Generalized Online Sparse Gaussian Processes with Application to Persistent Mobile Robot Localization

Kian Hsiang Low[1], Nuo Xu[1], Jie Chen[2], Keng Kiat Lim[1], and Etkin Bariş Özgül[1]

[1] Nat'l Univ. of Singapore, Singapore
{lowkh,xunuo,kengkiat,ebozgul}@comp.nus.edu.sg
[2] Singapore-MIT Alliance for Research and Technology, Singapore
chenjie@smart.mit.edu

Abstract. This paper presents a novel online sparse *Gaussian process* (GP) approximation method [3] that is capable of achieving *constant* time and memory (i.e., independent of the size of the data) per time step. We theoretically guarantee its predictive performance to be equivalent to that of a sophisticated offline sparse GP approximation method. We empirically demonstrate the practical feasibility of using our online sparse GP approximation method through a real-world persistent mobile robot localization experiment.

1 Introduction

Gaussian process (GP) models are a rich class of Bayesian non-parametric models that can perform probabilistic regression by providing Gaussian predictive distributions with formal measures of the predictive uncertainty. Unfortunately, the expressive power of a full GP model comes at a cost of poor scalability (i.e., cubic time) in the size of the data, which hinders its practical use for performing real-time predictions necessary in many time-critical applications and decision support systems (e.g., ocean sensing, traffic monitoring, geographical information systems) that need to process and analyze huge quantities of data streaming in over time (e.g., in astronomy, internet traffic, meteorology, surveillance). When the data stream is expected to be (possibly indefinitely) long, it is also computationally impractical to repeatedly use existing offline sparse GP approximation methods [2] or online GP model [1] for training at each time step because they incur, respectively, linear and quadratic time in the data size per time step.

This paper presents a novel online sparse GP approximation method [3] (Section 3) that, in contrast to existing works mentioned above, is capable of achieving *constant* time and memory (i.e., independent of the size of the data/observations) per time step. We provide a theoretical guarantee on its predictive performance to be equivalent to that of the offline sparse *partially independent training conditional* (PITC) approximation method. Our proposed method [3] generalizes the sparse online GP model of [1] by relaxing its conditional independence assumption significantly, hence potentially improving the predictive performance. We empirically demonstrate the practical feasibility of using our generalized online sparse GP approximation method [3] through a real-world persistent mobile robot localization experiment described in Section 4.

T. Calders et al. (Eds.): ECML PKDD 2014, Part III, LNCS 8726, pp. 499–503, 2014.

2 Background

A Gaussian process (GP) model can be used to perform probabilistic regression as follows: Let $\mathcal{X}$ be a set representing the input domain such that each input $x \in \mathcal{X}$ denotes a d-dimensional feature vector and is associated with a realized output value z_x (random output variable Z_x) if it is observed (unobserved). Let $\{Z_x\}_{x\in\mathcal{X}}$ denote a GP, that is, every finite subset of $\{Z_x\}_{x\in\mathcal{X}}$ has a multivariate Gaussian distribution. The GP is fully specified by its *prior* mean $\mu_x \triangleq \mathbb{E}[Z_x]$ and covariance $\sigma_{xx'} \triangleq \text{cov}[Z_x, Z_{x'}]$ for all $x, x' \in \mathcal{X}$. Supposing a column vector $z_{\mathcal{D}}$ of realized outputs is observed for some set $\mathcal{D} \in \mathcal{X}$ of inputs, the full GP model can exploit these observations to predict the unobserved measurement for any input $x \in \mathcal{X} \setminus \mathcal{D}$ as well as provide its predictive uncertainty using a Gaussian predictive distribution $p(z_x|x, \mathcal{D}, z_{\mathcal{D}}) = \mathcal{N}(\mu_{x|\mathcal{D}}, \sigma_{xx|\mathcal{D}})$ with the following *posterior* mean and variance, respectively:

$$\mu_{x|\mathcal{D}} \triangleq \mu_x + \Sigma_{x\mathcal{D}}\Sigma_{\mathcal{D}\mathcal{D}}^{-1}(z_{\mathcal{D}} - \mu_{\mathcal{D}}) \quad \text{and} \quad \sigma_{xx|\mathcal{D}} \triangleq \sigma_{xx} - \Sigma_{x\mathcal{D}}\Sigma_{\mathcal{D}\mathcal{D}}^{-1}\Sigma_{\mathcal{D}x} \tag{1}$$

where $\mu_{\mathcal{D}}$ is a column vector with mean components $\mu_{x'}$ for all $x' \in \mathcal{D}$, $\Sigma_{x\mathcal{D}}$ is a row vector with covariance components $\sigma_{xx'}$ for all $x' \in \mathcal{D}$, $\Sigma_{\mathcal{D}x}$ is the transpose of $\Sigma_{x\mathcal{D}}$, and $\Sigma_{\mathcal{D}\mathcal{D}}$ is a matrix with components $\sigma_{x'x''}$ for all $x', x'' \in \mathcal{D}$.

The key limitation hindering the practical use of the full GP model is that computing (1) requires inverting the covariance matrix $\Sigma_{\mathcal{D}\mathcal{D}}$, which incurs $\mathcal{O}(|\mathcal{D}|^3)$ time and $\mathcal{O}(|\mathcal{D}|^2)$ memory. To improve its scalability, the sparse *partially independent training conditional* (PITC) [2] approximation method is the most general form of a class of reduced-rank covariance matrix approximation methods in [2] exploiting the notion of a support set $\mathcal{S} \subset \mathcal{X}$. PITC computes a Gaussian predictive distribution of the unobserved measurement for any $x \in \mathcal{X} \setminus \mathcal{D}$ with the following posterior mean and variance:

$$\mu_{x|\mathcal{D}}^{\text{PITC}} \triangleq \mu_x + \Gamma_{x\mathcal{D}}(\Gamma_{\mathcal{D}\mathcal{D}} + \Lambda)^{-1}(z_{\mathcal{D}} - \mu_{\mathcal{D}}) \quad \text{and} \quad \sigma_{xx|\mathcal{D}}^{\text{PITC}} \triangleq \sigma_{xx} - \Gamma_{x\mathcal{D}}(\Gamma_{\mathcal{D}\mathcal{D}} + \Lambda)^{-1}\Gamma_{\mathcal{D}x} \tag{2}$$

where $\Gamma_{\mathcal{A}\mathcal{A}'} = \Sigma_{\mathcal{A}\mathcal{S}}\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\Sigma_{\mathcal{S}\mathcal{A}'}$ for all $\mathcal{A}, \mathcal{A}' \subset \mathcal{X}$ and Λ is a block-diagonal matrix constructed from the N diagonal blocks of $\Sigma_{\mathcal{D}\mathcal{D}|\mathcal{S}}$, each of which is a matrix $\Sigma_{\mathcal{D}_n\mathcal{D}_n|\mathcal{S}}$ for $n = 1, \cdots, N$ where $\mathcal{D} = \bigcup_{n=1}^{N} \mathcal{D}_n$. The covariance matrix $\Sigma_{\mathcal{D}\mathcal{D}}$ in (1) is approximated by a reduced-rank matrix $\Gamma_{\mathcal{D}\mathcal{D}}$ summed with the resulting sparsified residual matrix Λ in (2). So, computing either $\mu_{x|\mathcal{D}}^{\text{PITC}}$ or $\sigma_{xx|\mathcal{D}}^{\text{PITC}}$ (2), which requires inverting the approximated covariance matrix $\Gamma_{\mathcal{D}\mathcal{D}} + \Lambda$, incurs $\mathcal{O}(|\mathcal{D}|(|\mathcal{S}|^2 + (|\mathcal{D}|/N)^2))$ time and $\mathcal{O}(|\mathcal{S}|^2 + (|\mathcal{D}|/N)^2)$ memory. The sparse *fully independent training conditional* (FITC) approximation method is a special case of PITC where Λ is a diagonal matrix constructed from $\sigma_{x'x'|\mathcal{S}}$ for all $x' \in \mathcal{D}$ (i.e., $N = |\mathcal{D}|$).

3 Generalized Online Sparse GP (GOSGP) Approximation

The key idea of our GOSGP approximation method [3] is to summarize the newly gathered data/observations at regular time intervals/slices, assimilate the summary information of the new data with that of all the previously gathered data/observations, and then exploit the resulting assimilated summary information to compute a Gaussian predictive distribution of the unobserved measurement for any input. Let $x_{1:t-1} \triangleq \{x_1, \ldots, x_{t-1}\}$

denote a set of inputs from time steps 1 to $t-1$, each time slice n span time steps $(n-1)\tau+1$ to $n\tau$ for some user-defined slice size $\tau \in \mathbb{Z}^+$, and the number of time slices available thus far up until time step t be denoted by N (i.e., $N\tau < t$).

Definition 1 (Slice Summary). *Given a support set $\mathcal{S} \subset \mathcal{X}$, a subset $\mathcal{D}_n \triangleq x_{(n-1)\tau+1:n\tau} \in x_{1:t-1}$ of inputs associated with time slice n, and the column vector $z_{\mathcal{D}_n} = z_{(n-1)\tau+1:n\tau}$ of corresponding realized measurements, the slice summary of time slice n is defined as a tuple $(\mu^n_{\circledS}, \Sigma^n_{\circledS})$ for $n = 1, \ldots, N$ where $\mu^n_{\circledS} \triangleq \Sigma_{\mathcal{S}\mathcal{D}_n}\Sigma^{-1}_{\mathcal{D}_n\mathcal{D}_n|\mathcal{S}}(z_{\mathcal{D}_n} - \mu_{\mathcal{D}_n})$ and $\Sigma^n_{\circledS} \triangleq \Sigma_{\mathcal{S}\mathcal{D}_n}\Sigma^{-1}_{\mathcal{D}_n\mathcal{D}_n|\mathcal{S}}\Sigma_{\mathcal{D}_n\mathcal{S}}$ such that $\mu_{\mathcal{D}_n}$ is defined in a similar manner as $\mu_{\mathcal{D}}$ in* (1) *and $\Sigma_{\mathcal{D}_n\mathcal{D}_n|\mathcal{S}}$ is a posterior covariance matrix with components $\sigma_{xx'|\mathcal{S}}$ for all $x, x' \in \mathcal{D}_n$, each of which is defined in a similar way as* (1).

Definition 2 (Assimilated Summary). *Given $(\mu^n_{\circledS}, \Sigma^n_{\circledS})$, the assimilated summary $(\mu^n_{@}, \Sigma^n_{@})$ of time slices 1 to n is updated from the assimilated summary $(\mu^{n-1}_{@}, \Sigma^{n-1}_{@})$ of time slices 1 to $n-1$ using $\mu^n_{@} \triangleq \mu^{n-1}_{@} + \mu^n_{\circledS}$ and $\Sigma^n_{@} \triangleq \Sigma^{n-1}_{@} + \Sigma^n_{\circledS}$ for $n = 1, \ldots, N$ where $\mu^0_{@} \triangleq 0$ and $\Sigma^0_{@} \triangleq \Sigma_{\mathcal{S}\mathcal{S}}$.*

Remark 1. After constructing and assimilating $(\mu^n_{\circledS}, \Sigma^n_{\circledS})$ with $(\mu^{n-1}_{@}, \Sigma^{n-1}_{@})$ to form $(\mu^n_{@}, \Sigma^n_{@})$, $\mathcal{D}_n = x_{(n-1)\tau+1:n\tau}$, $z_{\mathcal{D}_n} = z_{(n-1)\tau+1:n\tau}$, and $(\mu^n_{\circledS}, \Sigma^n_{\circledS})$ (Definition 1) are no longer needed and can be removed from memory. As a result, at time step t where $N\tau + 1 \leq t \leq (N+1)\tau$, only $(\mu^N_{@}, \Sigma^N_{@})$, $x_{N\tau+1:t-1}$, and $z_{N\tau+1:t-1}$ have to be kept in memory, thus requiring only constant memory (i.e., independent of t).

Remark 2. The slice summaries are constructed and assimilated at a regular time interval of τ, specifically, at time steps $N\tau + 1$ for $N \in \mathbb{Z}^+$.

Theorem 1. *Given $\mathcal{S} \subset \mathcal{X}$ and $(\mu^N_{@}, \Sigma^N_{@})$, our GOSGP approximation method computes a Gaussian predictive distribution $p(z_t|x_t, \mu^N_{@}, \Sigma^N_{@}) = \mathcal{N}(\widetilde{\mu}_{x_t}, \widetilde{\sigma}_{x_t x_t})$ of the measurement for any $x_t \in \mathcal{X}$ at time step t (i.e., $N\tau + 1 \leq t \leq (N+1)\tau$) where*

$$\widetilde{\mu}_{x_t} \triangleq \mu_{x_t} + \Sigma_{x_t\mathcal{S}}\left(\Sigma^N_{@}\right)^{-1}\mu^N_{@} \text{ and } \widetilde{\sigma}_{x_t x_t} \triangleq \sigma_{x_t x_t} - \Sigma_{x_t\mathcal{S}}\left(\Sigma^{-1}_{\mathcal{S}\mathcal{S}} - \left(\Sigma^N_{@}\right)^{-1}\right)\Sigma_{\mathcal{S}x_t}. \tag{3}$$

If $t = N\tau + 1$, $\widetilde{\mu}_{x_t} = \mu^{\mathrm{PITC}}_{x_t|x_{1:t-1}}$ and $\widetilde{\sigma}_{x_t x_t} = \sigma^{\mathrm{PITC}}_{x_t x_t|x_{1:t-1}}$.

Remark 1. Theorem 1 implies that our GOSGP approximation method [3] is in fact equivalent to an online learning formulation/variant of the offline PITC (Section 2). Supposing $\tau < |\mathcal{S}|$, the $\mathcal{O}(t|\mathcal{S}|^2)$ time incurred by offline PITC can then be reduced to $\mathcal{O}(\tau|\mathcal{S}|^2)$ time (i.e., time independent of t) incurred by GOSGP [3] at time steps $t = N\tau + 1$ for $N \in \mathbb{Z}^+$ when slice summaries are constructed and assimilated. Otherwise, GOSGP [3] only incurs $\mathcal{O}(|\mathcal{S}|^2)$ time per time step.

Remark 2. The above equivalence result allows the structural property of GOSGP [3] to be elucidated using that of offline PITC: The measurements $Z_{\mathcal{D}_1}, \ldots, Z_{\mathcal{D}_N}, Z_{x_t}$ between different time slices are assumed to be conditionally independent given $Z_{\mathcal{S}}$. Such an assumption enables the data gathered during each time slice to be summarized independently of that in other time slices. Increasing slice size τ (i.e., less frequent assimilations of larger slice summaries) relaxes this conditional independence assumption (hence, potentially improving the predictive performance), but incurs more time at time steps when slice summaries are constructed and assimilated (see Remark 1).

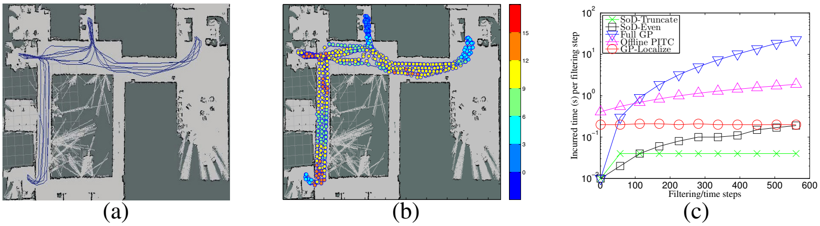

Fig. 1. (a) Pioneer 3-DX mobile robot trajectory of about 280 m in SMART FM IRG office/lab generated by AMCL package in ROS, along which (b) 561 relative light (%) observations/data are gathered at locations denoted by small colored circles. (c) Graphs of incurred time (s) per time step vs. number of time steps comparing different GP localization algorithms.

Remark 3. Since offline PITC generalizes offline FITC, our GOSGP approximation method [3] generalizes the online learning variant of FITC (i.e., $\tau = 1$) [1].

When $N\tau+1 < t \leq (N+1)\tau$ (i.e., before the next slice summary of time slice $N+1$ is constructed and assimilated), the most recent observations (i.e., $\mathcal{D}' \triangleq x_{N\tau+1:t-1}$ and $z_{\mathcal{D}'} = z_{N\tau+1:t-1}$), which are often highly informative, are not used to update $\widetilde{\mu}_{x_t}$ and $\widetilde{\sigma}_{x_t x_t}$ (3). This may hurt the predictive performance when τ is large. To resolve this, we exploit incremental update formulas of Gaussian posterior mean and variance [3] to update $\widetilde{\mu}_{x_t}$ and $\widetilde{\sigma}_{x_t x_t}$ with the most recent observations, thereby yielding a Gaussian predictive distribution $p(z_t|x_t, \mu^N_{@}, \Sigma^N_{@}, \mathcal{D}', z_{\mathcal{D}'}) = \mathcal{N}(\widetilde{\mu}_{x_t|\mathcal{D}'}, \widetilde{\sigma}_{x_t x_t|\mathcal{D}'})$ where

$$\widetilde{\mu}_{x_t|\mathcal{D}'} \triangleq \widetilde{\mu}_{x_t} + \widetilde{\Sigma}_{x_t\mathcal{D}'}\widetilde{\Sigma}^{-1}_{\mathcal{D}'\mathcal{D}'}(z_{\mathcal{D}'} - \widetilde{\mu}_{\mathcal{D}'}) \text{ and } \widetilde{\sigma}_{x_t x_t|\mathcal{D}'} \triangleq \widetilde{\sigma}_{x_t x_t} - \widetilde{\Sigma}_{x_t\mathcal{D}'}\widetilde{\Sigma}^{-1}_{\mathcal{D}'\mathcal{D}'}\widetilde{\Sigma}_{\mathcal{D}'x_t} \quad (4)$$

such that $\widetilde{\mu}_{\mathcal{D}'}$ is a column vector with mean components $\widetilde{\mu}_x$ (i.e., defined similarly to (3)) for all $x \in \mathcal{D}'$, $\widetilde{\Sigma}_{x_t\mathcal{D}'}$ is a row vector with covariance components $\widetilde{\sigma}_{x_t x}$ (i.e., defined similarly to (3)) for all $x \in \mathcal{D}'$, $\widetilde{\Sigma}_{\mathcal{D}'x_t}$ is the transpose of $\widetilde{\Sigma}_{x_t\mathcal{D}'}$, and $\widetilde{\Sigma}_{\mathcal{D}'\mathcal{D}'}$ is a matrix with covariance components $\widetilde{\sigma}_{xx'}$ (i.e., defined similarly to (3)) for all $x, x' \in \mathcal{D}'$.

Theorem 2. *Computing* (4) *incurs* $\mathcal{O}(\tau|\mathcal{S}|^2)$ *time at time steps* $t = N\tau+1$ *for* $N \in \mathbb{Z}^+$ *and* $\mathcal{O}(|\mathcal{S}|^2)$ *time otherwise. It requires* $\mathcal{O}(|\mathcal{S}|^2)$ *memory at each time step.*

So, GOSGP [3] incurs constant time and memory (i.e., independent of t) per time step.

4 Experiments and Discussion

In contrast to existing localization algorithms that train the GP observation model of a Bayes filter offline, GOSGP [3] is used to learn it *online* for persistent robot localization and the resulting algorithm is called *GP-Localize* [3]. The *adaptive Monte Carlo localization* (AMCL) package in the *Robot Operating System* (ROS) is run on a Pioneer 3-DX mobile robot mounted with a SICK LMS200 laser rangefinder to determine its trajectory (Fig. 1a) and the 561 locations at which the relative light measurements are taken using a weather board (Fig. 1b); these locations are assumed to be ground truth. For empirical evaluation of GP-Localize with other real-world datasets, refer to [3].

The localization performance/error (i.e., distance between the robot's estimated and true locations) and scalability of GP-Localize are compared to that of two sparse GP localization algorithms [3]: (a) The *Subset of Data (SoD)-Truncate* method uses $|\mathcal{S}| = 10$

most recent observations (i.e., compared to $|\mathcal{D}'| < \tau = 10$ most recent observations considered by GOSGP [3] besides the assimilated summary) as training data at each time step while (b) the *SoD-Even* method uses $|\mathcal{S}| = 40$ observations (i.e., compared to the support set of $|\mathcal{S}| = 40$ possibly unobserved locations selected *prior* to localization and exploited by GOSGP [3]) evenly distributed over the time of localization. The scalability of GP-Localize is further compared to that of GP localization algorithms employing full GP and offline PITC. GP-Localize, SoD-Truncate, and SoD-Even achieve, respectively, localization errors of 2.1 m, 5.4 m, and 4.6 m averaged over all 561 time steps and 3 runs. Fig. 1c shows the time incurred by GP-Localize, SoD-Truncate, SoD-Even, full GP, and offline PITC at each time step. GP-Localize is clearly much more scalable (i.e., constant time) than full GP and offline PITC. Though it incurs slightly more time than SoD-Truncate and SoD-Even, it can localize significantly better.

Acknowledgments. This work was supported by Singapore-MIT Alliance for Research and Technology Subaward Agreement No. 41 R-252-000-527-592.

References

1. Csató, L., Opper, M.: Sparse online Gaussian processes. Neural Comput. 14, 641–669 (2002)
2. Quiñonero-Candela, J., Rasmussen, C.E.: A unifying view of sparse approximate Gaussian process regression. JMLR 6, 1939–1959 (2005)
3. Xu, N., Low, K.H., Chen, J., Lim, K.K., Özgül, E.B.: GP-Localize: Persistent mobile robot localization using online sparse Gaussian process observation model. In: Proc. AAAI (2014)