

Gaussian-Dirichlet Random Fields for Inference over High Dimensional Categorical Observations

John E. San Soucie¹ Heidi M. Sosik² and Yogesh Girdhar³

Abstract—We propose a generative model for the spatio-temporal distribution of high dimensional categorical observations. These are commonly produced by robots equipped with an imaging sensor such as a camera, paired with an image classifier, potentially producing observations over thousands of categories. The proposed approach combines the use of Dirichlet distributions to model sparse co-occurrence relations between the observed categories using a latent variable, and Gaussian processes to model the latent variable’s spatio-temporal distribution. Experiments in this paper show that the resulting model is able to efficiently and accurately approximate the temporal distribution of high dimensional categorical measurements such as taxonomic observations of microscopic organisms in the ocean, even in unobserved (held out) locations, far from other samples. This work’s primary motivation is to enable deployment of informative path planning techniques over high dimensional categorical fields, which until now have been limited to scalar or low dimensional vector observations.

I. INTRODUCTION

The field of autonomous robotics has drastically increased the ability of scientists to explore remote environments [1]. Sites of research interest in outer space and the deep ocean are characterized by being extremely inhospitable to human life and difficult to reach from the Earth’s surface via radio-frequency communications. But by making higher-level decisions on-board, autonomous agents reduce the need for real-time communications with a human operator. The effectiveness of any autonomous agent, in terms of valuable information acquired per unit resource, depends largely on the plan it is programmed to follow. But remote environments are often under-explored, and thus the optimal plan is not known before the mission starts. Therefore, autonomous agents require plans that adapt to the environments they encounter, or informative path planning (IPP).

Robots performing adaptive sampling or IPP missions typically operate over measurements of interest that are low dimensional. For example, in the context of underwater robotic missions, IPP has been used to adaptively sample physical [2], [3], [4], chemical [5], [6], [7], [8], biological [9], [10], and general visual signals [11] in the ocean (for more details, see [12]). IPP techniques require a spatial model

*This work was supported by a grant from the Simons Foundation (561126, HMS) and NSF-NRI Award Number 1734400

¹J. San Soucie is with the Mechanical Engineering department at the Massachusetts Institute of Technology and the Applied Ocean Physics and Engineering Department at the Woods Hole Oceanographic Institution johnes@mit.edu

²H. Sosik is with the Biology Department at the Woods Hole Oceanographic Institution hsosik@whoi.edu

³Y. Girdhar is with the Applied Ocean Physics and Engineering Department at the Woods Hole Oceanographic Institution yogi@whoi.edu

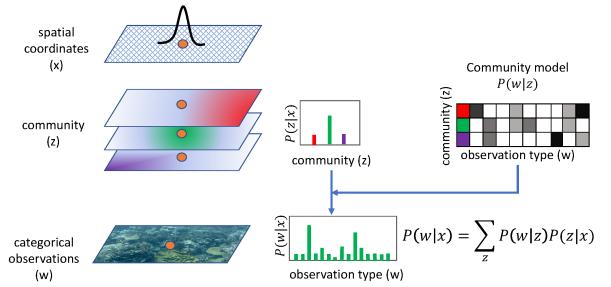


Fig. 1: Overview of the Gaussian-Dirichlet Random Field model. Categorical observations, such as observations of phytoplankton taxa, are factored into the product of a community model and spatiotemporal distributions for each community. The community model, which is the distribution of taxa in each community, is modeled with a Dirichlet prior; and the spatial distribution of each community is modeled using a Gaussian process.

for the information that is being sampled [13]. Gaussian processes (GPs) [14], [15], which are capable of modeling black-box scalar or vector functions, are a common choice of probabilistic observation model for IPP. GP models enable IPP algorithms to produce smooth estimates over future observations, along with uncertainties, which can then be used to predict expected trajectory reward and to balance the exploration-versus-exploitation trade-off using Bayesian Optimisation (BO) [16].

Recent advances in machine learning and artificial intelligence have enabled development of new types of sensing systems that effectively measure high-dimensional categorical data. These systems combine traditional sensors (cameras, probes, etc.) with deep neural networks to classify scalar, vector, and image-based measurements in real time [17], [18]. Examples of such systems include classifiers trained using datasets [19] produced by the Imaging FlowCytobot (IFCB) [20], that can produce taxonomic observations of phytoplankton with over 100 categories, and classifiers trained on the ImageNet dataset [21], which can produce categorical observations over 10,000 categories. IPP techniques based on GP models do not directly generalize to such high-dimensional categorical observations as common embedding methods (such as one-hot encoding [22]) embed a K -categorical variable in K dimensions, leading to exponential space complexity. Hierarchical Bayesian models designed for spatio-temporal data, such as HDP-ROST [23], can efficiently model distributions of high dimensional categorical

observations. But these models do not provide smoothly varying probability distribution estimates at arbitrary locations in space and time, which are needed for BO based IPP. Thus, there is a need for a spatio-temporal model of the distribution of high-dimensional categorical data compatible with IPP.

In this paper, we introduce the Gaussian-Dirichlet Random Field (GDRF), a hierarchical generative topic model for the spatial distribution of categorical observations. As shown in fig. 1, GDRFs factor the probability distribution over observation categories w into conditional distributions for w given a latent topic z , and conditional distributions of the latent topics given the observation's location x :

$$P(w|x) = \sum_z P(w|z)P(z|x) \quad (1)$$

Since w and z are both categorical random variables, (1) remains unusable for IPP. Therefore, we further factor $P(z|x)$ by introducing a latent Gaussian random field μ_i for each topic, normalized to a distribution via a link function f_i :

$$P(z_i|x) = f_i(\mu_1(x), \dots, \mu_K(x)) \quad (2)$$

To learn the word-topic model and the latent μ_i , we combine Gibbs sampling with variational inference. An immediate consequence of our choice of factorization is that the topics in a GDRF model are scientifically meaningful in that they capture the latent spatiotemporal structures relating different observation categories.

II. RELATED WORK

Latent Dirichlet Allocation (LDA), introduced in [24] is a topic model [25] originally designed for text documents. LDA models the relationships between words and documents using a set of latent topics that are linked to both the words and documents via Dirichlet distributions. Formally, given a corpus of M documents $\{d_1, \dots, d_M\}$, each with N_i words out of a vocabulary of size W , we can generate a topic model for K topics as follows:

$$\begin{aligned} \Theta_{d_i} &\sim \text{Dirichlet}(\alpha) \\ \Phi_{z_k} &\sim \text{Dirichlet}(\beta) \\ z_{j,d_i} &\sim \text{Cat}(\Theta_{d_i}) \\ w_{j,d_i} &\sim \text{Cat}(\Phi_{z_{j,d_i}}) \end{aligned} \quad (3)$$

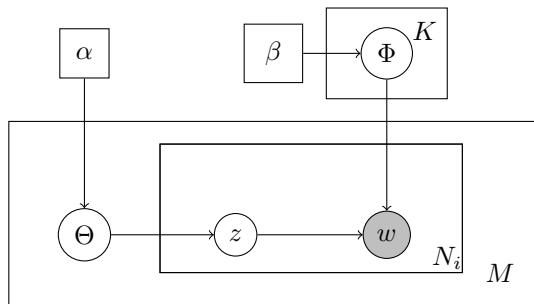


Fig. 2: The graphical model for LDA

The graphical model for LDA is shown in Fig. 2. To perform inference, we can do Gibbs sampling for the topic z_i assigned to a particular word w_i in document d_i [26]:

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{w_i} + \beta}{n_{-i,j} + W\beta} \frac{n_{-i,j}^{d_i} + \alpha}{n_{-i,j}^{d_i} + K\alpha} \quad (4)$$

where $n_{-i,j}^{w_i} + 1$ is the number of words with label w_i assigned to topic j , $n_{-i,j} + 1$ is the total number of words assigned to topic j , $n_{-i,j}^{d_i} + 1$ is the number of words in documents d_i assigned topic j , and $n_{-i,j}^{d_i} + 1$ is the number of words in document d_i . LDA has been applied to model natural scenes [27] and human actions [28], in addition to text corpora such as electronic health records [29], Twitter posts [30], and historical documents [31]. Spatial LDA [32] (SLDA) extends the LDA model to account for spatial structures in image data by adding priors over the location of a spatial word.

The Real-time Online Spatiotemporal Topic model (ROST), introduced in [23], models spatiotemporally distributed categorical data by discretizing an N -dimensional world and treating cells in that world as documents. Letting $G(d_i)$ be the spatiotemporal neighborhood of cell d_i , a hypothetical generative model for ROST simply replaces the distribution of topics for a document with the distribution of topics in the neighborhood of a cell:

$$\Theta_{G(d_i)} \sim \text{Dirichlet}(\alpha) \quad (5)$$

The Gibbs sampling step for inference similarly replaces the count of topics in a document $n_{-i,j}^{d_i}$ with the count of topics in the neighborhood of a $n_{-i,j}^{G(d_i)}$.

ROST has been used for robots displaying unsupervised curious behavior [33], [23], multi-robot topic modeling [34] and phytoplankton ecological modeling [35].

Formally, a Gaussian process [15] is a set of random variables $\{Z(x_i)\}$ defined on some possibly infinite indexing set $X = \{x_i\}$ such that for any finite subset $Y \subseteq \{Z(x_i)\}$, $Y \sim \mathcal{N}(\mu, \Sigma)$. The function $\Sigma : X \times X \rightarrow \mathbb{R}_{\geq 0}$ is called the *kernel* or *covariance* function of the GP, and specifies the structure of relationships between different points. Many common kernel functions on \mathbb{R}^D are stationary ($\Sigma = \Sigma(x - x')$) and isotropic ($\Sigma = \Sigma(|x - x'|)$), including the Matérn kernel function with $\nu = 3/2$:

$$k_{3/2}(r) = \sigma \left(1 + \frac{\sqrt{3}r}{\ell} \right) \exp \left\{ -\frac{\sqrt{3}r}{\ell} \right\} \quad (6)$$

In (6), ℓ represents the *length scale* of the Gaussian process, and σ represents a scale parameter for the kernel.

The geostatistics literature has used Kriging and similar spatial optimal linear prediction tools since the 1950s [36], [37]. In Kriging, the value of a scalar field is modeled as a random field $Z(x)$. The value of $Z(x_0)$, an unobserved location, is estimated from a weighted sum of n sampled locations:

$$\hat{Z}(x_0) = \sum_{i=1}^n \lambda_i Z(x_i) \quad (7)$$

The weights that provide the minimum variance unbiased estimator $\hat{Z}(x_0)$ can be calculated given only the sampled values and the covariance function for the random field. Standard Kriging is a form of Gaussian Process regression [14]. For categorical random fields, geostatisticians use variations on indicator Kriging [38], [39]. Methods from the Kriging family have been used for adaptive sampling [40] and marine visual data analysis [41].

There are many examples in the literature of IPP algorithms which utilize GPs as a model for observed scalar fields in marine environments. Binney, Krause, and Sukhatme [42] demonstrate a graph-based submodular optimization IPP algorithm with an objective function that can handle temporal nonstationarity and along-path sample collection (as opposed to sample collection at waypoints). Das et al. [43] use GPs to model both spatial scalar fields and the relationship between the scalar field's variables and organism abundance. They tested their model with two adaptive sampling strategies. Suryan and Tokek [44] develop a fast GP regression informative path planning algorithm over spatial fields. Berget et al. [45], Fossum and Eidviken et al. [46], and Fossum and Fragoso et al. [10] implement GP regression on AUVs and demonstrate the viability of simple IPP algorithms in real-world scenarios. Flaspohler et al. [47] introduce a plume-finding algorithm, which locates maxima of phenomena modeled by GPs. In all of these works, GPs model scalar fields.

III. GAUSSIAN-DIRICHLET RANDOM FIELDS

We begin with several preliminary definitions. A GDRF is defined on an *indexing set* $\mathbf{X} = \{x_1, x_2, \dots\}$, representing points in the world on which the model is defined. For example, a GDRF on a two-dimensional $A \times B$ grid has as its indexing set $\mathbf{X} = \{1, 2, \dots, A\} \times \{1, 2, \dots, B\}$. We will generally refer to the indexing set itself as the *world*, and call members of the world *locations*.

Words w_i and *topics* z_i are W - and K -categorical variables, respectively. The *mean latent log probabilities* (MLLPs) μ_j are Gaussian random fields defined on the world X . MLLPs are transformed to a probability distribution via a *link function* $f_j : \mathbb{R}^K \rightarrow [0, 1]$, where $\sum_j f_j(\mu_1, \dots, \mu_K) = 1$. For this paper, we exclusively use the softmax link function $f_j(\mu_1, \dots, \mu_K) = \exp(\mu_j) / \sum_k \exp(\mu_k)$. Finally, the generative model contains several hyperparameters: β is the Dirichlet parameter controlling the word distribution for each topic, and M_i and Σ_i are respectively the mean and covariance function of the Gaussian process from which μ_i are drawn.

Given a set of N (not necessarily unique) members of the indexing set, $\{x_1, x_2, \dots, x_N\}$, latent probabilities for K topics, as well as topics and words, are given by:

$$\begin{aligned} \mu_j &\sim \mathcal{N}(M_i, \Sigma_i), \quad j \in [1..K] \\ \Phi_z &\sim \text{Dirichlet}(\beta), \quad z \in \{z_1, \dots, z_K\} \\ z_i &\sim f(\mu_1(x_i), \dots, \mu_K(x_i)), \quad i \in [1..N] \\ w_i &\sim \Phi_{z_i}, \quad i \in [1..N] \end{aligned} \quad (8)$$

The graphical model for GDRFs is given in Fig. 3.

IV. APPROXIMATE INFERENCE OF GAUSSIAN-DIRICHLET RANDOM FIELDS

Assume we have collected a set of N categorical observations $\{w_i\}$ associated with N locations in the world $\{x_i\}$. We can decompose the learning of a GDRF into two steps: learning the word-topic model, and learning the latent log topic probabilities. We learn the word-topic model via Gibbs sampling. This model is similar to ROST in many ways, which is itself a spatiotemporal version of LDA. In ROST, the spatio-temporal world is discretized into cells, and the prior distribution of topics in a cell is defined by the distribution of topics in the Von Neumann neighborhood of the cell. Every word in a cell has the same prior topic distribution, independent of its exact location within the cell.

For GDRFs, the Gaussian Processes underlying the model allow us to consider topic densities, as opposed to counts, in $P(z|x)$. Normalizing $n_{-i,j}^{G(d_i)}$, the number of times topic j is observed in the neighborhood of cell d_i , by the hypervolume $V(G(d_i))$ of cell d_i , we get an approximation for the mean topic density in the neighborhood of d_i :

$$P(z_i = j|x) = \frac{\frac{n_{-i,j}^{G(d_i)}}{V(G(d_i))} + \frac{\alpha}{V(G(d_i))}}{\frac{n_{-i,:}^{G(d_i)}}{V(G(d_i))} + \frac{K\alpha}{V(G(d_i))}} \quad (9)$$

Finally, since the Dirichlet concentration parameter α can also be viewed as a smoothing “pseudocount”, we can factor it into a scale times the hypervolume of the neighborhood, $\alpha \rightarrow \alpha V(G(d_i))$. Then, in the limit as $V(G(d_i))$ approaches zero, we get our Gibbs sampling distribution for GDRFs:

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}, x_i) \propto \frac{n_{-i,j}^{w_i} + \beta}{n_{-i,:}^{w_i} + W\beta} \frac{\rho_j(x_i) + \alpha}{\rho(x_i) + K\alpha}. \quad (10)$$

Here $\rho_j(x)$ represent the density of topic j at location x , while $\rho(x)$ represents the observation density at location x . In GDRF, α is a pseudo-density, with the same smoothing properties as the LDA and ROST models.

After sampling a topic for each observation, we have collected a set of N categorical topics $\{z_i\}$, associated with N locations in the world $\{x_i\}$. We can use these to do approximate variational inference on the Gaussian processes. In our generative model, we let $P(z_i|x) = \exp(\mu_i) / \sum_j \exp(\mu_j)$. $P(z_i|x)$ represents the topic probability at location x . We can use the new topics from the Gibbs sampling $\{z_i\}$ to calculate

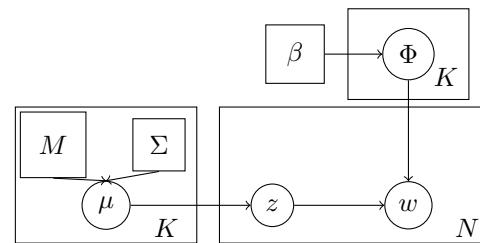


Fig. 3: The graphical model for GDRFs

approximate $\rho_j(x_i)$ by discretizing the world. Substituting in the expression from the Gibbs sampling distribution,

$$\frac{\rho_j(x_i) + \alpha}{\rho(x_i) + T\alpha} = \exp(\mu_j(x_i)) / \sum_k \exp(\mu_k(x_i)) \quad (11)$$

or

$$\log(\rho_j(x_i) + \alpha) = \mu_j(x_i) + C. \quad (12)$$

Because the softmax transformation is shift invariant, we can take C to be zero. Our training inputs are the locations of the observations $\{x_i\}$, while our training targets for each Gaussian process are $\{\log(\rho_j(x_i) + \alpha)\}$. We aim to maximize the sum of the evidence lower bound (ELBO) for each GP using stochastic variational inference [48]. SVI implementation was drastically simplified by using GPytorch [49], a GP library built on top of the Python library Pytorch [50], which offers simplified interfaces for automatic differentiation and GPU acceleration. After a training step of the GPs, we can calculate ρ_j using the same equation for the training targets. The training procedure for GDRFs is given in algorithm 1.

Algorithm 1 GDRF Inference

```

1: while true do
2:   for  $i = 1$  to  $N$  do
3:      $z_i \sim P(z_i = j | \mathbf{z}, \mathbf{w})$ 
4:   end for
5:   Update  $\Phi$  according to  $\{(w_i, z_i)\}$ 
6:   Update  $\rho_j$  according to  $\{(x_i, z_i)\}$ 
7:   for  $j = 1$  to  $K$  do
8:      $Y_j = \log(\rho_j(\mathbf{x}) + \alpha)$ 
9:      $\mathcal{L}_j(X, Y) = \text{ELBO}[\mathcal{G}_j(X), Y_j]$ 
10:    Update  $M_j$  and  $\Sigma_j$  according to  $\nabla \mathcal{L}_j$ 
11:     $\rho_j(x) = f_j(G_1(x), \dots, G_K(x))$ 
12:   end for
13: end while

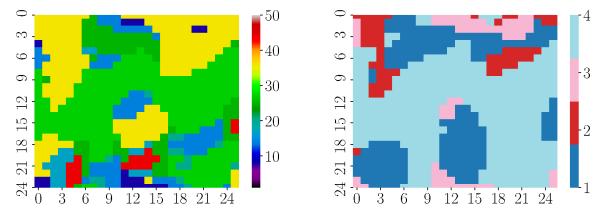
```

In practice, the two learned components of the GDRF model (the word-topic distributions Φ and the GPs) converge at different speeds. Empirical training accuracy and time to convergence improved when allowing the Gibbs sampler and the variational inference steps to run multiple times before proceeding. Theoretically, N_i should be chosen so each half of the algorithm converges at an equal rate during each global iteration. The Gibbs sampler is hyperparameter-free, while SVI has a single additional hyperparameter: the learning rate λ . Therefore, if T_1 is the mixing time of the Gibbs sampler, and T_2/λ is the number of SVI iterations until convergence of the GPs, $N_1/T_1 = \lambda N_2/T_2$. We found $N_1 = 50$, $N_2 = 5$, $\lambda = 0.25$ allowed each half of the model sufficient time to learn from the other half, without allowing either half to converge to an undesired local optimum.

V. RESULTS

A. Simulated Data

In order to demonstrate GDRF's ability to learn the latent topic structure underlying a set of observations, we tested both GDRF inference and ROST on a simulated dataset,



(a) Max Likelihood Word

(b) Max Likelihood Topic

Fig. 4: Forward simulation of the GDRF generative model as described in Eq. (8) and Fig. 3. The world used here is a 26×26 lattice from 0 to 25 in X and Y . The means of each GP are zero, and the GPs use a Matérn kernel [15] with a length scale of 2.5 in each dimension, scaled by 5. The Dirichlet parameter β is 0.1



Fig. 5: An example of a visually and spatially heterogeneous underwater scene. Image captured at Bellairs Research Institute, Barbados, January 2019.

shown in Fig. 4. We drew 10,000 random observations on a 26×26 grid from a GDRF generative model with 4 topics, a vocabulary of size 50, a β of 0.1, zero GP mean, and a Matérn kernel [15] of length scale $\ell = 2.5$ and overall scale $\sigma = 5$. These simulation parameters were chosen because they produced maximum likelihood word distributions, as shown in Fig. 4a, which contain realistic spatial heterogeneity. For comparison, see fig. 5, which contains coral reef structures.

Observations drawn from the simulation shown in Fig. 4 were used as inputs for both the GDRF model and the ROST model. The maximum likelihood topics for both models (shown in Fig. 6b) are visually similar to the ground truth maximum likelihood topics in Fig. 4. But the GDRF's inferred maximum likelihood topics in Fig. 6b have a much stronger resemblance to the ground truth in Fig. 6b than do ROST's in Fig. 6d (modulo the swapping of topic labels and colors). In addition, ROST learns a factorization in which three of the four topics are always most likely, while GDRF's factorization correctly models the observations with four topics. The GDRF model also captures small scale variation (on the order of a couple of spatial cells) in the latent topic field that ROST fails to pick up. Finally, note that the inputs to both of these models are the observations, as in Fig. 4a, but the similarity between the ground-truth

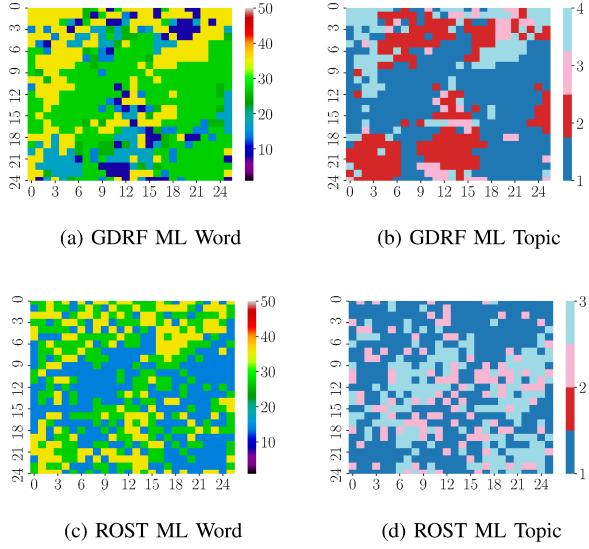


Fig. 6: GDRF and ROST inference on the data from Fig. 4.

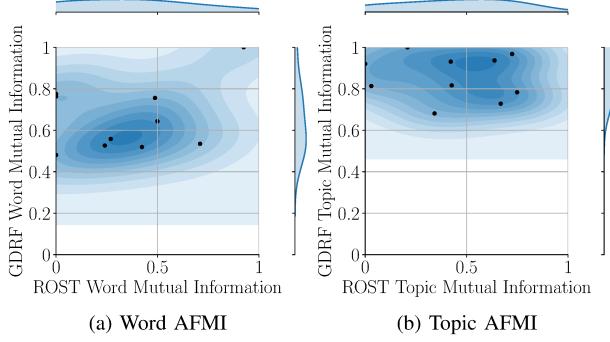


Fig. 7: Approximate Fractional Mutual Information (AFMI) for GDRF and ROST on 10 random data sets. 7b: Topic AFMI. 7a: Word AFMI. ROST AFMI is on the X-axis, while GDRF AFMI is on the Y-axis.

maximum likelihood topics in Fig. 4b and the GDRF’s inferred maximum likelihood topics in Fig. 6b is the product of a completely unsupervised training process.

To numerically evaluate the ability of these two models to capture complex spatial heterogeneity, we generated 10 simulation of 10,000 observations on an 11×11 grid, from GDRFs with 3 topics, a vocabulary of size 15, $\beta = 0.1$, $M = 0$, and Matérn kernels with $\sigma = 5$ and $\ell = 12.5$. We trained both GDRFs and ROST on each simulation, and we computed an approximate fractional mutual information (AFMI) score between the inferred and ground truth maximum likelihood topics and words for both models:

$$\text{AFMI}(M_{\text{model}}, M_{\text{gt}}) = \frac{I(M_{\text{model}}, M_{\text{gt}})}{I(M_{\text{gt}}, M_{\text{gt}})} \quad (13)$$

$$I(X, Y) = \sum_x \sum_y P(x, y) \log \left(\frac{P(x, y)}{P(x)P(y)} \right) \quad (14)$$

Here the marginal probabilities $P(M = i)$ are proportional to the number of cells in the maximum likelihood map M with category i , and the joint probabilities $P(M_a = i, M_b = j)$ are proportional to the number of cells with topic i in map a and topic j in map b . Note that AFMI scores range from zero to one, with zero representing two maps with no mutual information and one representing two identical maps.

The resultant AFMI scores are shown in Fig. 7. GDRF has consistently high AFMI scores for its topic maps in Fig. 7b, and slightly lower AFMI scores for its word maps Fig. 7a. In contrast, ROST has AFMI scores that vary along the entire scale for both topics and words. In both cases, GDRFs almost always get higher AFMI scores than ROST. The GDRF model is capable of learning complex, heterogeneous latent structures from densely sampled 2D categorical fields, and it learns those structures better than ROST.

B. Phytoplankton Taxa Dataset

The Martha’s Vineyard Coastal Observatory (MVCO) is a research station off the coast of Martha’s Vineyard, Massachusetts, providing long time series of oceanographic and meteorological measurements [51]. In addition to typical oceanographic sensors, the MVCO is equipped with an Imaging FlowCytobot[52], which takes high-throughput images of phytoplankton. These images are both selectively expert-annotated, and automatically labelled using a machine learning-based classifier [20].

We used a random subset of 100,000 images from three years of MVCO phytoplankton data, from 2013 to 2016, labeled by taxon using a random forest classifier, as training inputs for GDRF and ROST. The results are shown in Fig. 8. Both GDRFs and ROST generate inferred word distributions (Figs. 8c and 8e) that are visually similar to the observed distribution of phytoplankton taxa (Fig. 8a). But Fig 8b shows GDRF produces a distribution with a lower mean KL-divergence from the observed distribution, 0.064, than ROST, 0.095. This shows GDRF learns a more accurate model for the observations than ROST.

We demonstrate GDRF’s ability to extrapolate by comparing the model’s performance after training with and without withheld subsets of the data. The three years of phytoplankton taxon data are divided into 500 equal-time chunks. A held-out window five chunks wide (approximately 10 days) is slid across the dataset, and the model is trained without that chunk of data. The mean KL divergence between the model and the observation distribution within this window is calculated, and fig. 9 plots the log ratio between this window-mean divergence in the held-out and non-held-out models. On this figure, a ratio of 1 implies that both the model trained with held-out data and the model trained without held-out data predicted taxon distributions equally as far from the observed distribution (as measured using KL divergence); a ratio above 1 implies that the model trained with held-out data does worse. The mean ratio between the models is 1.07. The model trained without data in a particular region predicts a taxon distribution that is less accurate than a model trained to reproduce the observed distribution. But the magnitude of

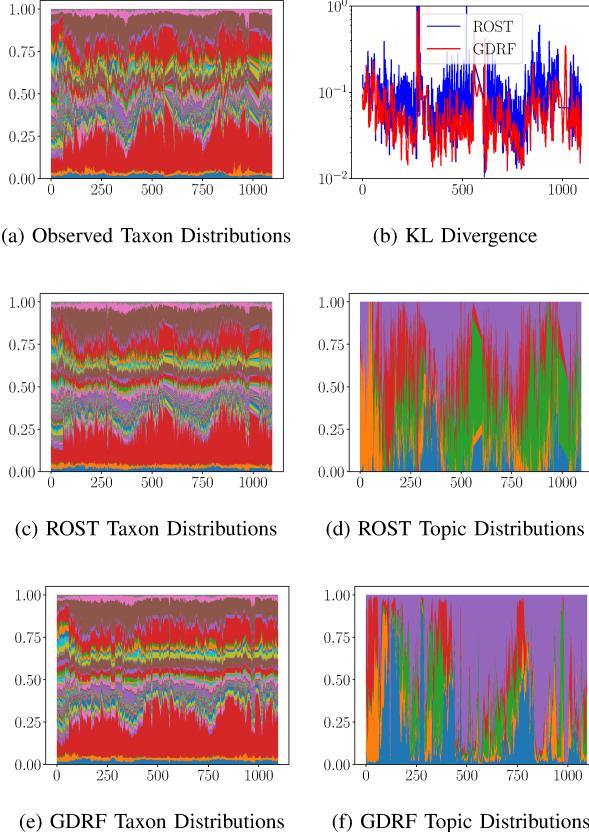


Fig. 8: MVCO phytoplankton taxa dataset. Ground truth taxon observations are produced by running phytoplankton images through a random forest classifier. The output labels are used as training data for both models. 8a: Observed phytoplankton taxon distributions. 8b: Mean KL Divergence between ROST (blue) or GDRF (red) inferred taxon distribution and ground-truth observed taxon distributions. 8c: Inferred taxon distributions for ROST model. 8d: Inferred latent topic distributions for ROST model. 8e: Inferred taxon distributions from GDRF model. 8f: Inferred latent topic distributions for GDRF model. The X-axis represents time in days. The Y-axis represents fraction of total probability.

the inaccuracy is relatively low, with no held-out windows exhibiting more than a factor of 2 increase in KL divergence from observations. Since the GDRF model trained without any held-out data performs well when compared to ROST, predictions from a GDRF model in unobserved regions or periods of time are likely accurate enough to use for adaptive sampling and informative path planning.

VI. CONCLUSION

The GDRF model provides both a generative process for spatially heterogeneous categorical observation fields, and a method for extrapolating from categorical observations underpinned by semantically meaningful Gaussian processes. By factoring the high-dimensional distribution over observation categories into a low-dimensional distribution over

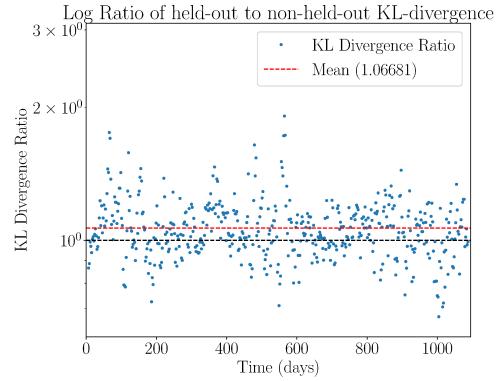


Fig. 9: Comparison of GDRF accuracy with and without held-out data. For each data point, a roughly 10 day window of MVCO data was held out of the training data set, to produce the held-out model. The average KL divergence between the model and the ground-truth in the window was calculated. The y-axis represents the ratio between this held-out KL divergence from ground truth and the non-held-out model's average KL divergence from ground truth within the same window. Values above one imply that the non-held-out model has a smaller KL divergence from ground truth than the held-out model.

topics, GDRFs are able to take advantage of the expressiveness of Gaussian process models for categorical data. We have demonstrated that GDRFs are capable of extracting meaningful structure from 2D spatial fields and interpolating on data sets with temporal gaps.

The effectiveness of GDRF inference on the MVCO phytoplankton taxon dataset demonstrates the viability of GDRFs for modeling biological systems in the ocean. We have also experimented with neural network classifiers capable of distinguishing over 140 phytoplankton taxa. Future work will explore the use of GDRFs to model 2-D and 3-D phytoplankton fields, labeled by neural network.

In this paper, we described a batch training procedure for learning GDRF models. Future work will develop techniques for doing online inference, enabling many applications such as IPP for mapping of map categorical fields autonomously.

REFERENCES

- [1] J. G. Bellingham and K. Rajan, "Robotics in remote and hostile environments," *Science*, vol. 318, no. 5853, pp. 1098–1102, 2007.
- [2] N. A. Cruz and A. C. Matos, "Adaptive sampling of thermoclines with autonomous underwater vehicles," in *MTS/IEEE Seattle, OCEANS 2010*, 2010.
- [3] Y. Zhang, M. A. Godin, J. G. Bellingham, and J. P. Ryan, "Using an autonomous underwater vehicle to track a coastal upwelling front," *IEEE Journal of Oceanic Engineering*, vol. 37, no. 3, pp. 338–347, 2012.
- [4] Y. Zhang, A. B. Baggeroer, and J. G. Bellingham, "Spectral-feature classification of oceanographic processes using an autonomous underwater vehicle," *IEEE Journal of Oceanic Engineering*, vol. 26, no. 4, pp. 726–741, 10 2001.
- [5] Z. A. Saigol, "Automated Planning for Hydrothermal Vent Prospecting Using AUVs : RSMG Report 8," Ph.D. dissertation, University of Birmingham, 2010.

[6] J. Farrell, S. Pang, W. Li, and R. Arrieta, "Chemical plume tracing experimental results with a REMUS AUV," in *Oceans 2003. Celebrating the Past ... Teaming Toward the Future (IEEE Cat. No.03CH37492)*. IEEE, 2003, pp. 962–968.

[7] R. Camilli, B. Bingham, M. Jakuba, H. Singh, and J. Whelan, "Integrating in-situ chemical sampling with AUV control systems," in *Ocean '04 - MTS/IEEE Techno-Ocean '04: Bridges across the Oceans - Conference Proceedings*, vol. 1. IEEE, 2004, pp. 101–109.

[8] M. Jakuba, J. Kinsey, D. Yoerger, and R. Camilli, "Exploration of the Gulf of Mexico Oil Spill with the Sentry Autonomous Underwater Vehicle," in *IROS2011 Workshop on Robotics for Environmental Monitoring*, no. September, 2011.

[9] M. A. Godin, Y. Zhang, J. P. Ryan, T. T. Hoover, and J. G. Bellingham, "Phytoplankton bloom patch center localization by the Tethys Autonomous Underwater Vehicle," in *OCEANS'11 MTS/IEEE KONA*. IEEE, 9 2011, pp. 1–6.

[10] T. O. Fossum, G. M. Fragozo, E. J. Davies, J. E. Ullgren, R. Mendes, G. Johnsen, I. Ellingsen, J. Eidsvik, M. Ludvigsen, and K. Rajan, "Toward adaptive robotic sampling of phytoplankton in the coastal ocean," *Science Robotics*, vol. 4, no. 27, p. eaav3041, 2 2019.

[11] S. Jamieson, How, Jonathan P., and Y. Girdhar, "Active Reward Learning for Co-Robotic Vision Based Exploration in Bandwidth Limited Environments," in *IEEE International Conference on Robotics and Automation*, 2020.

[12] J. Hwang, N. Bose, and S. Fan, "AUV Adaptive Sampling Methods: A Review," *Applied Sciences*, vol. 9, no. 15, p. 3145, 8 2019.

[13] J. Binney and G. S. Sukhatme, "Branch and bound for informative path planning," in *Proceedings - IEEE International Conference on Robotics and Automation*. Institute of Electrical and Electronics Engineers Inc., 2012, pp. 2147–2154.

[14] C. K. I. Williams and C. E. Rasmussen, "Gaussian Processes for Regression," in *Advances in Neural Information Processing Systems* 8, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. MIT Press, 1996, pp. 514–520.

[15] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. Cambridge: The MIT Press, 2005.

[16] R. Marchant, F. Ramos, and S. Sanner, "Sequential Bayesian Optimisation for Spatial-Temporal Monitoring," *International Conference on Uncertainty in Artificial Intelligence*, pp. 553–562, 2014.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *CoRR*, vol. abs/1506.0, 6 2015.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[19] E. C. Orenstein, O. Beijbom, E. E. Peacock, and H. M. Sosik, "WHOI-Plankton- A Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification," Tech. Rep., oct 2015.

[20] H. M. Sosik and R. J. Olson, "Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry," *Limnology and Oceanography: Methods*, vol. 5, no. 6, pp. 204–216, jun 2007.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, dec 2015.

[22] E. C. Garrido-Merchán and D. Hernández-Lobato, "Predictive Entropy Search for Multi-objective Bayesian Optimization with Constraints," 2016.

[23] Y. Girdhar and G. Dudek, "Modeling curiosity in a mobile robot for long-term autonomous exploration and monitoring," *Autonomous Robots*, vol. 40, no. 7, pp. 1267–1278, 2016.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[25] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, p. 77, 4 2012.

[26] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. SUPPL. 1, pp. 5228–5235, 2004.

[27] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, vol. II. IEEE Computer Society, 2005, pp. 524–531.

[28] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 9 2008.

[29] G. Hripcak and D. J. Albers, "Next-generation phenotyping of electronic health records," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 117–121, 1 2013.

[30] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[31] D. J. Newman and S. Block, "Probabilistic topic decomposition of an eighteenth-century american newspaper," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 6, pp. 753–767, 4 2006.

[32] X. Wang and E. Grimson, "Spatial Latent Dirichlet Allocation," in *Advances in Neural Information Processing Systems* 20, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 1577–1584.

[33] Y. Girdhar, P. Giguère, and G. Dudek, "Autonomous adaptive exploration using realtime online spatiotemporal topic modeling," *International Journal of Robotics Research*, vol. 33, no. 4, pp. 645–657, 2014.

[34] K. Doherty, G. Flaspohler, N. Roy, and Y. Girdhar, "Approximate Distributed Spatiotemporal Topic Models for Multi-Robot Terrain Characterization," in *IEEE International Conference on Intelligent Robots and Systems*, 2018, pp. 3730–3737.

[35] A. Kalmbach, H. M. Sosik, and Y. Girdhar, "Learning Seasonal Phytoplankton Communities with Topic Models," in *OCEANS MTS/IEEE Anchorage*, 2017.

[36] G. Matheron, "Principles of geostatistics," *Economic Geology*, vol. 58, no. 8, pp. 1246–1266, 12 1963.

[37] C. Noel, "the Origins of Kriging," *Mathematical Geology*, vol. 22, no. 3, pp. 47–55, 1990.

[38] A. G. Journel, "Nonparametric estimation of spatial distributions," *Journal of the International Association for Mathematical Geology*, vol. 15, no. 3, pp. 445–468, 1983.

[39] R. A. Olea, *Geostatistics for Engineers and Earth Scientists*. Boston, MA: Springer US, 1999.

[40] J. P. Fentanes, A. Badiee, T. Duckett, J. Evans, S. Pearson, and G. Cielniak, "Kriging-Based Robotic Exploration for Soil Moisture Mapping Using a Cosmic-Ray Sensor," 11 2018.

[41] K. Jerosch, M. Schlieter, and R. Pesch, "Spatial analysis of marine categorical information using indicator kriging applied to georeferenced video mosaics of the deep-sea Håkon Mosby Mud Volcano," *Ecological Informatics*, vol. 1, no. 4, pp. 391–406, 2006.

[42] J. Binney, A. Krause, and G. S. Sukhatme, "Optimizing waypoints for monitoring spatiotemporal phenomena," *International Journal of Robotics Research*, vol. 32, no. 8, pp. 873–888, 7 2013.

[43] J. Das, J. Harvey, F. Py, H. Vathsangam, R. Graham, K. Rajan, and G. S. Sukhatme, "Hierarchical probabilistic regression for AUV-based adaptive sampling of marine phenomena," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2013, pp. 5571–5578.

[44] V. Suryan and P. Tokekar, "Learning a Spatial Field with Gaussian Process Regression in Minimum Time," in *The 13th International Workshop on the Algorithmic Foundations of Robotics*, 2018.

[45] G. E. Berget, T. O. Fossum, T. A. Johansen, J. Eidsvik, and K. Rajan, "Adaptive Sampling of Ocean Processes Using an AUV with a Gaussian Proxy Model," *IFAC-PapersOnLine*, vol. 51, no. 29, pp. 238–243, 2018.

[46] T. O. Fossum, J. Eidsvik, I. Ellingsen, M. O. Alver, G. M. Fragozo, G. Johnsen, R. Mendes, M. Ludvigsen, and K. Rajan, "Information-driven robotic sampling in the coastal ocean," *Journal of Field Robotics*, vol. 35, no. 7, pp. 1101–1121, 10 2018.

[47] G. Flaspohler, V. Preston, A. P. M. Michel, Y. Girdhar, and N. Roy, "Information-Guided Robotic Maximum Seek-and-Sample in Partially Observable Continuous Environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3782–3789, 10 2019.

[48] M. D. Hoffman and D. M. Blei, "Structured stochastic variational inference," in *Journal of Machine Learning Research*, vol. 38, 2015, pp. 361–369.

[49] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, "GPytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration," in *Advances in Neural Information Processing Systems*, vol. 2018-Decem, 9 2018, pp. 7576–7586.

[50] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in *NeurIPS*, 2017.

[51] T. C. Austin, J. B. Edson, W. R. McGillis, M. Purcell, R. A. Petit, M. K. McElroy, C. W. Grant, J. Ware, and S. K. Hurst, “A network-based telemetry architecture developed for the Martha’s Vineyard Coastal Observatory,” *IEEE Journal of Oceanic Engineering*, vol. 27, no. 2, pp. 228–234, 4 2002.

[52] R. J. Olson and H. M. Sosik, “A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot,” *Limnology and Oceanography: Methods*, vol. 5, no. 6, pp. 195–203, 6 2007.