The Rényi Gaussian Process

Xubo Yue, Raed Al Kontar

1 Introduction

The Rényi α-divergence variational inference (RVI) has been introduced in the work of (Li and Turner, 2016). In their work, the parameters are approximated by the Monte Carlo (MC) method. We recognize that applying the RVI to the Gaussian process ($\mathcal{GP}$) can yield a closed-form lower bound. This new lower bound can be viewed as the convex combination of the sparse $\mathcal{GP}$ and the exact $\mathcal{GP}$. This combination allows us to control the smoothness of lower bound and achieves regularization purpose. It is more general than the traditional sparse variational $\mathcal{GP}$ regression. However, this inference is not sparse anymore due to the exact covariance component. Fortunately, the recent work from Wang et al. (2019) allows fast estimation of parameters. In this document, we provide a detailed inference procedure and provide a natural extension of the convergence result based on the previous work from Burt et al. (2019). We will provide computation details and applications in the future work.

Main Results We showed that with probability at least $1-\delta$, the Rényi α-divergence between the variational distribution and the true posterior is bounded. This bound converges to 0 as we have more data observation. Specifically,

$$D_\alpha[q||p] \leq \frac{1}{\delta}\frac{\alpha}{2(1-\alpha)}\log\left[1+\frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)N\sum_{m=M+1}^{\infty}\lambda_m+2Nv\epsilon]}{N}\right]^N + \alpha\frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m+2Nv\epsilon}{2\delta\sigma_\epsilon^2}\frac{\|\boldsymbol{y}\|^2}{\sigma_\epsilon^2}.$$

As shown in this equation, α plays an important role in controlling rate of convergence. Please refer to Sec. 3 for detailed information.

2 The Rényi Gaussian Processes and Variational Inference

Traditional variational inference is seeking to minimize the Kullback-Leibler (KL) divergence between the variational density $q(\boldsymbol{\theta})$ and the intractable posterior $p(\boldsymbol{\theta}|\mathcal{D})$, where $\boldsymbol{\theta}$ is a vector of parameters and $\mathcal{D}$ is the dataset. This minimization problem in turns yields a tractable evidence lower bound (ELBO) of the marginal log-likelihood function of data $\log p(\mathcal{D})$. The Rényi's α-divergence is a more general distance measure than the KL divergence. In this work, we want to explore the Rényi divergence based $\mathcal{GP}$.

2.1 Rényi Divergence

The Rényi's α-divergence between two distributions p and q on a random variable $\boldsymbol{\theta}$ is defined as

$$D_\alpha[p||q] = \frac{1}{\alpha-1}\log\int p(\boldsymbol{\theta})^\alpha q(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta}, \alpha \in (0,1).$$

This divergence contains a rich family of distance measure such as KL-divergence. Besides, the domain of α can be extended to $\alpha < 0$ and $\alpha > 1$.

Claim 1. $\lim_{\alpha\to 1} D_\alpha[p||q] = KL[p||q]$.

Therefore, KL-divergence is a special case of α-divergence. It is well-known that KL-divergence yields a popular ELBO. Therefore, it would be interesting to derive a similar bound using the α-divergence. Let $\mathcal{D} = \boldsymbol{y}$ (our data). Starting from $VR(q||p) := \log p(\boldsymbol{y}) - D_\alpha[q(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{\mathcal{Z}})||p(\boldsymbol{f}, \boldsymbol{U}, \boldsymbol{y}|\boldsymbol{\mathcal{Z}})]$, we will reach the variational Rényi (VR) bound (Li and Turner, 2016). This form is defined as

$$\mathcal{L}_\alpha(q; \boldsymbol{y}) := \frac{1}{1-\alpha}\log\mathbb{E}_q\left[\left(\frac{p(\boldsymbol{f}, \boldsymbol{U}, \boldsymbol{y}|\boldsymbol{\mathcal{Z}})}{q(\boldsymbol{f}, \boldsymbol{U}|\boldsymbol{\mathcal{Z}})}\right)^{1-\alpha}\right] = VR(q||p),$$

where $\boldsymbol{f}$ is a Gaussian process, $\boldsymbol{\mathcal{Z}}$ is the pseudo-input and $\boldsymbol{U}$ is the latent variable.

Claim 2. $\mathcal{L}_0(q; \boldsymbol{y}) = \log P(\boldsymbol{y})$.

Denote by the $\mathcal{L}_{VI}$ as the ELBO, we have the following claim.

Claim 3. $\mathcal{L}_{VI} = \lim_{\alpha\to 1}\mathcal{L}_\alpha(q; \boldsymbol{y}) \leq \mathcal{L}_{\alpha_+}(q; \boldsymbol{y}) \leq \log P(\boldsymbol{y}) \leq \mathcal{L}_{\alpha-}(q; \boldsymbol{y}), \forall\alpha_+ \in (0,1), \alpha_- < 0.$

We leave proof into appendix.

2.2 The Variational Rényi Lower Bound

Our bound is

$$\mathcal{L}_\alpha(q;\boldsymbol{y}) = \log \mathcal{N}(\boldsymbol{0}, \sigma_\epsilon^2 I + (1-\alpha)[\boldsymbol{K_{f,f}}] + \alpha \boldsymbol{Q}) + \log C_x,$$

where

$$\begin{aligned} C_x &= |\boldsymbol{I} + \frac{1-\alpha}{\sigma_\epsilon^2}(\boldsymbol{K_{f,f}} - \boldsymbol{Q})|^{\frac{-\alpha}{2(1-\alpha)}} \\ &\approx \left\{1 + \frac{1-\alpha}{\sigma_\epsilon^2}\mathrm{Tr}(\boldsymbol{K_{f,f}} - \boldsymbol{Q}) + \mathcal{O}(\frac{(1-\alpha)^2}{\sigma_\epsilon^4})\right\}^{\frac{-\alpha}{2(1-\alpha)}}, \end{aligned}$$

$\boldsymbol{K}$ is the covariance matrix and $\boldsymbol{Q} = \boldsymbol{K_{f,U}}\boldsymbol{K_{U,U}^{-1}}\boldsymbol{K_{U,f}}$. It can be seen that the new lower bound is the convex combination of components from sparse $\mathcal{GP}$ ($\boldsymbol{Q}$) and components from exact $\mathcal{GP}$ ($\boldsymbol{K_{f,f}}$). We can also see that α plays an important role in model regularization.

We also derive a data-dependent upper bound similar to Titsias (2014). See appendix for details.

3 Convergence Analysis

In this section, we will derive some convergence results based on recent works from Titsias (2014); Huggins et al. (2018); Burt et al. (2019). We will provide some extensions to those works. Due to space limit, we move all proofs into appendix.

Theorem 4. *Suppose N data points are drawn i.i.d from input distribution $p(\boldsymbol{x})$ and $k(\boldsymbol{x},\boldsymbol{x}) \leq v, \forall \boldsymbol{x} \in \mathcal{X}$. Sample M inducing points from the training data with the probability assigned to any set of size M equal to the probability assigned to the corresponding subset by an ϵ k-Determinantal Point Process (k-DPP) (Belabbas and Wolfe, 2009) with $k = M$. If $\boldsymbol{y}$ is distributed according to a sample from the prior generative model, with probability at least $1-\delta$,*

$$\begin{aligned} VR[q||p] \leq \alpha \frac{(M+1)N\sum_{m=M+1}^\infty \lambda_m + 2Nv\epsilon}{2\delta\sigma_\epsilon^2} + \\ \frac{1}{\delta}\frac{\alpha}{2(1-\alpha)} \log\left[1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)N\sum_{m=M+1}^\infty \lambda_m + 2Nv\epsilon]}{N}\right]^N. \end{aligned}$$

where λ_m are the eigenvalues of the integral operator $\mathcal{K}$ associated to kernel, k and $p(\boldsymbol{x})$.

As $\alpha \to 1$, we obtain the bound for the KL divergence.

Theorem 5. *Suppose N data points are drawn i.i.d from input distribution $p(\boldsymbol{x})$ and $k(\boldsymbol{x}, \boldsymbol{x}) \leq v, \forall \boldsymbol{x} \in \mathcal{X}$. Sample M inducing points from the training data with the probability assigned to any set of size M equal to the probability assigned to the corresponding subset by an ϵ k-Determinantal Point Process (k-DPP) (Belabbas and Wolfe, 2009) with $k = M$. With probability at least $1 - \delta$,*

$$D_\alpha[q||p] \leq \frac{1}{\delta} \frac{\alpha}{2(1-\alpha)} \log \left[1 + \frac{1-\alpha}{\sigma_\epsilon^2} \frac{[(M+1)N \sum_{m=M+1}^{\infty} \lambda_m + 2Nv\epsilon]}{N}\right]^N + \alpha \frac{(M+1)N \sum_{m=M+1}^{\infty} \lambda_m + 2Nv\epsilon}{2\delta\sigma_\epsilon^2} \frac{\|\boldsymbol{y}\|^2}{\sigma_\epsilon^2}$$

where $C = N \sum_{m=M+1}^{\infty} \lambda_m$ and λ_m are the eigenvalues of the integral operator $\mathcal{K}$ associated to kernel, k and $p(\boldsymbol{x})$.

As $\alpha \to 1$, we reach the bound for the KL divergence.

4 Consequences

4.1 Smooth Kernel

We will provide a convergence result with SE kernel. For SE kernel, we have $\lambda_m = v\sqrt{2a/A}B^{m-1}$, where $a = 1/(4\sigma_\epsilon^2)$, $b = 1/(2\ell^2)$, $c = \sqrt{a^2 + 2ab}$, $A = a+b+c$ and $B = b/A$. ℓ is the length parameter, v is signal variance and σ_ϵ is the noise parameter. We can obtain $\sum_{m=M+1}^{\infty} \lambda_m = \frac{v\sqrt{2a}}{(1-B)\sqrt{A}} B^M$ (Burt et al., 2019).

Corollary 6. *Suppose $\|\boldsymbol{y}\|^2 \leq RN$, where R is a constant. Fix $\gamma > 0$ and take $\epsilon = \frac{\delta\sigma_\epsilon^2}{vN^{\gamma+2}}$. Assume the input data is normally distributed and regression in performed with a SE kernel. With probability $1 - \delta$,*

$$VR[q||p] \leq 2\alpha \frac{R}{\sigma_\epsilon^2} \frac{1}{N^\gamma} + \frac{1}{\delta} \frac{\alpha}{2(1-\alpha)} \log \left[1 + (1-\alpha)(\frac{4\delta}{N^{\gamma+2}})\right]^N,$$

when inference is performed with $M = \frac{(3+\gamma)\log N + \log D}{\log(B^{-1})}$. where $D = \frac{v\sqrt{2a}}{a\sqrt{A}\sigma_\epsilon^2\delta(1-B)}$.

Proof. From Burt et al. (2019), we know $\frac{C(M+1)}{2\delta\sigma_\epsilon^2} < \frac{1}{N^{\gamma+1}}$. By Theorem 5, we will obtain

$$\begin{aligned}\mathrm{VR}[q||p] &\leq 2\alpha\frac{R}{\sigma_\epsilon^2}\frac{1}{N^\gamma} + \frac{1}{\delta}\frac{\alpha}{2(1-\alpha)}\log\left[1+(1-\alpha)(\frac{2\delta}{N^{\gamma+2}}+\frac{2\delta}{N^{\gamma+2}})\right]^N \\ &< 2\alpha\frac{R}{\sigma_\epsilon^2}\frac{1}{N^\gamma} + \alpha(\frac{2}{N^{\gamma+1}}) = \frac{\alpha}{N^\gamma}(\frac{2R}{\sigma_\epsilon^2}+\frac{2}{N}),\end{aligned}$$

□

As $N \to \infty$, $\mathrm{VR}[q||p] \to 0$. As $\alpha \to 1$, we obtain $\frac{1}{N^\gamma}(\frac{2R}{\sigma_\epsilon^2}+\frac{2}{N})$.

4.2 Non-smooth Kernel

For the Matérn $k+\frac{1}{2}$, $\lambda_m \asymp \frac{1}{m^{2k+2}}$. We can obtain $\sum_{m=M+1}^{\infty}\lambda_m = \mathcal{O}(\frac{1}{M^{2k+1}})$ by the following claim.

Claim 7. $\sum_{m=M+1}^{\infty}\lambda_m = \mathcal{O}(\frac{1}{M^{2k+1}})$.

Proof. It is easy to see that $\sum_{m=1}^{\infty}\lambda_m = \zeta(2k+2)$, where ζ is a Riemann zeta function. By the Euler-Maclaurin sum formula, we have the generalized harmonic number (Woon, 1998)

$$\sum_{m=1}^{M}(\frac{1}{m})^{2k+2} = \zeta(2k+2) + \frac{1}{-2k-1}M^{-2k-1} + \frac{1}{2}M^{-2k-2} - \frac{2k+2}{12}M^{-2k-3} + \mathcal{O}(M^{-2k-4}).$$

Therefore,

$$\sum_{m=M+1}^{\infty}\lambda_m = \mathcal{O}(\frac{1}{M^{2k+1}}) = -\frac{1}{-2k-1}M^{-2k-1} - \frac{1}{2}M^{-2k-2} + \frac{2k+2}{12}M^{-2k-3} - \mathcal{O}(M^{-2k-4}) = \mathcal{O}(\frac{1}{M^{2k+1}}).$$

□

Let $\sum_{m=M+1}^{\infty}\lambda_m \leq A\frac{1}{M^{2k+1}}$. Then by Theorem 5, we have

$$\begin{aligned}&\alpha\frac{(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon}{2\delta\sigma_\epsilon^2}\frac{\|\boldsymbol{y}\|^2}{\sigma_\epsilon^2} \\ &\leq \alpha\frac{(M+1)NA\frac{1}{M^{2k+1}} + 2Nv\epsilon}{2\delta\sigma_\epsilon^2}\frac{RN}{\sigma_\epsilon^2} \\ &= \frac{\alpha R}{2\delta\sigma_\epsilon^4}(\frac{(M+1)N^2A}{M^{2k+1}} + 2N^2v\epsilon).\end{aligned}$$

In order to let $\lim_{N\to\infty} \frac{(M+1)N^2}{M^{2k+1}} \to 0$, we require $M = N^p$. Therefore,

$$\frac{(M+1)N^2 A}{M^{2k+1}} = \frac{(N^p+1)N^2 A}{N^{(2k+1)p}} \leq \frac{A}{N^{2kp-2}}.$$

Let $2kp - 2 \geq \gamma$, then $p \geq \frac{\gamma+2}{2k}$. Therefore, we have

$$\frac{\alpha R}{2\sigma_\epsilon^4}(\frac{(M+1)N^2 A}{M^{2k+1}} + 2N^2 v\epsilon) \leq \frac{\alpha R}{N^\gamma \sigma_\epsilon^2} + \frac{\alpha R A}{2\delta\sigma_\epsilon^4 N^\gamma}.$$

Another term in the bound can also be simplified as

$$\begin{aligned}&\frac{\alpha}{2(1-\alpha)} \log\left[1 + \frac{1-\alpha}{\sigma_\epsilon^2}\frac{[(M+1)N\sum_{m=M+1}^{\infty}\lambda_m + 2Nv\epsilon]}{N}\right]^N \\ &\leq \frac{\alpha N}{2(1-\alpha)} \log\left[1 + (1-\alpha)(\frac{A}{\sigma_\epsilon^2 N^{\gamma+2}} + \frac{2\delta}{\sigma_\epsilon^2 N^{\gamma+2}})\right].\end{aligned}$$

References

Belabbas, M.-A. and Wolfe, P. J. (2009). Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences*, 106(2):369–374.

Burt, D. R., Rasmussen, C. E., and Van Der Wilk, M. (2019). Rates of convergence for sparse variational gaussian process regression. *arXiv preprint arXiv:1903.03571*.

Huggins, J. H., Campbell, T., Kasprzak, M., and Broderick, T. (2018). Scalable gaussian process inference with finite-data mean and variance guarantees. *arXiv preprint arXiv:1806.10234*.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081.

Titsias, M. K. (2014). Variational inference for gaussian and determinantal point processes.

Wang, K. A., Pleiss, G., Gardner, J. R., Tyree, S., Weinberger, K. Q., and Wilson, A. G. (2019). Exact gaussian processes on a million data points. *arXiv preprint arXiv:1903.08114*.

Woon, S. (1998). Generalization of a relation between the riemann zeta function and bernoulli numbers. *arXiv preprint math.NT/9812143*.