}}{\partial \mathbf{d}}$ is monotonous along PQ . The zero-crossing Z can be found using, for example, binary search. Notice that $\mathbf{g}|_Z$ satisfies both $\mathbf{g}|_Z \perp PQ$ and $\mathbf{g}|_Z \perp OZ$, so $\mathbf{g}|_Z = \mathbf{0}$, i.e., Equations (21) and (22) hold at Z . Therefore, Z is the solution point for (α_l, γ_l) .

Otherwise, if $\Delta \leq 0$, the gradient at Q is zero or downward as in Figure A1b. By concavity of E_{l-1} , the endpoint Q is the optimal point in the first quadrant. γ_l is set to 0, and α is updated by

$$\alpha_l = \frac{N}{\mathbb{E}_{l-1}[\mathbf{f}^T \mathbf{f}]} \quad (\text{A7})$$

In this case, $p(\mathbf{f}; \alpha_l, \gamma_l) = \mathcal{N}(\mathbf{f} | \mathbf{0}, \alpha_l^{-1} \mathbf{I}_N)$, signal values on different nodes become i.i.d., and only the total energy of the graph signal is restricted.

Appendix C. Proof of Lemma 1

Proof. Let $\{m_{(r),t}\}_{r=1}^N$ be the order statistics of $\{m_{i,t}\}_{i \in \mathcal{V}}$ satisfying $m_{(1),t} \leq \dots \leq m_{(N),t}$, and denote the sorted vertex indices by $\{i_{(r),t}\}_{r=1}^N$. Consider $\eta_{i,t} = \frac{m_{i,t}}{t}$, and $\eta_{(r),t} = \frac{m_{(r),t}}{t}$. Define

$$R = \min \left\{ r \mid \exists \delta > 0, \liminf_{t \rightarrow \infty} \eta_{(r),t} > \delta \right\} - 1. \quad (\text{A8})$$

Note that $\liminf_{t \rightarrow \infty} \eta_{(r),t}$ always exists as $\eta_{i,t}$ is bounded in $[0, 1]$. Moreover, the set is not empty as $\eta_{(N),t} = \max_{i \in \mathcal{V}} \eta_{i,t} \geq \frac{1}{N}$. So $0 \leq R \leq N - 1$. Suppose that $R \geq 1$. Let $\rho_t = \eta_{(R),t}$. We are to show that $\liminf_{t \rightarrow \infty} \rho_t \geq \underline{\rho}$ for some $\underline{\rho} > 0$, which contradicts the definition of R . Then we will arrive at an assertion $R = 0$, and lemma 1 is proved.

Define $\mathbf{P}_t = \frac{1}{t} \mathbf{C}_t = \frac{1}{t} \mathbf{A}'_t + \beta'_t \sum_{i \in \mathcal{V}} \eta_{i,t} \mathbf{e}_i \mathbf{e}_i^T$, where $\mathbf{A}'_t = \alpha'_t \mathbf{I}_N + \gamma'_t \mathbf{L}^{2n}$. The sampling strategy Equation (25) can also be rewritten as

$$\mathbf{s}_{t+1} = \arg \max_{i \in \mathcal{V}} \mathbf{e}_i^T \mathbf{P}_t^{-1} \mathbf{e}_i. \quad (\text{A9})$$

Define $\mathcal{I}_t = \{i_{(r),t} \mid r \geq R + 1\}$, and $\mathcal{I}_t^c = \{i_{(r),t} \mid r \leq R\}$. By definition, there exists $\delta > 0$ and $T < \infty$, for all $t > T$ and $i \in \mathcal{I}_t$, it holds $\eta_{i,t} > \delta$. As $\mathbf{A}'_t$ is positive definite,

$$\forall t > T, \forall i \in \mathcal{I}_t, \mathbf{e}_i^T \mathbf{P}_t^{-1} \mathbf{e}_i \leq \mathbf{e}_i^T \left(\beta'_t \sum_{j \in \mathcal{V}} \eta_{j,t} \mathbf{e}_j \mathbf{e}_j^T \right)^\dagger \mathbf{e}_i = \frac{1}{\beta'_t \eta_{i,t}} < \frac{1}{\beta'_t \delta}, \quad (\text{A10})$$

where $(\cdot)^\dagger$ denotes the pseudo-inverse of a matrix.

Consider $i \in \mathcal{I}_t^c$. Define

$$\mathbf{Q}_t^{(1)} = \beta'_t (1 - \rho_t) \sum_{i \in \mathcal{I}_t} \mathbf{e}_i \mathbf{e}_i^T, \quad (\text{A11})$$

$$\mathbf{Q}_t^{(2)} = \frac{1}{t} \mathbf{A}'_t + \beta'_t \rho_t \sum_{i \in \mathcal{V}} \mathbf{e}_i \mathbf{e}_i^T. \quad (\text{A12})$$

$$\text{As } \mathbf{P}_t \preceq \mathbf{Q}_t^{(1)} + \mathbf{Q}_t^{(2)} = \frac{1}{t} \mathbf{A}'_t + \beta'_t \left(\sum_{i \in \mathcal{I}_t} \mathbf{e}_i \mathbf{e}_i^T + \rho_t \sum_{i \in \mathcal{I}_t^c} \mathbf{e}_i \mathbf{e}_i^T \right),$$

$$\begin{aligned} \max_{i \in \mathcal{I}_t^c} \mathbf{e}_i^T \mathbf{P}_t^{-1} \mathbf{e}_i &\geq \max_{i \in \mathcal{I}_t^c} \mathbf{e}_i^T (\mathbf{Q}_t^{(1)} + \mathbf{Q}_t^{(2)})^{-1} \mathbf{e}_i \\ &= \max_{i \in \mathcal{I}_t^c, \mathbf{z} \in \mathbb{R}^N} 2\mathbf{e}_i^T \mathbf{z} - \mathbf{z}^T (\mathbf{Q}_t^{(1)} + \mathbf{Q}_t^{(2)}) \mathbf{z} \\ &\geq \max_{i \in \mathcal{I}_t^c, \mathbf{z} \in \mathcal{N}(\mathbf{Q}_t^{(1)})} 2\mathbf{e}_i^T \mathbf{z} - \mathbf{z}^T \mathbf{Q}_t^{(2)} \mathbf{z}, \end{aligned} \quad (\text{A13})$$

where $\mathcal{N}(\mathbf{Q}_t^{(1)})$ is the null space of $\mathbf{Q}_t^{(1)}$. Denote the maximum value in the last line, also a lower bound of $\max_{i \in \mathcal{I}_t^c} \mathbf{e}_i^T \mathbf{P}_t^{-1} \mathbf{e}_i$, as L_t , and the maximizers as i_t^* , $\mathbf{z}_t^*$. On one hand, when i is set to i_t^* , we have

$$L_t = \max_{\mathbf{z} \in \mathcal{N}(\mathbf{Q}_t^{(1)})} 2\mathbf{e}_{i_t^*}^T \mathbf{z} - \mathbf{z}^T \mathbf{Q}_t^{(2)} \mathbf{z}. \quad (\text{A14})$$

By Lagrange multiplier method,

$$\mathbf{z}_t^* = (\mathbf{Q}_t^{(2)})^{-1} (\mathbf{I}_N - \mathbf{Q}_t^{(1)} (\mathbf{Q}_t^{(1)} (\mathbf{Q}_t^{(2)})^{-1} \mathbf{Q}_t^{(1)})^\dagger \mathbf{Q}_t^{(1)} (\mathbf{Q}_t^{(2)})^{-1}) \mathbf{e}_{i_t^*}, \quad (\text{A15})$$

$$L_t = 2\mathbf{e}_{i_t^*}^T \mathbf{z}_t^* - (\mathbf{z}_t^*)^T \mathbf{Q}_t^{(2)} \mathbf{z}_t^* = \mathbf{e}_{i_t^*}^T \mathbf{z}_t^* = (\mathbf{z}_t^*)^T \mathbf{Q}_t^{(2)} \mathbf{z}_t^*. \quad (\text{A16})$$

On the other hand, when $\mathbf{z}$ is fixed at $\mathbf{z}_t^*$, we know

$$\mathbf{e}_{i_t^*}^T \mathbf{z}_t^* = \max_{i \in \mathcal{I}_t^c} \mathbf{e}_i^T \mathbf{z}_t^*. \quad (\text{A17})$$

We further assert that

$$\mathbf{e}_{i_t^*}^T \mathbf{z}_t^* = \max_{i \in \mathcal{I}_t^c} |\mathbf{e}_i^T \mathbf{z}_t^*|. \quad (\text{A18})$$

Otherwise, there exists $j \in \mathcal{I}_t^c$ such that $-\mathbf{e}_j^T \mathbf{z}_t^* > \mathbf{e}_{i_t^*}^T \mathbf{z}_t^*$. This will lead to a contradiction $2\mathbf{e}_j^T (-\mathbf{z}_t^*) - (-\mathbf{z}_t^*)^T \mathbf{Q}_t^{(2)} (-\mathbf{z}_t^*) > L_t$. Therefore,

$$\begin{aligned} (\mathbf{e}_{i_t^*}^T \mathbf{z}_t^*)^2 &= \max_{i \in \mathcal{I}_t^c} (\mathbf{e}_i^T \mathbf{z}_t^*)^2 = \max_{i \in \mathcal{I}_t^c} (\mathbf{z}_t^*)^T \mathbf{e}_i \mathbf{e}_i^T \mathbf{z}_t^* \\ &\geq \frac{1}{R} (\mathbf{z}_t^*)^T \sum_{i \in \mathcal{I}_t^c} \mathbf{e}_i \mathbf{e}_i^T \mathbf{z}_t^* \\ &= \frac{1}{R} (\mathbf{z}_t^*)^T \left(\sum_{i \in \mathcal{I}_t^c} \mathbf{e}_i \mathbf{e}_i^T + \sum_{i \in \mathcal{I}_t} \mathbf{e}_i \mathbf{e}_i^T \right) \mathbf{z}_t^* \\ &= \frac{1}{R} \|\mathbf{z}_t^*\|^2, \end{aligned} \quad (\text{A19})$$

where the last but one line involves $\mathbf{z}_t^* \in \mathcal{N}(\mathbf{Q}_t^{(1)})$. Therefore,

$$L_t = \mathbf{e}_{i_t^*}^T \mathbf{z}_t^* \geq \frac{1}{\sqrt{R}} \|\mathbf{z}_t^*\|. \quad (\text{A20})$$

At the same time, according to the Courant–Fischer theorem, along with the positive definiteness of $\mathbf{Q}_t^{(2)}$,

$$L_t = (\mathbf{z}_t^*)^T \mathbf{Q}_t^{(2)} \mathbf{z}_t^* \leq \lambda_{\max}(\mathbf{Q}_t^{(2)}) \|\mathbf{z}_t^*\|^2 < \text{tr}(\mathbf{Q}_t^{(2)}) \|\mathbf{z}_t^*\|^2, \quad (\text{A21})$$

where $\lambda_{\max}(Q_i^{(2)})$ denotes the maximum eigenvalue of $Q_i^{(2)}$, and $\text{tr}(Q_i^{(2)})$ is the trace. Equations (A20) and (A21) together they give

$$\|z_t^*\| > \frac{1}{\sqrt{R} \text{tr}(Q_i^{(2)})} = \frac{1}{\sqrt{R} \left(\frac{1}{t} \text{tr}(A_t') + N\beta_t' \rho_t \right)}. \quad (\text{A22})$$

Substituting it back to (A20),

$$\max_{i \in \mathcal{I}_t^c} e_i^T P_t^{-1} e_i \geq L_t \geq \frac{1}{\sqrt{R}} \|z_t^*\| > \frac{1}{R \left(\frac{1}{t} \text{tr}(A_t') + N\beta_t' \rho_t \right)}. \quad (\text{A23})$$

Let $\rho^* = \frac{\delta}{NR}$, and $\rho_t^* = \rho^* - \frac{1}{t} \frac{\text{tr}(A_t')}{N\beta_t'}$. According to the sampling strategy Equation (A9), from previous results Equations (A10) and (A23), for $t > T$, $\rho_t < \rho_t^*$ implies $s_{t+1} \in \mathcal{I}_t^c$. Define $\tilde{\rho}_t = \frac{1}{R} \sum_{i \in \mathcal{I}_t^c} \eta_{i,t}$.

$$\begin{aligned} \rho_{t+1} &\geq \tilde{\rho}_{t+1} = \frac{Rt\tilde{\rho}_t + 1}{R(t+1)} = \tilde{\rho}_t + \frac{1 - R\tilde{\rho}_t}{R(t+1)} \\ &> \tilde{\rho}_t + \frac{(N-R)\delta}{R(t+1)} \\ &\geq \frac{\rho_t}{R} + \frac{(N-R)\delta}{R(t+1)}. \end{aligned} \quad (\text{A24})$$

By induction, this lower bound of $\rho_{t+\tau}$ increases with τ

$$\rho_{t+\tau} > \frac{\rho_t}{R} + \frac{(N-R)\delta}{R} \sum_{k=1}^{\tau} \frac{1}{t+k} \quad (\text{A25})$$

until $\rho_{t+\tau} \geq \rho_{t+\tau}^*$. Suppose that $\rho_{t-1} \geq \rho_{t-1}^*$ but $\rho_t < \rho_t^*$ at some $t > T$. Then,

$$\rho_t \geq \frac{t-1}{t} \rho_{t-1} \geq \frac{t-1}{t} \rho_{t-1}^*. \quad (\text{A26})$$

If $\lim_{t \rightarrow \infty} \frac{1}{t} \frac{\text{tr}(A_t')}{N\beta_t'} = 0$, then $\lim_{t \rightarrow \infty} \rho_t^* = \rho^*$. We can obtain $\liminf_{t \rightarrow \infty} \rho_t \geq \lim_{t \rightarrow \infty} \frac{(t-1)\rho_{t-1}^*}{Rt} = \frac{\rho^*}{R} = \frac{\delta}{NR^2} > 0$, which contradicts the definition.

Therefore, the only possibility is $R = 0$, and Lemma 1 is proved. $\square$

The authors of Pronzato [28] have proved a similar statement in sequential DOE for nonlinear LS regression. In contrast, the proposed algorithm is an uncertainty sampling process for a Bayesian estimation problem, which makes our proposition and proof different in some way from theirs.

Appendix D. Proof of Theorem 2

Proof. Let $\mathcal{U}_\epsilon = \{f \in \mathbb{R}^N \mid \forall i \in \mathcal{V}, |f_i - f_i^*| < \epsilon\}$ be a “cubic” neighborhood around the true signal f^* , and its complement $\mathcal{U}_\epsilon^c = \mathbb{R}^N \setminus \mathcal{U}_\epsilon$. We are to prove that $P(\mathcal{U}_\epsilon^c | \mathbf{\Psi}_{1:t}, \mathbf{y}_{1:t}; \alpha_t', \gamma_t', \beta_t')$ converges to 0 with high probability for arbitrary $\epsilon > 0$.

Here, for brevity, γ_t' is absorbed into α_t' . We can get a fractional form of the target

$$P(\mathcal{U}_\epsilon^c | \mathbf{\Psi}_{1:t}, \mathbf{y}_{1:t}; \alpha_t', \beta_t') = \frac{N_t}{D_t} = \frac{\int_{\mathcal{U}_\epsilon^c} \frac{p(\mathbf{y}_{1:t} | \mathbf{\Psi}_{1:t}, f; \beta_t')}{p(\mathbf{y}_{1:t} | \mathbf{\Psi}_{1:t}, f^*; \beta^*)} p(f; \alpha_t') df}{\int_{\mathbb{R}^N} \frac{p(\mathbf{y}_{1:t} | \mathbf{\Psi}_{1:t}, f; \beta_t')}{p(\mathbf{y}_{1:t} | \mathbf{\Psi}_{1:t}, f^*; \beta^*)} p(f; \alpha_t') df}, \quad (\text{A27})$$

where f^* is the true signal, and β^* is the true noise precision.

First look at the numerator. According to Corollary 2, as $t \rightarrow \infty$, $m_{i,t} \rightarrow \infty$. We have

$$\begin{aligned} \frac{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f; \beta'_t)}{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f^*; \beta^*)} &\stackrel{(a)}{=} \exp\left(-\sum_{i=1}^N m_{i,t} \left(\frac{1}{m_{i,t}} \sum_{j=1}^{m_{i,t}} \ln \frac{p(y_{i,j}|f_i^*; \beta^*)}{p(y_{i,j}|f_i; \beta'_t)}\right)\right) \\ &\stackrel{(b)}{\rightarrow} \exp\left(-\sum_{i=1}^N m_{i,t} KL(p(y_i|f_i^*; \beta^*) || p(y_i|f_i; \beta'_t))\right), \end{aligned} \quad (A28)$$

where $y_{i,j}$ denotes the j -th observation of vertex i ; $KL(p||q)$ is the Kullback–Leibler divergence between two distributions; (a) takes into account the independence of different observations in $\mathbf{y}$ conditioned on $\mathbf{\Psi}$, f , and β ; and (b) obeys the Law of Large Numbers.

For any $f \in \mathcal{U}_\epsilon^c$, there exists $k \in \mathcal{V}$, such that $|f_k - f_k^*| > \epsilon$. Regardless of the value of β'_t ,

$$\begin{aligned} KL(p(y_k|f_k^*; \beta^*) || p(y_k|f_k; \beta'_t)) &= \frac{1}{2} \left(\beta'_t (f_k - f_k^*)^2 + \frac{\beta'_t}{\beta^*} - \ln \frac{\beta'_t}{\beta^*} - 1 \right) \\ &> \frac{1}{2} \ln(1 + \beta^* \epsilon^2). \end{aligned} \quad (A29)$$

For $i \neq k$, $KL(p(y_i|f_i^*; \beta^*) || p(y_i|f_i; \beta'_t)) \geq 0$ by definition. Moreover, from Corollary 1, there exists $\delta > 0$, such that $\frac{m_{k,t}}{t} > \delta$ for all large t . Thus, for $\epsilon_1 = \frac{1}{2} \delta \ln(1 + \beta^* \epsilon^2) > 0$,

$$\begin{aligned} \frac{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f; \beta'_t)}{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f^*; \beta^*)} &< \exp\left(-\frac{1}{2} m_{k,t} \ln(1 + \beta^* \epsilon^2)\right) \\ &< \exp(-t\epsilon_1) \end{aligned} \quad (A30)$$

for all $f \in \mathcal{U}_\epsilon^c$ for sufficiently large t with high probability.

Therefore, when t becomes large enough, no matter what values α'_t have, the numerator

$$N_t = \int_{\mathcal{U}_\epsilon^c} \frac{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f; \beta'_t)}{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f^*; \beta^*)} p(f; \alpha'_t) df < \exp(-t\epsilon_1) \quad (A31)$$

with high probability.

Then, come to the denominator. As α'_t , β'_t are estimated by MML,

$$D_t \geq \tilde{D}_t(\alpha, \beta^*) = \int_{\mathbb{R}^N} \frac{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f; \beta^*)}{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f^*; \beta^*)} p(f; \alpha) df, \quad (A32)$$

where α is allowed to be chosen arbitrarily.

Consider arbitrary $\epsilon_2 > 0$. Analogous to above, as $t \rightarrow \infty$ and $m_{i,t} \rightarrow \infty$, we have

$$\begin{aligned} \exp(t\epsilon_2) \frac{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f; \beta^*)}{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f^*; \beta^*)} &= \exp\left(\sum_{i=1}^N m_{i,t} \left(\epsilon_2 - \frac{1}{m_{i,t}} \sum_{j=1}^{m_{i,t}} \ln \frac{p(y_{i,j}|f_i^*; \beta^*)}{p(y_{i,j}|f_i; \beta^*)}\right)\right) \\ &\rightarrow \exp\left(\sum_{i=1}^N m_{i,t} (\epsilon_2 - KL(p(y_i|f_i^*; \beta^*) || p(y_i|f_i; \beta^*)))\right). \end{aligned} \quad (A33)$$

Define $\mathcal{X}_{\epsilon_2/2} = \left\{ f \in \mathbb{R}^N \mid \forall i \in \mathcal{V}, KL(p(y_i|f_i^*; \beta^*) || p(y_i|f_i; \beta^*)) < \frac{\epsilon_2}{2} \right\}$. For all $f \in \mathcal{X}_{\epsilon_2/2}$, when t grows large enough,

$$\exp(t\epsilon_2) \frac{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f; \beta^*)}{p(\mathbf{y}_{1:t}|\mathbf{\Psi}_{1:t}, f^*; \beta^*)} > \exp\left(t \frac{\epsilon_2}{2}\right) \quad (A34)$$

with high probability.

Note that the term inside integration is positive. We shrink the range of integration and obtain

$$\begin{aligned} \exp(t\epsilon_2)\tilde{D}_t(\alpha, \beta^*) &> \exp(t\epsilon_2) \int_{\mathcal{X}_{\epsilon_2/2}} \frac{p(\mathbf{y}_{1:t}|\Psi_{1:t}, \mathbf{f}; \beta^*)}{p(\mathbf{y}_{1:t}|\Psi_{1:t}, \mathbf{f}^*; \beta^*)} p(\mathbf{f}; \alpha) d\mathbf{f} \\ &> \exp\left(t\frac{\epsilon_2}{2}\right) \int_{\mathcal{X}_{\epsilon_2/2}} p(\mathbf{f}; \alpha) d\mathbf{f}. \end{aligned} \quad (\text{A35})$$

We can always choose an α such that $\int_{\mathcal{X}_{\epsilon_2/2}} p(\mathbf{f}; \alpha) d\mathbf{f} > 0$ since $p(\mathbf{f}; \alpha)$ is Gaussian. Then, we get $\exp(t\epsilon_2)\tilde{D}_t(\alpha, \beta^*) \rightarrow \infty$ as $t \rightarrow \infty$. Therefore, the denominator

$$D_t \geq \tilde{D}_t(\alpha, \beta^*) > \exp(-t\epsilon_2) \quad (\text{A36})$$

for all large t with high probability, where ϵ_2 can be any positive number.

Pick $0 < \epsilon_2 < \epsilon_1$. For $\epsilon_3 = \epsilon_1 - \epsilon_2 > 0$,

$$P(\mathcal{U}_\epsilon^c | \Psi_{1:t}, \mathbf{y}_{1:t}; \alpha'_t, \beta'_t) = \frac{N_t}{D_t} < \exp(-t\epsilon_3) \quad (\text{A37})$$

for all sufficiently large t with high probability.

We finally reach a conclusion that $P(\mathcal{U}_\epsilon^c | \Psi_{1:t}, \mathbf{y}_{1:t}; \alpha'_t, \beta'_t)$ converges to 0 with high probability as $t \rightarrow \infty$ for any positive ϵ . $\square$

The proposed algorithm is more complex than classical empirical Bayes method discussed in Petrone et al. [27], because in our setting, hyperparameters exist not only in signal prior but also in observation model, and i.i.d. complete observations are substituted by partial ones selected by strategy. That is why a more complicated proof of consistent signal estimation in our algorithm is provided here.

References

- Shuman, D.I.; Narang, S.K.; Frossard, P.; Ortega, A.; Vandergheynst, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.* **2013**, *30*, 83–98. [\[CrossRef\]](#)
- Sandryhaila, A.; Moura, J.M.F. Discrete signal processing on graphs. *IEEE Trans. Signal Process.* **2013**, *61*, 1644–1656. [\[CrossRef\]](#)
- Shuman, D.I.; Vandergheynst, P.; Frossard, P. Chebyshev polynomial approximation for distributed signal processing. In Proceedings of the 2011 International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS), Barcelona, Spain, 27–29 June 2011.
- Shi, X.; Feng, H.; Zhai, M.; Yang, T.; Hu, B. Infinite impulse response graph filters in wireless sensor networks. *IEEE Signal Process. Lett.* **2015**, *22*, 1113–1117. [\[CrossRef\]](#)
- Loukas, A.; Simonetto, A.; Leus, G. Distributed autoregressive moving average graph filters. *IEEE Signal Process. Lett.* **2015**, *22*, 1931–1935. [\[CrossRef\]](#)
- Chen, S.; Varma, R.; Sandryhaila, A.; Kovačević, J. Discrete signal processing on graphs: Sampling theory. *IEEE Trans. Signal Process.* **2015**, *63*, 6510–6523. [\[CrossRef\]](#)
- Anis, A.; Gadde, A.; Ortega, A. Efficient sampling set selection for bandlimited graph signals using graph spectral proxies. *IEEE Trans. Signal Process.* **2016**, *64*, 3775–3789. [\[CrossRef\]](#)
- Narang, S.K.; Gadde, A.; Ortega, A. Signal processing techniques for interpolation in graph structured data. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 5445–5449.
- Sakiyama, A.; Tanaka, Y.; Tanaka, T.; Ortega, A. Efficient sensor position selection using graph signal sampling theory. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 6225–6229.
- Jabłoński, I. Graph signal processing in applications to sensor networks, smart grids, and smart cities. *IEEE Sens. J.* **2017**, *17*, 7659–7666. [\[CrossRef\]](#)
- Huang, W.; Bolton, T.A.W.; Medaglia, J.D.; Bassett, D.S.; Ribeiro, A.; Ville, D.V.D. A graph signal processing perspective on functional brain imaging. *Proc. IEEE* **2018**, *106*, 868–885. [\[CrossRef\]](#)
- Cheung, G.; Magli, E.; Tanaka, Y.; Ng, M.K. Graph spectral image processing. *Proc. IEEE* **2018**, *106*, 907–930. [\[CrossRef\]](#)
- Dinesh, C.; Cheung, G.; Bajić, I.V. Point cloud denoising via feature graph Laplacian regularization. *IEEE Trans. Image Process.* **2020**, *29*, 4143–4158. [\[CrossRef\]](#)

14. Gadde, A.; Anis, A.; Ortega, A. Active semi-supervised learning using sampling theory for graph signals. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014; pp. 492–501.
15. Anis, A.; El Gamal, A.; Avestimehr, A.S.; Ortega, A. A sampling theory perspective of graph-based semi-supervised learning. *IEEE Trans. Inf. Theory* **2019**, *65*, 2322–2342. [\[CrossRef\]](#)
16. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Palais des Congrès Neptune, Toulon, France, 24–26 April 2017.
17. Chen, S.; Varma, R.; Singh, A.; Kovačević, J. Signal recovery on graphs: Fundamental limits of sampling strategies. *IEEE Trans. Signal Inf. Process. Netw.* **2016**, *2*, 539–554. [\[CrossRef\]](#)
18. Puy, G.; Tremblay, N.; Gribonval, R.; Vandergheynst, P. Random sampling of bandlimited signals on graphs. *Appl. Comput. Harmon. Anal.* **2018**, *44*, 446–475. [\[CrossRef\]](#)
19. Xie, X.; Feng, H.; Jia, J.; Hu, B. Design of sampling set for bandlimited graph signal estimation. In Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, 14–16 November 2017; pp. 653–657.
20. Lin, S.; Xie, X.; Feng, H.; Hu, B. Active sampling for approximately bandlimited graph signals. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5441–5445.
21. Perraudin, N.; Ricaud, B.; Shuman, D.I.; Vandergheynst, P. Global and local uncertainty principles for signals on graphs. *APSIPA Trans. Signal Inf. Process.* **2018**, *7*, e3. [\[CrossRef\]](#)
22. Xie, X.; Yu, J.; Feng, H.; Hu, B. Bayesian design of sampling set for bandlimited graph signals. In Proceedings of the 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Ottawa, ON, Canada, 11–14 November 2019.
23. Gadde, A.; Ortega, A. A probabilistic interpretation of sampling theory of graph signals. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 3257–3261.
24. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006; ISBN 9780387310732.
25. Settles, B. *Active Learning*; Morgan & Claypool: San Rafael, CA, USA, 2012; ISBN 9781608457250.
26. Bernardo, J.M.; Smith, A.F.M. *Bayesian Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1994; ISBN 0471924164.
27. Petrone, S.; Rousseau, J.; Scricciolo, C. Bayes and empirical Bayes: Do they merge? *Biometrika* **2014**, *101*, 285–302. [\[CrossRef\]](#)
28. Pronzato, L. One-step ahead adaptive D-optimal design on a finite design space is asymptotically optimal. *Metrika* **2010**, *71*, 219–238. [\[CrossRef\]](#)
29. Oppenheim, A.V.; Schaffer, R.W. *Discrete-Time Signal Processing*, 3rd ed.; Prentice Hall: Englewood Cliffs, NJ, USA, 2009; ISBN 9780136024552.
30. Kay, S.M. *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*; Prentice Hall: Englewood Cliffs, NJ, USA, 1993; ISBN 9780133457117.
31. Sakiyama, A.; Tanaka, Y.; Tanaka, T.; Ortega, A. Eigendecomposition-free sampling set selection for graph signals. *IEEE Trans. Signal Process.* **2019**, *67*, 2679–2692. [\[CrossRef\]](#)
32. Watts, D.J.; Strogatz, S.H. Collective dynamics of ‘small-world’ networks. *Nature* **1998**, *393*, 440–442. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Barabási, A.L.; Albert, R. Emergence of scaling in random networks. *Science* **1999**, *286*, 509–512. [\[CrossRef\]](#)
34. Perraudin, N.; Paratte, J.; Shuman, D.; Martin, L.; Kalofolias, V.; Vandergheynst, P.; Hammond, D.K. GSPBOX: A toolbox for signal processing on graphs. *arXiv* **2016**, arXiv:1408.5781v2 [cs.IT].
35. Integrated Surface Database (ISD) by National Climatic Data Center (NCDC) of the USA. Available online: <https://www.ncdc.noaa.gov/isd> (accessed on 11 December 2020).