

Streaming Gaussian Dirichlet Random Fields for Spatial Predictions of High Dimensional Categorical Observations

John E. San Soucie^{1,2}(✉), Heidi M. Sosik², and Yogesh Girdhar²

¹ Massachusetts Institute of Technology, Cambridge MA, USA

² Woods Hole Oceanographic Institution, Woods Hole MA, USA
 {jsansoucie,hsosik,ygirdhar}@whoi.edu

Abstract. We present the Streaming Gaussian Dirichlet Random Field (S-GDRF) model, a novel approach for modeling a stream of spatiotemporally distributed, sparse, high-dimensional categorical observations. The proposed approach efficiently learns global and local patterns in spatiotemporal data, allowing for fast inference and querying with a bounded time complexity. Using a high-resolution data series of plankton images classified with a neural network, we demonstrate the ability of the approach to make more accurate predictions compared to a Variational Gaussian Process (VGP), and to learn a predictive distribution of observations from streaming categorical data. S-GDRFs open the door to enabling efficient informative path planning over high-dimensional categorical observations, which until now has not been feasible.

Keywords: Machine learning · Adaptive sampling · Field robotics

1 Introduction and Background

1.1 Introduction

Recent advances in in-situ sensing and autonomy have drastically improved the resolving power available for field and marine robotics and lowered deployment and operation costs. In-situ imaging systems can have temporal resolutions ranging from milliseconds to minutes, while continuous sampling on autonomous [16] or towed [4] platforms allows for resolving sub-meter-scale phenomena on missions lasting from hours to days. These systems produce many orders of magnitude more data than their predecessors.

Marine autonomy poses a unique set of challenges as compared to typical terrestrial autonomy. With radio-frequency communications attenuated by water, marine autonomous systems must be capable of a high degree of autonomy without continuous access to GPS localization. Additionally, harsh environments, payload restrictions, and energy constraints limit the kinds and lengths of missions available for scientific data collection. These and other limitations incentivize the use of adaptive sampling techniques that allow scientists to collect more relevant data with the same time and energy budgets.

Adaptive sampling requires a model for how data are distributed in the world. For example, for the scalar field maximum-seeking [24], a Gaussian process (GP) can model the spatial distribution of the scalar field and allow for querying and prediction. Additionally, in settings with communications constraints, that model must be trained from streaming data entirely on-board the autonomous platform. Low cost and low energy computers have been increasingly deployed on underwater vehicles. However, development of streaming statistical models for data from in-situ marine imaging systems has lagged behind the rapid progress made in imaging and computing hardware. Thus in-situ sensors generally collect data following brute-force, pre-planned missions.

Building on the Gaussian Dirichlet Random Field (GDRF) model [22], a previously introduced probabilistic model for spatiotemporally distributed, sparse, high-dimensional categorical data, here we present a streaming inference approach that we call streaming GDRF (S-GDRF). The S-GDRF model builds on the success of existing topic models in capturing the latent co-occurrence patterns in sparse, high-dimensional categorical data, while enabling inference and querying through the entire world. Our main contribution is a novel sub-sampling approach to doubly-stochastic black-box variational inference, which enables fully streaming data processing that is bounded in time and linear in space for a large category of missions.

1.2 Background

Gaussian processes. GPs are a natural choice of belief model for field robotics, especially when performing adaptive sampling. They can naturally encode uncertainty in complex, dynamic field conditions [6], and prediction allows for online Bayesian optimization approaches to path planning [7, 20]. High-dimensional categorical observations, such as the output of a neural network classifying camera images, pose a challenge when applying existing Bayesian optimization techniques to adaptive sampling. Many adaptive sampling strategies compute expected reward rollouts only over scalar observation fields [3], or at best low-dimensional categorical fields [8].

Topic models, a type of Bayesian graphical model, represent the distribution of categorical observations by factoring the distributions with latent or unobserved “topics”. The Latent Dirichlet Allocation (LDA) model captures topics with semantic meaning, organized by co-occurring clusters of words, in collections of text documents [5]. The Real-time Online Spatiotemporal Topic (ROST) model generalizes the LDA model to spatiotemporally distributed categorical observations, with a fast inference procedure suitable for embedded computation [11, 10]. The ROST model has been used to represent distributions of corals and seafloor types from robotic surveys of coral reefs [13], and topics learned from a ROST model have been previously shown to capture meaningful co-occurrence relationships from phytoplankton observation data [14].

The GDRF model, a combination of GPs and topic models into a single unified framework, was originally developed to model spatiotemporally distributed

categorical data (e.g. plankton image classifications) while allowing for interpolation (e.g. time series with gaps, or path planning applications) [22]. The original GDRF used a mixed, offline inference algorithm. Existing streaming variational inference algorithms are generally limited to specific prior-likelihood combinations, such as exponential family models [1, 15]. In the context of streaming Markov Chain Monte Carlo inference, different approaches to subsampling from past data have been shown to impact convergence properties of the model [9]. Here we use a similar subsampling approach, but replace the Markov Chain Monte Carlo algorithm with a black-box variational inference algorithm to learn an approximate posterior [18].

Between GPs, topic models, and their unification in GDRFs, none provides a suitable belief model for adaptive sampling over high-dimensional categorical data. This work seeks to fill this gap — the lack of predictive models necessary for computing expected reward rollouts in informative path planning tasks — by providing a belief model for spatiotemporally distributed, sparse, high-dimensional categorical observations.

2 Methods

2.1 Technical Approach

Gaussian-Dirichlet Random Fields. At a high level, a GDRF is a smooth, continuous model for categorical data which factors observations into latent co-occurrence patterns. A GDRF models the distribution of categorical data (“observations”) in a spatiotemporal world by introducing latent communities, each of which represents a distribution over observation types. The relationship between communities and observations is modeled with Dirichlet distributions, while the distribution of communities in space-time is modeled by GPs. Formally, a GDRF is a continuous Bayesian model defined on a (potentially infinite) set of points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots\}$, each of which we call *locations* in the *world* $\mathbf{X}$. *Observations* w_i and *communities* z_i are W - and K -categorical variables, respectively. The community spatial distributions are computed by passing a set of GPs μ_j , representing mean latent log probability fields for each community through a *link function* $f_j : \mathbb{R}^K \rightarrow [0, 1]$, where $\sum_j f_j(\mu_1, \dots, \mu_K) = 1$, e.g. the *softmax* function. The Dirichlet-distributed global community distributions Φ are multiplied by the spatial probabilities to get the observation distributions (fig. 1).

Streaming inference for S-GDRFs. S-GDRFs augment the GDRF model with an inference procedure capable of handling streaming observations, as well as a sparse inducing-point approximation to allow for computationally tractable streaming. The GDRF prior and likelihood functions do not permit a simple conjugate posterior, so inference of the posterior to a given prior $p(x, z)$ on observations x and latent variables z must be approximate. We utilize stochastic black-box variational inference (BBVI) [18], with a key alteration to the final step of selecting observations to calculate the loss $\mathcal{L}$. After sampling latent variables from the approximate posterior, BBVI involves calculating the log probability of

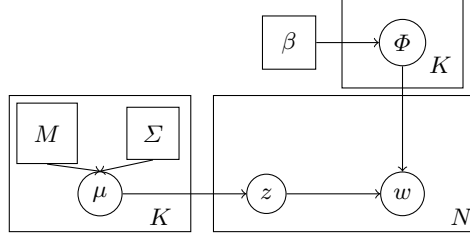


Fig. 1: The graphical model for GDRFs (reproduced from [22])

the observations. For the S-GDRF model, we allow the inference step to choose N_s points from any of the previously observed data, with a custom probability distribution $\pi_t^{N_s}(x)$ that can be weighted to bias random selection of recent observations. Training iterations can occur between observation of new data points, and the number of such iterations can vary if desired. The sparse inducing point approximation[25] reduces the complexity of covariance matrix inversions from $O(n^3)$ to $O(nm^2)$ with m inducing points. The selection of a finite collection of n_s past observations for calculating the loss allows the upper bound to be fixed at $O(n_s m^3)$ prior to training, even on an infinite stream of data.

Evaluation of GDRF predictive power We compare the S-GDRF model to a single-GP-per-category regression model, trained entirely offline, which we refer to as the Variational Gaussian Process (VGP) model. To evaluate the predictive power of each model, we define the predictive Kullback-Liebler divergence (PKL) metric: for a model trained on the first N data points in a stream, we query the observation distribution at all future locations to be observed. The PKL metric M_{PKL} for a single future observation is the KL divergence between the predicted and observed distribution of observation categories:

$$M_{PKL}(p_i(w_{i+1:}|x_{i+1:})) = D_{KL}(p_i(w_{i+1:}, x_{i+1:}) || p_{\text{emp}}(w_{i+1:}, x_{i+1:})) \quad (1)$$

where p_i is a model trained on the first i observations, $w_{i+1:}$ are the following observations, $x_{i+1:}$ are their locations, and p_{emp} is the empirical distribution of observations.

2.2 Experiments

Temporal data. We compared the SGDRF and VGP approaches using a dataset collected on a July 2021 research cruise on the *R/V Endeavor* (EN688) and classified via neural network (fig. 2). As part of the Northeast U.S. Shelf Long-Term Ecological Research (NES-LTER) project, the cruise sampled along an established north/south transect ranging from Martha’s Vineyard in the north to the continental slope beyond the Mid-Atlantic Bight in the south, punctuated by latitudinally sparse segments with high longitudinal resolution (fig. 2). An Imaging FlowCytobot (IFCB) [16], an automated imaging-in-flow cytometer, was

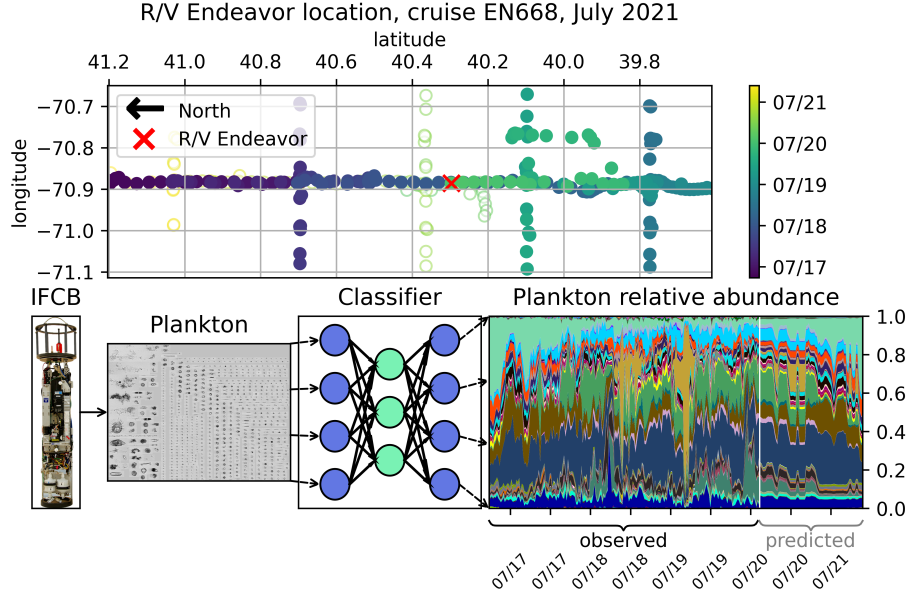


Fig. 2: In July 2021 aboard the R/V Endeavor, an Imaging FlowCytobot was used to take high-throughput, high-resolution images of plankton in surface seawater. Post-cruise, all images were classified with a neural network-based classifier. The S-GDRF model enables prediction of this kind of spatiotemporally distributed, sparse, high-dimensional categorical data.

connected to the ship’s underway seawater system and used to collect thousands of images of microscopic plankton in surface seawater, approximately once every 25 minutes. After the cruise, a pre-trained convolutional neural network classifier sorted the images into over 150 different classes [17]. We trained both an S-GDRF model and a series of VGP models on this data in a streaming fashion, in order to evaluate their predictive power.

Spatial data. Images of Joel’s Shoal Reef, St. John, US Virgin Islands were taken with an underwater robotic system in November 2022 [12]. These were stitched together into a single mosaic of the reef, and 15436 simple visual features [21] were extracted in a 130x130 cell grid (fig. 3a). A lawnmower trajectory over this grid was simulated, and observations from this trajectory were fed into an S-GDRF model to evaluate scalability and performance on two-dimensional data (fig. 3b).

3 Results

1-dimensional S-GDRF prediction. A mid-training snapshot of plankton predicted and actual relative abundances highlights how predictions from S-GDRF

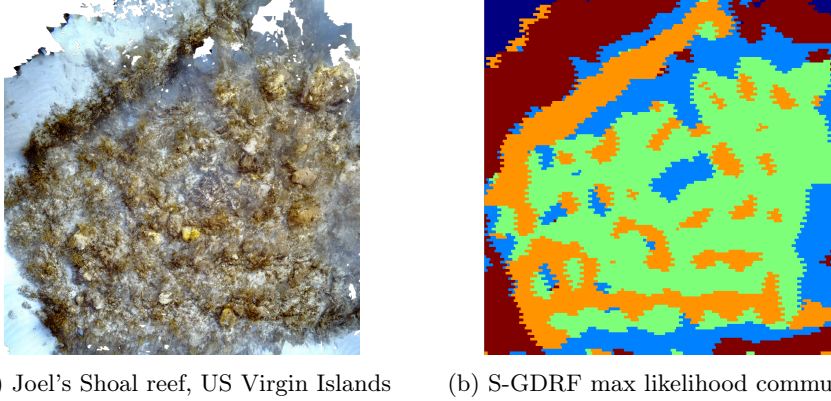


Fig. 3: S-GDRF inference on the Joel's Shoal coral head shows reasonable labelings for pixels in a 130x130 grid with 15436 unique extracted features.

are more useful than from VGP (fig. 4). The S-GDRF model tends to take a community-mean-field approach to prediction in out-of-coverage regions, while the VGP model generally makes a noisy and uninformative prediction. For observed and in-coverage data, the VGP model captures much of the high-frequency variability seen in the observations. In contrast, the S-GDRF model effectively smooths the observation distributions. Turning to predictive capability, the S-GDRF model has a consistently lower median PKL metric than the VGP model across the entire cruise, as well as lower variance (fig. 5). Median VGP PKL drops rapidly as the coverage fraction increases, reaching a reasonably low value once the coverage fraction is above 0.8. However, VGP variance remains high throughout the data series.

2-dimensional S-GDRF inference. After running a simulated lawnmower trajectory over the Joel's shoal reef, the final inferred maximum likelihood topics (fig. 3b) visually capture the major structures in the coral reef (fig. 3a). On average, training proceeded at 7.68 iterations per second on an NVIDIA A6000 GPU.

4 Discussion

Advantages over sparse and streaming GPs. The sparse GP[23], with its data-independent inducing points, opened the door for streaming GP inference algorithms which handle new data points as they arrive in a computationally efficient manner. Many streaming GP algorithms (e.g. [19]) extend standard GP inference, and only work for Gaussian likelihoods. Streaming variational GP inference algorithms can use non-Gaussian likelihoods [2]. However, these methods still require a strictly Gaussian stochastic process (as opposed to a more general

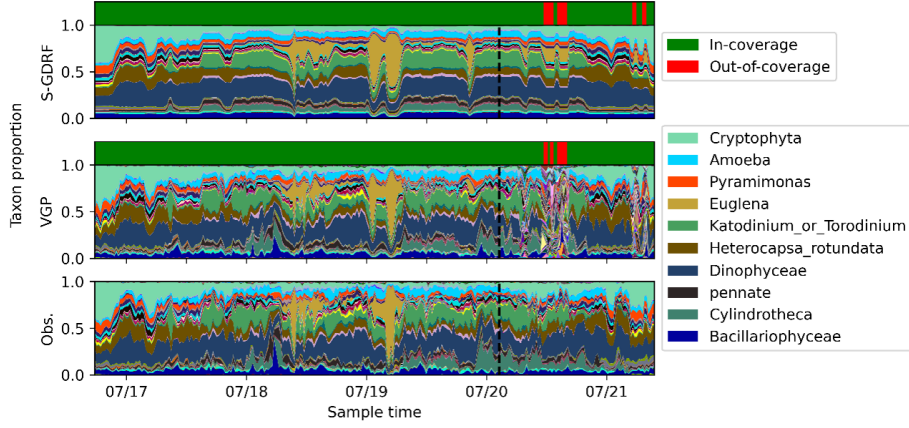


Fig. 4: Time-series plots of the distribution of plankton, with model fits (left of the dashed line) and predictions (right of the dashed line) in out-of-coverage areas (more than one kernel lengthscale from any datapoint left of the black line). Due to the large number of plankton taxa, only the top ten taxa (by mean observed relative abundance) are marked in the legend.

Bayesian structure), and even the most efficient models have not been demonstrated to work with hundreds or thousands of categories in a GP classifier. In contrast, the S-GDRF factorizes the observation model and uses a GP to only model the low-dimensional latent representation of the data. Our experiments show that the S-GDRF is able to accurately capture the distribution of sparse, high-dimensional categorical data, while being extremely computationally efficient. The coral reef experiment (fig. 3) demonstrated efficient (0.13 seconds per iteration) inference with over 15000 observation categories, as S-GDRF time and memory complexity generally scale with the amount of data and the number of latent GPs. In fact, no comparison to a strictly GP-based model (such as a VGP) is computationally feasible for these data. The low computational cost of S-GDRF, albeit on workstation hardware, supports its use on the embedded computers available for in-situ inference.

S-GDRF outperforms VGP. The S-GDRF model outperforms VGP according to the prediction metric, especially for out-of-coverage data. The S-GDRF latent variables likely aid in prediction here. The smaller number of GPs increases robustness to noise, while at the same time the Dirichlet distributions pool global information about relevant co-occurrence patterns. Even if no useful information exists about the distribution of communities far away from any observations, knowing the kinds of co-occurrence patterns that can mix to produce observations provides better predictions. The simulated results presented here support

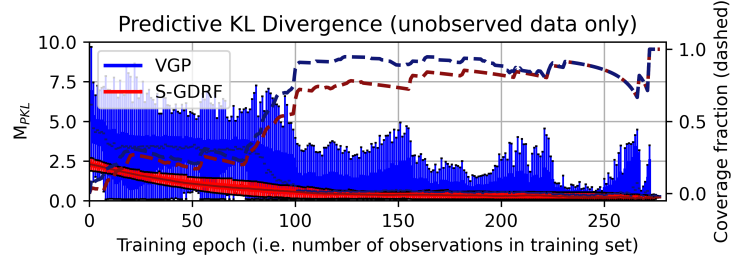


Fig. 5: S-GDRF and VGP predictive Kullback-Liebler divergence (PKL) metric, for the 1-D temporal dataset. The x-axis is the number of training data, while the y-axis represents the PKL metric, i.e. the KL divergence between model predictions on all other data and observations. The median and quartile PKL for all unobserved (i.e. non-training) data is displayed with solid lines. The dashed lines represent the coverage fraction, i.e. what proportion of the unobserved data points lie within one GP kernel lengthscale of a previously observed datapoint.

the use of S-GDRF models for streaming inference on autonomous platforms, and specifically as a belief model for informative path planning over categorical observations. The S-GDRF model outperformed a VGP model with more parameters and a much larger computational budget, training on the entire streaming dataset in minutes. A single refinement step, with a 25×25 uniform inducing point grid, takes less than a second on state-of-the-art hardware. This implies real-time computation is feasible on CUDA-capable embedded systems. The Appendix in the supplementary material contains a demo underscoring this, in which a 15000-category S-GDRF is trained on a 130×130 cell grid on desktop hardware.

5 Conclusion

The S-GDRF model, a streaming Bayesian model for spatiotemporally distributed, sparse, high-dimensional categorical data, was proposed as a belief model for robotic path planning over categorical data. This model was demonstrated to outperform a single-GP-per-observation model at streaming prediction tasks on a 1-dimensional temporal dataset. Evaluation on a 2-dimensional spatial dataset highlighted both the scalability of the S-GDRF model in terms of the number of observation categories and the computational efficiency of the S-GDRF inference algorithm. These results demonstrate how S-GDRFs can enable efficient informative path planning over high-dimensional categorical observations. Future work includes deployment of the S-GDRF algorithm on in-situ robotic platforms for real-time belief modeling and path planning.

Acknowledgement. This project was funded in part by NSF (NRI Award No. 2133029 to YG, LTER Award No. 1655686 to HMS), the Simons Foundation (grant 561126 to HMS), and the NDSEG Fellowship Program (to JESS).

Bibliography

- [1] Broderick T, Boyd N, Wibisono A, Wilson AC, Jordan MI (2013) Streaming variational bayes. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K (eds) *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol 26
- [2] Bui TD, Nguyen C, Turner RE (2017) Streaming sparse gaussian process approximations. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol 30
- [3] Chen W, Liu L (2022) Informative planning in the presence of outliers. In: 2022 International Conference on Robotics and Automation (ICRA), pp 5311–5318, DOI 10.1109/ICRA46639.2022.9812267
- [4] Cowen RK, Guigand CM (2008) In situ ichthyoplankton imaging system (isiis): system design and preliminary results. *Limnology and Oceanography: Methods* 6(2):126–132
- [5] David M Blei, Ng, Andrew Y, Jordan, Michael I (2003) Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022
- [6] Duckworth P, Lacerda B, Hawes N (2021) Time-bounded mission planning in time-varying domains with semi-mdps and gaussian processes. In: Kober J, Ramos F, Tomlin C (eds) *Proceedings of the 2020 Conference on Robot Learning*, PMLR, *Proceedings of Machine Learning Research*, vol 155, pp 1654–1668
- [7] Flaspohler G, Preston V, Michel AP, Girdhar Y, Roy N (2019) Information-guided robotic maximum seek-and-sample in partially observable continuous environments. *IEEE Robotics and Automation Letters* 4(4):3782–3789
- [8] Gilhuly B, Sadeghi A, Yedemellat P, Rezaee K, Smith SL (2022) Looking for trouble: Informative planning for safe trajectories with occlusions. In: 2022 International Conference on Robotics and Automation (ICRA), pp 8985–8991, DOI 10.1109/ICRA46639.2022.9811994
- [9] Girdhar Y (2014) Unsupervised Semantic Perception, Summarization, and Autonomous Exploration, for Robots in Unstructured Environments List of Tables. PhD thesis, McGill University
- [10] Girdhar Y, Dudek G (2015) Gibbs Sampling Strategies for Semantic Perception of Streaming Video Data. Tech. rep., McGill University, URL <http://arxiv.org/abs/1509.03242>, arXiv:1509.03242 [cs]
- [11] Girdhar Y, Giguère P, Dudek G (2014) Autonomous adaptive exploration using realtime online spatiotemporal topic modeling. *The International Journal of Robotics Research* 33(4):645–657
- [12] Girdhar Y, McGuire N, Cai L, Jamieson S, McCammon S, Claus B, Soucie JES, Todd JE, Mooney TA (2023) Curee: A curious underwater robot for ecosystem exploration. In: 2023 International Conference on Robotics and Automation (ICRA)

- [13] Jamieson S, Fathian K, Khosoussi K, How JP, Girdhar Y (2021) Multi-Robot Distributed Semantic Mapping in Unfamiliar Environments through Online Matching of Learned Representations. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, Xi'an, China, pp 8587–8593, DOI 10.1109/ICRA48506.2021.9561934
- [14] Kalmbach A, Girdhar Y, Sosik HM, Dudek G (2017) Phytoplankton hotspot prediction with an unsupervised spatial community model. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, Singapore, Singapore, pp 4906–4913, DOI 10.1109/ICRA.2017.7989568
- [15] McInerney J, Ranganath R, Blei D (2015) The population posterior and bayesian modeling on streams. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R (eds) *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol 28
- [16] Olson RJ, Sosik HM (2007) A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot: In situ imaging of nano- and microplankton. *Limnology and Oceanography: Methods* 5(6):195–203, DOI 10.4319/lom.2007.5.195
- [17] Orenstein EC, Beijbom O, Peacock EE, Sosik HM (2015) WHOI-Plankton-A Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification. Tech. rep., Woods Hole Oceanographic Institute
- [18] Ranganath R, Gerrish S, Blei D (2014) Black box variational inference. In: *Artificial intelligence and statistics*, PMLR, pp 814–822
- [19] Ranganathan A, Yang MH, Ho J (2010) Online sparse gaussian process regression and its applications. *IEEE Transactions on Image Processing* 20(2):391–404
- [20] Rothfuss J, Koenig C, Rupenyan A, Krause A (2023) Meta-learning priors for safe bayesian optimization. In: Liu K, Kulic D, Ichnowski J (eds) *Proceedings of The 6th Conference on Robot Learning*, PMLR, *Proceedings of Machine Learning Research*, vol 205, pp 237–265
- [21] Rublee E, Rabaud V, Konolige K, Bradski G (2011) Orb: An efficient alternative to sift or surf. In: 2011 International Conference on Computer Vision, pp 2564–2571, DOI 10.1109/ICCV.2011.6126544
- [22] San Soucie JE, Sosik HM, Girdhar Y (2020) Gaussian-dirichlet random fields for inference over high dimensional categorical observations. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 2924–2931
- [23] Snelson E, Ghahramani Z (2005) Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems* 18
- [24] Srinivas N, Krause A, Kakade SM, Seeger MW (2012) Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory* 58(5):3250–3265, DOI 10.1109/TIT.2011.2182033
- [25] Titsias M (2009) Variational learning of inducing variables in sparse gaussian processes. In: *Artificial intelligence and statistics*, PMLR, pp 567–574