Supplementary Material for Scalable Nonparametric Bayesian Inference on Point Processes with Gaussian Processes

Yves-Laurent KOM SAMO YVES-LAURENT.KOMSAMO@ENG.OX.AC.UK
Stephen Roberts SJROB@ROBOTS.OX.AC.UK
Department of Engineering Science and Oxford-Man Institute, University of Oxford

Appendix

A. There exists a Cox process with an a.s. $\mathcal{C}^\infty$ intensity coinciding with any finite dimensional prior.

In this section we prove the proposition below.

Proposition .1 *Let $\mathbb{Q}$ be an $(n+1)$ dimensional continuous probability distribution whose density has support $\bigotimes_{i=1}^{n+1}]0, +\infty[$, and let $x_1, \dots, x_n$ be n points on a compact domain $\mathcal{S} \subset \mathbb{R}^d$. There exists an almost surely non-negative and $\mathcal{C}^\infty$ stochastic process λ on $\mathcal{S}$ such that*

$$\left(\lambda(x_1), ..., \lambda(x_n), \int_{\mathcal{S}} \lambda(x) dx\right) \sim \mathbb{Q}.$$

Proof Let

$$(y_1, \dots, y_n, I) \sim \mathbb{Q}$$

and

$$(y_1(\omega), \dots, y_n(\omega), I(\omega))$$

a random draw. Let us denote $x^j, j \leq d$ the j-th coordinate of $x \in \mathbb{R}^d$. We consider the family of functions parametrized by $\alpha \in \mathbb{R}$:

$$\begin{aligned} f(\omega, x, \alpha) = &\exp\left(\alpha \sum_{j=1}^{d} \prod_{l=1}^{n} (x^j - x_l^j)^2\right) \qquad (1) \\ &\times \sum_{l=1}^{n} y_l(\omega) \frac{1}{d} \sum_{j=1}^{d} \prod_{k \neq l} \left(\frac{x^j - x_k^j}{x_l^j - x_k^j}\right)^2. \end{aligned}$$

We note that $\forall \alpha, x_i, \ f(\omega, x_i, \alpha) = y_i(\omega)$. Let us define the polynomial

$$P(x) = \sum_{l=1}^{n} y_l(\omega) \frac{1}{d} \sum_{j=1}^{d} \prod_{k \neq l} \left(\frac{x^j - x_k^j}{x_l^j - x_k^j}\right)^2.$$

As P is continuous, it is bounded on the compact $\mathcal{S}$, and reaches its bounds. Thus we have

$$\exists \, m_p, M_p \geq 0, \text{ s.t. } \forall x \in \mathcal{S}, \ 0 \leq m_p \leq P(x) \leq M_p.$$

Similarly, if we define

$$R(\alpha, x) = \exp\left(\alpha \sum_{j=1}^{d} \prod_{l=1}^{n} (x^j - x_l^j)^2\right) = R(1, x)^\alpha,$$

it follows that

$$\exists \, m_q, M_q > 1, \text{ s.t. } \forall x \in \mathcal{S}, \ 1 < m_q \leq R(1, x) \leq M_q.$$

Hence,

$$m_p m_q^\alpha \mu(\mathcal{S}) \leq \int_{\mathcal{S}} f(\omega, x, \alpha) dx \leq M_p M_q^\alpha \mu(\mathcal{S}). \qquad (2)$$

Moreover, we note that $\alpha \to \int_{\mathcal{S}} f(\omega, x, \alpha) dx$ is continuous on $\mathbb{R}$ as its restriction to any bounded interval is continuous (by dominated convergence theorem). Furthermore, given that $m_q, M_q > 1$, it follows from Equation (2) that

$$\lim_{\alpha \to +\infty} \int_{\mathcal{S}} f(\omega, x, \alpha) dx = +\infty$$

and

$$\lim_{\alpha \to -\infty} \int_{\mathcal{S}} f(\omega, x, \alpha) dx = 0.$$

Hence, by intermediate value theorem,

$$\forall \, I(\omega) > 0, \ \exists \alpha^*(\omega) \text{ s.t. } I(\omega) = \int_{\mathcal{S}} f(\omega, x, \alpha^*(\omega)) dx.$$

Finally, let us define the stochastic process λ on $\mathcal{S}$ as

$$\omega \to \lambda(\omega, x) := f(\omega, x, \alpha^*(\omega)).$$

To summarise,

$$\forall \, x_i, \lambda(\omega, x_i) := f(\omega, x_i, \alpha^*(\omega)) = y_i(\omega),$$

$$I(\omega) = \int_{\mathcal{S}} \lambda(\omega, x) dx,$$

and

$$(y_1, \dots, y_n, I) \sim \mathbb{Q} :$$

this implies $\left(\lambda(x_1), ..., \lambda(x_n), \int_{\mathcal{S}} \lambda(x) dx\right) \sim \mathbb{Q}$. Finally,

$$\forall \, x \in \mathcal{S}, \ \lambda(\omega, x) \geq 0, \text{ and } \forall \, \omega, \ x \to \lambda(\omega, x) \text{ is } \mathcal{C}^\infty,$$

which concludes our proof.

B. Proof of convergence of Algorithm 1

The idea behind the proof is to show that the sequence of maximum utility

$$u_k = \max_{s \in \mathcal{S}} \tilde{\mathcal{U}}(\{s'_1, ..., s'_{k-1}\} \cup \{s\})$$

is positive, increasing and upper-bounded and thus converges to a strictly positive limit. This would then imply that

$$\frac{u_{k+1} - u_k}{u_k} \underset{k \to \infty}{\longrightarrow} 0$$

and subsequently that

$$\forall\, 0 < \alpha < 1, \exists\, k_{\text{lim}} \in \mathbb{N} \text{ s.t. } \forall\, k > k_{\text{lim}}, \frac{u_{k+1} - u_k}{u_k} < \alpha$$

or in other words Algorithm 1 always stops in finite time.

To show that $\forall k > 0$, $u_k > 0$, we note that $\Sigma^*_{\mathcal{D}'\mathcal{D}'}(\tilde{\theta}_i)$ is a covariance matrix and as such it is positive definite. It follows that $\Sigma^{*-1}_{\mathcal{D}'\mathcal{D}'}(\tilde{\theta}_i)$ is also positive definite. We further note that the j-th diagonal term of $\Sigma^*_{\mathcal{D}\mathcal{D}'}(\tilde{\theta}_i)\Sigma^{*-1}_{\mathcal{D}'\mathcal{D}'}(\tilde{\theta}_i)\Sigma^{*T}_{\mathcal{D}\mathcal{D}'}(\tilde{\theta}_i)$ can be written as $x_j^T \Sigma^{*-1}_{\mathcal{D}'\mathcal{D}'}(\tilde{\theta}_i) x_j$ where x_j is the j-th column of $\Sigma^{*T}_{\mathcal{D}\mathcal{D}'}(\tilde{\theta}_i)$. Hence, by virtue of the positive definitiveness of $\Sigma^{*-1}_{\mathcal{D}'\mathcal{D}'}(\tilde{\theta}_i)$, the diagonal terms of $\Sigma^*_{\mathcal{D}\mathcal{D}'}(\tilde{\theta}_i)\Sigma^{*-1}_{\mathcal{D}'\mathcal{D}'}(\tilde{\theta}_i)\Sigma^{*T}_{\mathcal{D}\mathcal{D}'}(\tilde{\theta}_i)$ are all positive, which proves that the utility function $\tilde{\mathcal{U}}$ is positive, and subsequently that $\forall k > 0, u_k > 0$.

To show that $(u_k)_{k \in \mathbb{N}^*}$ is upper-bounded, we note that the matrix

$$C_{i\mathcal{D}'} = \Sigma^*_{\mathcal{D}\mathcal{D}}(\tilde{\theta}_i) - \Sigma^*_{\mathcal{D}\mathcal{D}'}(\tilde{\theta}_i)\Sigma^{*-1}_{\mathcal{D}'\mathcal{D}'}(\tilde{\theta}_i)\Sigma^{*T}_{\mathcal{D}\mathcal{D}'}(\tilde{\theta}_i)$$

where the notation is as per the rest of the paper, is an auto-covariance matrix, and as such has positive diagonal elements. Hence,

$$\text{Tr}(\Sigma^*_{\mathcal{D}\mathcal{D}}(\tilde{\theta}_i)) \geq \text{Tr}(\Sigma^*_{\mathcal{D}\mathcal{D}'}(\tilde{\theta}_i)\Sigma^{*-1}_{\mathcal{D}'\mathcal{D}'}(\tilde{\theta}_i)\Sigma^{*T}_{\mathcal{D}\mathcal{D}'}(\tilde{\theta}_i))$$

and finally

$$\forall\, k \in \mathbb{N}^*, u_k \leq \frac{1}{N} \sum_{i=1}^{N} \text{Tr}(\Sigma^*_{\mathcal{D}\mathcal{D}}(\tilde{\theta}_i)).$$

Moreover, we note that showing that $(u_k)_{k \in \mathbb{N}^*}$ is increasing is equivalent to showing that $(v_k)_{k \in \mathbb{N}^*}$ with

$$v_k = \min_{s \in \mathcal{S}} \frac{1}{N} \sum_{i=1}^{N} \text{Tr}(C_{i\{s'_1,...,s'_{k-1}\} \cup \{s\}})$$

is decreasing. We recall that $C_{i\{s'_1,...,s'_{k-1}\} \cup \{s\}}$ is the covariance matrix of the values of the stationary Gaussian Process of our model at the data points, conditioned on its values at $\{s'_1, ..., s'_{k-1}\} \cup \{s\}$.

It follows from the law of iterated expectations that $C_{i\{s'_1,...,s'_{k-1}\} \cup \{s\}}$ could also be seen as the covariance matrix of the values of a conditional Gaussian Process at the data points, [1] conditioned on its value at s. Hence,

$$C_{i\{s'_1,...,s'_{k-1}\} \cup \{s\}} =$$
$$C_{i\{s'_1,...,s'_{k-1}\}} - \frac{1}{\hat{\Sigma}_{ss}(\tilde{\theta}_i)} \hat{\Sigma}_{\mathcal{D}\{s\}}(\tilde{\theta}_i)\hat{\Sigma}^T_{\mathcal{D}\{s\}}(\tilde{\theta}_i)$$

where $\hat{\Sigma}_{XY}$ denotes the covariance matrix between the values of the conditional GP at points in X and at points in Y. In particular, $\hat{\Sigma}_{ss}(\tilde{\theta}_i)$ is a positive scalar. What's more the diagonal elements of $\hat{\Sigma}_{\mathcal{D}\{s\}}(\tilde{\theta}_i)\hat{\Sigma}^T_{\mathcal{D}\{s\}}(\tilde{\theta}_i)$ are all non-negative. Hence,

$$\forall s \in \mathcal{S}, \text{Tr}(C_{i\{s'_1,...,s'_{k-1}\} \cup \{s\}}) \leq \text{Tr}(C_{i\{s'_1,...,s'_{k-1}\}})$$

and averaging over the set of hyper-parameters θ_i and taking the min we get

$$\forall\, k \geq 2, v_k \leq v_{k-1}$$

which concludes the proof.

C. Proof of the rate of convergence of Algorithm 1 and that u_f in Algorithm 1 converges to $\frac{1}{N}\sum_{i=1}^{N} \textbf{Tr}(\Sigma^*_{\mathcal{D}\mathcal{D}}(\tilde{\theta}_i))$ as α goes to 0

The key idea of this proof is to note as previously shown that no set of inducing points has a utility greater than $w_\infty := \frac{1}{N}\sum_{i=1}^{N} \text{Tr}(\Sigma^*_{\mathcal{D}\mathcal{D}}(\tilde{\theta}_i))$, but that any set of inducing points that includes $\mathcal{D}$ has a utility equal to w_∞.

Let $\{s'_1, ..., s'_k\}$ be points selected after k iterations of Algorithm 1, and let us denote by $\{u_1, ..., u_k\}$ the maximum utilities after the corresponding iterations as usual. Let us denote by

$$\tilde{s}_k = \underset{s \in \mathcal{D}}{\text{argmax}}\, \tilde{\mathcal{U}}(\{s'_1, ..., s'_{k-1}\} \cup \{s\})$$

the best candidate *in the data set* to be the k-th inducing point after $k-1$ iterations of our algorithm. As previously mentioned, $\{s'_1, ..., s'_{k-1}\} \cup \mathcal{D}$ is a set of inducing points with perfect utility. Therefore, if we select the data points as inducing points after $\{s'_1, ..., s'_{k-1}\}$, their contribution to the overall utility will be $w_\infty - u_{k-1}$. If we further constrain our choice of $\mathcal{D}$ as additional inducing points to start with $\tilde{s}_k$ then the incremental utility of choosing $\tilde{s}_k$ will be

[1] The conditional GP is defined as the stationary Gaussian Process in our model is conditioned on its values at the points $\{s'_1, ..., s'_{k-1}\}$

at least $\frac{w_\infty - u_{k-1}}{n}$, where n is the data size as usual. This is because $\tilde{s}_k$ is the best choice for the k-th inducing point in $\mathcal{D}$ after having picked $\{s'_1, ..., s'_{k-1}\}$ and because the incremental utility of choosing an inducing point is higher earlier (when little is known about the GP) than later (when more is known about the GP). What's more, by definition, the incremental utility of choosing s'_k after $\{s'_1, ..., s'_{k-1}\}$ is higher than that of choosing $\tilde{s}_k$ after $\{s'_1, ..., s'_{k-1}\}$. Hence,

$$u_k - u_{k-1} \geq \frac{w_\infty - u_{k-1}}{n}.$$

Let us denote by w_k the sequence satisfying

$$w_0 = u_0, \forall\, k \in \mathbb{N}^* w_k - w_{k-1} = \frac{w_\infty - w_{k-1}}{n}.$$

It can be shown (by induction on k) that

$$\forall\, k \in \mathbb{N}^* w_k \leq u_k.$$

Moreover, we note that

$$w_k - w_\infty = (1 - \frac{1}{n})(w_{k-1} - w_\infty).$$

Hence

$$w_k = w_\infty + (1 - \frac{1}{n})^k (w_0 - w_\infty),$$

which proves that the sequence w_k converges linearly to w_∞ with rate $1 - \frac{1}{n}$.

On one hand, we have shown that the sequence u_k converges and is upper-bounded by w_∞, hence its limit is smaller than w_∞:

$$u_\infty := \lim_{k \to \infty} u_k \leq w_\infty.$$

On the other hand, we have shown that $\forall\, k \in \mathbb{N}^*\ w_k \leq u_k$ which implies

$$w_\infty \leq u_\infty.$$

Hence,

$$w_\infty = u_\infty = \frac{1}{N} \sum_{i=1}^{N} \text{Tr}(\Sigma^*_{\mathcal{D}\mathcal{D}}(\tilde{\theta}_i)).$$

As w_k is upper-bounded by u_k and both sequences converge to the same limit, u_k, and subsequently Algorithm 1, converge at least as fast as w_k.

In regards to the second statement of our proposition, we have that

$$\lim_{\alpha \to 0} u_f(\alpha) = \lim_{k \to \infty} u_k = \frac{1}{N} \sum_{i=1}^{N} \text{Tr}(\Sigma^*_{\mathcal{D}\mathcal{D}}(\tilde{\theta}_i)).$$