Deep Gaussian processes for weakly supervised learning: tumor mutational burden (TMB) prediction

Sunho Park, Hongming Xu, Tae Hyun Hwang*
Department of Quantitative Health Sciences
Cleveland Clinic
Cleveland, OH 44195
{parks,xu3,hwangt}@ccf.org

Saehoon Kim
AITRICS, Korea
shkim@aitrics.com

1 Introduction

Deep Gaussian process (DGP) model [5], a hierarchical composition of multiple Gaussian processes, can provide a more flexible prior distribution over functions than a single Gaussian process (GP) can. In this work, we propose a DGP based classification method for tumor mutational burden (TMB) prediction from histopathology whole slide images (WSIs). TMB (a quantitative measurement of the number of mutations in a patient's tumor), which can be fully assessed by next-generation sequencing (NGS) technology, has been suggested as a biomarker to predict a patient's treatment response to immunotherapy in several cancer types, including lung and bladder cancer [3, 6]. However, one key challenge is that not all patients would have adequate fresh tumor tissue samples or would be able to undergo biopsy to get tumor tissues. On the other hand, WSIs are widely available, as a standard diagnosis tool, for most cancer patients. We hypothesize that the morphological image features calculated from whole slide images could be used to predict TMB status.

Prediction of TBM from WSIs is naturally formulated as a weakly supervised learning problem due to the nature of WSIs: the typical size of a WSI might be 100,000x100,000 pixels, and its storage size can range from a few hundred megabytes to several gigabytes. In most cases, a WSI is divided into multiple (non-overlapping) small image patches (e.g., 256x256 in our experiment), and each image patch is processed independently. Prediction of TMB from WSIs can be done similarly: we make a prediction using a feature vector calculated from each image patch of an image and aggregate all prediction results from the image to make a final decision. This approach can be understood as *the instance-level approach* [10] in multiple instance learning (MIL): an image including multiple image patches corresponds to a bag in MIL. However, TMB prediction from WSIs can be distinguished from the standard MIL in the sense that a TBM-low labeled image (corresponding to a negative bag in MIL) might contain image patches that are close to typical TMB-high image patches (corresponding to positive instances in MIL). To handle this uniqueness of the TMB prediction problem, we use mean pooling to aggregate prediction results from an image, instead of max pooling. Another challenge is the high computational complexity due to the nature of WSIs. Although the number of WSIs might be small, the number of total image patches can be massive.

We propose a DGP based classification model in the weakly supervised learning setting and provide an efficient inference algorithm to train the model based on Black-box α-divergence (BB-α) approaches [8, 12]. We evaluate a DGP at each image patch of an image and make a final prediction for the image by aggregating all the prediction results through mean pooling. We train the model's parameters in the framework of power expectation and propagation (power EP). However, instead of using the iterative message update method, we directly optimize the energy function which can be derived by tying local factors (cite approximations). With different α values, the inference algorithm includes variational inference ($\alpha = 0$) or EP ($\alpha = 1$) as a special case. We test our DGP based TMB prediction model on The Cancer Genomic Atlas (TCGA) bladder cancer WSI dataset.

*Corresponding author

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

2 Deep Gaussian process classifier for weakly supervised problems

2.1 Problem definition

We formulate the TMB prediction from WSIs as a binary classification problem in the weakly supervised learning setting. For the ith image, $\boldsymbol{X}_i = \left[\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ...\boldsymbol{x}_{iN_i}\right] \in \mathbb{R}^{D\times N_i}$ denotes a set of feature vectors computed from its all N_i image patches and $\boldsymbol{x}_{ij} \in \mathcal{X} \subset \mathbb{R}^D$. All training input data is denoted by $\boldsymbol{X} = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_N\}$, where N is the total number of training images. For each image, we generate its label based on its actual TMB value accessed by NGS technique from the patient's tumor tissues: we assign $y_i = 1$ if the ith sample is TMB-high or $y_i = -1$ if it is TMB-low.

Assume that there is a score function to estimate the TMB value at each individual image patch of an image, i.e., $f : \mathcal{X} \mapsto \mathbb{R}$, and that the final prediction of TMB for the image is made by aggregating the function values evaluated at all the image patches of the image through mean pooling:

$$p(y_i = 1|\overline{f}_i) = \Phi\left(\overline{f}_i\right), \tag{1}$$

where Φ is a standard cumulative Gaussian function (the likelihood of the probit regression) and $\overline{f}_i$ is the mean of the function values evaluated at all the patches of the ith image, i.e., $\frac{1}{N_i}\sum_{j=1}^{N_i} f(\boldsymbol{x}_{ij})$.

2.2 Model formulation

Our main idea is to model the score function f using a DGP with L layers: the prior over the score function f is defined as the GP prior from the last (Lth) layer of the DGP. The probabilistic representation of this L-layered DGP can be defined in the following recursive way [5] (for simplicity, the latent input and the output of the lth layer are assumed to be scalar, i.e., $h_{ij}^{l-1}, h_{ij}^{l} \in \mathbb{R}$):

$$p(f^l \mid \theta^l) = \mathcal{GP}(f^l \mid \mathbf{0}, \kappa^l(\cdot,\cdot)), \quad \text{and } l = 1, ..., L \tag{2}$$

$$p(\boldsymbol{h}^l|f^l, \boldsymbol{h}^{l-1}, \sigma_l^2) = \prod_{i=1}^{N}\prod_{j}^{N_i}\mathcal{N}(h_{ij}^l \mid f^l(h_{ij}^{l-1}), \sigma_l^2), \tag{3}$$

where $\boldsymbol{h}^l = [h_{11}^l, ..., h_{1N_1}^l, ..., h_{ij}^l, ...h_{N1}^l, ..., h_{NN_N}^l]^\top$. Note that $h_{ij}^0 = \widetilde{\boldsymbol{x}}_{ij}$, where $\widetilde{\boldsymbol{x}}_{ij}$ is a linear projection of the high dimensional input point $\boldsymbol{x}_{ij}$, i.e., $\widetilde{\boldsymbol{x}}_{ij} = \boldsymbol{W}\boldsymbol{x}_{ij}$ and $\boldsymbol{W} \in \mathbb{R}^{\tilde{D}\times D}$ ($\tilde{D} \ll D$), and $h_{ij}^1 = f^1(\widetilde{\boldsymbol{x}}_{ij})$. Sparse approximation [14] is often used to reduce the computational complexity of full GP models. Let us define inducing variables $\boldsymbol{u} = \{\boldsymbol{u}^1, ..., \boldsymbol{u}^L\}$, where $\boldsymbol{u}^l \in \mathbb{R}^M$ are inducing variables evaluated at inducing points $\boldsymbol{Z}^l \triangleq \{\boldsymbol{z}_1^l, ..., \boldsymbol{z}_M^l\}$ in the lth layer (here, $\boldsymbol{z}_m^l \in \mathbb{R}$). The sparse approximate version of the above DGP model is defined in as follows.

$$p(\boldsymbol{u}^l|\theta^l) = \mathcal{N}(\boldsymbol{u}^l \mid \mathbf{0}, \boldsymbol{K}_{uu}^l), \quad l = 1, ..., L \tag{4}$$

$$p(\boldsymbol{h}^l|f^l, \boldsymbol{h}^{l-1}, \sigma_l^2) = \mathcal{N}(\boldsymbol{h}^l|\boldsymbol{K}_{fu}^l(\boldsymbol{K}_{uu}^l)^{-1}\boldsymbol{u}^l, \text{diag}[\boldsymbol{Q}_{ff}^l] + \sigma_l^2\boldsymbol{I}), \tag{5}$$

where $\boldsymbol{K}_{uu}^l = \kappa(\boldsymbol{Z}^l, \boldsymbol{Z}^l)$, $\boldsymbol{K}_{fu}^l = \kappa(\boldsymbol{h}^{l-1}, \boldsymbol{Z}^l)$ and $\boldsymbol{Q}_{ff}^l = \boldsymbol{K}_{fu}^l - \boldsymbol{K}_{fu}^l(\boldsymbol{K}_{uu}^l)^{-1}\boldsymbol{K}_{uf}^l$. Here, κ is a covariance function with parameters θ^l, and an automatic relevance determination (ARD) covariance function is used in every layer. With the likelihood defined in (1), the probabilistic graphical model of our DGP (L=3) classification model in the weakly supervised learning is depicted in Figure 1.

2.3 Inference with Black-box α-divergence (BB-α)

We train the model, including the inducing variables $\boldsymbol{u} = \{\boldsymbol{u}^l\}_{l=1}^L$, the inducing points $\boldsymbol{Z} = \{\boldsymbol{Z}^l\}_{l=1}^L$ and the ARD covariance parameters $\boldsymbol{\theta} = \{\theta^l\}_{l=1}^L$, in the framework of power EP. We approximately calculate the posterior distribution over the inducing variables, i.e., $q(\boldsymbol{u}) \approx p(\boldsymbol{u}|\boldsymbol{X}, \boldsymbol{y}) \propto p_0(\boldsymbol{u})\prod_{i=1}^N p(y_i|\boldsymbol{u}, \boldsymbol{X}_i)$, where $p_0(\boldsymbol{u}) = \prod_{l=1}^L \mathcal{N}(\boldsymbol{u}^l|0, \boldsymbol{K}_{uu}^l)$. We assume the approximate posterior $q(\boldsymbol{u})$ to be also Gaussian, i.e., $q(\boldsymbol{u}^l) = \mathcal{N}(\boldsymbol{u}^l|\boldsymbol{m}^l, \boldsymbol{\Sigma}^l)$, where $l = 1, ..., L$. Since the number of total image patches in the training data can be massive, the standard power EP [13], which is a batch algorithm, might not be applicable to TMB prediction from WSIs. In this work, we consider BB-α approaches [8, 12] which directly optimize the free energy function (derived in the framework of power-EP) using stochastic gradient descent. In particular, we apply the approximate version of

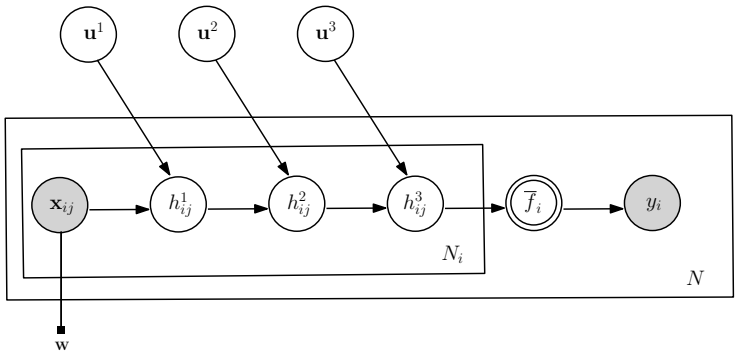

Figure 1: Probabilistic graphical model of the DGP classifier in the weakly supervised learning (L=3). The double-lined represents a deterministic node (the mean pooling) and h_{ij}^3 is the same as $f(x_{ij})$.

BB-α proposed in [12] to infer $q(\boldsymbol{u})$, due to its simpler form compared to the original version [8]. The BB-α energy function of our model with a mini-batch samples $\mathcal{S}$ can be given by

$$\mathcal{L}_\alpha(q) = KL[q||p_0] - \frac{N}{\alpha|\mathcal{S}|}\sum_{i\in\mathcal{S}} \log \mathbb{E}_q\Big[p^\alpha(y_i|\boldsymbol{X}_i, \overline{f}_i)\Big], \tag{6}$$

where the first term is Kullback-Leibler (KL) divergence between two distributions and can be analytically calculated in our case (q and p_0 are both Gaussian). The objective function (6) can be understood as a regularized loss minimization: the KL term and the second expectation term correspond to regularization and data fitting, respectively. Note that the expectation in the second term can be approximately computed using the probabilistic back-propagation [9] as in [4], where a Gaussian is propagated through a layer, this non-Gaussian output distribution is approximated again as a Gaussian before being fed to the next layer and these steps are repeated until reaching the last layer. Detailed derivations to compute this expectation are provided in Appendix. Finally, the parameters $\boldsymbol{Z}$ and $\boldsymbol{\theta}$ are also jointly optimized with the approximate posterior q as in [7].

3 Experimental results

We tested our method on the TCGA bladder cancer dataset. A cohort of 386 bladder cancer patients with 457 diagnostic H&E stained WSIs were downloaded from the TCGA data portal. Based on [1] which shows the link between the TMB status of the patients and their immunotherapy response in urothelial carcinoma (the patients are divided into 4 quartile groups according to their TMB values and the top quartile group shows better survival outcomes compared to the others), we selected the top 25% TMB-high patients as the positive group and the bottom 25% TMB-low patients as the negative group. We calculated features from each image patch of a WSI with a deep neural network trained on ImageNet [16]. The following are some statistics of the data. The number of the total WSIs is $N = 189$ (#positive: 94 and #negative: 95), the average number of image patches per WSI is 495.354, and the dimensionality of the input features is $D = 2,048$.

#layers (L)	$\alpha = 1$ (EP)	$\alpha = 0.5$	$\alpha = 10^{-6}$ (VI)
L=1	0.723 (0.016)	0.723 (0.019)	0.724 (0.016)
L=2	0.722 (0.016)	0.723 (0.021)	0.720 (0.016)
L=3	0.736 (0.017)	0.740 (0.014)	**0.745** (0.028)

Table 1: The predictive performance (AUC) of our method with different α values, 0, 0.5 and 10^{-6} (≈ 0) on the TCGA bladder cancer WSIs data. We reported the mean and standard deviation (in parentheses) of the AUC values from 10 experiments (5-fold CV). There is clear improvement in the performance compared to the base-line method, SVM+PCA [16], which scored 0.649 (0.013).

For our DGP model, we set the number of inducing points in each layer to 50, the dimensionality of the linear projection $\tilde{D}$ to 64 and the dimensionality of the output of each layer, i.e., h_{ij}^l, to 5 when

$L > 1$. To ensure the positive definite of the covariance matrix of the approximate posterior, e.g., $\boldsymbol{\Sigma}^l$ in $q(\boldsymbol{u}^l) = \mathcal{N}(\boldsymbol{u}^l|\boldsymbol{m}^l, \boldsymbol{\Sigma}^l)$, it was assumed to be a form of $\boldsymbol{\Sigma}^l = \boldsymbol{V}\boldsymbol{V}^\top + c\boldsymbol{I}$, where $\boldsymbol{V} \in \mathbb{R}^{50\times 20}$ and $c = 10^{-3}$. We randomly drew 20 samples for each mini-batch set $|\mathcal{S}| = 20$. The KL term in the objective function (6) is considerably larger compared to the second term (data fitting) for a dataset which consists of a small number training samples. For simplicity, we ignored the KL term and used the early stopping technique to prevent overfitting (the algorithm was stopped at 500 iterations). We also tried a schedule for the KL term (a time varying constant γ was multiplied by the KL term, and γ was set to 0 until the iteration reached 200 and 1e-4 later), but the results were not significantly different. We implemented our DGP model in Python with TensorFlow.

We compared our method with a baseline method proposed in [16], where feature vectors from all the image patches of an image are first combined into a single feature vector and then this combined vector is fed into a SVM (with a rbf kernel) classifier after reducing the input dimensionality using PCA. We refer to this approach as SVM+PCA. We evaluated the performance of each method in terms of area under ROC curve (AUC). We repeated random 5 fold cross-validation (CV) 10 times and report the mean and standard deviation of AUC values (for our method, we considered $\alpha = 1.0, 0.5, 0.0$ and $L = 1, 2, 3$ in Table 1). We can see that in our experiments the DGP model, no matter the α value and number of layers, always outperforms SVM+PCA. Although the performance is not significantly different across the cases, the DGP models with $L = 3$ showed the best performance.

4 Conclusion

We have proposed a DGP-based classification model for TMB prediction from WSIs. The prediction problem is naturally formed in the weakly supervised learning setting because a super resolution WSI has to be decomposed into multiple small image patches and a final prediction on the image is made by aggregating prediction results from all the image patches of the image. We apply a DGP to model the score function which evaluates TMB status of each image patch of an image. Using the mean pooling aggregation function, we could easily calculate the expectation of the likelihood of each sample with respect to the approximate posterior distributions using a forward pass of probabilistic propagation, which leads an efficient inference algorithm based on BB-α approaches. We have shown from the TCGA bladder cancer WSI data that our method outperforms the base-line method.

The current version of the paper provides preliminary experiment results: we include experimental results from only one data set. We plan to test our method on WSIs obtained from different cancer types and to compare our method with other base-line methods, including deep neural networks or multiple instance learning methods. In addition, we currently built the model from predefined input features, not directly from image pixels. However, the image features calculated from the pretrained neural network might not be relevant to TMB prediction. One possible solution is to extend our model to include convolutional structures (layers) so that the model can learn discriminative features for TMB prediction directly from images, as similarly done in [2, 15, 11].

Acknowledgments

Saehoon Kim was supported by grant funded by 2019 IT Promotion fund of the Korea government (MSIT) (CONnected Network for EMS Comprehensive Technical-support using Artificial Intelligence ('CONNECT-AI')).

References

[1] Rosenberg JE Powles T Petrylak DP Bellmunt J Loriot Y Necchi A Hoffman-Censits J Perez-Gracia JL Dawson NA van der Heijden MS Dreicer R Srinivas S Retz MM Joseph RW Drakaki A Vaishampayan UN Sridhar SS Quinn DI Durán I Shaffer DR Eigl BJ Grivas PD Yu EY Li S Kadel EE 3rd Boyd Z Bourgon R Hegde PS Mariathasan S Thåström A Abidoye OO Fine GD Bajorin DF Balar AV, Galsky MD. Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial. *Lancet*, 16 Suppl 1:67–76, 2017.

[2] Kenneth Blomqvist, Samuel Kaski, and Markus Heinonen. Deep convolutional Gaussian processes. *ArXiv*, abs/1810.03052, 2018.

[3] R. L. Brown, S. D. nad Warren, E. A. Gibb, S. D. Martin, J. J. Spinelli, B. H. Nelson, and R. A. Holt. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome research*, 25:743–50, 2014.

[4] Thang D. Bui, José Miguel Hernández-Lobato, Daniel Hernández-Lobato, Yingzhen Li, and Richard E. Turner. Deep gaussian processes for regression using approximate expectation propagation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[5] Andreas Damianou and Neil Lawrence. Deep gaussian processes. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 31, pages 207–215, Scottsdale, Arizona, USA, 29 Apr–01 May 2013.

[6] Aaron M. Goodman, Shumei Kato, Lyudmila Bazhenova, Sandip P. Patel, Garrett M. Frampton, Vincent Miller, Philip J. Stephens, Gregory A. Daniels, and Razelle Kurzrock. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Molecular Cancer Therapeutics*, 16(11):2598–2608, 2017.

[7] Daniel Hernandez-Lobato and Jose Miguel Hernandez-Lobato. Scalable gaussian process classification via expectation propagation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51 of *Proceedings of Machine Learning Research*, pages 168–176, Cadiz, Spain, May 2016. PMLR.

[8] Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernandez-Lobato, and Richard Turner. Black-box alpha divergence minimization. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, volume 48, pages 1511–1520, New York, New York, USA, 20–22 Jun 2016.

[9] José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 2015.

[10] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[11] Vinayak Kumar, Vaibhav Singh, P. K. Srijith, and Andreas C. Damianou. Deep Gaussian Processes with Convolutional Kernels. *ArXiv*, abs/1806.01655, 2018.

[12] Yingzhen Li and Yarin Gal. Dropout inference in Bayesian neural networks with alpha-divergences. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2052–2061, 2017.

[13] Thomas Minka. Power ep. Technical report, Microsoft Research Cambridge, 2004.

[14] Joaquin Quiñonero Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.*, 6:1939–1959, December 2005.

[15] Mark van der Wilk, Carl Edward Rasmussen, and James Hensman. Convolutional Gaussian Processes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 2845–2854, USA, 2017. Curran Associates Inc. event-place: Long Beach, California, USA.

[16] Hongming Xu, Sunho Park, Sung Hak Lee, and Tae Hyun Hwang. Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients. *bioRxiv*, 2019.

Appendix

How to compute the expectation in (6)

The expectation in the second term in (6) can be written as follows

$$\mathbb{E}_q\Big[p^\alpha(y_i|\boldsymbol{X}_i,\overline{f}_i)\Big] = \int p^\alpha(y_n|\overline{f}_i)\int p(\boldsymbol{f}_i^L|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u}\,d\boldsymbol{f}_i^L, \tag{7}$$

where $\boldsymbol{f}_i^L$ are the final function values evaluated at all the image patches of the ith image, i.e., $\boldsymbol{f}_i^L = [f_{11}^L, ..., f_{1N_i}^L]^\top$, where $f_{11}^L = f^L(h_{11}^{L-1})$, and f^L is the final output function defined in the last layer

of the L-layered DGP. With abuse of notation, we can redefine the mean pooling as $\overline{f}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} f_{ij}^L$. We first consider the inner integration in (7), and assume that the marginal distribution over $\boldsymbol{f}_i^L$ (after integrating the including variables $\boldsymbol{u}$ out) is fully factorized, i.e., $q(\boldsymbol{f}_i^L) = \prod_{j=1}^{N_i}\mathcal{N}(f_{ij}^L|m_{ij}^L, v_{ij}^L)$. Note that the distribution of each factor f_{ij}^L is approximated by a Gaussian whose mean and variance, $m_i^L j$ and $v_i^L j$, are calculated using the exact same method in [4] which is based on the forward pass of probabilistic propagation (for more details, please see Section 5 in [4]). Once all the means and variances, $\{m_{ij}^L\}$ and $\{v_{ij}^L\}$, are calculated, the approximate marginal distribution over $\overline{f}_i$ can be given as $q(\overline{f}_i) = \mathcal{N}(\overline{f}_i|\overline{m}_i^L, \overline{v}_i^L)$, where $\overline{m}_i^L = \frac{1}{N_i}\sum_{j=1}^{N_i} m_{ij}^L$ and $\overline{v}_i = \frac{1}{N_i^2}\sum_{j=1}^{N_i} v_{ij}^L$. As a result, the expectation (6) can be calculated as follows

$$\mathbb{E}_q\Big[p^\alpha(y_i|\boldsymbol{X}_i, \overline{f}_i)\Big] \approx \int p^\alpha(y_n|\overline{f}_i)q(\overline{f}_i)d\overline{f}_i = \int \Phi^\alpha\left(y_i\overline{f}_i\right)\mathcal{N}(\overline{f}_i|\overline{m}_i, \overline{v}_i)d\overline{f}_i. \tag{8}$$

When $\alpha = 1$, the last integration allows a closed form, i.e., $\mathbb{E}_q\Big[p^\alpha(y_i|\boldsymbol{X}_i, \overline{f}_i)\Big] \approx \Phi\Big(\frac{y_i\overline{m}_i}{\sqrt{\overline{v}_i+1}}\Big)$. For $\alpha \neq 1$, we need numerical methods to approximate the integration in (8). However, it involves just a 1D integration, and there are many efficient numerical tools available, such as Gauss-Hermit quadrature.