Asymptotic results in nonparametric Bayesian function estimation

Kirichenko, A.

References

Adams, R. P., Murray, I. and MacKay, D. J. (2009). Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 9–16. ACM. 9, 71, 72, 73, 74, 88, 90

Ando, R. K. and Zhang, T. (2007). Learning on graph with Laplacian regularization. *Advances in neural information processing systems* **19**, 25. 43, 44

Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.* **31**(2), 536–559. Dedicated to the memory of Herbert E. Robbins. 19, 20

Belitser, E., Serra, P. and van Zanten, J. H. (2015). Rate-optimal Bayesian intensity smoothing for inhomogeneous Poisson processes. *J. Statist. Plann. Inference* **166**, 24–35. 71, 72, 75, 76

Belkin, M., Matveeva, I. and Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pp. 624–638. Springer, Berlin. 43, 44

Berger, J. O. (1993). *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. Springer-Verlag, New York. Corrected reprint of the second (1985) edition. 6, 16

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**(3), 647–671. 19

Borgs, C., Chayes, J. T., Cohn, H. and Zhao, Y. (2014). An l^p theory of sparse graph convergence i: limits, sparse random graph models, and power law distributions. *arXiv:1401.2906* . 22

Brooks, S., Gelman, A., Jones, G. L. and Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL. 16

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24**(6), 2384–2398. 15

Castillo, I., Kerkyacharian, G. and Picard, D. (2014). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* **158**(3-4), 665–710. 44, 46

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England journal of medicine* **357**(4), 370–379. 5

Chung, F. (2014). From quasirandom graphs to graph limits and graphlets. *Adv. in Appl. Math.* **56**, 135–174. 22

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York. Revised reprint of the 1991 edition, A Wiley-Interscience Publication. 23

Cvetković, D. S., Rowlinson, P. and Simić, S. (2010). An introduction to the theory of graph spectra **75**, xii+364. 24, 26

de Jonge, R. and van Zanten, J. H. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.* **38**(6), 3300–3320. 20

de Jonge, R. and van Zanten, J. H. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electron. J. Stat.* **6**, 1984–2001. 20

de Jonge, R. and van Zanten, J. H. (2013). Semiparametric Bernstein–von Mises for the error standard deviation. *Electron. J. Stat.* **7**, 217–243. 47

Diaconis, P. and Freedman, D. (1986a). On the consistency of Bayes estimates. *Ann. Statist.* **14**(1), 1–67. With a discussion and a rejoinder by the authors. 16

Diaconis, P. and Freedman, D. (1986b). On the consistency of Bayes estimates. *Ann. Statist.* **14**(1), 1–67. With a discussion and a rejoinder by the authors. 16, 72

DiMatteo, I., Genovese, C. R. and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**(4), 1055–1071. 71

Doob, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, pp. 23–27. Centre National de la Recherche Scientifique, Paris. 16

Dunker, T., Lifshits, M. A. and Linde, W. (1998). Small deviation probabilities of sums of independent random variables. In *High dimensional probability (Oberwolfach, 1996)*, volume 43 of *Progr. Probab.*, pp. 59–74. Birkhäuser, Basel. 56

Efromovich, S. (2011). Nonparametric regression with responses missing at random. *J. Statist. Plann. Inference* **141**(12), 3744–3752. 50, 54

Efromovich, S. and Pinsker, M. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica* **6**(4), 925–942. 19

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach*. Probability and Mathematical Statistics, Vol. 1. Academic Press, New York-London. 16

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34**, 1386–1403. 16

Freedman, D. A. (1965). On the asymptotic behavior of Bayes estimates in the discrete case. II. *Ann. Math. Statist.* **36**, 454–456. 16

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**(2), 500–531. 17, 19, 50

Ghosal, S., Lember, J. and van der Vaart, A. (2008). Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.* **2**, 63–89. 20

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.* **35**(1), 192–223. 17, 18, 50, 64, 66

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press. 6, 16, 19

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29**(5), 1233–1263. 50, 63

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. Springer Series in Statistics. Springer-Verlag, New York. 6, 16

Gilks, W. R., Richardson, S. and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press. 16

Gugushvili, S. and Spreij, P. (2013). A note on non-parametric Bayesian estimation for Poisson point processes. *ArXiv preprint arXiv:1304.7353* . 71

Hartog, J. and van Zanten, J. H. (2016). Nonparametric bayesian label prediction on a graph. *arXiv preprint arXiv:1612.01930* . 49

Huang, J., Ma, S., Li, H. and Zhang, C.-H. (2011). The sparse Laplacian shrinkage estimator for high-dimensional regression. *Ann. Statist.* **39**(4), 2021–2046. 43

Ibragimov, I. A. and Has′minskiĭ, R. Z. (1981). *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York-Berlin. Asymptotic theory, Translated from the Russian by Samuel Kotz. 17

Johnson, R. and Zhang, T. (2007). On the effectiveness of laplacian normalization for graph semi-supervised learning. *Journal of Machine Learning Research* **8**(4). 44

Kolaczyk, E. D. (2009). *Statistical analysis of network data*. Springer Series in Statistics. Springer, New York. Methods and models. 29

Kottas, A. and Sansó, B. (2007). Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *J. Statist. Plann. Inference* **137**(10), 3151–3163. 71

Kruijer, W., Rousseau, J. and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4**, 1225–1257. 20

Kutoyants, Y. A. (1998). *Statistical inference for spatial Poisson processes*, volume 134 of *Lecture Notes in Statistics*. Springer-Verlag, New York. 72, 75

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1**, 38–53. 17

Lember, J. and van der Vaart, A. (2007). On universal Bayesian adaptation. *Statist. Decisions* **25**(2), 127–152. 20

Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25**(6), 2512–2546. 19

Li, W. V. and Shao, Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications **19**, 533–597. 60

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition. 50

Liu, X., Zhao, D., Zhou, J., Gao, W. and Sun, H. (2014). Image interpolation via graph-based Bayesian label propagation. *IEEE Trans. Image Process.* **23**(3), 1084–1096. 21

Lovász, L. (2012). *Large networks and graph limits*, volume 60 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI. 22

Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *J. Combin. Theory Ser. B* **96**(6), 933–957. 22

Merris, R. (1998). Laplacian graph eigenvectors. *Linear Algebra Appl.* **278**(1-3), 221–236. 32, 33

Mohar, B. (1991a). Eigenvalues, diameter, and mean distance in graphs. *Graphs Combin.* **7**(1), 53–64. 25

Mohar, B. (1991b). The Laplacian spectrum of graphs pp. 871–898. 25, 26

Møller, J., Syversveen, A. R. and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Statist.* **25**(3), 451–482. 71

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24**(6), 2399–2430. 15

Prakasa Rao, B. L. S. (1983). *Nonparametric functional estimation*. Probability and Mathematical Statistics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York. 5

Rasmussen, C. E. and Williams, C. K. I. (2006). Gaussian processes for machine learning pp. xviii+248. 46

Reynaud-Bouret, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields* **126**(1), 103–153. 72

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics* **27**(3), 832–837. 5

Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.* **38**(1), 146–180. 20

Schwartz, L. (1965). On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **4**, 10–26. 16

Sharan, R., Ulitsky, I. and Shamir, R. (2007). Network-based prediction of protein function. *Molecular systems biology* **3**(1), 88. 21

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29**(3), 687–714. 17

Smola, A. J. and Kondor, R. (2003). Kernels and regularization on graphs. In *Learning theory and kernel machines*, pp. 144–158. Springer. 43, 44

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. 8, 13, 14, 15, 34, 35, 37, 39, 40

van der Vaart, A. W. and van Zanten, J. H. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36**(3), 1435–1463. 20, 47, 48, 72

van der Vaart, A. W. and van Zanten, J. H. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections* **3**, 200–222. 55, 59, 77

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37**(5B), 2655–2675. 19, 20, 72, 76, 78, 79 van der Vaart, A. W. and van Zanten, J. H. (2011). Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.* **12**, 2095–2119. 72

Wasserman, L. (2004). *All of statistics*. Springer Texts in Statistics. Springer-Verlag, New York. A concise course in statistical inference. 6, 12

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393**(6684), 440–442. 27

Wood, D. (1992). The computation of polylogarithms. Technical Report 15-92*, University of Kent, Computing Laboratory, University of Kent, Canterbury, UK. 60

Zhu, X. and Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation. 21

Zhu, X., Ghahramani, Z., Lafferty, J. et al. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, volume 3, pp. 912–919. 43

Summary

Asymptotic results in nonparametric Bayesian function estimation

Nonparametric models are widely used in practical applications and involve at least one infinite-dimensional parameter of interest which is commonly a function or a measure. A standard example of such a model would be a density function estimation problem, in which the goal is to recover the density function corresponding to a certain distribution by looking at the data drawn from it.

In the last decades Bayesian nonparametric methods became extremely popular for numerous reasons, such as their philosophical appeal and conceptual simplicity. However, until recently there was little fundamental mathematical understanding of such procedures. The essence of the Bayesian paradigm is the philosophical idea that the parameter of interest in a statistical model does not have one true value, but it is rather perceived as a random object itself. The evidence about the true state of the world is then expressed in terms of degrees of belief. In Bayesian statistics in order to get an estimator one has to put a prior distribution on a parameter space. This can be interpreted as one's prior opinion about the nature of the data. Central in the Bayesian framework is the posterior distribution which can be viewed as an updated version of the prior belief given the evidence.

It is sensible to evaluate the robustness of Bayesian procedures used in nonparametric settings in order to gain insight about which priors to use and how their tuning affects the performance of the procedures. One way to do it is to take an asymptotic approach and to study the Bayesian procedures from frequentist perspective by assuming that there exists a true underlying parameter and studying how well it could be approximated by the estimator as the data sample grows. With the growth of the data sample a good estimator should get in smaller and smaller neighbourhoods of the truth. Moreover, it is desirable for the estimator to be able to have a fast "learning" rate (in terms of the size of this neighbourhood) for every parameter in the class. However, it is well known that this rate is bounded from below by a minimax rate.

This thesis is focused on function estimation problems. For such problems the minimax rate, among other things, depends on the smoothness of the target function. For instance, in a particular case of a regression problem on the interval $[0, 1]$ it is clear that the rougher the target function is, the more data points we need in order to be able to accurately approximate the behaviour of the function. In most real world applications the smoothness of the function is not known in advance, so it is desirable to develop adaptive procedures that can adapt to all levels of smoothness in the parameter class.

In this thesis we develop Bayesian procedures and study their asymptotic behaviour in the context of two different statistical settings. First, we discuss regression and binary classification problems on large graphs, where the goal is to estimate a smooth function on the vertices of a graph. The second problem that we consider is the problem of intensity estimation of inhomogeneous Poisson process from a realised point pattern.

Chapter 1 introduces the main notions and concepts of Bayesian nonparametric statistics. In Chapter 2 we present a framework for studying the performance of methods for nonparametric function estimation on large graphs. We propose assumptions on the geometry of the underlying graph and the regularity of the function formulated in terms of the Laplacian of the graph. We also derive minimax rates for the regression and classification problems on large graph within the introduced framework. In Chapter 3 we exhibit nonparametric Bayes function estimation procedures in the graph setting. We study rescaled Gaussian priors and we show that the procedures based on these priors achieve good convergence rates and that they are rate–adaptive. We present the results for the cases of full and missing observations. Chapter 4 is dedicated to estimating the intensity function of an inhomogeneous Poisson process. We study the Bayesian procedure developed by Adams et al. [2009]. We show that their SGCP approach to learning intensity functions enjoys very favourable theoretical properties, provided the priors on the hyperparameters are chosen appropriately.

Samenvatting

Asymptotische resultaten in het niet-parametrisch Bayesiaans schatten van functies

Niet-parametrische modellen worden veel gebruikt in praktische toepassingen en brengen ten minste één oneindig-dimensionale parameter met zich mee, meestal in de vorm van een functie of een maat. Een standaard voorbeeld is het schatten van een onbekende dichtheidsfunctie aan de hand van trekkingen uit een bepaalde verdeling.

In de laatste decennia zijn niet-parametrische Bayesiaanse modellen bijzonder populair geworden om verschillende redenen. Zo zijn deze modellen vaak conceptueel eenvoudig en bevatten ze een zekere filosofische aantrekkingskracht. Tot voor kort waren er weinig fundamentele resultaten voor de bijbehorende schattingsprocedures. De essentie van het Bayesiaanse denkkader is het idee dat de parameter die men wil schatten in het statistische model niet een vaste, echte waarde heeft, maar zelf ook een toevalsvariabele is. De waarde van deze parameter in de echte wereld wordt uitgedrukt in waarschijnlijkheidstermen. Om binnen de Bayesiaanse statistiek tot een schatter te komen dient men eerst een a-prioriverdeling te kiezen op de parameterruimte. Het is gebruikelijk dat deze wordt geïnterpreteerd als a-priori vermoeden over de parameter. Centraal in het Bayesiaanse raamwerk is de a-posteriori verdeling, die kan worden geïnterpreteerd als een herziene versie van het a-priori vermoeden, gegeven de observaties.

Het ligt voor de hand om de robuustheid van niet-parametrische Bayesiaanse procedures te evalueren om inzicht te krijgen in welke a-prioriverdelingen bruikbaar zijn en hoe hun afstelling de prestatie beïnvloedt. Een manier om dit te doen, is een asymptotische aanpak waarin de Bayesiaanse procedures vanuit een frequentistisch oogpunt bestudeerd worden. Dit wordt gedaan door aan te nemen dat er een echte waarde bestaat voor de onderliggende parameter en te onderzoeken hoe goed deze benaderd kan worden naarmate de steekproef groter wordt. Voor een goede schatter geldt dat deze, naarmate de steekproef groter wordt, zich concentreert rond de echte waarde. Daarnaast is het wenselijk dat de schatter zich snel concentreert voor elke parameter in een bepaalde klasse. Het is bekend dat deze snelheid begrensd wordt door de zogenoemde “minimax rate”.

Dit proefschrift behandelt het schatten van functies. Voor dergelijke problemen hangt de minimax rate onder andere af van de gladheid van de te schatten functie. Bijvoorbeeld, in het geval van regressie op het interval $[0, 1]$ is het duidelijk dat naarmate de te schatten functie minder glad is, er meer data nodig is om het gedrag van de functie nauwkeurig te benaderen. Normaal gesproken is de gladheid van deze functie onbekend en derhalve is het wenselijk om adaptieve procedures te ontwikkelen die zich aan kunnen passen aan verschillende gladheidsgraden.

In dit proefschrift ontwikkelen we Bayesiaanse procedures en bestuderen we hun asymptotische gedrag in twee verschillende statistische contexten. Eerst behandelen we regressie en binaire classificatie op grote grafen, met als doel een gladde functie op de knooppunten van de graaf te schatten. Ten tweede kijken we naar het schatten van de intensiteitsfunctie van een inhomogeen Poisson proces.

Hoofdstuk 1 introduceert de belangrijkste begrippen en concepten van niet-parametrisch Bayesiaanse statistiek. In hoofdstuk 2 presenteren we een raamwerk voor het bestuderen van de prestaties van niet-parametrische methodes om functies te schatten op grote grafen. We presenteren aannames over de geometrie van de onderliggende graaf en de regulariteit van de functie geformuleerd in termen van de Laplaciaan van de graaf. Daarnaast leiden we minimax rates af voor regressie- en classificatieproblemen voor grote grafen binnen het geïntroduceerde raamwerk. In hoofdstuk 3 worden niet-parametrische Bayesiaanse functieschatters gepresenteerd voor grafen. We bestuderen geschaalde Gaussische a-prioriverdelingen en laten zien dat procedures gebaseerd op deze a-prioriverdelingen goede convergentiesnelheden hebben en dat zij zich aanpassen aan de onbekende gladheid van de functie. We presenteren de resultaten voor zowel het geval van volledige en van onvolledige observaties. Hoofdstuk 4 behandelt het schatten van de intensiteitsfunctie van een inhomogeen Poisson proces. We bestuderen de Baysiaanse procedure ontwikkeld door Adams et al. [2009] en laten zien dat hun SGCP aanpak erg gunstige theoretische eigenschappen heeft, gegeven dat de a-prioriverdeling van de hyperparameters juist gekozen zijn.

Author contributions

This thesis is based on the following two published articles and one article in preparation:

[1] Kirichenko, A. and van Zanten, J. H. (2015). Optimality of Poisson processes intensity learning with Gaussian processes. *J. Mach. Learn. Res.* **16**, 2909–2919.

[2] Kirichenko, A. and van Zanten, J. H. (2017). Estimating a smooth function on a large graph by Bayesian Laplacian regularisation. *Electron. J. Stat.* **11**(1), 891–915.

[3] Kirichenko, A. and van Zanten, J. H. (2017). Minimax lower bounds for function estimation problems on graphs. *In preparation.*

Each of the authors equally contributed to each of the articles.

Acknowledgments

I am thankful to all the people who supported me during my time as a PhD student at University of Amsterdam. Foremost, I want to express my deepest gratitude to my supervisor, Harry van Zanten. He was a constant source of inspiration and it was a great honour for me to work with him during the last four years. I appreciate his immediate availability for me at times when I needed help as well as the space that he gave me to conduct the research at my own pace. Harry?s cheerful character and genuine eagerness to do mathematics made working with him a very enjoyable experience.

I wish to express my sincere appreciation to NWO (De Nederlandse Organisatie voor Wetenschappelijk Onderzoek) that funded this PhD work. I am grateful to the members of my defence committee, Bert van Es, Subhashis Ghoshal, Mathisca de Gunst, Michel Mandjes, and Sindo Nunez Queija, for taking their time to read this thesis. I also thank the director of KdVI Eric Opdam and my co-promoter Bas Kleijn.

I am very grateful to Yakov Nikitin who introduced me to the fascinating field of mathematical statistics, guided me during my master's program, and supported me in finding a PhD position.

Furthermore, I would like to thank all my friends and colleagues who made my time in Amsterdam so pleasurable. I am especially grateful to my paranymphs, Jarno and Ana. Jarno's enthusiasm and dedication to every activity he is engaged in made the atmosphere at the university lively. He always made time for me and supported me. Ana is a person that I can turn to when I am struggling. Her constant optimism, deeply caring character, and her unconditional faith in me helped me to get through the most difficult times. I am also thankful to my officemate Jan for his kindness, passion for mathematics, and for all the delightful conversations we had over the years.

This dissertation has greatly benefitted from my numerous talks with Botond whose mathematical abilities and life philosophy I deeply admire. I would like to thank him for always being available if I had any questions and for his insightful comments on my work. I thank Denis for being there for me to discuss mathematical computations, polish my presentations, improve my writing skills, and assist in practically every other matter that I ever needed help with. I am also grateful to Aram for supporting me during the final stage of writing my PhD, for commenting on my writing, and for motivating me in every aspect of my life. I would like to mention Claudia, who inspired me greatly with her joyful spirit and inexhaustible energy. Her perspective on the challenges of academia and her useful advice on academic writing encouraged me to overcome the difficulties I encountered on my way.

I thank Kiyoshi Takahase Segundo for providing a picture for the cover of my thesis and Vladimir Levitin for its design.

Finally, I thank my family for their unconditional love and support. Words cannot express the amount of help I received from my parents, and I will be forever be grateful for all the sacrifices they made on my behalf. They have always been a great example of diligence, hard work, and loving kindness. Their keen passion for education and constant self development gave me the strength to enroll and complete my PhD.