A Statistical Taylor Theorem and Extrapolation of Truncated Densities

Constantinos Daskalakis COSTIS@CSAIL.MIT.EDU
Massachusetts Institute of Technology

Vasilis Kontonis KONTONIS@WISC.EDU
University of Wisconsin-Madison

Christos Tzamos TZAMOS@WISC.EDU
University of Wisconsin-Madison

Manolis Zampetakis MZAMPET@BERKELEY.EDU
University of California, Berkeley

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We show a statistical version of Taylor's theorem and apply this result to non-parametric density estimation from truncated samples, which is a classical challenge in Statistics Woodroofe (1985); Stute (1993). The single-dimensional version of our theorem has the following implication: "For any distribution P on $[0, 1]$ with a smooth log-density function, given samples from the conditional distribution of P on $[a, a + \varepsilon] \subset [0, 1]$, we can efficiently identify an approximation to P over the *whole* interval $[0, 1]$, with quality of approximation that improves with the smoothness of P."

To the best of knowledge, our result is the first in the area of non-parametric density estimation from truncated samples, which works under the hard truncation model, where the samples outside some survival set S are never observed, and applies to multiple dimensions. In contrast, previous works assume single dimensional data where each sample has a different survival set S so that samples from the whole support will ultimately be collected.

From a technical point of view, a central challenge that we face is to bound the extrapolation error of multivariate polynomial approximation. Our main technical contribution is to show a novel way to prove strong bounds on the extrapolation error of our algorithms invoking only well-studied *anti-concentration* theorems, which we believe that it will have applications beyond truncated statistics. [1]

Keywords: non-parametric density estimation, truncated statistics, extrapolation error

Acknowledgments

Constantinos Daskalakis was supported by NSF Awards IIS-1741137, CCF-1617730, and CCF-1901292, by a Simons Investigator Award, by the Simons Collaboration on the Theory of Algorithmic Fairness, by a DSTA grant, and by the DOE PhILMs project (No. DE-AC05-76RL01830). Christos Tzamos and Vasilis Kontonis were supported by the NSF grant CCF-2008006.

1. Extended abstract. Full version appears as [arXiv reference,2106.15908], https://arxiv.org/abs/2106.15908

References

Ibrahim Ahamada and Emmanuel Flachaire. Non-parametric econometrics. *OUP Catalogue*, 2010.

Andrew R. Barron and Chyong-Hwa Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, 19(3):1347–1369, 1991. ISSN 00905364. URL http://www.jstor.org/stable/2241953.

Andrew R Barron, Lhszl Gyorfi, and Edward C van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE transactions on Information Theory*, 38(5):1437–1454, 1992.

Behrouz Behmardi, Raviv Raich, and Alfred O Hero. Entropy estimation using the principle of maximum entropy. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2008–2011. IEEE, 2011.

Shalev Ben-David, Adam Bouland, Ankit Garg, and Robin Kothari. Classical lower bounds from quantum upper bounds. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 339–349, 2018.

Axel Börsch-Supan and Vassilis A Hajivassiliou. Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of econometrics*, 58(3):347–368, 1993.

Zdravko I. Botev, Joseph F. Grotowski, and Dirk P. Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916–2957, 2010.

Clément L Canonne, Anindya De, and Rocco A Servedio. Learning from satisfying assignments under continuous distributions. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 82–101. SIAM, 2020.

Stéphane Canu and Alex Smola. Kernel methods and the exponential family. *Neurocomputing*, 69 (7-9):714–720, 2006.

Anthony Carbery and James Wright. Distributional and ℓ^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical research letters*, 8(3):233–248, 2001.

Xiaohong Chen, Yanqin Fan, and Viktor Tsyrennikov. Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association*, 101(475):1228–1240, 2006.

A Clifford Cohen. *Truncated and censored samples: theory and applications*. CRC press, 1991.

Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *the 59th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2018.

Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Computationally and statistically efficient truncated regression. In *Conference on Learning Theory*, pages 955–960, 2019.

Leslaw Gajek. On the minimax value in the scale model with truncated data. *The Annals of Statistics*, 16(2):669–677, 1988.

Mariano Gasca and Thomas Sauer. Polynomial interpolation in several variables. *ADV. COMPUT. MATH*, 12:377–410, 2000.

Kaan Gokcesu and Suleyman S Kozat. Online density estimation of nonstationary sources using exponential family of distributions. *IEEE transactions on neural networks and learning systems*, 29(9):4473–4478, 2017.

Irving J Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, 34(3):911–934, 1963.

James J Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. NBER, 1976.

Andrew Ilyas, Manolis Zampetakis, and Daskalakis Constantinos. A theoretical and practical framework for regressionand classification from truncated samples. In *AISTATS 2020*, 2020.

Vasilis Kontonis, Christos Tzamos, and Manolis Zampetakis. Efficient truncated statistics with unknown truncation. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1578–1595. IEEE, 2019.

Tze Leung Lai and Zhiliang Ying. Estimating a distribution function with truncated and censored data. *The Annals of Statistics*, pages 417–442, 1991.

Qi Li and Jeffrey Scott Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 2007.

Gangadharrao S Maddala. Limited dependent variable models using panel data. *Journal of Human resources*, pages 307–338, 1987.

Alexander Schrijver (auth.) Martin Grötschel, László Lovász. *Geometric Algorithms and Combinatorial Optimization*. Algorithms and Combinatorics 2. Springer-Verlag Berlin Heidelberg, 2 edition, 1993. ISBN 9780387136240,038713624X,354013624X,3540152857,3540170960. URL http://gen.lib.rus.ec/book/index.php?md5=70edd72d6da66b28a18839be9c2b1f9a.

Daniel McDonald. Minimax density estimation for growing dimension. In *Artificial Intelligence and Statistics*, pages 194–203, 2017a.

Daniel J McDonald. Minimax density estimation for growing dimension. *arXiv preprint arXiv:1702.08895*, 2017b.

Jerzy Neyman. Smooth test for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4):149–199, 1937.

WJ Padgett and Diane T McNichols. Nonparametric density estimation from censored data. *Communications in Statistics-Theory and Methods*, 13(13):1581–1611, 1984.

James Renegar. On the computational complexity of approximating solutions for real algebraic formulae. *SIAM J. Comput.*, 21(6):1008–1025, 1992a. doi: 10.1137/0221060. URL https://doi.org/10.1137/0221060.

James Renegar. On the computational complexity and geometry of the first-order theory of the reals, part III: quantifier elimination. *J. Symb. Comput.*, 13(3):329–352, 1992b. doi: 10.1016/S0747-7171(10)80005-7. URL https://doi.org/10.1016/S0747-7171(10)80005-7.

David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Jeffrey S Simonoff. *Smoothing methods in statistics*. Springer Science & Business Media, 2012.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18, 2017.

Winfried Stute. Almost sure representations of the product-limit estimator for truncated data. *The Annals of Statistics*, 21(1):146–156, 1993.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008a.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008b.

Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC press, 1994.

Shaojun Wang, Russell Greiner, and Shaomin Wang. Consistency and generalization bounds for maximum entropy density estimation. *Entropy*, 15(12):5439–5463, 2013.

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

Michael Woodroofe. Estimating a distribution function with truncated data. *The Annals of Statistics*, 13(1):163–177, 1985.

Ximing Wu. Exponential series estimator of multivariate densities. *Journal of Econometrics*, 156 (2):354–366, 2010.