Approximate Inference for Non-parametric Bayesian Hawkes Processes and Beyond

Rui Zhang

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

February 2022

Except where otherwise indicated, this thesis is my own original work.

Rui Zhang
28 February 2022

To my parents Zike Zhang, Cuihua Qi, and my fiancée Xiaoyi Chen.

Chair of Panel

Lexing Xie
Professor of Computer Science, The Australian National University
Canberra, ACT, Australia

Primary Supervisor

Christian Walder
Activity Leader and Senior Researcher, Data61 in CSIRO
Canberra, ACT, Australia

Associate Supervisor

Marian-Andrei Rizoiu
Lecturer of Computer Science, The University of Technology Sydney
Sydney, NSW, Australia

Acknowledgments

I would like to express my sincere gratitude to the people supporting me during the doctoral journey:

1. My supervisors: Prof. Lexing Xie, Dr. Christian Walder and Dr. Marian-Andrei Rizoiu. It is an extraordinary honor for me to have Lexing, Christian and Andrei to be my advisors. They have immense knowledge in various fields and a strong dedication to research, which motivates me to become a good researcher. They provide patient and continuous guidance and support of my research, which helped this thesis to become a reality.

2. The Research School of Computer Science at ANU and Data61 in CSIRO. I would like to thank both institutes for providing me the PhD scholarships. Their healthy workspace and sufficient resources are crucial for me to develop research skills and explore fields of interest.

3. Members and Alumni of the ANU Computational Media Lab: Dongwoo Kim, Quyu Kong, Swapnil Mishra, Umanga Bista, Siqi Wu, and many others. Much of my work benefits from the discussions with them. I am grateful that I can work with a group of supportive talents.

4. External collaborators: Krikamol Muandet and Bernhard Schölkopf at Max Planck Institute for Intelligent Systems, Masaaki Imaizumi at the University of Tokyo and Edwin Bonilla at Data61 in CSIRO. Their rigorous research spirits have made a profound influence on me.

5. National Computational Infrastructure (NCI). I acknowledge the computational support from NCI which is a very helpful research platform to my PhD.

6. My parents Zike Zhang and Cuihua Qi, my fiancée Xiaoyi Chen and my relatives. I would like to thank my relatives for their support and care. Most importantly, I would like to express deep thanks to my parents Zike Zhang and Cuihua Qi and my fiancée Xiaoyi Chen. They gave me unconditional love and encouragement throughout my PhD.

Abstract

The Hawkes process has been widely applied to modeling self-exciting events including neuron spikes, earthquakes and tweets. To avoid designing parametric triggering kernels, the non-parametric Hawkes process has been proposed, in which the triggering kernel is in a non-parametric form. However, inference in such models suffers from poor scalability to large-scale datasets and sensitivity to uncertainty in the random finite samples. To deal with these issues, we employ Bayesian non-parametric Hawkes processes and propose two kinds of efficient approximate inference methods based on existing inference techniques. Although having worked as the cornerstone of probabilistic methods based on Gaussian process priors, most of existing inference techniques approximately optimize standard divergence measures such as the Kullback-Leibler (KL) divergence, which lacks the basic desiderata for the task at hand, while chiefly offering merely technical convenience. In order to improve them, we further propose a more advanced Bayesian inference approach based on the Wasserstein distance, which is applicable to a wide range of models. Apart from these works, we also explore a robust frequentist estimation method beyond the Bayesian field. Efficient inference techniques for the Hawkes process will help all the different applications that it already has, from earthquake forecasting, finance to social media. Furthermore, approximate inference techniques proposed in this thesis have the potential to be applied to other models to improve robustness and account for uncertainty.

More specifically, we first develop an efficient non-parametric Bayesian estimation of the triggering kernel of Hawkes processes based on Gibbs sampling. Our method considers a Gaussian process modulated triggering kernel and is developed based on the cluster representation of Hawkes processes. Utilizing the finite support assumption of the Hawkes process, we efficiently sample random branching structures and thus, we split the Hawkes process into clusters of Poisson processes. We derive two algorithms — a block Gibbs sampler and a maximum a posteriori estimator based on expectation maximization — and we show that our methods have a linear time complexity, both theoretically and empirically. On synthetic data, we show our methods to be able to infer flexible Hawkes triggering kernels. On two large-scale Twitter diffusion datasets, we show that our methods outperform the current state-of-the-art in goodness-of-fit and that the time complexity is linear in the size of the dataset. We also observe that on diffusions related to online videos, the learned kernels reflect the perceived longevity for different content types such as music or pet videos.

Secondly, we propose a new non-parametric Bayesian Hawkes process in which the triggering kernel is modeled as a squared sparse Gaussian process and a novel

variational inference schema is proposed for model optimization. We again employ the branching structure of the Hawkes process so that maximization of evidence lower bound (ELBO) is tractable by the expectation-maximization algorithm. We propose a tighter ELBO which improves the fitting performance. Further, we accelerate the novel variational inference scheme to linear time complexity by leveraging the finite support assumption of the triggering kernel. Different from prior acceleration methods, ours enjoys higher efficiency. Finally, we exploit synthetic data and two large social media datasets to evaluate our method. We show that our approach outperforms state-of-the-art non-parametric frequentist and Bayesian methods. We validate the efficiency of our accelerated variational inference scheme and the practical utility of our tighter ELBO for model selection. We observe that the tighter ELBO exceeds the common one in model selection.

Thirdly, we develop a new approximate inference method for Gaussian process models which overcomes the technical challenges arising from abandoning those convenient divergences. Our method—dubbed Quantile Propagation (QP)—is similar to expectation propagation (EP) but minimizes the L_2 Wasserstein distance instead of the KL divergence. The Wasserstein distance exhibits all the required properties of a distance metric, while respecting the geometry of the underlying sample space. We show that QP matches quantile functions rather than moments as in EP and has the same mean update but a smaller variance update than EP, thereby alleviating EP's tendency to over-estimate posterior variances. Crucially, despite the significant complexity of dealing with the Wasserstein distance, QP has the same favorable locality property as EP, and thereby admits an efficient algorithm. Experiments on classification and Poisson regression show that QP outperforms both EP and variational Bayes.

Finally, we propose a simple and robust framework for the estimation of conditional moment restriction (CMR) models which include the Hawkes process. The framework is developed based on a kernelized CMR known as a maximum moment restriction (MMR) and applied to nonlinear instrumental variable (IV) regression, which we are particularly interested in. The MMR is formulated by maximizing the interaction between the residual and the instruments belonging to a unit ball in a reproducing kernel Hilbert space (RKHS). The MMR allows us to reformulate the IV regression as a single-step empirical risk minimization problem, where the risk depends on the reproducing kernel on the instrument and can be estimated by a U-statistic or V-statistic. This simplification not only eases the proofs of consistency and asymptotic normality in both parametric and non-parametric settings, but also results in easy-to-use algorithms with an efficient hyper-parameter selection procedure. We demonstrate the advantages of our framework over existing ones using experiments on both synthetic and real-world data.

Publications and Software

Published and under-reviewed papers are summarized as follows. The software developed in these works is also provided.

- **Zhang**, R.; Muandet, K.; Schölkopf, B.; and Imaizumi, M., 2021b. Instrument space selection for kernel maximum moment restriction. *arXiv preprint*, (2021). Submitted to NeurIPS 2021. **Paper**: https://arxiv.org/abs/2106.03340. **Software**: https://github.com/RuiZhang2016/Instrument-Space-Selection-for-Kernel-Maximum-Moment-Restriction. (Not included)

- **Zhang**, R.; Imaizumi, M.; Schölkopf, B.; and Muandet, K., 2021a. Maximum moment restriction for instrumental variable regression. *arXiv preprint*, (2021). Submitted to NeurIPS 2021. **Paper**: https://arxiv.org/abs/2010.07684. **Software**: https://github.com/RuiZhang2016/MMRIV. (In Chapter 6)

- **Zhang**, R.; Walder, C.; Bonilla, E. V.; Rizoiu, M.-A.; and Xie, L., 2020a. Quantile propagation for wasserstein-approximate gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, 21566–21578. Curran Associates, Inc. **Paper**: https://papers.nips.cc/paper/2020/hash/f5e62af885293cf4d511ceef31e61c80-Abstract.html. **Software**: https://github.com/RuiZhang2016/Quantile-Propagation-for-Wasserstein-Approximate-Gaussian-Processes. (In Chapter 5)

- **Zhang**, R.; Walder, C.; and Rizoiu, M.-A., 2020b. Variational inference for sparse gaussian process modulated hawkes proces. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*. New York, U.S.A. **Paper**: https://arxiv.org/abs/1905.10496. **Software**: https://github.com/RuiZhang2016/Variational-Inference-for-SGP-Modulated-Hawkes-Process. (In Chapter 4)

- **Zhang**, R.; Walder, C.; Rizoiu, M.-A.; and Xie, L., 2019. Efficient non-parametric bayesian hawkes processes. In *International Joint Conference on Artificial Intelligence (IJCAI 2019)*. Macao, China. **Paper**: http://arxiv.org/abs/1810.03730. **Software**: https://github.com/RuiZhang2016/Efficient-Nonparametric-Bayesian-Hawkes-Processes. (In Chapter 3)

Contents

List of Figures

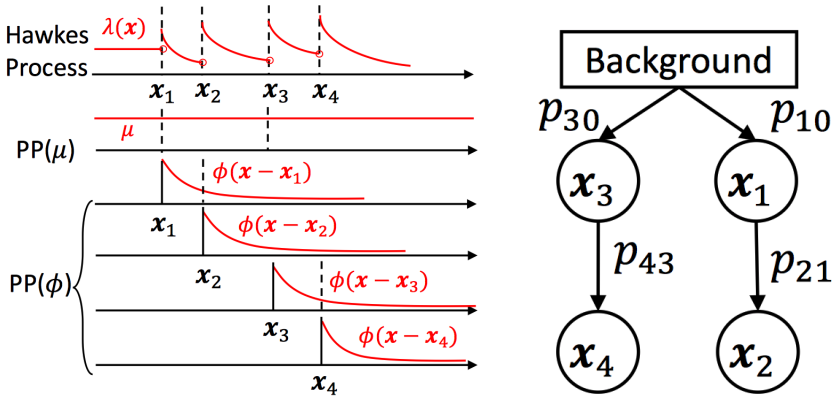
List of Tables

Chapter 1

Introduction

Hawkes (point) processes [Hawkes, 1971] have been widely used to model self-exciting event sequences in which existing events increase the likelihood of occurrence of future events. An example of self-exciting data is that on the online social media, attractive tweets can easily diffuse by being retweeted [Simma and Jordan, 2010], and retweeting promotes these tweets being seen and retweeted by more people. As another example, in the seismic activities, an intensive earthquake tends to trigger a series of aftershocks, and aftershocks possibly cause more aftershocks [Zhao et al., 2015]. In addition, many finance data also have the self-excitement characteristic [Bacry et al., 2015] — we recommend related works therein for more self-exciting cases.

The Hawkes process is suitable for self-exciting events because it explicitly models the self-exciting interactions between events. Specifically, every event in the Hawkes process is assumed to be triggered by either a previous event or the exterior stimulus. To quantify the former triggering factor, the model uses a triggering kernel function $\phi(\cdot)$, and for the latter, it employs a background intensity function $\mu(\cdot)$. ϕ and μ measure the occurrence rate of events, and are non-negative-valued functions. The aggregate occurrence rate is the sum of μ and ϕ of every previous event, which is usually denoted as a $\lambda(\cdot)$ function. Consequently, we can understand a Hawkes process as a cluster of Poisson processes [Hawkes and Oakes, 1974]. In the cluster view, a Poisson process with an intensity μ (denoted as PP(μ)) generates immigrant points which arrive in the system from the outside, and every existing point triggers offspring points, which are generated internally through self-excitement, via a PP(ϕ). Points are therefore structured into clusters where each cluster contains either a point and its direct offspring or the background process (an example is shown in Figure 1.1(a)). Connecting all points using the triggering relations yields a tree structure, which is called the branching structure (an example is shown in Figure 1.1(b) corresponding to Figure 1.1(a)). With the branching structure, we can decompose the Hawkes process into a cluster of Poisson processes. The triggering kernel ϕ is shared among all cluster Poisson processes relating to a Hawkes process, and it determines the overall behavior of the process. Consequently, designing the kernel functions is of utmost importance for employing the Hawkes process to a new application, and its study has attracted much attention.



(a) Poisson cluster process (b) Branching structure

Figure 1.1: The cluster representation of a Hawkes process. In (a), a Hawkes process with a decaying triggering kernel $\phi(\cdot)$ has intensity $\lambda(x)$ which increases after each new point (vertical dash line) is generated. It can be viewed as a cluster of Poisson processes: PP(μ) and PP($\phi(x - x_i)$) associated with each x_i. Figure (b) presents the branching structure of the Hawkes process in (a) and it reflects the triggering relationships between points. Here, an edge $x_i \rightarrow x_j$ means that x_i triggers x_j, and its probability is denoted as p_{ji}.

1.1 Research Questions

Most of the recent works employing Hawkes Processes [Bao et al., 2015; Filimonov and Sornette, 2015; Lallouache and Challet, 2016; Mishra et al., 2016; Rizoiu et al., 2017] design parametric kernels modeling a predetermined subset of social processes such as the limited length of collective memory [Wu and Huberman, 2007], or preferential attachment [Barabási, 2005]. Manually designing parametric kernels is an expensive process and may not generalize well to other applications. An open question is **(Q1) can we design non-parametric solutions for the kernel function?**

There are many non-parametric estimations of Hawkes triggering kernels, such as the works of Lewis and Mohler [2011]; Zhou et al. [2013]; Bacry and Muzy [2016]; Eichler et al. [2017]. These are all frequentist methods and among them, the Wiener-Hopf equation based method [Bacry and Muzy, 2016] takes the advantage of quadrature approximation for the integrals in the equation and obtains linear time complexity in estimating the triggering kernel, while it is sensitive to the employed quadrature points. A different class of estimation methods are based on the Euler-Lagrange equation [Lewis and Mohler, 2011; Zhou et al., 2013]. Similarly, these methods require discretizing the input domain and as a result, they face a problem of poorly scaling with the dimension of the domain. The same problem is also faced by Eichler et al. [2017]'s discretization based method. Besides, frequentist methods do not model uncertainty over the learned triggering kernels and tend to be sensitive to the point process realizations. A second open question is **(Q2) can we design continuous and more robust non-parametric estimations, which account for the variance and the noise in the observed real life data?**

The Bayesian inference for the Hawkes process has also been studied, including the work of Rasmussen [2013]; Linderman and Adams [2014]; Linderman and Adams [2015]. These works require either constructing a parametric triggering kernel [Rasmussen, 2013; Linderman and Adams, 2014] or discretizing the input domain to scale with the data size [Linderman and Adams, 2015]. To overcome the aforementioned shortcomings of discretization, Donnet et al. [2018] propose a continuous non-parametric Bayesian Hawkes process and resort to an unscalable Markov chain Monte Carlo (MCMC) approach to the posterior distribution. The third open question is **(Q3) can we design efficient inference for non-parametric Bayesian Hawkes processes?**

1.2 Thesis Contributions

In the thesis, we answer the above three questions by proposing four solutions:

(i) the Laplace Bayesian Hawkes process [**Zhang** et al., 2019], in Section 1.2.1,
(ii) the variational Bayesian Hawkes process [**Zhang** et al., 2020b], in Section 1.2.2,
(iii) the quantile propagation [**Zhang** et al., 2020a], in Section 1.2.3,
(iv) the kernel maximum moment restriction estimation [**Zhang** et al., 2021a], in Section 1.2.4.

The first two works, namely, (i) the Laplace Bayesian Hawkes process and (ii) the variational Bayesian Hawkes process, are proposed as direct solutions to all three questions (Q1∼3). They are applications of existing approximation methods for Bayesian inference. Compared with the former, the latter employs the more complicated and advanced variational inference method, and enjoys faster training speed.

The last two works (iii) and (iv) study robust estimation methods beyond the field of the Hawkes process and are more general solutions to Q2∼3. The work of quantile propagation explores a new approximation method for Bayesian inference. This method is similar to the expectation propagation (EP) algorithm [Opper and Winther, 2000] in the use of iterative local updates but employs the L_2 Wasserstein distance instead of the Kullback-Leibler (KL) divergence. Due to the locality of the computations, it is simple for our method and EP to parallelize and distribute, leading to more efficiency than variational inference [Li et al., 2015] especially on large-scale data [Gelman et al., 2017]. However unlike variational inference which minimizes the lower bound of the log model evidence, EP and QP correspond to no explicit global objective functions being minimized, and there is lack of works providing a clear understanding of the iterative updates, making EP and our QP behave more in a more complicated way than variational inference. Consequently, applying the new method to the Bayesian Hawkes process is challenging and non-trivial, and left for future work. It is observed that our method outperforms EP on the Gaussian process binary classification tasks, so we expect it to offer efficient and accurate performance for the Bayesian Hawkes process on large scale data.

Different from the above Bayesian solutions (i∼iii), the final one (iv) proposes a frequentist estimation method which has an advantage of robustness and simplicity.

This method is a generalization of maximum likelihood estimation and the method of moments that are commonly employed by existing frequentist estimation methods [Lewis and Mohler, 2011; Zhou et al., 2013; Bacry and Muzy, 2016; Eichler et al., 2017]. Different from obtaining robustness via modeling uncertainty in the Bayesian methods, the method has the spirit of adversarial learning. It considers a reproducing kernel Hilbert space (RKHS) weighted average of the loss function values on data points and estimates the model parameters by optimizing the maximal or the worst average value. This idea is general and applicable to a wide range of estimation tasks, including estimation of the Hawkes process (as elaborated in Section 2.4), and we demonstrate its effectiveness on instrumental variable regression, which we are particularly interested in.

1.2.1 Laplace Bayesian Hawkes Process

In the first solution, we exploit block Gibbs sampling [Ishwaran and James, 2001] to iteratively sample the latent branching structure, the background intensity μ and the triggering kernel ϕ. In each iteration, the point data are decomposed as a cluster of Poisson processes based on the sampled branching structure. The posterior μ and ϕ are estimated using the resulting cluster processes. Our framework is close to the stochastic Expectation-Maximization (EM) algorithm [Celeux and Diebolt, 1985] where posterior μ and ϕ are estimated [Lloyd et al., 2015; Walder and Bishop, 2017] in the M-step and random samples of μ and ϕ are drawn. We adapt the approach of the recent non-parametric Bayesian estimation for Poisson process intensities, termed Laplace Bayesian Poisson process [Walder and Bishop, 2017], to estimate the posterior ϕ given the sampled branching structure. We utilize the finite support assumption of the Hawkes Process to speed up sampling and to compute the probability of the branching structure. We theoretically show our method to be of linear time complexity. Furthermore, we explore the connection with the EM algorithm [Dempster et al., 1977] and develop a second variant of our method, as an approximate EM algorithm. We empirically show that our method enjoys linear time complexity and can infer known analytical kernels, i.e., exponential and sinusoidal kernels. On two large-scale social media datasets, our method outperforms the current state-of-the-art non-parametric approaches and the learned kernels reflect the perceived longevity for different content types. We propose a new acceleration trick based on the finite support assumption of the triggering kernel. The new trick enjoys higher efficiency than previous methods and accelerates the variational inference schema to linear time complexity per iteration.

1.2.2 Variational Bayesian Hawkes Process

The second solution proposes the first sparse Gaussian process modulated Hawkes process which employs a novel variational inference scheme, enjoys a linear time complexity per iteration and scales to large real world data. Our method is inspired by the variational Bayesian Poisson process (VBPP) [Lloyd et al., 2015] which provides Bayesian non-parametric inference only for the whole intensity of the Hawkes process

rather than its components: the background intensity μ and the triggering kernel ϕ. Thus, the VBPP loses the internal reactions between points, and developing the variational Bayesian non-parametric inference for the Hawkes process is non-trivial and more challenging than the VBPP. In this work, we adapt the VBPP for the Hawkes process and term the new approach the variational Bayesian Hawkes process (VBHP). We employ a sparse Gaussian process modulated triggering kernel and a Gamma distributed background intensity and propose a new variational inference scheme for such models. Specifically, we employ the branching structure of the HP so that maximization of the evidence lower bound (ELBO) is tractable by the expectation-maximization algorithm, and we contribute a tighter ELBO which improves the fitting performance of our model. We propose a new acceleration trick based on the finite support assumption of the triggering kernel. The new trick enjoys higher efficiency than prior methods and accelerates the variational inference schema to linear time complexity per iteration. We empirically show that VBHP provides more accurate predictions than state-of-the-art methods on synthetic data and on two large online diffusion datasets. We validate the linear time complexity and faster convergence of our accelerated variational inference scheme compared to the Gibbs sampling method, and the practical utility of our tighter ELBO for model selection, which outperforms the common one in model selection.

1.2.3 Quantile Propagation for Gaussian Process Models

Beyond the area of the Hawkes process, we develop an approximate inference algorithm for Gaussian process models with factorized likelihoods based on minimization of the L_2 Wasserstein distance. Our approach employs a Gaussian likelihood to approximate the non-Gaussian likelihood. In order to optimize Gaussian likelihoods, we avoid directly minimizing the global L^2 Wasserstein distance between true and approximate joint posteriors, due to its computational and analytical intractability in high dimensional spaces. Instead, our method iteratively minimizes local L^2 Wasserstein distances, like EP. As a result, our method matches two quantile functions different from moment matching in EP, thus named quantile propagation. We further derive updating formulas for the mean and the variance of the Gaussian likelihood. The estimation of the mean is equal to EP's, while the variance is less than EP's, hence alleviating EP's deficiency of over-estimating variances [Minka, 2005; Heess et al., 2013; Hernández-Lobato et al., 2016]. We show that the optimal approximate Gaussian likelihood enjoys an economical parameterization like EP, i.e., relying on a single latent variable instead of all of them. This property allows our method or EP to significantly reduce memory consumption by a factor N (the number of data) and computation time via optimizing much less ($O(1)$, vs $O(N^2)$ for the full parameterization) parameters in each local update. We regard both methods as approximate coordinate descent algorithms to a KL divergence and a L_2 Wasserstein objective function respectively, under the same approximation assumption. In the experiment part, we evaluate EP and our method via Gaussian process binary classification on a number of real world datasets. Our results show that our method can outperform

EP in both predictive accuracy and uncertainty quantification, which validates our method alleviating over-estimation of the variance. We will apply this method to the Bayesian Hawkes process in the future and consider parallelizing or distributing the local updates. We expect it to perform efficiently and accurately on large-scale data.

1.2.4 Kernel Maximum Moment Restriction Estimation

The last solution proposes a simple framework for the nonlinear instrumental variable regression, which is a more difficult and general problem than the model parameter estimation for the Hawkes process. The framework is based on a kernelized conditional moment restriction known as a kernel maximum moment restriction (KMMR). The KMMR is formulated by maximizing the interaction between the residual and the instruments belonging to a unit ball in an RKHS. The KMMR allows us to reformulate the IV regression as a single-step empirical risk minimization problem, where the risk depends on the reproducing kernel on the instrument and can be estimated by a U-statistic or V-statistic. This simplification not only eases the proofs of consistency and asymptotic normality in both parametric and non-parametric settings, but also results in easy-to-use algorithms with an efficient hyper-parameter selection procedure. We demonstrate the advantages of our framework over existing methods using experiments on both synthetic and real-world data.

1.3 Broader Impact

The proposed inference techniques for Hawkes processes in the thesis have an advantage of a linear time complexity. Most of existing applications of Hawkes processes, from earthquake forecasting, finance to social media, take quadratic time complexity to estimate model parameters. Our algorithms thus will help to develop efficient applications. The approximate inference techniques also have the potential to be applied to other models to improve their robustness and account for uncertainty.

1.4 Thesis Outline

The subsequent chapters provide details of solutions to the proposed research questions. We first introduce prerequisites of the solutions and review related developments of them in Chapter 2. Then we elaborate on the four proposed methods in Chapters 3, 4, 5 and 6 respectively, which are summarized concisely in Chapter 7. Chapter 7 also provides some interesting future research directions.

Chapter 2

Preliminaries and Related Work

The works in the thesis are on Hawkes processes, approximate inference for Gaussian processes and robust frequentist estimation. In this chapter, we introduce the preliminaries of them and review the related works in the three research areas separately. Specifically, in Section 2.1, we first introduce the Poisson and Hawkes processes. We then review the Gaussian process model and the approximate Bayesian inference approaches used in the thesis in Section 2.2. In Section 2.3 and 2.4, we introduce the Wasserstein distance and the conditional moment restriction respectively. Then, we review the related work on estimation of Hawkes process, on expectation propagation and Wasserstein distance, and on condition moment restriction based estimation, in Sections 2.5, 2.6 and 2.7 respectively.

2.1 Poisson and Hawkes Processes

2.1.1 Poisson Processes

The Poisson (point) process assumes that points in any bounded subregion of the domain are independent of those in other subregions. More specifically, for any subset of the d-dimensional real space $\mathcal{T} \in \mathbb{R}^d$, other intervals $\mathcal{T}'$ disjoint from $\mathcal{T}$ do not affect the occurrence of events in $\mathcal{T}$, and the probability of "the event count in $\mathcal{T}$, $N(\mathcal{T})$, being equal to n" is determined by the Poisson distribution:

$$P\{N(\mathcal{T}) = n\} \coloneqq \frac{\Lambda(\mathcal{T})^n}{n!} e^{-\Lambda(\mathcal{T})}, \quad \Lambda(\mathcal{T}) \coloneqq \int_{\mathcal{T}} \lambda(x)\, \mathrm{d}x,$$

where $\Lambda(\mathcal{T})$ is the expected number of points in $\mathcal{T}$, and the function $\lambda(x)$ is known as the intensity and modulates the occurrence rate of events at x. The log-likelihood of $\mathcal{D} \coloneqq \{x_i\}_{i=1}^N$ given λ is [Rubin, 1972]:

$$\log p(\mathcal{D}|\lambda) = \sum_{i=1}^{N} \log \lambda(x_i) - \int_{\Omega} \lambda(x)\, \mathrm{d}x,$$

where Ω is the sample domain of $\{x_i\}_{i=1}^N$. The log-likelihood of any stochastic point process, such as the Hawkes process, has the same form. We recommend [Daley and Vere-Jones, 2003] for a detailed introduction to stochastic point processes.

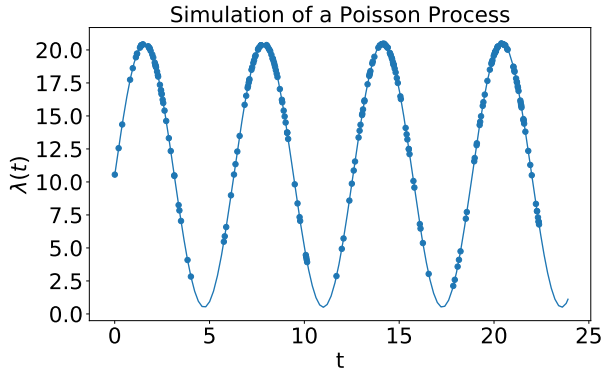


Figure 2.1: A Poisson process realization (points). The intensity function $\lambda(t) = 10\sin(t) + 10.5$ (solid line) is used to simulate points and higher function values generates more points.

2.1.2 Hawkes Processes

The *Hawkes process* [Hawkes, 1971] is a self-exciting point process, in which the occurrence of a point increases the arrival rate $\lambda(\cdot)$ of new points. Given a set of ordered points $\mathcal{D} = \{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}^d$, the intensity at x conditioned on given points is written as:

$$\lambda(x) = \mu + \sum_{x_i < x} \phi(x - x_i),$$

where $\mu > 0$ is background intensity, commonly considered as a constant, and $\phi : \mathbb{R}^d \to [0, \infty)$ is the triggering kernel. We consider $d = 1$ for a concise presentation and extension to $d > 1$ follows by the same development procedure. In the thesis, we are interested in the branching structure of the Hawkes process. As introduced in Chapter 1, each point x_i has a parent that we represent by the below one-hot vector. The index is up to $i - 1$ as the points are ordered.

$$\boldsymbol{b}_i = [b_{i0}, b_{i1}, \cdots, b_{i,i-1}]^T. \tag{2.1}$$

Each element b_{ij} is binary, $b_{ij} = 1$ represents that x_i is triggered by x_j ($0 \le j \le i - 1$, x_0: the background), and $\sum_{j=0}^{i-1} b_{ij} = 1$. A *branching structure* B is the set of $\boldsymbol{b}_i$ and is determined by $B = \{\boldsymbol{b}_i\}_{i=1}^N$. We define the probability of $b_{ij} = 1$ as (see e.g. [Lewis and Mohler, 2011])

$$p_{ij} \coloneqq p(b_{ij} = 1) = \begin{cases} \phi(x_i - x_j)/\lambda(x_i), & 0 < j \le j - 1 \\ \mu/\lambda(x_i), & j = 0 \end{cases}, \tag{2.2}$$

then the probability of B is expressed as below and it is clear $\sum_{j=0}^{i-1} p_{ij} = 1$ for all i.

$$p(B) = \prod_{i=1}^{N} \prod_{j=0}^{i-1} p_{ij}^{b_{ij}}.$$

Log-likelihood Given Branching Structure. Given the branching structure B, the Hawkes process can be viewed as a cluster of Poisson processes, namely, $\mathrm{PP}(\mu)$ and $\{\mathrm{PP}(\phi(x - x_i))\}_{i=1}^N$ as introduced in Chapter 1. Consequently, the log likelihood of the Hawkes process becomes a sum of log-likelihoods of Poisson processes. The data domain of $\mathrm{PP}(\mu)$ equals that of the Hawkes process, which we denote as Ω, and we denote the data domain of $\mathrm{PP}(\phi(\boldsymbol{x} - \boldsymbol{x}_i))$ by $\Omega_i \subset \Omega$. As a result, the log-likelihood of data points and the branching structure is calculated by:

$$\log p(\mathcal{D}, B|\mu, \phi) = \sum_{i=1}^{N} \Big(\sum_{j=1}^{i-1} b_{ij} \log \phi_{ij} + b_{i0} \log \mu \Big) - \sum_{i=1}^{N} \int_{\Omega_i} \phi - \mu |\mathcal{T}|, \tag{2.3}$$

where $|\Omega| \coloneqq \int_\Omega 1 \, \mathrm{d}x$, $\phi_{ij} \coloneqq \phi(x_i - x_j)$ and $\int_{\Omega_i} \phi \coloneqq \int_{\Omega_i} \phi(x) \, \mathrm{d}x$. Note that the Lebesgue measure is considered here.

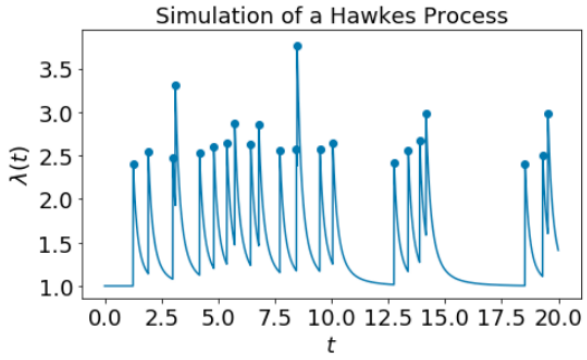


Figure 2.2: A Hawkes process realization (points). The intensity function $\lambda(t) = 1 + \sum_{t_i<t} 0.1e^{-(t-t_i)} + 0.6e^{-3(t-t_i)} + 0.7e^{-7(t-t_i)}$ (solid line) is used to simulate points.

2.2 Gaussian Processes

2.2.1 Exact Gaussian Processes

The Gaussian process (GP) is defined as a distribution over a function f, denoted as $\mathcal{GP}(f)$, such that for any non-empty set of $\boldsymbol{x} \coloneqq \{x_1, \cdots, x_N\}$ in the domain of f, function values $\boldsymbol{f} \coloneqq \{f(x_i)\}_{i=1}^N$ jointly have a Gaussian distribution, i.e.,

$$p(\boldsymbol{f}|\theta) = \mathcal{N}(\boldsymbol{\mu_x}, \boldsymbol{K_x}), \tag{2.4}$$

where the mean vector $\boldsymbol{\mu_x}$ and the covariance matrix $\boldsymbol{K_x}$ are the evaluation of a mean function $\mu(\cdot)$ and a covariance function $k(\cdot, \cdot)$ on $\boldsymbol{x}$, that is, $\boldsymbol{\mu_x} \coloneqq \{\mu(x_i)\}_{i=1}^N$ and $\boldsymbol{K_x} \coloneqq [k(x_i, x_j)]_{i,j=1}^N$, and θ is the set of parameters of μ and k. The covariance function defines the covariance between two function values,

$$k(x, y) = \mathrm{Cov}(f(x), f(y)),$$

and any covariance function k can be represented in terms of the eigenvalues $\{\lambda_i\}_{i=1}^K$ and eigenfunctions $\{e_i\}_{i=1}^K$ according to Mercer's theorem [Mercer, 1909]

$$k(x, y) = \sum_{i=1}^{K} \lambda_i e_i(x) e_i(y),$$

where $\{e_i\}_{i=1}^K$ are chosen to be orthonormal in $L^2(\Omega, \nu)$ for some sample space Ω with the measure ν and $K = \infty$ if k is non-degenerate. A popular approach to determine θ is to maximize the log marginal likelihood given a set of observations $\mathcal{D}$,

$$\widehat{\theta} = \operatorname*{argmax}_{\theta} \log p(\mathcal{D}|\theta) := \log \int p(\mathcal{D}|\boldsymbol{f}, \theta) p(\boldsymbol{f}|\theta) \, \mathrm{d}\boldsymbol{f}. \tag{2.5}$$

With the optimized parameters $\widehat{\theta}$, the posterior distribution of $\boldsymbol{f}$ is obtained as

$$p(\boldsymbol{f}|\mathcal{D}, \widehat{\theta}) = \frac{p(\mathcal{D}|\boldsymbol{f}, \widehat{\theta}) p(\boldsymbol{f}|\widehat{\theta})}{\int p(\mathcal{D}|\boldsymbol{f}, \widehat{\theta}) p(\boldsymbol{f}|\widehat{\theta}) \, \mathrm{d}\boldsymbol{f}}. \tag{2.6}$$

This exact GP model often suffers from computational and analytical intractability. Specifically, computing the exact posterior requires to store and invert an $(N \times N)$-matrix, which consumes $O(N^2)$ memory and $O(N^3)$ time. Both impede the method from scaling up to large problems. The analytical intractability originates from the non-conjugate Gaussian prior $p(\boldsymbol{f}|\theta)$ for the likelihood $p(\mathcal{D}|\boldsymbol{f}, \theta)$ and as a results, the integral in Equation (2.6) (as well as Equation (2.5)) has no closed-form expression. To overcome the two issues, the sparse GP model (Section 2.2.2) and approximate Bayesian inference (Section 2.2.3, 2.2.4, 2.2.5) are often employed.

2.2.2 Sparse Gaussian Processes

The sparse GP [Quiñonero-Candela and Rasmussen, 2005; Titsias, 2009] is proposed to reduced the computational burden. With the same assumption on function distributions as that of the GP, the sparse GP introduces inducing points to approximate distributions of function values at any point, which is realized based on the Bayes' rules for Gaussian variables. As a result, the computational time and memory are reduced to $O(NM^2)$ and $O(NM)$ respectively, where M is the number of inducing points.

More specifically, given a set of M inducing points $\boldsymbol{z} := \{z_1, \cdots, z_M\} \subset \Omega$ and corresponding function values $\boldsymbol{u} := \{f(z_i)\}_{i=1}^M$, $\boldsymbol{f}$ evaluated at $\boldsymbol{x}$ have a joint Gaussian distribution

$$\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{u} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_z \end{pmatrix}, \begin{pmatrix} \boldsymbol{K}_x & \boldsymbol{K}_{xz} \\ \boldsymbol{K}_{zx} & \boldsymbol{K}_z \end{pmatrix} \right)$$

where $\boldsymbol{K}_x$ and $\boldsymbol{K}_z$ are the covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{z}$ respectively, as defined in Equation (2.4), and $\boldsymbol{K}_{xz}$ and $\boldsymbol{K}_{zx}$ are cross-covariance matrices between $\boldsymbol{x}$ and $\boldsymbol{z}$, which are obtained as the evaluations of the covariance function on $\boldsymbol{x}$ and $\boldsymbol{z}$, namely,

$\boldsymbol{K}_{xz} = \{k(x_i, z_j)\}_{i,j=1}^{N,M}$ and $\boldsymbol{K}_{xz}^\top = \boldsymbol{K}_{zx}$. The conditional distribution of $\boldsymbol{f}$ given $\boldsymbol{u}$ can then be expressed as

$$p(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{K}_{xz}\boldsymbol{K}_z^{-1}(\boldsymbol{u} - \boldsymbol{\mu}_z), \boldsymbol{K}_x - \boldsymbol{K}_{xz}\boldsymbol{K}_z^{-1}\boldsymbol{K}_{zx}).$$

We omit the condition on θ from time to time for a neat presentation. The quantity of interest is the posterior distribution of $\boldsymbol{f}$, which can be calculated based on the posterior distribution of $\boldsymbol{u}$ as

$$p(\boldsymbol{f}|\mathcal{D}) = \int p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u}|\mathcal{D})\ \mathrm{d}\boldsymbol{u},$$

where $\mathcal{D}$ is the observation. The posterior distribution of $\boldsymbol{u}$ often has no closed-form expressions and the Sparse GP model employs a Gaussian distribution $q(\boldsymbol{u})$ for $\boldsymbol{u}$, which is an approximation to the posterior distribution of $\boldsymbol{u}$. Suppose that $q(\boldsymbol{u})$ has a Gaussian form

$$q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{m}, \boldsymbol{S}),$$

and then the optimal $\boldsymbol{m}$ and $\boldsymbol{S}$ of $q(\boldsymbol{u})$ can be obtained by using variational inference for the marginal likelihood (Equation (2.5)) [Titsias, 2009]. With the optimal $q(\boldsymbol{u}) \approx p(\boldsymbol{u}|\mathcal{D})$, we obtain a distribution over $\boldsymbol{f}$, which is Gaussian again and is the approximation of interest to the posterior distribution of $\boldsymbol{f}$,

$$\begin{aligned}
q(\boldsymbol{f}) &= \int p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})\ \mathrm{d}\boldsymbol{u} \\
&= \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{K}_{xz}\boldsymbol{K}_z^{-1}(\boldsymbol{m} - \boldsymbol{\mu}_z), \boldsymbol{K}_x - \boldsymbol{K}_{xz}\boldsymbol{K}_z^{-1}(\boldsymbol{K}_z - \boldsymbol{S})\boldsymbol{K}_z^{-1}\boldsymbol{K}_{zx}) \\
&\approx \int p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u}|\mathcal{D})\ \mathrm{d}\boldsymbol{u} \\
&= p(\boldsymbol{f}|\mathcal{D}).
\end{aligned}$$

2.2.3 Laplace Approximation

Laplace approximation [Rasmussen and Williams, 2005, section 3.4], [Bishop, 2006, section 4.4] is a widely-used method and it provides a Gaussian approximation to the analytical intractable posterior probability. Suppose that the analytically intractable probability has the following form,

$$p(\boldsymbol{f}) = Z^{-1}h(\boldsymbol{f}), \quad Z \coloneqq \int h(\boldsymbol{f})\ \mathrm{d}\boldsymbol{f},$$

where Z is the unknown normalization constant. First, the method finds a mode of $p(\boldsymbol{f})$, which is a point $\boldsymbol{f}_0$ satisfying

$$\nabla h(\boldsymbol{f})|_{\boldsymbol{f}=\boldsymbol{f}_0} = \boldsymbol{0}.$$

Then, it considers a second-order Taylor expansion of $\ln h(\boldsymbol{f})$ centred on the mode $\boldsymbol{f}_0$,

$$\ln h(\boldsymbol{f}) \approx \ln h(\boldsymbol{f}_0) - \frac{1}{2}(\boldsymbol{f}-\boldsymbol{f}_0)^\top A(\boldsymbol{f}-\boldsymbol{f}_0), \quad A \coloneqq \nabla^2 \ln h(\boldsymbol{f})\big|_{\boldsymbol{f}=\boldsymbol{f}_0},$$

where the first-order term is omitted as it is equal to zero. Taking the exponential and normalization on both sides of the above equation, we obtain a Gaussian distribution approximating $p(\boldsymbol{f})$,

$$p(\boldsymbol{f}) \approx \frac{|A|^{1/2}}{(2\pi)^{N/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{f}_0)^\top A(\boldsymbol{f}-\boldsymbol{f}_0)\right\} = \mathcal{N}(\boldsymbol{f}_0, A^{-1}).$$

2.2.4 Variational Inference

Variational inference [Bishop, 2006, Section 10.1] is a more general approach than Laplace approximation because it allows non-Gaussian approximation to intractable distributions. Consider the joint distribution of observations and variables, $p(\mathcal{D}, \boldsymbol{f}|\theta)$. The variational approach introduces a variational distribution $q(\boldsymbol{f}|\theta')$ to approximate the posterior distribution $p(\boldsymbol{f}|\mathcal{D},\theta)$ and optimizes $q(\boldsymbol{f}|\theta')$ by maximizing a lower bound of the log-likelihood, known as the evidence lower bound (ELBO), which can be derived from the non-negative gap perspective:

$$\begin{aligned}
\log p(\mathcal{D}|\theta) &= \log \frac{p(\mathcal{D},\boldsymbol{f}|\theta)}{q(\boldsymbol{f}|\theta')} - \log \frac{p(\boldsymbol{f}|\mathcal{D},\theta)}{q(\boldsymbol{f}|\theta')} \\
&= \underbrace{\underbrace{\mathbb{E}_{q(\boldsymbol{f}|\theta')}\left[\log p(\mathcal{D}|\boldsymbol{f},\theta)\right]}_{\text{reconstruction term}} - \underbrace{\mathrm{KL}(q(\boldsymbol{f}|\theta')||p(\boldsymbol{f}|\theta))}_{\text{regularization term}}}_{\equiv \mathrm{ELBO}(q(\boldsymbol{f}),p(\mathcal{D}|\boldsymbol{f}),p(\boldsymbol{f}))} + \underbrace{\mathrm{KL}(q(\boldsymbol{f}|\theta')||p(\boldsymbol{f}|\mathcal{D},\theta))}_{\text{intractable (non-negative) gap}} \\
&\geq \mathrm{ELBO}(q(\boldsymbol{f}), p(\mathcal{D}|\boldsymbol{f}), p(\boldsymbol{f})),
\end{aligned} \tag{2.7}$$

where we omit θ and θ' in conditions. For notational convenience, we will often omit conditioning on θ and θ' hereinafter. Optimizing the ELBO w.r.t. θ' balances between the reconstruction error and the Kullback-Leibler (KL) divergence from the prior. Generally, the conditional $p(\mathcal{D}|\boldsymbol{f})$ is known, so is the prior. Thus, for an appropriate choice of q, it is easier to work with this lower bound than with the intractable posterior $p(\boldsymbol{f}|\mathcal{D})$. We also see that, due to the form of the intractable gap, if q is from a distribution family containing elements close to the true unknown posterior, then q will be close to the true posterior when the ELBO is close to the true likelihood. An alternative derivation applies Jensen's inequality [Jordan et al., 1999]. A vanilla ELBO objective can usually be straight-forwardly minimized by standard algorithms, such as gradient descent. In Chapter 4, we will show that the ELBO of our Bayesian Hawkes process model has extra constraints on certain variables, so we accordingly develop a variational-inference based two-step iterative optimization algorithm.

2.2.5 Expectational Propagation

In this subsection, we introduce the application of the expectational propagation (EP) algorithm to the GP models [Opper and Winther, 2000; Minka, 2001b,c]. Given $\mathcal{D} = \{y_i\}_{i=1}^N$ and $f_i \coloneqq f(x_i)$, suppose that the GP model has the factorized likelihood,

$$p(\mathcal{D}|\boldsymbol{f}) = \prod_{i=1}^{N} p(y_i|f_i).$$

Numerous problems take this form: binary classification [Williams and Barber, 1998], single-output regression with Gaussian likelihood [Matheron, 1963], Student's-t likelihood [Jylänki et al., 2011] or Poisson likelihood [Zou, 2004], and the warped GP [Snelson et al., 2004]. EP deals with the analytical intractability by using Gaussian approximations to the individual non-Gaussian likelihoods, namely,

$$p(y_i|f_i) \approx t_i(f_i) \equiv \widetilde{Z}_i \mathcal{N}(f_i|\widetilde{\mu}_i, \widetilde{\sigma}_i^2).$$

The function t_i is often called the *site function* and is specified by the *site parameters*: the scale $\widetilde{Z}_i$, the mean $\widetilde{\mu}_i$ and the variance $\widetilde{\sigma}_i^2$. Notably, it is sufficient to use univariate site functions given that the local update can be efficiently computed using the marginal distribution only [Seeger, 2005]. We refer to this as the *locality property*. Although in this thesis we employ a more complex L_2 Wasserstein distance, our approach retains this property, as we elaborate in Chapter 5.

Given the site functions, one can approximate the intractable posterior distribution $p(\boldsymbol{f}|\mathcal{D})$ using a Gaussian $q(\boldsymbol{f})$ as below,

$$\begin{aligned} q(\boldsymbol{f}|\mathcal{D}) = q(\mathcal{D})^{-1} p(\boldsymbol{f}) \prod_{i=1}^{N} t_i(f_i) \equiv \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ \boldsymbol{\mu} = \boldsymbol{\Sigma}(\boldsymbol{K}_{\boldsymbol{x}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}} + \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\mu}}), \;\; \boldsymbol{\Sigma} = (\boldsymbol{K}_{\boldsymbol{x}}^{-1} + \widetilde{\boldsymbol{\Sigma}}^{-1})^{-1}, \end{aligned} \tag{2.8}$$

where $\widetilde{\boldsymbol{\mu}}$ is the vector of $\widetilde{\mu}_i$, $\widetilde{\boldsymbol{\Sigma}}$ is diagonal with $\widetilde{\boldsymbol{\Sigma}}_{ii} = \widetilde{\sigma}_i^2$; $\log q(\mathcal{D})$ is the log approximate model evidence expressed as below and further employed to optimize the GP hyperparameters:

$$\log\, q(\mathcal{D}) = \sum_{i=1}^{N} \log(\widetilde{Z}_i/\sqrt{2\pi}) - \frac{1}{2}\log|\boldsymbol{K}_{\boldsymbol{x}} + \widetilde{\boldsymbol{\Sigma}}| - \frac{1}{2}\widetilde{\boldsymbol{\mu}}^{\mathrm{T}}(\boldsymbol{K}_{\boldsymbol{x}} + \widetilde{\boldsymbol{\Sigma}})^{-1}\widetilde{\boldsymbol{\mu}}. \tag{2.9}$$

The core of EP is to optimize site functions $\{t_i(f_i)\}_{i=1}^N$. Ideally, one would seek to minimize the (global) KL divergence between the true and approximate posterior distributions $\mathrm{KL}(p(\boldsymbol{f}|\mathcal{D})\|q(\boldsymbol{f}))$, however this is intractable. Instead, EP is built based on the assumption that the global divergence can be approximated by the local divergence $\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|q(\boldsymbol{f}))$, where $\widetilde{q}(\boldsymbol{f}) \propto q^{\setminus i}(\boldsymbol{f})p(y_i|f_i)$ and $q^{\setminus i}(\boldsymbol{f}) \propto q(\boldsymbol{f})/t_i(f_i)$ are referred to as the tilted and cavity distributions, respectively. Note that the cavity distribution is Gaussian while the tilted distribution is usually not. The local divergence can be simplified from multi-dimensional to univariate, $\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|q(\boldsymbol{f})) = \mathrm{KL}(\widetilde{q}(f_i)\|q(f_i))$

(detailed in Appendix C.7), and $t_i(f_i)$ is optimized by minimizing it.

The minimization is realized by projecting the tilted distribution $\widetilde{q}(f_i)$ onto the Gaussian family, with the projected Gaussian denoted as

$$\mathrm{proj}_{\mathrm{KL}}(\widetilde{q}(f_i)) := \underset{\mathcal{N}}{\operatorname{argmin}} \mathrm{KL}(\widetilde{q}(f_i) \| \mathcal{N}(f_i)).$$

Then the projected Gaussian is used to update $t_i(f_i) \propto \mathrm{proj}_{\mathrm{KL}}(\widetilde{q}(f_i))/q^{\setminus i}(f_i)$. The mean and the variance of $\mathrm{proj}_{\mathrm{KL}}(\widetilde{q}(f_i)) \equiv \mathcal{N}(\mu^\star, \sigma^{\star 2})$ match the moments of $\widetilde{q}(f_i)$ and are used to update $t_i(f_i)$'s parameters:

$$\mu^\star = \mu_{\widetilde{q}_i}, \quad \sigma^{\star 2} = \sigma^2_{\widetilde{q}_i}, \tag{2.10}$$

$$\widetilde{\mu}_i = \widetilde{\sigma}_i^2 \left(\mu^\star (\sigma^\star)^{-2} - \mu_{\setminus i} \sigma_{\setminus i}^{-2} \right), \ \ \widetilde{\sigma}_i^{-2} = (\sigma^\star)^{-2} - \sigma_{\setminus i}^{-2}, \tag{2.11}$$

where $\mu_{\widetilde{q}_i}$ and $\sigma^2_{\widetilde{q}_i}$ are the mean and the variance of $\widetilde{q}(f_i)$, and $\mu_{\setminus i}$ and $\sigma^2_{\setminus i}$ are the mean and the variance of $q^{\setminus i}(f_i)$. We refer to the projection as the local update. Note that $\widetilde{Z}$ does not impact the optimization of $q(\boldsymbol{f})$ or the GP hyper-parameters $\boldsymbol{\theta}$, so we omit the update formula for $\widetilde{Z}$. We summarize EP in Algorithm 2 (Appendix).

2.3 Wasserstein Distance

We denote by $\mathcal{M}_+^1(\Omega)$ the set of all probability measures on Ω. We consider probability measures on the d-dimensional real space $\mathbb{R}^d$. The Wasserstein distance between two probability distributions $\xi, \nu \in \mathcal{M}_+^1(\mathbb{R}^d)$ can be intuitively defined as the cost of transporting the probability mass from one distribution to the other. We are particularly interested in the subclass of L_p Wasserstein distance, formally defined as follows.

Definition 1 (L_p Wasserstein distance)**.** *Consider the set of all probability measures on the product space $\mathbb{R}^d \times \mathbb{R}^d$, whose marginal measures are ξ and ν respectively, denoted as $U(\xi, \nu)$. The L_p Wasserstein distance between ξ and ν is defined as*

$$W_p^p(\xi, \nu) := \inf_{\pi \in U(\xi,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{z}\|_p^p \, d\pi(\boldsymbol{x}, \boldsymbol{z}),$$

where $p \in [1, \infty)$ and $\| \cdot \|_p$ is the L_p norm.

Like the KL divergence, the L_p Wasserstein distance it has a minimum of zero, achieved when the distributions are equivalent. *Unlike the KL*, however, it is a proper distance metric, and thereby satisfies the triangle inequality, and has the appealing property of symmetry.

A less fundamental property of the Wasserstein distance which we exploit for computational efficiency is:

Proposition 1. *[Peyré et al., 2019, Remark 2.30] The L_p Wasserstein distance between 1-d distribution functions ξ and $\nu \in \mathcal{M}_+^1(\mathbb{R})$ equals the L_p distance between the quantile*

functions,

$$W_p^p(\xi, \nu) = \int_0^1 \left| F_\xi^{-1}(y) - F_\nu^{-1}(y) \right|^p \, dy,$$

where $F_z : \mathbb{R} \to [0, 1]$ *is the cumulative distribution function (CDF) of* z*, defined as* $F_z(x) = \int_{-\infty}^{x} dz$*, and* F_z^{-1} *is the pseudo-inverse or quantile function, defined as* $F_z^{-1}(y) = \min_x\{x \in \mathbb{R} \cup \{-\infty\} : F_z(x) \geq y\}$.

Finally, the following translation property of the L_2 Wasserstein distance is central to our proof of locality in Chapter 5:

Proposition 2. *[Peyré et al., 2019, Remark 2.19] Consider the* L_2 *Wasserstein distance defined for* ξ *and* $\nu \in \mathcal{M}_+^1(\mathbb{R}^d)$*, and let* $f_{\boldsymbol{\tau}}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{\tau}$*,* $\boldsymbol{\tau} \in \mathbb{R}^d$*, be a translation operator. If* $\xi_{\boldsymbol{\tau}}$ *and* $\nu_{\boldsymbol{\tau}'}$ *denote the probability measures of translated random variables* $f_{\boldsymbol{\tau}}(\boldsymbol{x})$*,* $\boldsymbol{x} \sim \xi$*, and* $f_{\boldsymbol{\tau}'}(\boldsymbol{x})$*,* $\boldsymbol{x} \sim \nu$*, respectively, then*

$$W_2^2(\xi_{\boldsymbol{\tau}}, \nu_{\boldsymbol{\tau}'}) = W_2^2(\xi, \nu) - 2(\boldsymbol{\tau} - \boldsymbol{\tau}')^T(\boldsymbol{m}_\xi - \boldsymbol{m}_\nu) + \|\boldsymbol{\tau} - \boldsymbol{\tau}'\|_2^2,$$

where $\boldsymbol{m}_\xi$ *and* $\boldsymbol{m}_\nu$ *are means of* ξ *and* ν *respectively. In particular when* $\boldsymbol{\tau} = \boldsymbol{m}_\xi$ *and* $\boldsymbol{\tau}' = \boldsymbol{m}_\nu$*,* $\xi_{\boldsymbol{\tau}}$ *and* $\nu_{\boldsymbol{\tau}'}$ *become zero-mean measures, and*

$$W_2^2(\xi_{\boldsymbol{\tau}}, \nu_{\boldsymbol{\tau}'}) = W_2^2(\xi, \nu) - \|\boldsymbol{m}_\xi - \boldsymbol{m}_\nu\|_2^2.$$

2.4 Conditional Moment Restriction

Let (X, Z) be a random variable taking values in $\mathcal{X} \times \mathcal{Z}$ and Θ a parameter space. A conditional moment restriction (CMR) [Newey, 1993; Ai and Chen, 2003] can then be expressed as

$$\text{CMR}(\theta_0) = \mathbb{E}[\varphi_{\theta_0}(X) \mid Z] = \mathbf{0}, \quad P_Z - \text{almost surely (a.s.)} \tag{2.12}$$

for the true parameter $\theta_0 \in \Theta$. The function $\varphi_\theta(X)$ is a problem-dependent generalized residual function in $\mathbb{R}^q$ parameterized by θ. Intuitively, the CMR asserts that, for correctly specified models, the conditional mean of the generalized residual function is almost surely equal to zero. Many statistical models can be written as Equation (2.12) including nonparametric regression models where $X = (\tilde{X}, Y), Z = \tilde{X}$ and $\varphi_\theta(X) = Y - f(\tilde{X}; \theta)$; conditional quantile models where $X = (\tilde{X}, Y), Z = \tilde{X}$, and $\varphi_\theta(X) = \mathbb{1}\{Y < f(\tilde{X}; \theta)\} - \tau$ for the target quantile $\tau \in [0, 1]$; IV regression models where $X = (\tilde{X}, Y)$, Z is an IV, and $\varphi_\theta(X) = Y - f(\tilde{X}; \theta)$; the stochastic point process, where $X = (\tilde{X})$, Z is the observed region in the domain, and $\varphi_\theta(X) = \nabla_\theta \log p_\theta(X)$ with p_θ the probability model of the point process. More specifically in the point process case, the maximum likelihood estimator is written in general as below

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, \mathbb{E}_{X \sim p}[\log p_\theta(X)|Z] \implies \mathbb{E}_{X \sim p}[\nabla_\theta \log p_\theta(X)|Z]\big|_{\theta=\hat{\theta}} = \mathbf{0},$$

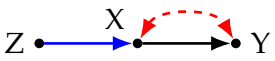


Figure 2.3: A causal graph depicting an instrumental variable Z that satisfies an exclusion restriction and unconfoundedness (there may be a confounder ε acting on X and Y, but it is independent of Z).

where p is the true probability of data, and as we note, $\hat{\theta}$ is often estimated as the solution to the first-order condition shown in the right part. Besides, the finite region Z where real-world data are collected is random and dependent on the time when we observe the data, so it is reasonable to view the point process as a CMR model.

2.4.1 Instrumental Variable Regression

We introduce details of instrumental variable (IV) regression on which we test our kernel maximum moment restriction estimation method. Following standard setting in the literature [Hartford et al., 2017; Lewis and Syrgkanis, 2018; Bennett et al., 2019; Singh et al., 2019; Muandet et al., 2020b], let X be a treatment (endogenous) variable taking value in $\mathcal{X} \subseteq \mathbb{R}^d$ and Y a real-valued outcome variable. The goal of IV regression is to estimate a function $f : \mathcal{X} \to \mathbb{R}$ from a structural equation model (SEM) of the form

$$Y = f(X) + \varepsilon, \quad X = t(Z) + g(\varepsilon) + \nu, \tag{2.13}$$

where we assume that $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\nu] = 0$. Unfortunately, as we can see from Equation (2.13), ε is correlated with the treatment X, i.e., $\mathbb{E}[\varepsilon|X] \neq 0$, and hence standard regression methods cannot be used to estimate f. This setting arises, for example, when there exist unobserved confounders (i.e., common causes) between X and Y.

To illustrate the problem, let us consider an example taken from Hartford et al. [2017] which aims at predicting sales of airline ticket Y under an intervention in price of the ticket X. However, there exist unobserved variables that may affect both sales and ticket price, e.g., conferences, COVID-19 pandemic, etc. This creates a correlation between ε and X in (2.13) that prevents us from applying the standard regression toolboxes directly on observational data.

Instrumental variable (IV). To address this problem, we assume access to an *instrumental* variable Z taking value in $\mathcal{Z} \subseteq \mathbb{R}^{d'}$. As we can see in (2.13), the instrument Z is associated with the treatments X, but not with the outcome Y, other than through its effect on the treatments. Formally, Z must satisfy (i) *Relevance:* Z has a causal influence on X, (ii) *Exclusion restriction:* Z affects Y only through X, i.e., $Y \perp\!\!\!\perp Z|X, \varepsilon$, and (iii) *Unconfounded instrument(s):* Z is independent of the error, i.e., $\varepsilon \perp\!\!\!\perp Z$. For example, the instrument Z may be the cost of fuel, which influences sales only via price. Intuitively, Z acts as a "natural experiment" that induces variation in X; see Figure 2.3.

2.5 Related Works of Hawkes Process Estimation

As mentioned in Chapter 1, the triggering kernel ϕ is important as it is shared and decides the class of the whole process. In this section, we review the estimation of the triggering kernel ϕ as well as topics related to it, including modeling of the intensity function and of the integral of the triggering/intensity function, and model parameter estimation methods and so on.

2.5.1 Parametric Frequentist Solutions

The Hawkes processes with the parametric triggering kernels are applied in a wide range of areas and the popularity can be explained by their simplicity and flexibility compared with other models. Mishra et al. [2016] employ the branching factor of the Hawkes process with the power-law kernel to predict popularity of tweets; Kurashima et al. [2018] predict human actions using a Hawkes process equipped with exponential, Weibull and Gaussian mixture kernels; online popularity unpredictability is explained using the Hawkes process with a variant of the exponential kernel by Rizoiu et al. [2018]; Xu et al. [2016] employ a sum of Gaussian kernels to discover the Granger causality for the Hawkes process. The model parameters can be easily determined by the maximum likelihood estimation [Ozaki, 1979]. However, most works employing Hawkes processes with parametric triggering kernels encode strong assumptions, and limit the expressivity of the models. Therefore, recent works design practical approaches to learn flexible representations of the optimal triggering kernel from data, as the following subsections.

2.5.2 Non-parametric Frequentist Solutions

A popular direction on learning the flexible triggering kernel functions is the non-parametric frequentist estimation. Basically, the triggering kernel function is assigned a non-parametric form which is mainly a piecewise function [Lewis and Mohler, 2011; Zhou et al., 2013; Bacry and Muzy, 2016; Eichler et al., 2017]. The form of the function is caused by discretization of the function domain and leads to poor scaling with the domain dimension and sensitivity to the choice of discretization. Moreover, they do not quantify the uncertainty of the learned triggering kernels, so are sensitive to randomness of finite data. In contrast, our methods require no discretization and are Bayesian, so has an advantage of scalability with the dimension of the domain and robustness to uncertainty in the finite samples.

Frequentist Estimation of Intensities. Learning the triggering kernels based on the maximum likelihood requires $O(N^2)$ computational time and to circumvent this problem, many frequentist methods focus on estimating the intensity function, which needs $O(N)$ time. These methods rely on modern machine learning models, such as the neural network [Du et al., 2016; Xiao et al., 2017a,b; Mei and Eisner, 2017; Omi et al., 2019; Zhang et al., 2020; Zuo et al., 2020] and the reproducing kernel Hilbert space [Flaxman et al., 2017] (RKHS). As no assumption is made on the type of interaction between points such as self-excitement, they allow the modelling of flexible point

processes. Early neural network approaches on modeling intensity functions rely on the recurrent neural network (RNN) or long short-term memory (LSTM) network [Du et al., 2016; Xiao et al., 2017b; Mei and Eisner, 2017], and estimate model parameters based on maximum likelihood. Training deep RNNs and LSTMs is notoriously difficult because of gradient explosion and gradient vanishing [Pascanu et al., 2013]. The inherently sequential nature of RNN and LSTM renders the impossibility to process all the events in parallel and limits the methods' ability to scale to large datasets. More recent works deal with this issue by employing the transformer or the self-attention mechanism [Zhang et al., 2020; Zuo et al., 2020]. Apart from the issues with models, most of the neural network approaches suffer from no analytical expression for the integral of the intensity function in the likelihood. Hence, the integral has to be approximated numerically, which leads to in-accurate parameter estimation. In order to circumvent this problem, Omi et al. [2019] directly model the integral by a RNN. As a result, the intensity function is obtained by differentiation of the integral, which is easier to compute than integration. Shchur et al. [2020b] present a new modeling method for the integral by exploiting the normalizing flows which later is employed to model the distribution of the inter-arrival time [Shchur et al., 2020a]. As shown by Walder and Bishop [2017], the integral can have a closed-form expression with kernel methods by explicitly constructing the eigen-components of the kernel function. Orthogonal to the afore-mentioned development, there is recent interest in parameter estimation methods beyond the popular maximum likelihood estimation, including the Wiener-hopf equation based method [Bacry and Muzy, 2016], the adversarial loss such as the wasserstein distance based loss [Xiao et al., 2017a], the least square loss [Eichler et al., 2017], the cumulants-based method [Achab et al., 2017], the reinforcement learning based method [Li et al., 2018]. Interestingly, it is unnecessary to estimate the triggering kernel or the intensity function in some applications, such as discovering the Granger causality, where it is sufficient to only estimate the integral of the triggering kernel [Achab et al., 2017]. We recommend the related work sections of these methods for more development on point processes. Different from works reviewed here, this thesis focuses on the non-parametric Bayesian estimation of the triggering kernel.

2.5.3 Bayesian Parametric and Non-parametric Solutions

The Bayesian non-parametric estimation for the Hawkes process has been studied, including the work of Rasmussen [2013]; Linderman and Adams [2014]; Linderman and Adams [2015]. These work require either constructing a parametric triggering kernel [Rasmussen, 2013; Linderman and Adams, 2014] or discretizing the input domain to scale with the data size [Linderman and Adams, 2015]. The shortcoming of discretization is just mentioned and to overcome it, Donnet et al. [2018] propose a continuous non-parametric Bayesian Hawkes process and resort to an unscalable Markov chain Monte Carlo (MCMC) estimator to the posterior distribution. We comparatively summarize a part of related works in 2.1. Compared with these works, our methods are Bayesian and non-parametric without requiring discretization of

Table 2.1: Non-parametric triggering kernel estimation.

Methods	Time Complexity	Bayesian	Continuous	Non-parametric
Zhou et al. [2013]	$O(n^3)$	×	✓	×
Xu et al. [2016]	$O(n^3)$	×	✓	×
Lewis and Mohler [2011]	$O(n^3)$	×	×	✓
Zhou et al. [2013]	$O(n^3)$	×	×	✓
Rasmussen [2013]	$O(n)$	✓	✓	×
Linderman and Adams [2015]	$O(n)$	✓	interval-censored	×
Donnet et al. [2018]	unspecified	✓	✓	✓
Ours	$O(n)$	✓	✓	✓

domains and has an advantage of a linear time complexity allowing it to be applied to large-scale real-world datasets.

2.6 Related Works of Expectation Propagation and Wasserstein Distance

The basis of the EP algorithm for GP models was first proposed by Opper and Winther [2000] and then generalized by Minka [2001b,c]. Power EP [Minka, 2004, 2005] is an extension of EP that exploits the more general α-divergence (with $\alpha = 1$ corresponding to the forward KL divergence in EP) and has been recently used in conjunction with GP pseudo-input approximations [Bui et al., 2017]. Although generally not guaranteed to converge locally or globally, Power EP uses fixed-point iterations for its local updates and has been shown to perform well in practice for GP regression and classification [Bui et al., 2017]. In comparison, our approach (quantile propagation) uses the L_2 Wasserstein distance, and like EP, it yields convex local optimizations for GP models with factorized likelihoods. This convexity benefits the convergence of the local update, and is retained even with the general L_p ($p \geq 1$) Wasserstein distance as shown in Theorem 2 (Chapter 5). Moreover, for the same class of GP models, both EP and our approach have the locality property [Seeger, 2005] and can be unified in the generic message passing framework [Minka, 2005].

Without the guarantee of convergence for the explicit global objective function, understanding EP has proven to be a challenging task. As a result, a number of works have instead attempted to directly minimize divergences between the true and approximate joint posteriors, for divergences such as the KL [Jordan et al., 1999; Dezfouli and Bonilla, 2015], Rényi [Li and Turner, 2016], α [Hernández-Lobato et al., 2016] and optimal transport divergences [Ambrogioni et al., 2018]. To deal with the infinity issue of the KL (and more generally the Rényi and α divergences) which arises from different distribution supports [Montavon et al., 2016; Arjovsky et al., 2017; Gulrajani et al., 2017], Hensman et al. [2014] employ the product of tilted distributions

as an approximation. A number of variants of EP have also been proposed, including the convergent double loop algorithm [Opper and Winther, 2005], parallel EP [Minka, 2001a], distributed EP built on partitioned datasets [Xu et al., 2014; Gelman et al., 2017], averaged EP assuming that all approximate likelihoods contribute similarly [Dehaene and Barthelmé, 2018], and stochastic EP which may be regarded as sequential averaged EP [Li et al., 2015].

The L_2 Wasserstein distance between two Gaussian distributions has a closed form expression [Dowson and Landau, 1982]. Detailed research on the Wasserstein geometry of the Gaussian distribution is conducted by Takatsu [2011]. Recently, this closed form expression has been applied to robust Kalman filtering [Shafieezadeh-Abadeh et al., 2018] and to the analysis of populations of GPs [Mallasto and Feragen, 2017]. A more general extension to elliptically contoured distributions is provided by Gelbrich [1990] and used to compute probabilistic word embeddings [Muzellec and Cuturi, 2018]. A geometric interpretation for the L_2 Wasserstein distance between any distributions [Benamou and Brenier, 2000] has already been exploited to develop approximate Bayesian inference schemes [El Moselhy and Marzouk, 2012]. Our approach is based on the L_2 Wasserstein distance but does not exploit these closed form expressions; instead we obtain computational efficiency by leveraging the EP framework and using the quantile function form of the L_2 Wasserstein distance for univariate distributions. We believe our work paves the way for further practical approaches to Wasserstein-distance-based Bayesian inference.

2.7 Related Works of Conditional Moment Restriction

We review conditional moment restriction (CMR) models [Newey, 1993; Ai and Chen, 2003; Dikkala et al., 2020] in this section and the models have a wide range of applications in causal inference, economics, and finance modeling, where for correctly-specified models the conditional mean of certain functions of data equals zero almost surely. This kind of models also appear in Mendelian randomization, a technique in genetic epidemiology that uses genetic variation to improve causal inference of a modifiable exposure on disease [Davey Smith and Ebrahim, 2003; Burgess et al., 2017a]. Rational expectation models [Muth, 1961], widely-used in macroeconomics, measures how available information is exploited to form future expectations by decision-makers as conditional moments [Muth, 1961]. Furthermore, CMRs have also gained popularity in the community of causal machine learning, leading to novel algorithms such as generalized random forests [Athey et al., 2019], double/debiased machine learning [Chernozhukov et al., 2018] and nonparametric IV regression [Bennett et al., 2019; Muandet et al., 2020b]; see also related works therein, as well as in offline reinforcement learning [Liao et al., 2021].

This thesis focuses on CMR based estimation with the application to IV regression (introduced in Section 2.4). Therefore, we mainly review literature on the CMR based estimation for IV regression.

Classical methods for IV regression often rely on a linearity assumption in which

a two-stage least squares (2SLS) is the most popular technique [Angrist et al., 1996]. The generalized method of moments (GMM) of Hansen [1982], which imposes the orthogonality restrictions, can also be used for the linear IV regression. For nonlinear regression, numerous methodologies have been developed in the field of nonparametric IV [Newey and Powell, 2003; Hall and Horowitz, 2005; Blundell et al., 2007; Horowitz, 2011] and recent machine learning [Hartford et al., 2017; Lewis and Syrgkanis, 2018; Bennett et al., 2019; Muandet et al., 2020b; Singh et al., 2019]. However, these estimators have complicated structures in general. The nonparametric IV is an ill-posed problem or requires estimating a conditional density. The machine learning methods needs two-stage estimation or minimax optimization. As a result, it is not easy to obtain an asymptotic distribution of estimation errors and to simply apply learning algorithms such as the stochastic gradient descent (SGD), e.g., see Daskalakis and Panageas [2018] for the limitation of SGD with minimax problems.

Several extensions of 2SLS and GMM exist for the nonlinear IV problem. In the two-stage approach, the function $f(x)$ has often been obtained by solving a Fredholm integral equation of the first kind $\mathbb{E}[Y|Z] = \int f(x)\,\mathrm{d}P(X|Z)$. In Newey and Powell [2003]; Blundell et al. [2007]; Horowitz [2011]; Chen and Pouzo [2012], linear regression is replaced by a linear projection onto a set of known basis functions. A uniform convergence rate of this approach is provided in Chen and Christensen [2018]. However, it remains an open question how to best choose the set of basis functions. In Hall and Horowitz [2005] and Darolles et al. [2011], the first-stage regression is replaced by a conditional density estimation of $\mathbb{P}(X|Z)$ using a kernel density estimator. Estimating conditional densities is a difficult task and is known to perform poorly in high dimension [Tsybakov, 2008].

The IV regression has also recently received attention in the machine learning community. Hartford et al. [2017] proposed to solve the integral equation by first estimating $P(X|Z)$ with a mixture of deep generative models on which the function $f(x)$ can be learned with another deep NNs. Instead of NNs, Singh et al. [2019] proposed to model the first-stage regression using the conditional mean embedding of $P(X|Z)$ [Song et al., 2009, 2013; Muandet et al., 2017] which is then used in the second-step kernel ridge regression. In other words, the first-stage estimation in Singh et al. [2019] becomes a vector-valued regression problem. The critical drawback of these algorithms is that they involve the intermediate first-stage regression which may not be of our primary interest. In an attempt to alleviate this drawback, Muandet et al. [2020b] and Liao et al. [2020] reformulate the two-stage procedure as a convex-concave saddle-point problem. The DualIV [Muandet et al., 2020b] has a quadratic objective function similar to ours, but its RKHS is applied on (Z, Y) which cannot be interpreted as a valid instrument. In contrast, the approach of Liao et al. [2020, Appendix F] is highly related to GMM and provides a dual reformulation of our method by using a RKHS for the inner maximization. Besides, the starting points of these works differ from ours. In both DualIV and Liao et al. [2020], they started from the population risk functional, whereas we start from the CMR. The fact that they arrive at the same objective highlights a deeper connection which requires further investigation.

Our work follows in spirit many GMM-based approaches for IV regression, namely,

Lewis and Syrgkanis [2018]; Bennett et al. [2019]; Muandet et al. [2020a]. We apply the MMR framework of Muandet et al. [2020a] to the IV regression problem, whereas Muandet et al. [2020a] only considers a conditional moment (CM) testing problem. In fact, this framework was initially inspired by Lewis and Syrgkanis [2018] and Bennett et al. [2019] which instead parametrize the instruments by deep NNs and Muandet et al. [2020b] which proposes the dual formulation of the two-stage procedure. By combining the GMM framework with RKHS functions, our objective function can be evaluated in closed-form. As a result, our IV estimate can be obtained by minimizing the empirical risk, as opposed to an adversarial optimization used in Lewis and Syrgkanis [2018] and Bennett et al. [2019]. Furthermore, unlike existing two-stage procedures [Angrist et al., 1996; Hartford et al., 2017; Singh et al., 2019], our algorithm does not require the first-stage regression. It is important to note that, concurrent to our work, Dikkala et al. [2020] extends the work of Lewis and Syrgkanis [2018] to an algorithm similar to our MMR-IV (RKHS) [Dikkala et al., 2020, Section 4]. Although both work employs RKHSs in the minimax frameworks, Dikkala et al. [2020] incorporate a Tikhonov regularization on h in (6.2) and resort to the representer theorem [Schölkopf et al., 2001a] to develop the analytical objective function, whereas we impose a unit-ball constraint which is a form of Ivanov regularization [Ivanov et al., 2002] and enables not to rely on such a theorem.

Beyond the IV regression, there are numerous prior studies that are related to ours, especially in policy evaluation, reinforcement learning, and causal inference. We leave the review of them to Appendix D.4.

2.8 Summary

In this chapter, we first review different technical prerequisites, including the Poisson and Hawkes processes, the Gaussian process model, the approximate Bayesian inference approaches, the Wasserstein distance and the conditional moment restriction respectively. We then review the related development of these techniques. The prerequisite are essential for the methods proposed in the following four chapters. In the next chapter, we present an efficient non-parametric Bayesian inference framework for estimation of the Hawkes process. The approach is developed based on the branching structure of the Hawkes process and exploit Gibbs sampling and Laplace approximation.

Chapter 3

Gibbs Sampling and Laplace Approximation Based Efficient Inference

In this chapter, we present a general framework for the efficient non-parametric Bayesian inference of Hawkes processes. The contents are categorized as:

(i) In Section 3.1, we first review a prerequisite of our work, the Laplace Bayesian Poisson process.

(ii) In Section 3.2, we propose an efficient non-parametric Bayesian framework for the Hawkes process by combining Gibbs sampling and the Laplace Bayesian Poisson process. We exploit block Gibbs sampling [Ishwaran and James, 2001] to iteratively sample the latent branching structure, the background intensity μ and the triggering kernel ϕ from their posterior distributions. In each iteration, the point data are decomposed as a cluster of Poisson processes based on the sampled branching structure and the posterior distributions of μ and ϕ are estimated using the resulting cluster processes. Our framework is close to the stochastic Expectation-Maximization (EM) algorithm [Celeux and Diebolt, 1985] where posterior μ and ϕ are estimated [Lloyd et al., 2015; Walder and Bishop, 2017] in the M-step and random samples of μ and ϕ are drawn. We adapt the approach of the recent non-parametric Bayesian estimation for Poisson process intensities, termed Laplace Bayesian Poisson process (LBPP) [Walder and Bishop, 2017], to estimate the posterior ϕ given the sampled branching structure. Especially in Section 3.2.4, we utilize the finite support assumption of the Hawkes Process to speed up sampling and computing the probability of the branching structure. We theoretically show our method to be of linear time complexity.

(iii) In Section 3.3, we furthermore explore the connection with the EM algorithm [Dempster et al., 1977] and develop a second variant of our method, as an approximate EM algorithm.

(iv) In Section 3.4, we empirically show our method enjoys linear time complexity and can infer known analytical kernels, i.e., exponential and sinusoidal kernels. On two large-scale social media datasets, our method outperforms the current

state-of-the-art non-parametric approaches and the learned kernels reflect the perceived longevity for different content types.

3.1 Laplace Bayesian Poisson Process

In this section, we introduce a prerequisite of our work, the Laplace Bayesian Poisson process (LBPP) [Walder and Bishop, 2017]. LBPP has been proposed for the non-parametric Bayesian estimation of the intensity of a Poisson process. To satisfy non-negativity of the intensity function, LBPP models the intensity function λ as a permanental process [Shirai and Takahashi, 2003], i.e., $\lambda = g \circ f$ where the link function $g(z) = z^2/2$ and $f(\cdot)$ obeys a Gaussian process (GP) prior. Alternative link functions include $\exp(\cdot)$ [Møller et al., 1998; Diggle et al., 2013] and $g(z) = \lambda^*(1+\exp(-z))^{-1}$ [Adams et al., 2009] where λ^* is constant.

The choice $g(z) = z^2/2$ has the analytical advantages; for some covariances the log-likelihood can be computed in closed form [Lloyd et al., 2015; Flaxman et al., 2017]. LBPP exploits the Mercer expansion [Mercer, 1909] of the GP covariance function $k(x,y) \equiv \mathrm{Cov}(f(x), f(y))$, namely,

$$k(x,y) = \sum_{i=1}^{K} \lambda_i e_i(x) e_i(y), \tag{3.1}$$

where for non-degenerate kernels, $K = \infty$. The eigenfunctions $\{e_i(\cdot)\}_{i=1}^K$ are chosen to be orthonormal in $L^2(\Omega, \mathcal{M})$ for some sample space Ω with the measure $\mathcal{M}$. $f(\cdot)$ can be represented as a linear combination of $e_i(\cdot)$ [Rasmussen and Williams, 2005, section 2.2] as below

$$f(\cdot) = \boldsymbol{\omega}^T \boldsymbol{e}(\cdot), \quad \boldsymbol{\omega} \sim \mathcal{N}(0, \Lambda), \tag{3.2}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_K)$ is a diagonal covariance matrix and $\boldsymbol{e}(\cdot) = \{e_1(\cdot)\}_{i=1}^K$ is a (column) vector of basis functions. Computing the posterior distribution of the intensity function λ is equivalent to estimating the posterior distribution of $\boldsymbol{\omega}$ which, in LBPP, is approximated by a normal distribution (as known as Laplace approximation introduced in Section 2.2.3). That is

$$\log p(\boldsymbol{\omega}|\boldsymbol{x}, \Omega, k) \approx \log \mathcal{N}(\boldsymbol{\omega}|\hat{\boldsymbol{\omega}}, Q),$$

where $\boldsymbol{x} := \{x_i\}_{i=1}^N$ is a set of point data, Ω the sample space and k the Gaussian process kernel function. $\hat{\omega}$ is selected as the mode of the true posterior and Q the negative inverse Hessian of the true posterior at $\hat{w}$:

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\mathrm{argmax}} \log p(\boldsymbol{\omega}|\boldsymbol{x}, \Omega, k), \tag{3.3}$$

$$Q^{-1} = -\partial_{\boldsymbol{\omega}\boldsymbol{\omega}^T} \log p(\boldsymbol{\omega}|\boldsymbol{x}, \Omega, k)|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}}\,. \tag{3.4}$$

The approximate posterior distribution of $f(x)$ is expressed as the following normal

distribution [Rasmussen and Williams, 2005, section 2.2]:

$$f(x) \sim \mathcal{N}(\hat{\boldsymbol{\omega}}^T \boldsymbol{e}(x), \boldsymbol{e}(x)^T Q \boldsymbol{e}(x)) \equiv \mathcal{N}(\nu, \sigma^2). \tag{3.5}$$

Furthermore, the posterior distribution of $\lambda(x) = f(x)^2/2$ is a Gamma distribution:

$$\text{Gamma}(t|\alpha, \beta) \coloneqq \beta^\alpha t^{\alpha-1} e^{-\beta t} / \Gamma(\alpha), \tag{3.6}$$

where $\alpha = (\nu^2 + \sigma^2)^2 / (4\nu^2\sigma^2 + 2\sigma^4)$ and $\beta = (\nu^2 + \sigma^2) / (2\nu^2\sigma^2 + \sigma^4)$.

3.2 Inference via Sampling

We now detail our efficient non-parametric Bayesian estimation algorithm for Hawkes processes, which employs block Gibbs sampling to iteratively draw samples from the posterior distributions of μ (constant background intensity) and $\phi(\cdot)$ (triggering kernel). Our method starts with random μ_0 and $\phi_0(\cdot)$, and iterates by cycling through the following four steps (k is the iteration index):

(i) Calculate $p(B|\boldsymbol{x}, \phi_{k-1}, \mu_{k-1})$, the distribution of the branching structure B given the data $\boldsymbol{x}$, triggering kernel ϕ_{k-1}, and background intensity μ_{k-1} (see details in Section 3.2.1).
(ii) Sample a branching structure B_k from $p(B|\boldsymbol{x}, \phi_{k-1}, \mu_{k-1})$ (Section 3.2.1).
(iii) Estimate $p(\phi|B_k, \boldsymbol{x})$ (Section 3.2.3) and $p(\mu|B_k, \boldsymbol{x})$ (Section 3.2.2).
(iv) Sample a sample ϕ_k and μ_k from $p(\phi|B_k, \boldsymbol{x})$ and $p(\mu|B_k, \boldsymbol{x})$, respectively.

By standard Gibbs sampling arguments, the samples of ϕ and μ drawn in the step (iv) converge to the desired posterior, modulo the Laplace approximation for estimation of $p(\phi|B_k, \boldsymbol{x})$ in (iii). As the method is based on block Gibbs sampling [Ishwaran and James, 2001], we term it *Gibbs-Hawkes* in this chapter.

3.2.1 Distribution and Sampling of the Branching Structure

The branching structure B has a data structure of tree (as Figure 1.1(b)) and consists of independent triggering events. Therefore, we may sample a branching structure by sampling a parent for each x_i independently, where sampling exploits the probabilities of triggering events for x_i, namely, $\{p_{ij}\}_{j=0}^{i-1}$ as defined in Equation (2.2). The sampled branching structure separates a set of points into immigrants and offspring (introduced in Chapter 1). Immigrants can be regarded as a sequence generated from PP(μ), where PP(μ) is a Poisson process owning an intensity μ, and can be used to estimate the posterior distribution of μ.

The key property which we exploit in the subsequent Section 3.2.2 and Section 3.2.3 is the following. Denote by $\{x_k^{(i)}\}_{k=1}^{N_{x_i}}$, the N_{x_i} offspring is generated by point x_i. If such a sequence is *aligned* to an origin at x_i, yielding $S_{x_i} \coloneqq \{t_k^{(i)} - x_i\}_{k=1}^{N_i}$, then the aligned sequence is drawn from PP(ϕ) over [0, T-x_i] where $[0, T]$ is the sample domain of the Hawkes process. The posterior distribution of ϕ is estimated on all such aligned sequences.

3.2.2 Posterior Distribution of μ

Continuing from the observations in Section 3.2.1, note that if we are given a set of points $\{x_i\}_{i=1}^M$ ($M \leq N$) generated by PP(μ) over $\Omega = [0, T]$, the likelihood for $\{x_i\}_{i=1}^M$ is the Poisson likelihood, $p(\{x_i\}_{i=1}^M|\mu, \Omega) = e^{-\mu T}(\mu T)^M/M!$. For simplicity, we place a conjugate (Gamma) prior on μT, $\mu T \sim \text{Gamma}(\alpha, \beta)$; the Gamma-Poisson conjugate family conveniently gives the posterior distribution of μT, *i.e.*, $p(\mu T|\{x_i\}_{i=1}^M, \alpha, \beta) = \text{Gamma}(\alpha + M, \beta + 1)$. We choose the scale α and the rate β in the Gamma prior by making the mean of the Gamma posterior equal to N and the variance $M/2$, which is easily shown to correspond to $\alpha = M$ and $\beta = 1$. Finally, due to conjugacy we obtain the posterior

$$p(\mu|\{x_i\}_{i=1}^M, \alpha, \beta) = \text{Gamma}(2M, 2T).$$

3.2.3 Posterior Distribution of ϕ

We handle the posterior distribution of the triggering kernel ϕ given the branching structure in an analogous manner to the LBPP method of Walder and Bishop [2017]. That is, we assume that $\phi(\cdot) = f^2(\cdot)/2$ where $f(\cdot)$ is Gaussian process distributed as described in Section 3.1. In line with [Walder and Bishop, 2017], we consider the sample domain $[0, \pi]$ and the so-called *cosine kernel*,

$$k(x, y) = \sum_{\gamma \geq \mathbf{0}} \lambda_\gamma e_\gamma(x) e_\gamma(y), \tag{3.7}$$

$$\lambda_\gamma := 1/(a(\gamma^2)^m + b), \tag{3.8}$$

$$e_\gamma(x) := (2/\pi)^{1/2} \sqrt{1/2}^{[\gamma=0]} \cos(\gamma x). \tag{3.9}$$

Here, γ is a multi-index with non-negative (integral) values, $[\cdot]$ is the indicator function, a and b are parameters controlling the prior smoothness, and we let $m = 2$. The choice of m affects the rate of change or shape of λ_γ with γ and results in different priors: for large m λ_γ decreases rapidly with γ, giving a-priori preference to smoother functions. So do the parameters a and b. This basis is orthonormal w.r.t. the Lebesgue measure on $\Omega = [0, \pi]$. The expansion Equation (3.7) is an explicit kernel construction based on the Mercer expansion as per Equation (3.1), but other kernels may be used, for example by Nyström approximation of the Mercer decomposition [Flaxman et al., 2017].

As mentioned at the end of Section 3.2.1, by conditioning on the branching structure we may estimate ϕ by considering the *aligned* sequences. In particular, letting S_{x_i} denote the aligned sequence generated by x_i, the joint distribution of $\boldsymbol{\omega}$ and $\boldsymbol{S} := \{S_{x_i}\}_{i=1}^N$ is calculated as [Walder and Bishop, 2017],

$$\log p(\boldsymbol{\omega}, \boldsymbol{S}|\Omega, k) = \sum_{i=1}^N \sum_{\Delta t \in S_{x_i}} \log \frac{1}{2} \left(\boldsymbol{\omega}^T \boldsymbol{e}(\Delta t)\right)^2 - \frac{1}{2}\boldsymbol{\omega}^T(A + \Lambda^{-1})\boldsymbol{\omega} + C, \tag{3.10}$$

$$A := \sum_{i=1}^N \int_0^{T-x_i} \boldsymbol{e}(t)\boldsymbol{e}(t)^T \, \mathrm{d}t, \quad C := -\frac{1}{2} \log \left[(2\pi)^K |\Lambda|\right],$$

where K is the number of eigenfunctions and Λ is defined in Equation (3.2). Note that there is a subtle but important difference between the integral term above and that of Walder and Bishop [2017], namely, the limit of integration; closed-form expressions for the present case are provided in Section A.1 of the appendix. Putting the above equation into Equation (3.3) and Equation (3.4), and we obtain the mean $\hat{\boldsymbol{\omega}}$ and the covariance Q of the (Laplace) approximate log-posterior in $\boldsymbol{\omega}$:

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmax}} \ \log p(\boldsymbol{\omega}, \boldsymbol{S}|\Omega, k), \tag{3.11}$$

$$Q^{-1} = -\sum_{i=1}^{N} \sum_{\Delta t \in S_{x_i}} 2\boldsymbol{e}(\Delta t)\boldsymbol{e}(\Delta t)^T / (\hat{\boldsymbol{\omega}}^T \boldsymbol{e}(\Delta t))^2 + A + \Lambda^{-1}. \tag{3.12}$$

Then, the posterior ϕ is achieved by Equation (3.5) and (3.6).

3.2.4 Computational Complexity

For the LBPP method, constructing Equation (3.10) and (3.12) takes $O(N_o K^2)$ where K is the number of basis functions and N_o is the number of offspring. Optimizing $\boldsymbol{\omega}$ (Equation (3.11)) is a concave problem, which can be solved efficiently. If L-BFGS is used, $O(CK)$ will be taken to calculate the gradient on each $\boldsymbol{\omega}$ where C is the number of steps stored in memory. Computing Q requires inverting a $K \times K$ matrix, which is $O(K^3)$. As a result, the complexity of estimating the conditional probability $p(\phi|B)$ is $O((N_o + K)K^2)$. In terms of estimating $p(\mu|B)$ taking $O(1)$, the complexity of estimating $p(\mu|B)$ and $p(\phi|B)$ is linear to the number of data. The time taken to sample μ and ϕ is minor ($O(1)$ and $O(K)$ respectively), so estimation time dominates. Although the naive complexity for p_{ij} is $O(N^2)$, Halpin [2012] provides an optimized approach to reduce it to $O(N)$, which relies on the finite support assumption of Hawkes processes. **The finite support assumption** says that the value of the triggering kernel is negligible when the input is large [Halpin, 2012, p. 9]. As a result, the step of sampling branching structures can also be run in $O(N)$ and points with negligible impacts on another point are not sampled as its parents. Interestingly, in comparison with LBPP, while our model is in some sense more complex, it enjoys a more favorable computational complexity. In summary, we have the following complexities per iteration and in Section 3.4, we validate the complexity on both synthetic and real data.

Table 3.1: Time complexity.

Operation	$p(\mu\|B)$	p_{ij}	$p(\phi\|B)$	overall
Complexity	$O(1)$	$O(N)$	$O((N_o + K)K^2)$	$O((N + K)K^2)$

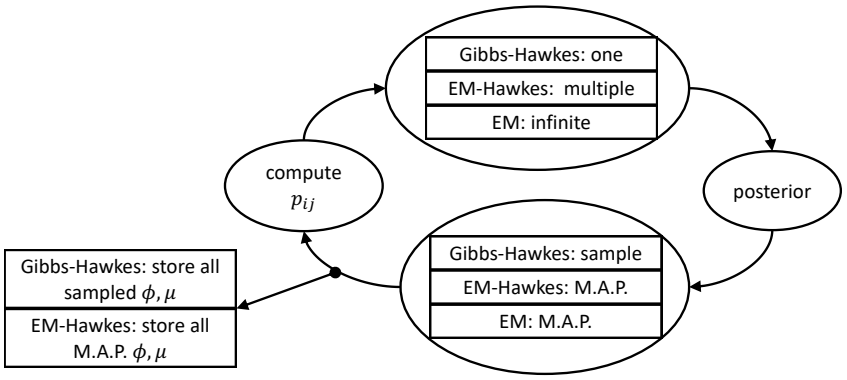


Figure 3.1: A visual summary of the Gibbs-Hawkes, EM-Hawkes and the EM algorithms. The differences between them are (1) the number of sampled branching structures and (2) selected ϕ and μ for p_{ij}. In contrast with with Gibbs-Hawkes, the EM-Hawkes method draws multiple branching structures at once and calculates p_{ij} using M.A.P. ϕ and μ. The EM algorithm is equivalent to sampling infinite branching structures and exploiting M.A.P. or constrained M.L.E. ϕ and μ to calculate p_{ij} (see Section 3.3).

3.3 Maximum-A-Posterior Estimation

We explore a connection between the sampler of Section 3.2 and the EM algorithm, which allows us to introduce an analogous but intermediate scheme between them. In contrast to the random sampler of Section 3.2, the proposed scheme employs a deterministic *maximum-a-posteriori* (M.A.P.) sampler.

3.3.1 Relationship to EM

At the very beginning of this chapter, we mentioned the connection between our method and the stochastic EM algorithm [Celeux and Diebolt, 1985]. The difference is in the M-step; to perform EM [Dempster et al., 1977], we need only modify our sampler by: (a) sampling infinite branching structures at each iteration, and (b) re-calculating the probability of the branching structure with the M.A.P. μ and ϕ, given the infinite set of branching structures. More specifically, maximizing the expected log posterior distribution to estimate M.A.P. μ and ϕ given infinite branching structures is equivalent to maximizing the EM objective in the M-step (see Section A.2 of the appendix for the detailed derivation). Finally, note that the above step (b) is identical to the E-step of the EM algorithm.

3.3.2 EM-Hawkes

Following the discussion above, we propose *EM-Hawkes*, an approximate EM algorithm variant of Gibbs-Hawkes proposed in Section 3.2. Specifically, at each iteration EM-Hawkes (a) samples a finite number of cluster assignments (to approximate the

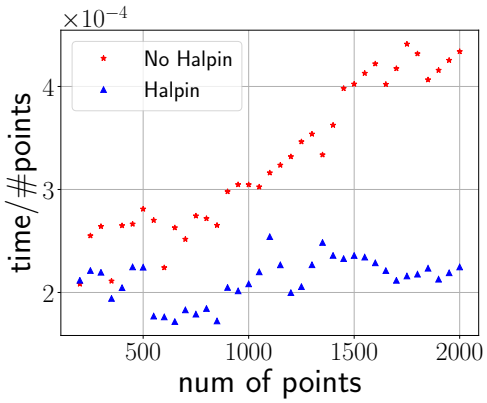


Figure 3.2: Computation time (seconds) for calculating p_{ij} and sampling branching structures, with and without Halpin's speed up.

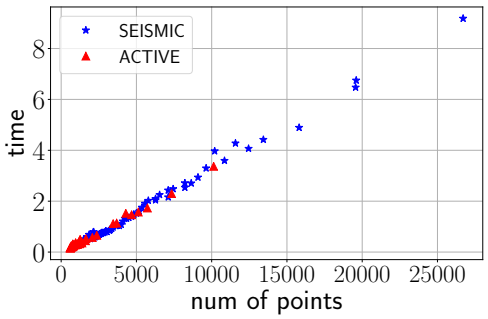


Figure 3.3: Running time (seconds) per iteration on ACTIVE and SEISMIC.

expected log posterior distribution), and (b) finds the M.A.P. triggering kernels and background intensities rather than sampling them as per block Gibbs sampling (the M-step of the EM algorithm). An overview of the Gibbs-Hawkes, EM-Hawkes and EM algorithm is illustrated in Figure 3.1.

Note that under our LBPP-like posterior, finding the most likely triggering kernel ϕ is intractable (see details in Appendix A.3). As an approximation, we take the element-wise mode of the *marginal distributions* of $\{\phi(x_i)\}_{i=1}^N$ to approximate the mode of the joint distribution of the $\{\phi(x_i)\}_{i=1}^N$.

3.4 Experiments

We now evaluate our proposed approaches — Gibbs-Hawkes and EM-Hawkes — and compare them to three baseline models, on synthetic data and on two large Twitter online diffusion datasets. The three baselines are:

(i) A naive parametric Hawkes equipped with a constant background intensity and an exponential (Exp) triggering kernel $\phi = a_1 a_2 \exp(-a_2 t)$, $a_1, a_2 > 0$, estimated by maximum likelihood.

(ii) Ordinary differential equation (ODE)-based non-parametric non-bayesian Hawkes [Zhou et al., 2013].

(iii) Wiener-Hopf (WH) equation based non-parametric non-bayesian Hawkes [Bacry and Muzy, 2016]. Codes of ODE based and WH based methods are publicly available [Bacry et al., 2017].

3.4.1 Synthetic Data

We employ two toy Hawkes processes to generate data, both having the same background intensity $\mu = 10$, and cosine (Equation (3.13)) and exponential (Equation (3.14)) triggering kernels respectively. Note that compared to the cosine triggering kernel, the exponential one has a larger L2 norm for its derivative, and the difference is designed to test the performance of the approaches in different situations. We can check that both triggering kernels have negligible values when the input is large in the domain which we will choose as $[0, \pi]$, so the finite support assumption is satisfied.

$$\phi_{\cos}(x) = \cos(3\pi x) + 1, \quad x \in [0,1]; \quad 0, \quad \text{otherwise}; \tag{3.13}$$

$$\phi_{\exp}(x) = 5\exp(-5x), \quad x \geq 0. \tag{3.14}$$

Prediction. For three baseline models and EM-Hawkes, the predictions μ_{pred} and ϕ_{pred} are taken to be the M.A.P. values, while for Gibbs-Hawkes we use the posterior mean.

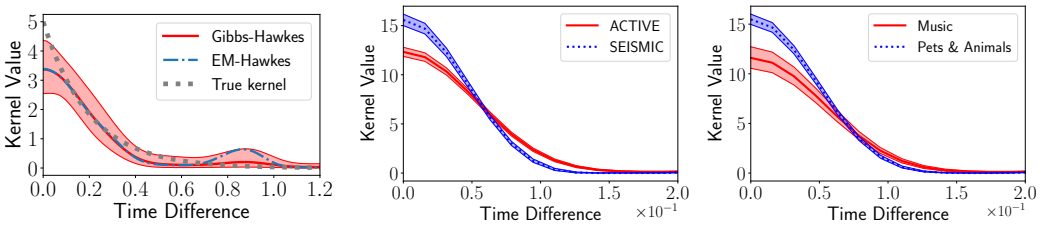
Evaluation. Each toy model generates 400 point sequences over $\Omega = [0, \pi]$, which are evenly split into 40 groups, 20 for training and 20 for test. Each of the three methods fit on each group, *i.e.*, summing log-likelihoods for 10 sequences (for the parametric Hawkes) or estimating the log posterior probability of the Hawkes process given 10 sequences (for Gibbs-Hawkes and EM-Hawkes) or fitting the superposition of 10 sequences [Xu et al., 2018]. Since the true models are known, we evaluate fitting results using the relative L2 distance between predicted and true μ and $\phi(\cdot)$: $d_{L2}(g_{\text{pred}}, g_{\text{true}}) = (\int_\Omega \left(g_{\text{pred}}(t) - g_{\text{true}}(t)\right)^2 dt)^{1/2} / (\int_\Omega (g_{\text{true}}(t))^2 dt)^{1/2}$.

Table 3.2: Empirical performance comparison between algorithms (columns) with different measures (rows). *Top:* relative L2 distance to known ϕ and μ, and AVG denoted the average of L2 errors of ϕ and μ. *bottom:* mean predictive log likelihood on real data. Bold numbers denote the best performance and the underlined numbers for the second best.

Data	Exp	ODE	WH	Gibbs	EM
$\phi_{\cos}$	0.661	0.553	1.000	0.338	0.318
$\mu_{\cos}$	0.069	0.071	1.739	0.078	0.119
$\text{AVG}_{\cos}$	0.365	0.312	1.370	**0.208**	<u>0.219</u>
$\phi_{\exp}$	0.120	0.610	1.000	0.147	0.140
$\mu_{\exp}$	0.086	0.309	4.631	0.103	0.204
$\text{AVG}_{\exp}$	**0.103**	0.460	2.816	<u>0.125</u>	0.172
ACTIVE	2.369	2.370	1.315	2.580	**2.592**
SEISMIC	3.335	3.357	2.131	3.576	**3.578**

Experimental Details. For Gibbs-Hawkes and EM-Hawkes, we must select parameters of the GP kernel (Equation (3.7), (3.8) and (3.9)). An arbitrary choice of them can lead to poor performance, and to this end, we apply the standard cross validation based on the log-likelihood. We choose the number of basis functions in $[8, 16, 32, 64, 128]$ and $a = b$ from $[0.2, 0.02, \cdots, 2 \times 10^{-8}]$. We found that having many

basis functions leads to a high fitting accuracy, but low speed. So, we use 32 basis functions which provides a suitable balance. In terms of kernel parameters a, b of Equation (3.8), we observed that large values return smooth triggering kernels which have a large distance to the ground truth, while small values result in non-smooth predictions which however have small log-likelihoods. As a result, the values $a, b = 0.002$ were chosen. 5000 iterations are run to fit each group and first 1000 are ignored (i.e. *burned-in*).



(a) Exponential synthetic data (b) ACTIVE vs. SEISMIC (c) Categories Music vs. Pets

Figure 3.4: Learned Hawkes triggering kernels using our non-parametric Bayesian approaches. Each red or blue area shows the estimated posterior distributions of ϕ, while the solid lines indicate the 10, 50 and 90 percentiles. In Figure (a), a synthetic dataset simulated using $\phi_{\text{exp}}(t)$ (in gray) is fit using Gibbs-Hawkes (in red) and EM-Hawkes (in blue); Figure (b) presents learning outcomes on Twitter data in ACTIVE (in red) and SEISMIC (in blue); Figure (c) presents learning outcomes on Twitter data associated with two categories in the ACTIVE set: Music (in red) and Pets & Animals (in blue).

Results. The top section of Table 3.2 shows the mean relative L2 distance between the learned and the true ϕ and μ on toy data. First, Gibbs-Hawkes and EM-Hawkes are the closest the models to the ground truth cosine model according to the average error values (AVG_{cos}). For the exponential simulation model, both approaches gain the second and the third lowest errors respectively among all methods, and as expected, the parametric Hawkes – which uses an exponential kernel – fits the model best. In contrast, the parametric model retrieves the cosine model worse because of its mismatch with the ground truth model. The learned triggering kernels for ϕ_{exp} and ϕ_{cos} by our approaches are shown in Figure 3.4(a) and Figure A.1 in the appendix. The ODE-based method performs unsatisfactorily on both simulation settings and it is observed that it performs better on $(\mu_{\text{cos}}, \phi_{\text{cos}})$ than on $(\mu_{\text{exp}}, \phi_{\text{exp}})$. We explain the second observation as that the regularization of the ODE-based method encourages those triggering kernels that have small L2 norms for their derivatives and the derivative of $\phi_{\text{exp}}(x)$ has a larger norm than that of $\phi_{\text{cos}}(x)$. Notably, tuning the hyper-parameters of the WH method is challenging, and Table 3.2 shows the best result obtained after a rather exhaustive experimentation. We speculate that the overall better performance of our approaches is due to the regularization induced by the prior distributions and less difficult hyper-parameter selection. In addition, we also note that EM-Hawkes always performs better at discovering triggering kernels than Gibbs-Hawkes and this observation also holds on the real-life data. Thus, we

conclude that generating multiple samples per iteration tends to improve modeling of the triggering kernels. In summary, compared with state-of-the-art methods, our approaches achieve better performances for data generated by kernels from several parametric classes; as expected, the parametric models are only effective for data generated from their own class.

Effect of Halpin's Procedure. In Section 3.2.4, we show that using Halpin's procedure reduces the complexity of calculating p_{ij} from quadratic to linear. We now empirically validate this speed-up. To distinguish between quadratic and linear complexity, we compute the ratio between running time and data size, shown in Figure 3.2. The ratio when using Halpin's procedure remains roughly constant as data size increases (the ratio increases linearly without the optimization), which implies that Halpin's procedure renders linear calculation of estimating p_{ij} and sampling branching structures. Later, we show the linear complexity of our method on real-world data.

3.4.2 Twitter Diffusion Data

We evaluate the performance of our two proposed approaches on two Twitter datasets that consist of retweet cascades. A retweet cascade contains an original tweet, together with its direct and indirect retweets. Current state of the art diffusion modeling approaches [Zhao et al., 2015; Mishra et al., 2016; Rizoiu et al., 2018] are based on the self-exciting assumption: users get in contact with online content, and then diffuse it to their friends, therefore generating a cascading effect. The two datasets we use have been employed in prior works and they are publicly available:

(i) ACTIVE [Rizoiu et al., 2018] owns 41k retweet cascades, each containing at least 20 (re)tweets with links to Youtube videos. It was collected in 2014 and each Youtube video (and therefore each cascade) is associated with a Youtube category, e.g., *Music* or *News*.
(ii) SEISMIC [Zhao et al., 2015] owns 166k randomly sampled retweet cascades, collected in from Oct 7 to Nov 7, 2011. Each cascade contains at least 50 tweets.

Setup. The temporal extent of each cascade is scaled to $[0, \pi]$, and assigned to either training or test data with equal probability. We bundle together groups of 30 cascades of similar size, and we estimate one Hawkes process for each bundle. Unlike for the synthetic dataset, for the retweet cascades dataset there is no *true* Hawkes process to evaluate against. Instead, we measure using log-likelihood how well the learned model generalizes to the test set. We use the same hyper-parameters values as for the synthetic data. Finally, we follow the prior works on these cascade datasets [Zhao et al., 2015; Rizoiu et al., 2018] by setting the background intensity μ as 0, because the cascade datasets contain only the information of triggering relationships.

Fitting Performance. For each dataset, we calculate the log-likelihood per event for each tweet cascade obtained by three baselines and our approaches (Table 3.2). Visibly, our proposed methods consistently outperform baselines, with EM-Hawkes performing slightly better than Gibbs-Hawkes (by 0.5% for ACTIVE and 0.06% for

SEISMIC). This seems to indicate that online diffusion is influenced by factors not captured by the parametric kernel, therefore justifying the need to learn the Hawkes kernels non-parametrically. As mentioned in the synthetic data part, the WH-based method has a disadvantage of hard-to-tune hyper-parameters, which leads to the worst performance among all methods.

Scalability. To validate the linear complexity of our method, we record running time per iteration of Gibbs-Hawkes on ACTIVE and SEISMIC in Figure 3.3. The running time rises linearly with the number of points increasing, in line with the theoretical analysis. Linear complexity makes our method scalable and applicable on large datasets.

Interpretation. We show in Figure 3.4(b) and 3.4(c) the learned kernels for information diffusions. We notice that the learned kernels appear to be decaying and long-tailed, in accordance with the prior literature. Figure 3.4(b) shows that the kernel learned on SEISMIC is decaying faster than the kernel learned on ACTIVE. This indicates that non-specific (i.e. random) cascades have a faster decay than video-related cascades, presumably due to the fact that Youtube videos stay longer in the human attention. This connection between the type of content and the speed of the decay seems further confirmed in Figure 3.4(c), where we show the learned kernels for two categories in ACTIVE: *Music* and *Pets & Animals*. Cascades relating to *Pets & Animals* have a faster decaying kernel than *Music*, most likely because Music is an ever-green content.

3.5 Summary

In this chapter, we provided the first non-parametric Bayesian inference procedure for the Hawkes process which requires no discretization of the input domain and has an advantage of a linear time complexity. Our method iterates between two steps. First, it samples the branching structure, effectively transforming the Hawkes process into a cluster of Poisson processes. Next, it estimates the Hawkes triggering kernel using a non-parametric Bayesian estimation of the intensity of the cluster Poisson processes. We provide both a full posterior sampler and an EM estimation algorithm based on our ideas. We demonstrated our approach can infer flexible triggering kernels on simulated data. On two large Twitter diffusion datasets, our method outperforms the state-of-the-art in held-out likelihood. Moreover, the learned non-parametric kernel reflects the intuitive longevity of different types of content. The linear complexity of our approach is corroborated on both the synthetic and real problems. The present framework is limited to the univariate unmarked Hawkes process and will be extended to marked multivariate Hawkes process.

The Bayesian Hawkes process model in this chapter is a Gaussian latent model so an advanced Laplace approximation based approach, i.e., the Integrated nested Laplace approximation (INLA) [Rue et al., 2009, 2017], is applicable. INLA is applicable to the case of univariate posterior marginals of latent variables and hyper-parameters. More specifically, we could apply INLA in each iteration of Gibbs

sampling by introducing an additional prior distribution for hyper-parameters $\boldsymbol{\theta}$. As per Rue et al. [2009], INLA first approximates indirectly the posterior distribution of $\boldsymbol{\theta}$, say $p(\boldsymbol{\theta}|\mathcal{D})$, and then approximates the posterior distribution of ω_i by integrating out $\boldsymbol{\theta}$, $p(\omega_i|\mathcal{D}) = \int p(\omega_i|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) \, \mathrm{d}\boldsymbol{\theta}$ (for algorithmic details see Rue et al. [2009]). Both steps apply Laplace approximation and the main difference from our Laplace approximation application includes: (1) INLA aims to model univariate marginal distributions while we focus on the joint distribution; (2) INLA's indirect approximation to the posterior $\boldsymbol{\theta}$ can result in non-Gaussian distributions, which enables more flexible modeling. (3) INLA considers a distribution for hyper-parameters, which increases computational burden: e.g. integrating out $\boldsymbol{\theta}$ in computing $p(\omega_i|\mathcal{D})$ often needs numerical integration. We may also apply INLA to the Bayesian Hawkes process without the Gibbs sampling framework, which needs deeper investigation.

In the next chapter, we will introduce a sparse Gaussian process modulated Hawkes process model and propose variational inference for it.

Chapter 4

Variational Inference for Sparse Gaussian Process Modulated Hawkes Process

In this chapter, we propose the sparse Gaussian process modulated Hawkes process which employs a novel variational inference schema, has an advantage of a linear time complexity per iteration and scales to large real-world data. Our method is inspired by the variational Bayesian Poisson process (VBPP) [Lloyd et al., 2015] which provides the Bayesian non-parametric inference only for the whole intensity of the Hawkes process without for its components: the background intensity μ and the triggering kernel ϕ. Thus, the VBPP loses the internal reactions between points, and developing the variational Bayesian non-parametric inference for the Hawkes process is non-trivial and more challenging than the VBPP. In this paper, we adapt the VBPP for the Hawkes process and term the new approach the variational Bayesian Hawkes process (VBHP). The structure of this chapter is as follows:

(i) In Section 4.1, we first review a prerequisite of our approach, LBPP.

(ii) In Section 4.2, we introduce a Bayesian non-parametric Hawkes process, which employs a sparse Gaussian process modulated triggering kernel and a Gamma distributed background intensity. We propose a new variational inference schema for such a model which is an EM-like two-step iterative algorithm. Specifically, we employ the branching structure of the Hawkes process so that maximization of the evidence lower bound (ELBO) is tractable. As a result, it introduces extra constraints to the ELBO. To deal with the constraints, we apply the expectation-maximization algorithm to minimize the ELBO.

(iii) In Section 4.3, we contribute a tighter ELBO which improves the fitting performance of our model.

(iv) In Section 4.4, we propose a new acceleration trick based on the finite support assumption of the triggering kernel. The new trick enjoys higher efficiency than prior methods and accelerates the variational inference schema to linear time complexity per iteration.

(v) In Section 4.5, we empirically show that VBHP provides more accurate predictions than state-of-the-art methods on synthetic data and on two large online

diffusion datasets. We validate the linear time complexity and faster convergence of our accelerated variational inference schema compared to the Gibbs sampling method, and the practical utility of our tighter ELBO for model selection, which outperforms the common one in model selection.

4.1 Variational Bayesian Poisson Process

Before presenting our approach, we first review Variational Bayesian Poisson process (VBPP) [Lloyd et al., 2015], which is a prerequisite of our method. VBPP applies the VI to the Bayesian Poisson process, which exploits the sparse Gaussian process (GP) to model the Poisson intensity. Specifically, VBPP uses a squared link function to map a sparse GP distributed function f to the Poisson intensity $\lambda(\cdot) = f^2(\cdot)$. The sparse GP employs the ARD kernel for given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^R$:

$$k(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \gamma \prod_{r=1}^{R} \exp\Big(- \frac{(x_r - y_r)^2}{2\alpha_r} \Big).$$

where γ and $\{\alpha_r\}_{r=1}^R$ are GP hyper-parameters. Let $\boldsymbol{u} \coloneqq (f(z_1), f(z_2), \cdots, f(z_M))$ where $\{z_i\}_{i=1}^M$ are inducing points. The prior and the approximate posterior distributions of $\boldsymbol{u}$ are Gaussian distributions,

$$p(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{0}, \boldsymbol{K}_{zz}) \qquad q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{m}, \boldsymbol{S}),$$

where $\boldsymbol{m}$ is the mean vector, and $\boldsymbol{K}_{zz}$ and $\boldsymbol{S}$ are the covariance matrices. Note both $\boldsymbol{u}$ and function evaluations f employ zero mean priors. Notations of VBPP are connected with those of VI (Section 2.2.4) in Table 4.1.

Table 4.1: Notations.

VBHP	VBPP	VI
$\mathcal{D} \coloneqq \{\boldsymbol{x}_i\}_{i=1}^N$	$\mathcal{D} \coloneqq \{\boldsymbol{x}_i\}_{i=1}^N$	$\mathcal{D}$
$B, \mu, f, \boldsymbol{u}$	$f, \boldsymbol{u}$	$\boldsymbol{f}$
$k_0, c_0, \{\alpha_i\}_{i=1}^R, \gamma$	$\{\alpha_i\}_{i=1}^R, \gamma$	θ
$k, c, \boldsymbol{m}, \boldsymbol{S}, \{\alpha_i\}_{i=1}^R, \gamma, \{\{q_{ij}\}_{j=0}^{i-1}\}_{i=1}^N$	$\boldsymbol{m}, \boldsymbol{S}, \{\alpha_i\}_{i=1}^R, \gamma$	θ'

Importantly, the variational joint distribution of f and $\boldsymbol{u}$ uses the exact conditional distribution $p(f|\boldsymbol{u})$, i.e.,

$$q(f, \boldsymbol{u}) \equiv p(f|\boldsymbol{u}) q(\boldsymbol{u}) \tag{4.1}$$

which in turn leads to the posterior GP:

$$\begin{aligned} q(f) &= \mathcal{N}(f|\nu, \Sigma), \\ \nu(\boldsymbol{x}) &\equiv \boldsymbol{K}_{xz} \boldsymbol{K}_{zz}^{-1} \boldsymbol{m}, \\ \Sigma(\boldsymbol{x}, \boldsymbol{x}') &\equiv \boldsymbol{K}_{xx'} + \boldsymbol{K}_{xz} \boldsymbol{K}_{zz}^{-1} (\boldsymbol{S} \boldsymbol{K}_{zz}^{-1} - I) \boldsymbol{K}_{zx'}. \end{aligned} \tag{4.2}$$

Then, the ELBO is obtained by using Equation (2.7):

$$\text{ELBO}(q(f,\boldsymbol{u}), p(\mathcal{D}|f,\boldsymbol{u}), p(f,\boldsymbol{u})) = \mathbb{E}_{q(f)}[\log p(\mathcal{D}|f)] - \text{KL}(q(\boldsymbol{u})||p(\boldsymbol{u})).$$

Note that the second term is the KL divergence between two multivariate Gaussian distributions, so is available in closed form. The first term turns out to be the expectation w.r.t. $q(f)$ of the log-likelihood $\log p(\mathcal{D}|f) = \sum_{i=1}^{N} \log f^2(\boldsymbol{x}_i) - \int_{\mathcal{T}} f^2$. The expectation of the integral part is relatively straight-forward to compute and the expectation of the other (data-dependent) part is available in almost closed-form with a hyper-geometric function.

4.2 Variational Bayesian Hawkes Process

4.2.1 Notations

To extend VBPP to Hawkes process, we introduce two more variables: the background intensity μ and the branching structure B, defined in Section 2.1.2. We assume that the prior distribution of μ is a Gamma distribution $p(\mu) = \text{Gamma}(\mu|k_0, c_0)$[1] and the posterior distribution is approximated by another Gamma distribution $q(\mu) = \text{Gamma}(\mu|k, c)$. Given a dataset $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^N$, the branching structure is expressed as $B = \{\boldsymbol{b}_i\}_{i=1}^N$ with $\boldsymbol{b}_i = \{b_{ij}\}_{j=0}^{i-1}$ defined around Equation (2.1). As $b_{ij} \in \{0, 1\}$ is a binary variable and there is only one b_{ij} equal to 1 among $\boldsymbol{b}_i = \{b_{ij}\}_{j=0}^{i-1}$ (one point has a unique parent), $\boldsymbol{b}_i$ hence has a categorical distribution. We let the variational posterior probability of $b_{ij} = 1$ be denoted as $q_{ij} = q(b_{ij} = 1)$ and there must be $\sum_{j=0}^{i-1} q_{ij} = 1$. As a result, the variational posterior probability of B is expressed as:

$$q(B) = \prod_{i=1}^{N} \prod_{j=0}^{i-1} q_{ij}^{b_{ij}}. \tag{4.3}$$

The same squared link function is adopted for the triggering kernel $\phi(\cdot) = f^2(\cdot)$, so are the priors for f and $\boldsymbol{u}$, namely $\mathcal{N}(f|\mathbf{0}, \mathbf{K}_{xx'})$ and $\mathcal{N}(\boldsymbol{u}|\mathbf{0}, \mathbf{K}_{zz'})$. More link functions such as $\exp(\cdot)$ are discussed by Lloyd et al. [2015]. Moreover, we use the same variational joint posterior on f and $\boldsymbol{u}$ as Equation (4.1). Consequently, we complete the variational joint distribution on all latent variables as below:

$$q(B, \mu, f, \boldsymbol{u}) \equiv q(B)q(\mu)p(f|\boldsymbol{u})q(\boldsymbol{u}), \tag{4.4}$$

and notations of VBHP are summarized in Table 4.1. The variational posterior probability of B (Equation (4.3)) has the same form as that of the true posterior, so it is a good approximation. The variational joint distribution $q(B, \mu, f, \boldsymbol{u})$ assumes independence between the modeled variables, which is not the case for the true joint posterior distribution, e.g. the branching structure B is closely dependent on the

[1] $\text{Gamma}(\mu|k_0, c_0) = \dfrac{1}{\Gamma(k_0)c_0^k} \mu^{k_0-1} e^{-x/c_0}$

background intensity μ and the function f. Thus such an approximation is not perfect. The approximation error is insignificant in practice and our experiments demonstrate the effectiveness of the proposed approximation.

Based on Equation (2.7) and (4.4), we obtain the ELBO for VBHP as below (see details in Section B.1 of the appendix). To differentiate with the tighter ELBO presented later in Section 4.3, we name the below one as the common ELBO (CELBO).

$$\begin{aligned}&\text{CELBO}(q(B,\mu,f,\boldsymbol{u}), p(\mathcal{D}|B,\mu,f,\boldsymbol{u}), p(B,\mu,f,\boldsymbol{u}))\\ &= \underbrace{\mathbb{E}_{q(B,\mu,f)}\Big[\log p(\mathcal{D},B|f,\mu)\Big]}_{\text{Data Dependent Expectation (DDE)}} + H_B - \text{KL}(q(\mu)||p(\mu)) - \text{KL}(q(\boldsymbol{u})||p(\boldsymbol{u})). \quad (4.5)\end{aligned}$$

where $H_B = -\sum_{i=1}^{N}\sum_{j=0}^{i-1} q_{ij}\log q_{ij}$ is the entropy of the variational posterior B. The KL terms are between gamma and Gaussian distributions, for which closed forms are provided in Section B.2 of the appendix.

4.2.2 Data Dependent Expectation

Now, we are left with the problem of computing the data dependent expectation (DDE) in Equation (4.5). The DDE is w.r.t. the variational posterior probability $q(B,\mu,f)$. From Equation (4.4), $q(B,\mu,f) = \int q(B,\mu,f,\mathbf{u})\,\mathrm{d}\mathbf{u} = q(B)q(\mu)q(f)$ and $q(f)$ is identical to Equation (4.2). As a result, we can compute the DDE w.r.t. $q(B)$ first, and then w.r.t. $q(\mu)$ and $q(f)$.

Expectation w.r.t. $q(B)$. From Equation (2.3), we easily obtain $\log p(\mathcal{D},B|f,\mu)$ by replacing ϕ with f^2, whereupon it is clear that only b_{ij} in $\log p(\mathcal{D},B|f,\mu)$ is dependent on B. Therefore, $\mathbb{E}_{q(B)}[\log p(\mathcal{D},B|f,\mu)]$ is computed as:

$$\mathbb{E}_{q(B)}[\log p(\mathcal{D},B|f,\mu)] = \sum_{i=1}^{N}\Big(\sum_{j=1}^{i-1} q_{ij}\log f_{ij}^2 + q_{i0}\log\mu\Big) - \sum_{i=1}^{N}\int_{\mathcal{T}_i} f^2 - \mu|\mathcal{T}|$$

where $f_{ij} \coloneqq f(\boldsymbol{x}_i - \boldsymbol{x}_j)$.

Expectation w.r.t. $q(f)$ and $q(\mu)$. We compute the expectation w.r.t. $q(f)$ and $q(\mu)$ by exploiting the expectation and the log expectation of the Gamma distribution: $\mathbb{E}_{q(\mu)}(\mu) = kc$ and $\mathbb{E}_{q(\mu)}(\log(\mu)) = \psi(k) + \log c$, and also the property $\mathbb{E}(x^2) = \mathbb{E}(x)^2 + \text{Var}(x)$:

$$\begin{aligned}\text{DDE} = \sum_{i=1}^{N}\Big[\sum_{j=1}^{i-1} q_{ij}\,\mathbb{E}_{q(f)}(\log f_{ij}^2) + q_{i0}\big(\psi(k)+\log c\big)\\ - \int_{\mathcal{T}_i}\mathbb{E}_{q(f)}^2(f) - \int_{\mathcal{T}_i}\text{Var}_{q(f)}(f)\Big] - kc|\mathcal{T}|,\end{aligned}$$

where ψ is the Digamma function. We provide closed form expressions for $\int_{\mathcal{T}_i}\mathbb{E}_{q(f)}^2(f)$

and $\int_{\mathcal{T}_i} \mathrm{Var}_{q(f)}(f)$ in Section B.3 of the appendix. As in VBPP,

$$\mathbb{E}_{q(f)}(\log f_{ij}^2) = -\widetilde{G}(-\nu_{ij}^2/(2\Sigma_{ij})) + \log(\Sigma_{ij}/2) - C$$

is available in the closed form with a hyper-geometric function, where $\nu_{ij} = \nu(\boldsymbol{x}_i - \boldsymbol{x}_j)$, $\Sigma_{ij} = \Sigma(\boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{x}_i - \boldsymbol{x}_j)$ (ν and Σ are defined in Equation (4.2)), $C \approx 0.57721566$ is the Euler-Mascheroni constant and $\widetilde{G}$ is defined as

$$\widetilde{G}(z) = {}_1F_1^{(1,0,0)}(0, 1/2, z),$$

that is, the partial derivative of the confluent hyper-geometric function ${}_1F_1$ w.r.t. the first argument. We compute $\widetilde{G}$ using the method of Ancarani and Gasaneo [2008] and implement $\tilde{G}$ and $\tilde{G}'$ by linear interpolation of a lookup table.

4.2.3 Predictive Distribution of ϕ

The predictive distribution of $f(\boldsymbol{x})$ depends on the posterior $\mathbf{u}$. We assume that the optimal variational distribution of $\mathbf{u}$ approximates the true posterior distribution, namely $q(\mathbf{u}|\mathcal{D}, \boldsymbol{\theta}'^*) = \mathcal{N}(\mathbf{u}|\boldsymbol{m}^*, \mathbf{S}^*) \approx p(\mathbf{u}|\mathcal{D}, \boldsymbol{\theta})$. Therefore, there is

$$\begin{aligned} q(f|\mathcal{D}, \boldsymbol{\theta}'^*) &= \int p(f|\boldsymbol{u}) q(\mathbf{u}|\mathcal{D}, \boldsymbol{\theta}'^*)\, \mathrm{d}\boldsymbol{\theta} \\ &\approx \int p(f|\boldsymbol{u}) p(\mathbf{u}|\mathcal{D}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} = p(f|\mathcal{D}, \boldsymbol{\theta}), \end{aligned}$$

and we thus use $q(f(\tilde{\boldsymbol{x}})|\mathcal{D}, \boldsymbol{\theta}'^*)$ as the the approximate predictive distribution which is calculated as

$$\begin{aligned} q(f(\tilde{\boldsymbol{x}})|\mathcal{D}, \boldsymbol{\theta}'^*) &= \int p(f(\tilde{\boldsymbol{x}})|\boldsymbol{u}) q(\boldsymbol{u}|\mathcal{D}, \boldsymbol{\theta}'^*)\, \mathrm{d}\boldsymbol{u} \\ &= \mathcal{N}(\boldsymbol{K}_{\tilde{x}z}\boldsymbol{K}_{zz}^{-1}\boldsymbol{m}^*, \boldsymbol{K}_{\tilde{x}\tilde{x}} - \boldsymbol{K}_{\tilde{x}z}\boldsymbol{K}_{zz'}^{-1}\boldsymbol{K}_{z\tilde{x}} + \boldsymbol{K}_{\tilde{x}z}\boldsymbol{K}_{zz'}^{-1}\mathbf{S}^*\boldsymbol{K}_{zz'}^{-1}\boldsymbol{K}_{z\tilde{x}}) \\ &\equiv \mathcal{N}(\tilde{v}, \tilde{\sigma}^2). \end{aligned}$$

Given the relation $\phi = f^2$, it is straightforward to derive the corresponding approximate posterior of $\phi(\tilde{\boldsymbol{x}})$

$$\phi(\tilde{\boldsymbol{x}}) \sim \mathrm{Gamma}(\tilde{k}, \tilde{c})$$

where the shape $\tilde{k} = (\tilde{v}^2 + \tilde{\sigma}^2)^2/[2\tilde{\sigma}^2(2\tilde{v}^2 + \tilde{\sigma}^2)]$ and the scale $\tilde{c} = 2\tilde{\sigma}^2(2\tilde{v}^2 + \tilde{\sigma}^2)/(\tilde{v}^2 + \tilde{\sigma}^2)$.

4.3 New Variational Inference Schema

We now propose a new variational inference (VI) schema which uses a tighter ELBO than the common one, *i.e.*

Theorem 1. *For VBHP, there is a tighter ELBO*

$$\underbrace{\mathbb{E}_{q(B,\mu,f)}\Big[\log p(\mathcal{D}, B|f,\mu)\Big] + H_B}_{\equiv \text{TELBO}} \leq \log p(\mathcal{D}).$$

Remark. *TELBO is tighter because it is equivalent to the CELBO (Equation* (4.5)*) except without subtracting non-negative KL divergences over μ and* $\mathbf{u}$*. Thus, it is easy to see that the gap between the TELBO and* $\log p(\mathcal{D})$ *is*

$$\log p(\mathcal{D}) - \text{TELBO} = -KL(q(\mu)||p(\mu)) - KL(q(\boldsymbol{u})||p(\boldsymbol{u})) + KL(q(B,\mu,f,\boldsymbol{u})||p(B,\mu,f,\boldsymbol{u}|\mathcal{D})).$$

The aggregate effect of the three KL terms is rather challenging to understand and needs further investigation. Other graphical models such as the variational Gaussian mixture model [Attias, 1999] have a similar TELBO. Later on, we propose a new VI schema based on the TELBO, where the TELBO will be applied to selecting the hyper-parameters as it provides a tighter approximation to the log marginal likelihood compared to CELBO.

Proof. With the variational posterior probability of the branching structure $q(B)$ defined in Equation (4.3) and through the Jensen's inequality, we have:

$$\log p(\mathcal{D}) \geq \sum_B q(B) \log p(\mathcal{D}, B) + H_B, \tag{4.6}$$

where H_B is the entropy of B defined in Equation (4.5). The term $\sum_B q(B) \log p(\mathcal{D}, B)$ can be understood as follows. Consider that infinite branching structures are drawn from $q(B)$ independently, say $\{B_i\}_{i=1}^{\infty}$. Given a branching structure B_i, the Hawkes process can be decomposed into a cluster of Poisson processes, denoted as $(\mathcal{D}, B_i)$, and the corresponding log-likelihood is $\log p(\mathcal{D}, B_i)$. Then, $\sum_B q(B) \log p(\mathcal{D}, B)$ is the mean of all log likelihoods $\{\log p(\mathcal{D}, B_i)\}_{i=1}^{\infty}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log p(\mathcal{D}, B_i) = \lim_{n\to\infty} \sum_B \frac{n_B}{n} \log p(\mathcal{D}, B) = \sum_B q(B) \log p(\mathcal{D}, B), \tag{4.7}$$

where n_B is the number of occurrences of branching structure B. Since all branching structures $\{B_i\}_{i=1}^{\infty}$ are i.i.d., the clusters of Poisson processes generated over $\{B_i\}_{i=1}^{\infty}$ should also be independent, i.e., $\{(\mathcal{D}, B_i)\}_{i=1}^{\infty}$ are i.i.d.. It follows that

$$\sum_{i=1}^{\infty} \log p(\mathcal{D}, B_i) = \log p(\{(\mathcal{D}, B_i)\}_{i=1}^{\infty}). \tag{4.8}$$

We compute the CELBO of $\log p(\{(\mathcal{D}, B_i)\}_{i=1}^{\infty})$ by making $\boldsymbol{z} = (\mu, f, \mathbf{u})$ and $\boldsymbol{x} = \{(\mathcal{D}, B_i)\}_{i=1}^{n}$ in Equation (2.7):

$$\begin{aligned} \log p(\{(\mathcal{D}, B_i)\}_{i=1}^{n}) \geq\ & E_{q(f,\mu)}\big[\log p(\{(\mathcal{D}, B_i)\}_{i=1}^{n}|f,\mu)\big] \\ & -\text{KL}(q(\mu)||p(\mu)) - \text{KL}(q(\boldsymbol{u})||p(\boldsymbol{u})). \end{aligned} \tag{4.9}$$

Further, we plug Equation (4.8) and (4.9) into Equation (4.7):

$$\begin{aligned}\text{Equation (4.7)} &= \lim_{n\to\infty}\frac{1}{n}\log p(\{(\mathcal{D}, B_i)\}_{i=1}^n)\\ &\overset{(a)}{\geq} \lim_{n\to\infty}\frac{1}{n}E_{q(f,\mu)}\big[\log p(\{(\mathcal{D}, B_i)\}_{i=1}^n|f,\mu)\big]\\ &\overset{(b)}{=} \lim_{n\to\infty}\sum_{n_B}\frac{n_B}{n}E_{q(f,\mu)}\big[\log p(\mathcal{D}, B|f,\mu)\big]\\ &= E_{q(f,\mu,B)}\big[\log p(\mathcal{D}, B|f,\mu)\big]\end{aligned}$$

where (a) is because the finite values of KL terms are divided by infinitely large n, and (b) is due to i.i.d. $(\mathcal{D}, B_i)$ and the variational posterior B being independent of f and μ. Finally, we plug the above inequality into Equation (4.6) and obtain the TELBO. □

4.3.1 New Optimization Schema for VBHP

To optimize the model parameters under constraints $\sum_{j=0}^{i-1} q_{ij} = 1$, we employ the expectation-maximization algorithm. Specifically, in the **E step**, all q_{ij} are optimized to maximize the CELBO, and in the **M step**, $\boldsymbol{m}$, $\boldsymbol{S}$, k and c are updated to increase the CELBO. We don't use the TELBO to optimize the variational distributions because it doesn't guarantee minimizing the KL divergence between variational and true posterior distributions. Instead, the TELBO is employed to select GP hyper-parameters:

$$\{\alpha_i^*\}_{i=1}^R, \gamma^* = \operatorname*{argmax}_{\{\alpha_i\}_{i=1}^R,\gamma} \text{TELBO}.$$

The TELBO bounds the marginal likelihood more tightly than CELBO, and is therefore expected to lead to a better predictive performance — an intuition which we empirically validate in Section 4.5.

The updating equations for q_{ij} are derived through maximization of Equation (4.5) under the constraints $\sum_{j=0}^{i-1} q_{ij} = 1$ for all i. This maximization problem is dealt with the Lagrange multiplier method, and yields the below updating equations:

$$q_{ij} = \begin{cases}\exp(\mathbb{E}_{q(f)}(\log f_{ij}^2))/A_i, & j > 0;\\ \theta\exp(\psi(k))/A_i, & j = 0,\end{cases}$$

where $A_i = \theta\exp(\psi(k)) + \sum_{j=1}^{i-1}\exp(\mathbb{E}_{q(f)}(\log f_{ij}^2))$ is the normalizer.

Furthermore, and similarly to VBPP, we fix the inducing points on a regular grid over $\mathcal{T}$. Despite the observation that more inducing points lead to better fitting accuracy [Lloyd et al., 2015; Snelson and Ghahramani, 2006], in the case of our more complex VBHP, more inducing points may cause slow convergence (Figure B.2(a) in the appendix) for some hyper-parameters, and therefore lead to poor performance in limited iterations. Generally, more inducing points improve accuracy at the expense of longer fitting time.

4.4 Acceleration Trick

4.4.1 Time Complexity Without Acceleration

In the E step of model optimization, updating q_{ij} requires computing the mean and the variance of all f_{ij}, which both take $O(M^3 + M^2N^2)$ with N points in the HP and M inducing points. Here, we omit the dimension of data R since normally $M > R$ for a regular grid of inducing points. Similarly, in the M step, computing the hypergeometric term requires the means and variances of all the f_{ij}. Finally, computation of the integral terms takes $O(M^3N)$. Thus, the total time complexity per iteration is $O(M^3N + M^2N^2)$.

4.4.2 Acceleration to Linear Time Complexity

To accelerate our VBHP, similarly to **Zhang** et al. [2019] we exploit the finite support assumption of the triggering kernel, assuming the kernel has negligible values for sufficiently large inputs. As a result, sufficiently distant pairs of points do not enter into the computations. This trick reduces possible parents of a point from all prior points to a set of neighbors. The number of relevant neighbors is bounded by a constant C and as a result the total time complexity is reduced to $O(CM^3N)$.

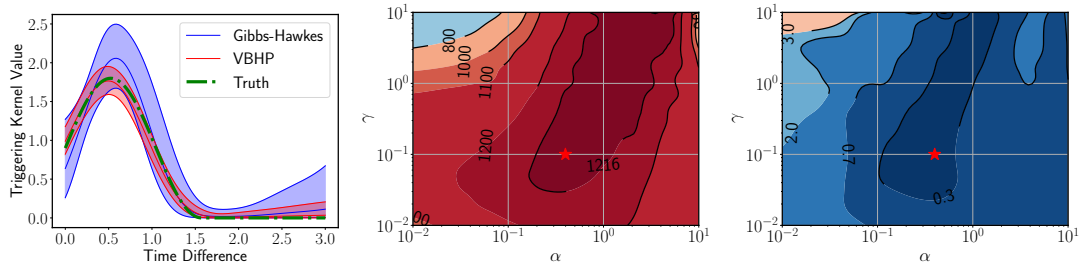
Specifically, we introduce a compact region $\mathcal{S} = \bigtimes_{r=1}^{R}[\mathcal{S}_r^{\min}, \mathcal{S}_r^{\max}] \subseteq \mathcal{T}$ so that $\phi(\boldsymbol{x}_i - \boldsymbol{x}_j) = 0$ and $q_{ij} = 0$ if $\boldsymbol{x}_i - \boldsymbol{x}_j \notin \mathcal{S}$. As a result, all terms related to $\boldsymbol{x}_i - \boldsymbol{x}_j \notin \mathcal{S}$ vanish. To choose a suitable $\mathcal{S}$, we again use the TELBO, taking the smallest $\boldsymbol{S}$ for which the TELBO doesn't drop significantly; we optimize $S_r^{\min}$ and $S_r^{\max}$ by grid search with other dimensions fixed (so that this step is run R times in total) and we optimize $S_r^{\min}$ after optimizing $S_r^{\max}$.

Rather than selecting pairs of points in each iteration in the manner of Halpin's trick [Halpin, 2012; **Zhang** et al., 2019], our method pre-computes those pairs, leading to gains in computational efficiency. The similar aspect is that both tricks have hyper-parameters to select to threshold the triggering kernel value. We employ the TELBO for hyper-parameter selection while frequentist methods use the cross validation.

4.5 Experiments

4.5.1 Evaluation

We employ two metrics: the first is the $\mathbf{L_2}$ distance (for cases with a known ground truth), which measures the difference between predictive and truth Hawkes kernels, formulated as $\mathrm{L}_2(\phi_{\text{pred}}, \phi_{\text{true}}) = (\int_{\mathcal{T}}(\phi_{\text{pred}}(\boldsymbol{x}) - \phi_{\text{true}}(\boldsymbol{x}))^2 \, \mathrm{d}\boldsymbol{x})^{0.5}$ and $\mathrm{L}_2(\mu_{\text{pred}}, \mu_{\text{true}}) = |\mu_{\text{pred}} - \mu_{\text{true}}|$; the second is the hold-out log likelihood (**HLL**), which describes how well the predictive model fits the test data, formulated as $\log p(\mathcal{D}_{\text{Test}} = \{\boldsymbol{x}_i\}_{i=1}^{N} | \mu, f) = \sum_{i=1}^{N} \log \lambda(\boldsymbol{x}_i) - \int_{\mathcal{T}} \lambda$. To calculate the HLL for each process, we generate a number of test sequences by every time randomly assigning each point of the original process to either a training or testing sequence with equal probability; HLLs of test sequences are normalized (by dividing test sequence length) and averaged.



(a) Variational posterior triggering kernel

(b) TELBO

(c) L2 distance for ϕ

Figure 4.1: The relationship between the log marginal likelihood and the L_2 distance. In (a), the true $\phi_{\sin}$ (dash green) is plotted with the median (solid) and the [0.1, 0.9] interval (filled) of the approximate posterior triggering kernel obtained by VBHP and Gibbs Hawkes (10 inducing points). It uses the maximum point of the TELBO (red star in (b)). In (c), the maximum point of the TELBO is marked. The maximum point overlaps with that of the CELBO. $[0, 1.4]$ is used as the support of the predictive triggering kernel and 10 inducing points are used.

4.5.2 Prediction

We use the pointwise mode of the approximate posterior triggering kernel as the prediction because it is computationally intractable to find the posterior mode at multiple point locations [**Zhang** et al., 2019]. Besides, we exploit the mode of the approximate posterior background intensity as the predictive background intensity.

4.5.3 Baselines

We use the following models as baselines.

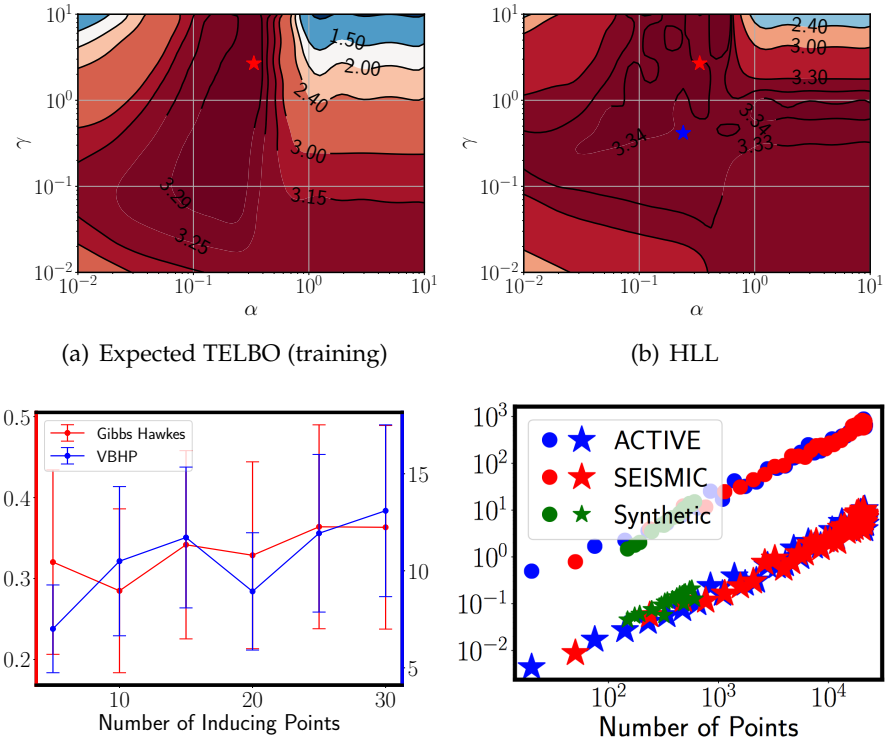
(i) A parametric Hawkes process equipped with the sum of exponential (**SumExp**) triggering kernel $\phi(x) = \sum_{i=1}^{K} a_1^i a_2^i \exp(-a_2^i x)$ and the constant background intensity.

(ii) The ordinary differential equation (**ODE**) based non-parametric non-Bayesian Hawkes process [Zhou et al., 2013]. The code is publicly available [Bacry et al., 2017].

(iii) Wiener-Hopf (**WH**) equation based non-parametric non-Bayesian Hawkes process [Zhou et al., 2013]. The code is publicly available [Bacry et al., 2017].

(iv) The Gibbs sampling based Bayesian non-parametric Hawkes process (**Gibbs Hawkes**) [**Zhang** et al., 2019].

For fairness, the ARD kernel is used by Gibbs Hawkes and corresponding eigenfunctions are approximated by Nyström method [Williams and Seeger, 2001], where regular grid points are used as VBHP. Different from batch training in [**Zhang** et al., 2019], all experiments are conducted on single sequences.



(a) Expected TELBO (training)

(b) HLL

(c) Time VS inducing points

(d) Time VS data size

Figure 4.2: Figure (a), (b): The relationship between the TELBO and the HLL. Figure (c), (d): Average fitting time (seconds) per iteration. In Figure (a), the maximum point is marked by the red star. In Figure (b), the maximum points of the TELBO and CELBO are marked by red and blue stars. Figure (c) is plotted on 50 processes. Figure (d) shows the fitting time of Gibbs Hawkes (star) and VBHP (circle) on 120 processes. 10 inducing points are used unless specified.

Table 4.2: Results on synthetic and real-world data (mean $\pm$ one standard variance). VBHP (C) and (T) use the CELBO and the TELBO to update the hyper-parameters respectively. Bold numbers denote the best performance.

Measure	Data	SumExp	ODE	WH	Gibbs Hawkes	VBHP (C)	VBHP (T)
L_2	Sin	ϕ:0.693$_{\pm 0.028}$	0.665$_{\pm 0.121}$	2.463$_{\pm 0.145}$	0.408$_{\pm 0.198}$	**0.152**$_{\pm 0.091}$	0.183$_{\pm 0.076}$
		μ:2.968$_{\pm 1.640}$	4.514$_{\pm 3.808}$	6.794$_{\pm 5.054}$	4.108$_{\pm 3.949}$	0.640$_{\pm 0.528}$	**0.579**$_{\pm 0.523}$
	Cos	ϕ:0.473$_{\pm 0.102}$	0.697$_{\pm 0.065}$	1.743$_{\pm 0.083}$	0.667$_{\pm 0.686}$	0.325$_{\pm 0.073}$	**0.292**$_{\pm 0.096}$
		μ:2.751$_{\pm 1.902}$	7.030$_{\pm 5.662}$	6.099$_{\pm 4.613}$	4.685$_{\pm 4.421}$	0.555$_{\pm 0.294}$	**0.515**$_{\pm 0.293}$
	Exp	ϕ:**0.133**$_{\pm 0.138}$	1.835$_{\pm 0.539}$	2.254$_{\pm 2.042}$	0.676$_{\pm 0.233}$	0.257$_{\pm 0.086}$	0.235$_{\pm 0.102}$
		μ:3.290$_{\pm 1.991}$	8.969$_{\pm 8.604}$	16.66$_{\pm 20.95}$	7.648$_{\pm 9.647}$	**0.471**$_{\pm 0.432}$	0.486$_{\pm 0.418}$
HLL	Sin	3.490$_{\pm 0.400}$	3.489$_{\pm 0.413}$	3.233$_{\pm 0.273}$	3.492$_{\pm 0.406}$	3.488$_{\pm 0.400}$	**3.497**$_{\pm 0.406}$
	Cos	3.874$_{\pm 0.544}$	3.872$_{\pm 0.552}$	3.613$_{\pm 0.373}$	3.871$_{\pm 0.562}$	3.876$_{\pm 0.541}$	**3.878**$_{\pm 0.548}$
	Exp	2.825$_{\pm 0.481}$	2.822$_{\pm 0.496}$	2.782$_{\pm 0.490}$	2.826$_{\pm 0.492}$	2.826$_{\pm 0.491}$	**2.829**$_{\pm 0.487}$
	ACTIVE	1.692$_{\pm 1.371}$	0.880$_{\pm 2.716}$	0.710$_{\pm 0.943}$	1.323$_{\pm 2.160}$	1.824$_{\pm 1.159}$	**1.867**$_{\pm 1.181}$
	SEISMIC	2.943$_{\pm 0.959}$	2.582$_{\pm 1.665}$	1.489$_{\pm 1.796}$	3.110$_{\pm 1.251}$	3.143$_{\pm 0.895}$	**3.164**$_{\pm 0.843}$

4.5.4 Synthetic Experiments

Synthetic Data. Our synthetic data are generated from three Hawkes processes over $\mathcal{T} = [0, \pi]$, whose triggering kernels are sin, cos and exp functions respectively, shown as below, and whose background intensities are the same $\mu = 10$:

$$\phi_{\sin}(x) = 0.9[\sin(3x) + 1], x \in [0, \pi/2]; \text{otherwise}, 0;$$
$$\phi_{\cos}(x) = \cos(2x) + 1, x \in [0, \pi/2]; \text{otherwise}, 0;$$
$$\phi_{\exp}(x) = 5\exp(-5x), x \in [0, \infty).$$

As a result, for any generated sequence, say $\{x_i\}_{i=1}^N$, $\mathcal{T}_i = [0, \pi - x_i]$ is used in the CELBO and the TELBO. We can check that all three triggering kernels have negligible values when the input is large, so the finite support assumption is satisfied.

Model Selection. As the marginal likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ is a key advantage of our method over non-Bayesian approaches [Zhou et al., 2013; Bacry and Muzy, 2016], we investigate its efficacy for model selection. Figure 4.1(b) shows the contour plot of the approximate log marginal likelihood (the TELBO) of a sequence. It is observed that the contour plot of the TELBO has agreement to the contour plots of $\mathrm{L}_2(\phi)$ (Figure 4.1(c)) — GP hyper-parameters with relatively high marginal likelihoods have relatively low L_2 errors. Figure 4.1(a) plots the posterior triggering kernel corresponding to the maximal approximate marginal likelihood. Similar agreement is also observed between the TELBO and the HLL (Figure 4.2(a), 4.2(b)). This demonstrates the practical utility of both the marginal likelihood itself and our approximation of it.

Evaluation. To evaluate VBHP on synthetic data, 20 sequences are drawn from each model and 100 pairs of train and test sequences drawn from each sample to compute the HLL. We select GP hyper-parameters of Gibbs Hawkes and of VBHP by maximizing approximate marginal likelihoods. Table 4.2 shows evaluations for baselines and VBHP (using 10 inducing points for trade-off between accuracy and time, so does Gibbs Hawkes) in both L_2 and HLL. VBHP achieves the best performance but is two orders of magnitudes slower than Gibbs Hawkes per iteration (shown as Figure 4.2(c) and 4.2(d)). The speed of VBHP is limited by its complicated implementation, such as the linear interpolation of the lookup table when computing $\tilde{G}$ as Section 4.2.2, and we could accelerate it with more advanced implementation techniques. Although the Gibbs Hawkes is based on the Markov Chain Monte Carlo algorithm and is expected to return more accurate results, it employs Laplace approximation per iteration and leads to approximation error, which is one of the causes of the inferior performance. The TELBO performs closely to the CELBO in the L_2 error and this is also reflected in Figure 4.1(c) where the maximum points of the TELBO and the CELBO overlap. In contrast, the TELBO consistently improves the performance of VBHP in the HLL, which is also reflected in Figure 4.2(b) where hyper-parameters selected by the TELBO tend to have a higher HLL. Interestingly, when the parametric model SumExp uses the same triggering kernel (a single exponential function) as the ground truth $\phi_{\exp}$, SumExp fits $\phi_{\exp}$ best in L_2 distance while due to learning on single sequences, the background intensity has relatively high errors. Although our method

is not aware of the parametric family of the ground truth, it performs well. Compared with non-parametric frequentist methods which have strong fitting capacity but suffer from noisy data and have difficulties with hyper-parameter selection, our Bayesian solution overcomes these disadvantages and achieves better performance.

4.5.5 Real World Experiments

Real World Data. We conclude our experiments with two large scale tweet datasets. **ACTIVE** [Rizoiu et al., 2018] is a tweet dataset which was collected in 2014 and contains ∼41k (re)tweet temporal point processes with links to Youtube videos. Each sequence contains at least 20 (re)tweets. **SEISMIC** [Zhao et al., 2015] is a large scale tweet dataset which was collected from October 7 to November 7, 2011, and contains ∼166k (re)tweet temporal point processes. Each sequence contains at least 50 (re)tweets.

Evaluation. Similarly to synthetic experiments, we evaluate the fitting performance by averaging HLL of 20 test sequences randomly drawn from each original datum. We scale all original data to $\mathcal{T} = [0, \pi]$ (leading to $\mathcal{T}_i = [0, \pi - x_i]$ used in the CELBO and the TELBO for a sequence $\{x_i\}_{i=1}^N$) and employ 10 inducing points to balance time and accuracy. The model selection is performed by maximizing the approximate marginal likelihood. The obtained results are shown in Table 4.2. Again, we observe similar predictive performance of VBHP: the TELBO performs better the CELBO; VBHP achieves best scores. This demonstrates our Bayesian model and novel VI schema are useful for flexible real life data.

Fitting Time. We further evaluate the fitting speed[2] of VBHP and Gibbs Hawkes on synthetic and real-world point processes, which is summarized in Figure 4.2(c) and 4.2(d). The fitting time is averaged over iterations and we observe that the increasing trends with the number of inducing points and with the data size are similar between Gibbs Hawkes and VBHP. Although VBHP is significantly slower than Gibbs Hawkes per iteration, VBHP converges faster, in 10∼20 iterations (Figure B.2 of the appendix), giving an average convergence time of 549 seconds for a sequence of 1000 events, compared to 699 seconds for Gibbs Hawkes. The slope of VBHP in Figure 4.2(d) is 1.04 (log-scale) and the correlation coefficient is 0.96, so we conclude that the fitting time is linear to the data size.

4.6 Conclusions

In this chapter, we presented a new Bayesian non-parametric Hawkes process whose triggering kernel is modulated by a sparse Gaussian process and background intensity is Gamma distributed. We provided a novel VI schema for such a model: we employed the branching structure so that the common ELBO is maximized by the expectation-maximization algorithm; we contributed a tighter ELBO which performs better in model selection than the common one. To address the difficulty with scaling with the

[2]The CPU we use is Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz and the language is Python 3.6.5.

data size, we utilize the finite support assumption of the triggering kernel to reduce the number of possible parents for each point. Different from prior acceleration methods, ours enjoys higher efficiency. On synthetic data and two large Twitter diffusion datasets, VBHP enjoys linear fitting time with the data size and fast convergence rate, and provides more accurate predictions than those of state-of-the-art approaches. The novel ELBO is also demonstrated to exceed the common one in model selection.

Comparison with the Approach in Chapter 3. The approach in Chapter 3 similarly exploits the Hawkes process branching structure and the finite support assumption of the triggering kernel, while it builds on Gibbs sampling. It considers only high-probability triggering relationships in computations for acceleration. However, those relationships are updated in each iteration, which is less efficient than our precomputing them. Besides, our variational inference schema enjoys faster convergence than that of the Gibbs sampling based method.

Both methods in this chapter and the last chapter are designed based on classical approximate Bayesian inference, we propose a new approximate inference technique which outperforms classical ones in the next chapter.

Chapter 5

Quantile Propagation for Wasserstein-Approximate Gaussian Processes

Approximate inference techniques are the cornerstone of probabilistic methods based on Gaussian process (GP) priors. Despite this, most work approximately optimizes standard divergence measures such as the Kullback-Leibler (KL) divergence, which lack the basic desiderata for the task at hand, while chiefly offering merely technical convenience. In this chapter, we develop an efficient approximate Bayesian scheme that minimizes a specific class of Wasserstein distances (WDs), which we refer to as the L_2 WD. Our method overcomes some of the shortcomings of the KL divergence for approximate inference with GP models. The contents of this chapter is organized as below:

(i) In Section 5.1, we first introduce the background and motivation of the new work.

(ii) In Section 5.2, we develop quantile propagation (QP), an approximate inference algorithm for models with GP priors and factorized likelihoods. Like the expectation propagation (EP) algorithm, QP does not directly minimize global distances between high-dimensional distributions. Instead, QP estimates a fully coupled Gaussian posterior by iteratively minimizing *local* divergences between two particular marginal distributions. As these marginals are univariate, QP boils down to an iterative quantile function matching procedure (rather than moment matching as in EP) — hence we term our inference scheme *quantile propagation*. We derive the updates for the approximate likelihood terms and show that while the QP mean estimates match those of EP, the variance estimates are lower for QP.

(iii) In Section 5.3, we show that like EP, QP satisfies the locality property, meaning that it is sufficient to employ *univariate* approximate likelihood terms, and that the updates can thereby be performed efficiently using only the marginal distributions. Consequently, although our method employs a more complex divergence than that of EP (L_2 WD vs KL), it has the same computational complexity, after the precomputation of certain (data independent) lookup

tables.

(iv) In Section 5.4, we employ eight real-world datasets and compare our method to EP and variational Bayes (VB) on the tasks of binary classification and Poisson regression. We find that in terms of predictive accuracy, QP performs similarly to EP but is superior to VB. In terms of predictive uncertainty, however, we find QP superior to both EP and VB, thereby supporting our claim that QP alleviates variance over-estimation associated with the KL divergence when approximating short-tailed distributions [Minka, 2005; Jylänki et al., 2011; Heess et al., 2013].

5.1 Introduction

Gaussian process (GP) models have attracted the attention of the machine learning community due to their flexibility and their capacity to measure uncertainty. They have been widely applied to learning tasks such as regression [Matheron, 1963], classification [Williams and Barber, 1998; Hensman et al., 2015] and stochastic point process modeling [Møller et al., 1998; **Zhang** et al., 2019]. However, exact Bayesian inference for GP models is intractable for all but the Gaussian likelihood function. To address this issue, various approximate Bayesian inference methods have been proposed, such as Markov Chain Monte Carlo [MCMC, see *e.g.* Neal, 1997], the Laplace approximation [Williams and Barber, 1998], variational inference [Jordan et al., 1999; Opper and Archambeau, 2009] and expectation propagation [Opper and Winther, 2000; Minka, 2001c].

The existing approach most relevant to this work is expectation propagation (EP), which approximates each non-Gaussian likelihood term with a Gaussian by iteratively minimizing a set of local forward Kullback-Leibler (KL) divergences. As shown by Gelman et al. [2017], EP can scale to very large datasets. However, EP is not guaranteed to converge, and is known to over-estimate posterior variances [Minka, 2005; Jylänki et al., 2011; Heess et al., 2013] when approximating a short-tailed distribution. By over-estimation, we mean that the approximate variances are larger than the true variances so that more distribution mass lies in the *ineffective* domain. Interestingly, many popular likelihoods for GPs results in short-tailed posterior distributions, such as Heaviside and probit likelihoods for GP classification and Laplacian, Student's t and Poisson likelihoods for GP regression.

The tendency to over-estimate posterior variances is an inherent drawback of the forward KL divergence for approximate Bayesian inference. More generally, several authors have pointed out that the KL divergence can yield undesirable results such as (but not limited to) over-dispersed or under-dispersed posteriors [Dieng et al., 2017; Li and Turner, 2016; Hensman et al., 2014].

As an alternative to the KL, optimal transport metrics—such as the Wasserstein distance [WD, Villani, 2008, §6]—have seen a recent boost of attention. The WD is a natural distance between two distributions, and has been successfully employed in tasks such as image retrieval [Rubner et al., 2000], text classification [Huang et al., 2016] and image fusion [Courty et al., 2016]. Recent work has begun to employ the

WD for inference, as in Wasserstein generative adversarial networks [Arjovsky et al., 2017], Wasserstein variational inference [Ambrogioni et al., 2018] and Wasserstein auto-encoders [Tolstikhin et al., 2017]. In contrast to the KL divergence, the WD is computationally challenging [Cuturi, 2013], especially in high dimensions [Bonneel et al., 2015], in spite of its intuitive formulation and excellent performance.

5.2 Quantile Propagation

We now propose our new approximation algorithm which, as summarized in Algorithm 2 (Appendix), employs an L_2 WD based projection rather than the forward KL divergence projection of EP. Although QP employs a more complex divergence, it has the same computational complexity as EP, with the following caveat. To match the speed of EP, it is necessary to precompute sets of (data independent) lookup tables. Once precomputed, the resulting updates are only a constant factor slower than EP — a modest price to pay for optimizing a divergence which is challenging *even to evaluate*. Appendix C.10 provides further details on the precomputation and use of these tables.

As stated in Proposition 1, minimizing $\mathrm{W}_2^2(\widetilde{q}(f_i), \mathcal{N}(f_i))$ is equivalent to minimizing the L_2 distance between quantile functions of $\widetilde{q}(f_i)$ and $\mathcal{N}(f_i)$, so we refer to our method as quantile propagation (QP). This section focuses on deriving local updates for the site functions and analyzing their relationships with those of EP. Later in Section 5.3, we show the locality property of QP, meaning that the site function $t(f)$ has a univariate parameterization and so the local update can be efficiently performed using marginals only.

5.2.1 Convexity of L_p Wasserstein Distance

We first show $\mathrm{W}_p^p(\widetilde{q}(f), \mathcal{N}(f|\mu, \sigma^2))$ to be strictly convex in μ and σ. Formally, we have:

Theorem 2. *Given two probability measures in $\mathcal{M}_+^1(\mathbb{R})$: a Gaussian $\mathcal{N}(\mu, \sigma^2)$ with mean μ and standard deviation $\sigma > 0$, and an arbitrary measure $\widetilde{q}$, $W_p^p(\widetilde{q}, \mathcal{N})$ is strictly convex in μ and σ.*

Proof. See Appendix C.4. □

5.2.2 Minimization of L_2 WD

An advantage of using the L_p WD with $p = 2$, rather than other choices of p, is the computational efficiency it admits in the local updates. As we show in this section, optimizing the L_2 WD yields neat analytical updates of the optimal $\mu^\star$ and $\sigma^\star$ that require only univariate integration and the CDF of $\widetilde{q}(f)$. In contrast, other L_p WDs lack convenient analytical expressions. Nonetheless, other L_p WDs may have useful properties, the study of which we leave to future work.

The optimal parameters $\mu^\star$ and $\sigma^\star$ corresponding to the L_2 WD $\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu, \sigma^2))$ can be obtained using Proposition 1. Specifically, we employ the quantile function reformulation of $\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu, \sigma^2))$, and zero its derivatives w.r.t. μ and σ. The results provided below are derived in section C.1:

$$\begin{aligned}\mu^\star &= \mu_{\widetilde{q}},\\ \sigma^\star &= \sqrt{2}\int_0^1 F_{\widetilde{q}}^{-1}(y)\mathrm{erf}^{-1}(2y-1)\,\mathrm{d}y = 1/\sqrt{2\pi}\int_{-\infty}^{\infty} e^{-[\mathrm{erf}^{-1}(2F_{\widetilde{q}}(f)-1)]^2}\,\mathrm{d}f. \end{aligned} \tag{5.1}$$

Interestingly, the update for μ matches that of EP, namely the expectation under $\widetilde{q}$. However, for the standard deviation we have the difficulty of deriving the CDF $F_{\widetilde{q}}$. If a closed form expression is available, we can apply numerical integration to compute the optimal standard deviation; otherwise, we may use sampling based methods to approximate it. In our experiments we employ the former.

5.2.3 Properties of the Variance Update

Given the update equations in the previous section, here we show that the standard deviation estimate of QP, denoted as σ_{QP}, is less or equal to that of EP, denoted as σ_{EP}, when projecting the same tilted distribution to the Gaussian space.

Theorem 3. *The variances of the Gaussian approximation to a univariate tilted distribution $\widetilde{q}(f)$ as estimated by QP and EP satisfy $\sigma_{QP}^2 \leq \sigma_{EP}^2$.*

Proof. See Appendix C.5. □

Corollary 3.1. *The variances of the site functions updated by EP and QP satisfy: $\widetilde{\sigma}_{QP}^2 \leq \widetilde{\sigma}_{EP}^2$, and the variances of the approximate posterior marginals satisfy $\sigma_{q,QP}^2 \leq \sigma_{q,EP}^2$.*

Proof. Since the cavity distribution is unchanged, we can calculate the variance of the site function as per Equation (2.11) and conclude that the variance of the site function also satisfies $\widetilde{\sigma}_{\mathrm{QP}}^2 \leq \widetilde{\sigma}_{\mathrm{EP}}^2$. Moreover as per the definition of the cavity distribution in Section 2.2.5, the approximate marginal distribution is proportional to the product of the cavity distribution and the site function $q(f_i) \propto q^{\setminus i}(f_i)t(f_i)$, which are two Gaussian distributions. By the product of Gaussians formula (Equation (2.11)), we know the variance of $q(f_i)$ estimated by EP as $\sigma_{q,\mathrm{EP}}^2 = (\widetilde{\sigma}_{\mathrm{EP}}^{-2} + \sigma_{\setminus i}^{-2})^{-1} = \sigma_{\mathrm{EP}}^2$ and similarly $\sigma_{q,\mathrm{QP}}^2 = \sigma_{\mathrm{QP}}^2$, where σ_{EP}^2 and σ_{QP}^2 are defined in Theorem 3. Thus, there is $\sigma_{q,\mathrm{QP}}^2 \leq \sigma_{q,\mathrm{EP}}^2$. □

Corollary 3.2. *The predictive variances of latent functions at $\boldsymbol{x}_*$ by EP and QP satisfy: $\sigma_{QP}^2(f(\boldsymbol{x}_*)) \leq \sigma_{EP}^2(f(\boldsymbol{x}_*))$.*

Proof. The predictive variance of the latent function was analyzed in [Rasmussen and Williams, 2005, Equation (3.61)]: $\sigma^2(f_*) = k_* - \boldsymbol{k}_*^{\mathrm{T}}(K + \widetilde{\Sigma})^{-1}\boldsymbol{k}_*$,

where we define $f_* = f(\boldsymbol{x}_*)$ and $k_* = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$, and let $\boldsymbol{k}_* = (k(\boldsymbol{x}_*, \boldsymbol{x}_i))_{i=1}^N$ be the (column) covariance vector between the test data $\boldsymbol{x}_*$ and the training data $\{\boldsymbol{x}_i\}_{i=1}^N$.

After updating parameters of the site function $t_i(f_i)$, the predictive variance can be written as (details in Appendix C.9):

$$\sigma^2_{\text{new}}(f_*) = k_* - \boldsymbol{k}_*^{\mathrm{T}} A \boldsymbol{k}_* + \boldsymbol{k}_*^{\mathrm{T}} \boldsymbol{s}_i \boldsymbol{s}_i^{\mathrm{T}} \boldsymbol{k}_* / [(\widetilde{\sigma}^2_{i,\text{new}} - \widetilde{\sigma}_i^2)^{-1} + A_{ii}],$$

where $\widetilde{\sigma}^2_{i,\text{new}}$ is the site variance updated by EP or QP, $A = (K + \widetilde{\Sigma})^{-1}$ and $\boldsymbol{s}_i$ is the i's column of A. Since $\widetilde{\sigma}^2_{i,\text{QP}} \leq \widetilde{\sigma}^2_{i,\text{EP}}$, we have $\sigma^2_{\text{QP}}(f_*) \leq \sigma^2_{\text{EP}}(f_*)$. □

Remark. *We compared variance estimates of EP and QP assuming the same cavity distribution. Proving analogous statements for the fixed points of the EP and QP algorithms is more challenging, however, and we leave this to future work, while providing empirical support for these analogous statements in Figure 5.1(a) and Figure 5.1(b).*

5.3 Locality Property

In this section we detail the central result on which our QP algorithm is based upon, which we refer to as the *locality property*. That is, the optimal site function t_i is defined only in terms of the single corresponding latent variable f_i, and thereby and similarly to EP, it admits a simple and efficient sequential update of each individual site approximation.

5.3.1 Review: Locality Property of EP

We provide a brief review of the locality property of EP for GP models; for more details see Seeger [2005]. We begin by defining the general site function $t_i(\boldsymbol{f})$ in terms of all of the latent variables, and the cavity and the tilted distributions as $q^{\backslash i}(\boldsymbol{f}) \propto p(\boldsymbol{f}) \prod_{j \neq i} \widetilde{t}_j(\boldsymbol{f})$ and $\widetilde{q}(\boldsymbol{f}) \propto q^{\backslash i}(\boldsymbol{f}) p(y_i|f_i)$, respectively. To update $t_i(\boldsymbol{f})$, EP matches a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{f})$ to $\widetilde{q}(\boldsymbol{f})$ by minimizing the KL divergence $\mathrm{KL}(\widetilde{q} \| \mathcal{N})$, which is further rewritten as (see details in Appendix C.6.1):

$$\mathrm{KL}(\widetilde{q} \| \mathcal{N}) = \mathrm{KL}(\widetilde{q}_i \| \mathcal{N}_i) + \mathbb{E}_{\widetilde{q}_i}\Big[\mathrm{KL}\big(q^{\backslash i}_{\backslash i|i} \| \mathcal{N}_{\backslash i|i}\big)\Big], \tag{5.2}$$

where and hereinafter, $\backslash i|i$ denotes the conditional distribution of $\boldsymbol{f}_{\backslash i}$ (taking f_i out of $\boldsymbol{f}$) given f_i, namely, $q^{\backslash i}_{\backslash i|i} = q^{\backslash i}(\boldsymbol{f}_{\backslash i}|f_i)$ and $\mathcal{N}_{\backslash i|i} = \mathcal{N}(\boldsymbol{f}_{\backslash i}|f_i)$. Note that $q^{\backslash i}_{\backslash i|i}$ and $\mathcal{N}_{\backslash i|i}$ in the second term in Equation (5.2) are both Gaussian, and so setting them equal to one another causes that term to vanish. Furthermore, it is well known that the term $\mathrm{KL}(\widetilde{q}_i \| \mathcal{N}_i)$ is minimized w.r.t. the parameters of $\mathcal{N}_i$ by matching the first and second moments of $\widetilde{q}_i$ and $\mathcal{N}_i$. Finally, according to the usual EP logic, we recover the site function $t_i(\boldsymbol{f})$ by dividing the optimal Gaussian $\mathcal{N}(\boldsymbol{f})$ by the cavity $q^{\backslash i}(\boldsymbol{f})$:

$$t_i(\boldsymbol{f}) \propto \mathcal{N}(\boldsymbol{f}) / q^{\backslash i}(\boldsymbol{f}) = \cancel{\mathcal{N}(\boldsymbol{f}_{\backslash i}|f_i)} \mathcal{N}(f_i) / (\cancel{q^{\backslash i}(\boldsymbol{f}_{\backslash i}|f_i)} q^{\backslash i}(f_i)) = \mathcal{N}(f_i) / q^{\backslash i}(f_i). \tag{5.3}$$

Here we can see the optimal site function $t_i(f_i)$ relies solely on the local latent variable f_i, so it is sufficient to assume a univariate expression for site functions. Besides, the

site function can be efficiently updated by using the marginals $\widetilde{q}(f_i)$ and $\mathcal{N}(f_i)$ only, namely, $t_i(f_i) \propto \left(\min_{\mathcal{N}_i} \mathrm{KL}(\widetilde{q}_i \| \mathcal{N}_i) \right) / q^{\setminus i}(f_i) \big)$.

5.3.2 Locality Property of QP

This section proves the locality property of QP, which turns out to be rather more involved to show than is the case for EP. We first prove the following theorem, and then follow the same procedure as for EP (Equation (5.3)).

Theorem 4. *Minimization of* $\mathrm{W}_2^2(\widetilde{q}(\boldsymbol{f}), \mathcal{N}(\boldsymbol{f}))$ *w.r.t.* $\mathcal{N}(\boldsymbol{f})$ *results in* $q^{\setminus i}(\boldsymbol{f}_{\setminus i} | f_i) = \mathcal{N}(\boldsymbol{f}_{\setminus i} | f_i)$.

Proof. See Appendix C.6. □

Theorem 5 (Locality Property of QP)**.** *For GP models with factorized likelihoods, QP requires only univariate site functions, and so yields efficient updates using only marginal distributions.*

Proof. We apply the same steps as in Equation (5.3) for the EP case to QP and we conclude that the site function $t_i(f_i) \propto \mathcal{N}(f_i) / q^{\setminus i}(f_i)$ relies solely on the local latent variable f_i. And as per Equation (C.15) (Appendix C.6), $\mathcal{N}(f_i)$ is estimated by $\min_{\mathcal{N}_i} \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i)$, so the local update only uses marginals and can perform efficiently. □

Benefits of the Locality Property. The locality property admits an analytically economic form for the site function $t_i(f_i)$, requiring a parameterization that depends on a single latent variable. In addition, this also yields a significant reduction in the computational complexity, as only marginals are involved in each local update. In contrast, if QP (or EP) had no such a locality property, estimating the mean and the variance would involve integrals w.r.t. high-dimensional distributions, with a significantly higher computational cost should closed form expressions be unavailable.

5.4 Experiments

In this section, we compare the QP, EP and variational Bayes [VB, Opper and Archambeau, 2009] algorithms for binary classification and Poisson regression. The experiments employ eight real world datasets and aim to compare relative accuracy of the three methods, rather than optimizing the absolute performance. The implementations of EP and VB in Python are publicly available [GPy, since 2012], and our implementation of QP is based on that of EP. For both EP and QP, we stop local updates, i.e., , the inner loop in Algorithm 2 (Appendix), when the root mean squared change in parameters is less than 10^{-6}. In the outer loop, the GP hyper-parameters are optimized by L-BFGS-B [Byrd et al., 1995] with a maximum of 10^3 iterations and a relative tolerance of 10^{-9} for the function value. VB is also optimized by L-BFGS-B with the same configuration. Parameters shared by the three methods are initialized to be the same.

5.4.1 Binary Classification

Benchmark Data. We perform binary classification experiments on the five real world datasets employed by Kuss and Rasmussen [2005]: Ionosphere (IonoS), Wisconsin Breast Cancer, Sonar [Dua and Graff, 2017], Leptograpsus Crabs and Pima Indians Diabetes [Ripley, 1996]. We use two additional UCI datasets as further evidence: Glass and Wine [Dua and Graff, 2017]. As the Wine dataset has three classes, we conduct binary classification experiments on all pairs of classes. We summarize the dataset size and data dimensions in Table 5.1.

Table 5.1: Results on benchmark datasets. The first three columns give dataset names, the number of instances m and the number of features n. The table records the test errors (TEs) and the negative test log-likelihoods (NTLLs). The top section is on the benchmark datasets employed by Kuss and Rasmussen [2005] and the middle section uses additional datasets. The bottom section shows Poisson regression results. * indicates that QP outperforms EP in more than 90% of experiments *consistently*.

			TE ($\times 10^{-2}$)			NTLL($\times 10^{-3}$)		
Data	m	n	EP	QP	VB	EP	QP	VB
IonoS	351	34	$\mathbf{7.9}_{\pm\mathbf{0.5}}$	$\mathbf{7.9}_{\pm\mathbf{0.5}}$	$18.9_{\pm 6.9}$	$\mathbf{215.9}_{\pm\mathbf{8.4}}$	$\mathbf{215.9}_{\pm\mathbf{8.5}}$	$337.4_{\pm 70.8}$
Cancer	683	9	$3.2_{\pm 0.2}$	$3.2_{\pm 0.2}$	$\mathbf{3.1}_{\pm\mathbf{0.2}}$	$\mathbf{88.2}_{\pm\mathbf{3.1}}$	$\mathbf{88.2}^{*}_{\pm\mathbf{3.1}}$	$88.9_{\pm 19.1}$
Pima	732	7	$\mathbf{20.3}_{\pm\mathbf{1.0}}$	$\mathbf{20.3}_{\pm\mathbf{1.0}}$	$21.9_{\pm 0.4}$	$424.7_{\pm 13.0}$	$\mathbf{424.0}^{*}_{\pm\mathbf{13.2}}$	$450.3_{\pm 2.6}$
Crabs	200	7	$\mathbf{2.7}_{\pm\mathbf{0.5}}$	$\mathbf{2.7}_{\pm\mathbf{0.5}}$	$3.7_{\pm 0.7}$	$64.4_{\pm 8.2}$	$\mathbf{64.3}_{\pm\mathbf{8.4}}$	$164.7_{\pm 7.5}$
Sonar	208	60	$\mathbf{14.0}_{\pm\mathbf{1.1}}$	$\mathbf{14.0}_{\pm\mathbf{1.1}}$	$25.7_{\pm 3.9}$	$306.7_{\pm 10.8}$	$\mathbf{306.2}^{*}_{\pm\mathbf{10.9}}$	$693.1_{\pm 0.0}$
Glass	214	10	$1.1_{\pm 0.4}$	$\mathbf{1.0}_{\pm\mathbf{0.4}}$	$2.6_{\pm 0.5}$	$29.5_{\pm 5.4}$	$\mathbf{29.0}^{*}_{\pm\mathbf{5.5}}$	$79.5_{\pm 6.3}$
Wine1	130	13	$\mathbf{1.5}_{\pm\mathbf{0.5}}$	$\mathbf{1.5}_{\pm\mathbf{0.5}}$	$1.7_{\pm 0.6}$	$48.0_{\pm 3.4}$	$\mathbf{47.4}^{*}_{\pm\mathbf{3.4}}$	$83.9_{\pm 5.2}$
Wine2	107	13	$\mathbf{0.0}_{\pm\mathbf{0.0}}$	$\mathbf{0.0}_{\pm\mathbf{0.0}}$	$\mathbf{0.0}_{\pm\mathbf{0.0}}$	$18.0_{\pm 1.2}$	$\mathbf{17.8}^{*}_{\pm\mathbf{1.2}}$	$26.7_{\pm 1.9}$
Wine3	119	13	$2.0_{\pm 1.0}$	$2.0_{\pm 1.0}$	$\mathbf{1.2}_{\pm\mathbf{0.7}}$	$52.1_{\pm 5.6}$	$\mathbf{51.8}^{*}_{\pm\mathbf{5.6}}$	$69.4_{\pm 5.0}$
Mining	112	1	$\mathbf{118.6}_{\pm\mathbf{27.0}}$	$\mathbf{118.6}_{\pm\mathbf{27.0}}$	$170.3_{\pm 15.9}$	$1606.8_{\pm 116.3}$	$\mathbf{1606.5}_{\pm\mathbf{116.3}}$	$2007.3_{\pm 119.8}$

Note: Wine1: Class 1 vs. 2. Wine2: Class 1 vs. 3. Wine3: Class 2 vs. 3.

Prediction. We predict the test labels using models optimized by EP, QP and VB on the training data. For a test input $\boldsymbol{x}_*$ with a binary target y_*, the approximate predictive distribution is written as: $q(y_*|\boldsymbol{x}_*) = \int_{-\infty}^{\infty} p(y_*|f_*)q(f_*)\,\mathrm{d}f_*$ where $f_* = f(\boldsymbol{x}_*)$ is the value of the latent function at $\boldsymbol{x}_*$. We use the probit likelihood for the binary classification task, which admits an analytical expression for the predictive distribution and results in a short-tailed posterior distribution. Correspondingly, the predicted label $\hat{y}_*$ is determined by thresholding the predictive probability at $1/2$.

Performance Evaluation. To evaluate the performance, we employ two measures: the test error (TE) and the negative test log-likelihood (NTLL). The TE and the NTLL quantify the prediction accuracy and uncertainty, respectively. Specifically, they are defined as $(\sum_{i=1}^{m} |y_{*,i} - \hat{y}_{*,i}|/2)/m$ and $-(\sum_{i=1}^{m} \log q(y_{*,i}|\boldsymbol{x}_{*,i}))/m$, respectively, for a set of test inputs $\{\boldsymbol{x}_{*,i}\}_{i=1}^{m}$, test labels $\{y_{*,i}\}_{i=1}^{m}$, and the predicted labels $\{\hat{y}_{*,i}\}_{i=1}^{m}$. Lower values indicate better performance for both measures.

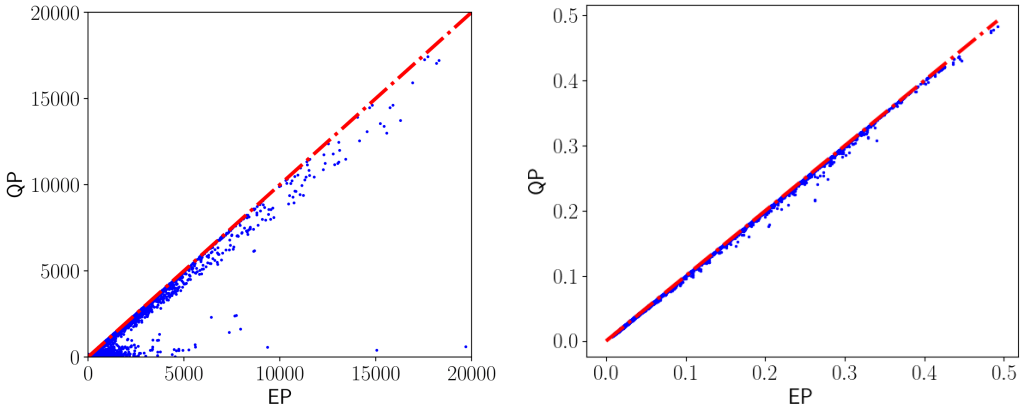
Experiment Settings. In the experiments, we randomly split each dataset into 10 folds, each time using 1 fold for testing and the other 9 folds for training, with features standardized to zero mean and unit standard deviation. We repeat this 100 times for a random seed ranging 0 through 99. As a result, there are a total of 1,000 experiments for each dataset. We report the average and the standard deviation of the above metrics over the 100 rounds.

Results. The evaluation results are summarized in Table 5.1. The top section presents the results on the datasets employed by Kuss and Rasmussen [2005], whose reported TEs match ours as expected. While QP and EP exhibit similar TEs on these datasets, QP is superior to EP in terms of the NTLL. VB under-performs both EP and QP on all datasets except `Cancer`. The middle section of Table 5.1 shows the results on additional datasets. The TEs are again similar for EP and QP, while QP has lower NTLLs. Again, VB performs worst among the three methods. To emphasize the difference between NTLLs of EP and QP, we mark with an asterisk those results in which QP outperforms EP in more than 90% of the experiments. Furthermore, we visualize the predictive variances of QP in comparison with those of EP in Figure 5.1(a)., which shows that the variances of QP are always less than or equal to those of EP, thereby providing empirical evidence of QP alleviating the over-estimation of predictive variances associated with the EP algorithm.

5.4.2 Poisson Regression

Data and Settings. We perform a Poisson regression experiment to further evaluate the performance of our method. The experiment employs the coal-mining disaster dataset [Jarrett, 1979] which has 190 data points indicating the time of fatal coal mining accidents in the United Kingdom from 1851 to 1962. To generate training and test sequences, we randomly assign every point of the original sequence to either a training or test sequence with equal probability, and this is repeated 200 times (random seeds $0, \cdots, 199$), resulting in 200 pairs of training and test sequences. We use the TE and the NTLL to evaluate the performance of the model on the test dataset. The NTLL has the same expression as that of the Binary classification experiment, but with a different predictive distribution $q(y_*|\boldsymbol{x}_*)$. The TE is defined slightly differently as $(\sum_{i=1}^{m} |y_{*,i} - \hat{y}_{*,i}|)/m$. To make the rate parameter of the Poisson likelihood non-negative, we use the square link function [Flaxman et al., 2017; Walder and Bishop, 2017], and as a result, the likelihood becomes $p(y|f^2)$. We use this link function because it is more mathematically convenient than the exponential function: the EP and QP update formulas, and the predictive distribution $q(y_*|\boldsymbol{x}_*)$ are available in Appendices C.3.2 and C.8, respectively.

Results. The means and the standard deviations of the evaluation results are reported in the last row of Table 5.1. Compared with EP, QP yields lower NTLL, which implies a better fitting performance of QP to the test sequences. We also provide the predictive variances in Figure 5.1(b)., in the variance of QP is once again seen to be less than or equal to that of EP. This experiment further supports our claim that QP alleviates the problem with EP of over-estimation of the predictive variance.



(a) Binary Classification (b) Poisson Regression

Figure 5.1: A scatter plot of the predictive variances of latent functions on test data, for EP and QP. The diagonal dash line represents equivalence. We see that the predictive variance of QP is always less than or equal to that of EP.

Finally, once again we find that both EP and QP outperform VB.

5.5 Conclusions

We have proposed QP as the first efficient L_2-WD based approximate Bayesian inference method for Gaussian process models with factorized likelihoods. Algorithmically, QP is similar to EP but uses the L_2 WD instead of the forward KL divergence for estimation of the site functions. When the likelihood factors are approximated by a Gaussian form we show that QP matches quantile functions rather than moments as in EP. Furthermore, we show that QP has the same mean update but a smaller variance than that of EP, which in turn alleviates the over-estimation by EP of the posterior variance in practice. Crucially, QP has the same favorable locality property as EP, and thereby admits efficient updates. Our experiments on binary classification and Poisson regression have shown that QP can outperform both EP and variational Bayes. Approximate inference with WD is promising but hard to compute, especially for continuous multivariate distributions. We believe our work paves the way for further practical approaches to WD-based inference.

Limitations and Future Work. Although we have presented properties and advantages of our method, it is still worth pointing out its limitations. First, our method does not provide a methodology for hyper-parameter optimization that is consistent with our proposed WD minimization framework. Instead, for this purpose, we rely on optimization of EP's marginal likelihood. We believe this is one of the reasons for the small performance differences between QP and EP. Furthermore, the computational efficiency of our method comes at the price of additional memory requirements and the look-up tables may exhibit instabilities on high-dimensional data. To overcome

these limitations, future work will explore alternatives to hyper-parameter optimization, improvements on numerical computation under the current approach and a variety of WD distances under a similar algorithm framework.

Applications to Bayesian Hawkes Processes. Applying QP to Bayesian Hawkes processes is a challenging task. This goal first requires solving a less difficult problem: using the EP or QP algorithm for Bayesian Poisson processes. We didn't explore QP for Poisson processes but for Poisson regression. Even for EP, the application to Poisson processes is missing. The main reason is that Poisson processes involve generally analytically intractable integrals in the likelihood and therefore are hard to handle. Further to this problem, applying QP to Bayesian Hawkes processes will face new problems, such as a high computational burden.

In the next chapter, we study a simple and robust frequentist estimation framework beyond the Bayesian field. The framework builds on the kernel maximum moment restriction and is applicable to a wide range of models.

Chapter 6

Kernel Maximum Moment Restriction for Instrumental Variable Regression

In this chapter, we propose a simple estimation framework for conditional moment restriction (CMR) models. We focus on instrumental variable (IV) regression models and design the method based on the kernelized maximum moment restriction (MMR). We organize the contents as below:

(i) In Section 6.1, we introduce the background of IV regression and present an overview of our proposed method.
(ii) In Section 6.2, we elaborate the prerequisites of our method: the detailed settings of IV regression, generalized method of moment which is widely applied to IV regression estimation and related to our method, and the kernelized MMR framework which is formulated by maximizing the interaction between the residual and the instruments belonging to a unit ball in a reproducing kernel Hilbert space (RKHS).
(iii) In Section 6.3, we propose the simple framework for IV regression estimation based on the kernelized MMR. Different from two-step optimization in most of existing methods, our method reformulates the IV regression estimation as a single-step empirical risk minimization problem, where the risk depends on the reproducing kernel on the instrument and can be estimated by a U-statistic or V-statistic. We then present two practical algorithms by considering two modern machine learning models, neural networks and kernel machines, in the framework.
(iv) In Section 6.4, we present an efficient hyper-parameter selection procedure.
(v) In Section 6.5, we analyze consistency and asymptotic normality of our estimator in both parametric and non-parametric settings.
(vi) In Section 6.6, we demonstrate the advantages of our framework over existing ones using experiments on both synthetic and real-world data.

6.1 Introduction

Instrumental variables (IV) have become standard tools for economists, epidemiologists, and social scientists to uncover causal relationships from observational data [Angrist and Pischke, 2008; Klungel et al., 2015]. Randomization of treatments or policies has been perceived as the gold standard for such tasks, but is generally prohibitive in many real-world scenarios due to time constraints or ethical concerns. When treatment assignment is not randomized, it is generally impossible to discern between the causal effect of treatments and spurious correlations that are induced by unobserved factors. Instead, the use of IV enables the investigators to incorporate *natural* variation through an IV that is associated with the treatments, but not with the outcome variable, other than through its effect on the treatments. In economics, for instance, the season-of-birth was used as an IV to study the return from schooling, which measures causal effect of education on labor market earning [Card, 1999]. In genetic epidemiology, the idea to use genetic variants as IVs, known as Mendelian randomization, has also gained increasing popularity [Burgess et al., 2017b, 2020].

To overcome these drawbacks, we propose a simple framework which views nonlinear IV regression as an empirical risk minimization (ERM) problem with U-statistic or V-statistic. This framework is based on (i) the maximum moment restriction (MMR) by Muandet et al. [2020a], which develops an equivalent form of moment conditions by a reproducing kernel Hilbert space (RKHS), and (ii) a U/V-statistics approximation technique [Serfling, 1980] for the form. We call our framework **MMR-IV**. Based on this formulation, we can solve the nonlinear IV problem by a single-step optimization with an empirical risk.

Our MMR-IV framework has the following advantages:

(i) A closed-form solution of the optimization problem is available. Also, for neural networks, we can simply apply common algorithms such as SGD.
(ii) We can interpret MMR-IV as an analogy for Gaussian processes, and consequently, we develop an efficient hyper-parameter selection procedure.
(iii) We prove consistency and asymptotic normality of estimators by MMR-IV in both parametric and non-parametric settings.

All of these advantages come from the empirical risk minimization form of MMR-IV. To the best of our knowledge, we do not find any other method that can achieve all of these. Further, our experiments show MMR-IV has better performance with synthetic data, and it provides appropriate interpretation of real data.

6.2 Preliminaries

Generalized method of moment (GMM). The aforementioned conditions imply that $\mathbb{E}[\varepsilon \mid Z] = 0$ for P_Z-almost all z. This is a *conditional moment restriction* (CMR) which we can use to estimate f [Newey, 1993]. For any measurable function h, the CMR implies a continuum of unconditional moment restrictions [Lewis and Syrgkanis, 2018;

Bennett et al., 2019][1]:

$$\mathbb{E}[(Y - f(X))h(Z)] = 0. \tag{6.1}$$

That is, there exists an infinite number of moment conditions, each of which is indexed by the function h. As a result, learning with Equation (6.1) is challenging. Although the asymptotic efficiency of the estimator can in principle improve when we consider increasingly many moment conditions, it was observed that the excessive number of moments can be harmful in practice [Andersen and Sorensen, 1996] because its finite-sample bias increases with the number of moment conditions [Newey and Smith, 2004]. Hence, traditional works in econometrics often select a finite number of moment conditions for estimation based on the generalized method of moments (GMM) [Hansen, 1982; Hall et al., 2005]. Unfortunately, an adhoc choice of moments can potentially lead to a loss of efficiency or even a loss of identification [Dominguez and Lobato, 2004]. For this reason, subsequent works advocate an incorporation of all moment restrictions simultaneously in different ways such as the method of sieves [de Jong, 1996; Donald et al., 2003] and a continuum of moment restrictions [Carrasco and Florens, 2000; Carrasco et al., 2007; Carrasco, 2012; Carrasco and Florens, 2014], among others.

One of the key questions in econometrics is which moment conditions should be used as a basis for estimating the function f [Donald and Newey, 2001; Hall, 2005]. In this work, we show that, for the purpose of estimating f, it is sufficient to restrict h to be within a unit ball of a RKHS of real-valued functions on Z.

6.2.1 Maximum Moment Restriction

Throughout this paper, we assume that h is a real-valued function on $\mathcal{Z}$ which belongs to a RKHS $\mathcal{H}_k$ endowed with a reproducing kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. The RKHS $\mathcal{H}_k$ satisfies two important properties: (i) for all $z \in \mathcal{Z}$ and $h \in \mathcal{H}_k$, we have $k(z, \cdot) \in \mathcal{H}_k$ and (ii) (reproducing property of $\mathcal{H}_k$) $h(z) = \langle h, k(z, \cdot)\rangle_{\mathcal{H}_k}$ where $k(z, \cdot)$ is a function of the second argument. Furthermore, we define $\Phi_k(z)$ as a *canonical* feature map of z in $\mathcal{H}_k$. It follows from the reproducing property that $k(z, z') = \langle \Phi_k(z), \Phi_k(z')\rangle_{\mathcal{H}_k}$ for any $z, z' \in \mathcal{Z}$, i.e., an inner product between the feature maps of z and z' can be evaluated implicitly through the kernel evaluation. Every positive definite kernel k uniquely determines the RKHS for which k is a reproducing kernel [Aronszajn, 1950]. For detailed exposition on kernel methods, see, e.g., Schölkopf and Smola [2002], Berlinet and Thomas-Agnan [2004], and Muandet et al. [2017].

Instead of (6.1), we define a risk in terms of a maximum moment restriction (MMR) [Muandet et al., 2020a]:

$$R_k(f) := \sup_{h \in \mathcal{H}_k, \|h\| \leq 1} (\mathbb{E}[(Y - f(X))h(Z)])^2. \tag{6.2}$$

That is, instead of considering all measurable functions h as instruments, we only restrict to functions that lie within a unit ball of the RKHS $\mathcal{H}_k$. The risk is then defined

[1] Also, IV assumptions imply ε is independent of Z, i.e., $\mathbb{E}[(Y - f(X))h(Z)] = 0$ for all measurable h.

as a maximum value of the moment restriction with respect to this function class. The benefits of our formulation are two-fold. First, it is computationally intractable to learn f from (6.1) using *all* measurable functions as instruments. By restricting the function class to a unit ball of the RKHS, the problem becomes computationally tractable, as will be shown in Lemma 1 below. Second, this restriction still preserves the consistency of parameter estimated using $R_k(f)$. In some sense, the RKHS is a sufficient class of instruments for the nonlinear IV problem (cf. Theorem 6).

A crucial insight for our approach is that the population risk $R_k(f)$ has an analytic solution.

Lemma 1 (Muandet et al. [2020a], Thm 3.3)**.** *Assume that* $\mathbb{E}[(Y - f(X))^2 k(Z, Z)] < \infty$. *Then, we have*

$$R_k(f) = \mathbb{E}[(Y - f(X))(Y' - f(X'))k(Z, Z')] \tag{6.3}$$

where (X', Y', Z') *is an independent copy of* (X, Y, Z).

We assume throughout that the reproducing kernel k is integrally strictly positive definite (ISPD).

Assumption 1. *The kernel* k *is continuous, bounded (i.e.,* $\sup_{z \in \mathcal{Z}} \sqrt{k(z, z)} < \infty$*) and satisfies the condition of integrally strictly positive definite (ISPD) kernels, i.e., for any function* g *that satisfies* $0 < \|g\|_2^2 < \infty$*, we have* $\int\int_{\mathcal{Z}} g(z)k(z, z')g(z') \, \mathrm{d}z \, \mathrm{d}z' > 0$.

The assumption is also related to the notion of characteristic and universal kernels; see, e.g., Simon-Gabriel and Schölkopf [2018]. More details on ISPD kernels are given in Appendix D.1. We further assume the identification for the minimizer of $R_k(f)$.

Assumption 2. *Consider the function space* $\mathcal{F}$ *and* $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} R_k(f)$. *Then for any* $g \in \mathcal{F}$ *with* $|\mathbb{E}[g(X)]| < \infty$, $\mathbb{E}[g(X) - f^*(X) \mid Z] = 0$ *implies* $g = f^*$.

A sufficient condition for identification follows from the *completeness* property of X for Z [D'Haultfoeuille, 2011], e.g., the conditional distribution of X given Z belongs to the exponential family [Newey and Powell, 2003]. Provided identification, it is straightforward to obtain consistency.

6.3 Our Method

We propose to learn f by minimizing $R_k(f)$ in (6.3). To this end, we define an optimal function f^* as a minimizer of the above population risk with respect to a function class $\mathcal{F}$ of real-valued functions on $\mathcal{X}$, i.e.,

$$f^* \in \operatorname{argmin}_{f \in \mathcal{F}} \; R_k(f).$$

It is instructive to note that population risk R_k depends on the choice of the kernel k. Based on Assumption 1 and Lemma 1, we obtain the following result, which is a special case of Muandet et al. [2020a, Theorem 3.2], showing that $R_k(f) = 0$ if and only if f satisfies the original conditional moment restriction (see Appendix D.2.2 for the proof).

Theorem 6. *Assume that the kernel k is integrally strictly positive definite (ISPD). Then, for any real-valued measurable function f, $R_k(f) = 0$ if and only if $\mathbb{E}[Y - f(X) \mid z] = 0$ for P_Z-almost all z.*

Theorem 6 holds as long as the kernel k belongs to a class of ISPD kernels. Hence, it allows for more flexibility in terms of the kernel choice. Moreover, it is not difficult to show that $R_k(f)$ is strictly convex in f (see Appendix D.2.3).

6.3.1 Empirical Risk Minimization with U, V-Statistics

The previous results pave the way for an empirical risk minimization (ERM) framework [Vapnik, 1998] to be used in our work. That is, given an i.i.d. sample $\{(x_i, y_i, z_i)\}_{i=1}^n \sim P^n(X, Y, Z)$ of size n, an empirical estimate of the risk $R_k(f)$ can be obtained as

$$\widehat{R}_U(f) := \sum_{1 \le i \ne j \le n}^n \frac{(y_i - f(x_i))(y_j - f(x_j))k(z_i, z_j)}{n(n-1)}, \tag{6.4}$$

which is in the form of U-statistic [Serfling, 1980, Section 5]. Alternatively, an empirical risk based on V-statistic can also be used, i.e.,

$$\widehat{R}_V(f) := \sum_{i,j=1}^n \frac{(y_i - f(x_i))(y_j - f(x_j))k(z_i, z_j)}{n^2}. \tag{6.5}$$

Both forms of empirical risk can be used as a basis for a consistent estimation of f. The advantage of (6.4) is that it is a minimum-variance unbiased estimator with appealing asymptotic properties, whereas (6.5) is a biased estimator of the population risk (6.3), i.e., $\mathbb{E}[\widehat{R}_V] \ne R_k$. However, the estimator based on V-statistics employs a full sample and hence may yield better estimate of the risk than the U-statistic counterpart.

Let $\boldsymbol{x} := [x_1, \ldots, x_n]^\top$, $\boldsymbol{y} := [y_1, \ldots, y_n]^\top$ and $\boldsymbol{z} := [z_1, \ldots, z_n]^\top$ be column vectors. Let $K_{\mathbf{z}}$ be the kernel matrix $K(\boldsymbol{z}, \boldsymbol{z}) = [k(z_i, z_j)]_{ij}$ evaluated on the instruments $\mathbf{z}$. Then, both (6.4) and (6.5) can be rewritten as

$$\widehat{R}_{V(U)}(f) = (\boldsymbol{y} - f(\boldsymbol{x}))^\top W_{V(U)} (\boldsymbol{y} - f(\boldsymbol{x})), \tag{6.6}$$

where $f(\boldsymbol{x}) := [f(x_1), \ldots, f(x_n)]^\top$ and $W_{V(U)} \in \mathbb{R}^{n \times n}$ is a symmetric weight matrix that depends on the kernel matrix $K_{\boldsymbol{z}}$. Specifically, the weight matrix W corresponding to (6.4) is given by $W_U = (K_{\boldsymbol{z}} - \operatorname{diag}(k(z_1, z_1), \ldots, k(z_n, z_n)))/(n(n-1))$ where $\operatorname{diag}(a_1, \ldots, a_n)$ denotes an $n \times n$ diagonal matrix whose diagonal elements are $a_1, \ldots, a_n$. As shown in Appendix D.2.5, W_U is indefinite and may cause problematic inferences. The weight matrix W for (6.5) is given by $W_V := K_{\boldsymbol{z}}/n^2$ which is a positive definite matrix for the ISPD kernel k. Finally, our objective (6.6) also resembles the well-known generalized least regression with correlated noise [Kariya and Kurata, 2004, Chapter 2] where the covariance matrix is the Z-dependent invertible matrix $W_{V(U)}^{-1}$.

Based on (6.4) and (6.5), we estimate f^* by minimizing the *regularized* empirical risk over a function class $\mathcal{F}$:

$$\hat{f}_{V(U)} \in \operatorname{argmin}_{f \in \mathcal{F}} \ \widehat{R}_{V(U)}(f) + \lambda \Omega(f)$$

where $\lambda > 0$ is a regularization constant satisfying $\lim_{n\to\infty} \lambda = 0$, and $\Omega(f)$ is the regularizer. Since $\hat{f}_U$ and $\hat{f}_V$ minimizes objectives which are regularized U-statistic and V-statistic, they can be viewed a specific form of M-estimators; see, e.g., Van der Vaart [2000, Ch. 5]. In this work, we focus on the V-statistic empirical risk and provide practical algorithms when $\mathcal{F}$ is parametrized by deep neural networks (NNs) and an RKHS of real-valued functions.

Kernelized GMM. We may view the objective (6.6) from the GMM perspective [Hall, 2005]. The assumption that the instruments Z are exogenous implies that $\mathbb{E}[\Phi_k(Z)\varepsilon] = 0$ where Φ_k denotes the canonical feature map associated with the kernel k. This gives us an infinite number of moments, $g(f) = \Phi_k(Z)(Y - f(X))$. Hence, we can write the sample moments as $\hat{g}(f) = (1/n)\sum_{i=1}^n \Phi_k(z_i)(y_i - f(x_i))$. The intuition behind GMM is to choose a function f that sets these moment conditions as close to zero as possible, motivating the objective function $J(f) := \|\hat{g}(f)\|^2_{\mathcal{H}_k} = \langle \hat{g}(f), \hat{g}(f)\rangle_{\mathcal{H}_k} = \frac{1}{n^2}\sum_{i,j=1}^n (y_i - f(x_i))\langle \Phi_k(z_i), \Phi_k(z_j)\rangle_{\mathcal{H}_k}(y_j - f(x_j)) = (\boldsymbol{y} - f(\boldsymbol{x}))^\top W_V(\boldsymbol{y} - f(\boldsymbol{x}))$. Hence, our objective (6.6) defined using V-statistic is a special case of the GMM objective when the weighting matrix is the identity operator. Carrasco et al. [2007, Ch. 6] shows that the optimal weighting operator is given in terms of the inversed covariance operator.

6.3.2 Practical MMR-IV Algorithms

A workflow of our algorithm based on $\widehat{R}_V$ is summarized in Algorithm 1; we leave the $\widehat{R}_U$ based method to future work to solve the inference issues caused by indefinite W_U. We provide examples of the class $\mathcal{F}$ in both parametric and non-parametric settings below.

Deep neural networks. In the parametric setting, the function class $\mathcal{F}$ can often be expressed as $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^m$ denotes a parameter space. We consider a very common nonlinear model in machine learning $f(x) = W_0\Phi(x) + b_0$ where $\Phi : x \longmapsto \sigma_h(W_h\sigma_{h-1}(\cdots\sigma_1(W_1x)))$ denotes a nonlinear feature map of a depth-h NN. Here, W_i for $i = 1, \ldots, h$ are parameter matrices and each σ_i denotes the entry-wise activation function of the i-th layer. In this case, $\theta = (b_0, W_0, W_1, \ldots, W_h)$. As a result, we can rewrite $\hat{f}_V$ in terms of their parameters as

$$\hat{\theta}_V \in \arg\min_{\theta\in\Theta} \ \widehat{R}_V(f_\theta) + \lambda\|\theta\|_2^2$$

where $f_\theta \in \mathcal{F}_\Theta$. We denote $\theta^* \in \arg\min_{\theta\in\Theta} \ R_k(f_\theta)$. In what follows, we refer to this algorithm as **MMR-IV (NN)**; see Algorithm 4 in Appendix D.5.1.

Kernel machines. In a non-parametric setting, the function class $\mathcal{F}$ becomes an infinite dimensional space. In this work, we consider $\mathcal{F}$ to be an RKHS $\mathcal{H}_l$ of real-valued functions on $\mathcal{X}$ with a reproducing kernel $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then, the regularized solution can be obtained by $\arg\min_{f\in\mathcal{H}_l} \ \widehat{R}_V(f) + \lambda\|f\|^2_{\mathcal{H}_l}$. As per the representer theorem, any optimal $\hat{f}$ admits a form $\hat{f}(x) = \sum_{i=1}^n \alpha_i l(x, x_i)$ for some $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ [Schölkopf et al., 2001b], and based on this representation, we

Algorithm 1 MMR-IV

Input: Dataset $D = \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$, kernel function k with parameters θ_k, a function class $\mathcal{F}$, regularization functional $\Omega(\cdot)$, and regularization parameter λ

Output: The estimate of f^* in $\mathcal{F}$.

1: Compute the kernel matrix $K = k(\boldsymbol{z}, \boldsymbol{z}; \theta_k)$.
2: Define the residual $\boldsymbol{\varepsilon}(f) = \boldsymbol{y} - f(\boldsymbol{x})$.
3: $\hat{f}_\lambda \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \ \frac{1}{n^2} \boldsymbol{\varepsilon}(f)^\top K \boldsymbol{\varepsilon}(f) + \lambda \Omega(\|f\|_{\mathcal{F}})$
4: **return** $\hat{f}_\lambda$

rewrite the objective as

$$\hat{f}_V = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} (\boldsymbol{y} - L\boldsymbol{\alpha})^\top W_V (\boldsymbol{y} - L\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^\top L \boldsymbol{\alpha}, \tag{6.7}$$

where $L = [l(x_i, x_j)]_{ij}$ is the kernel matrix on $\boldsymbol{x}$. For U-statistic version, the quadratic program (6.7) substitutes indefinite W_U for W_V, so it may not be positive definite. The value of λ needs to be sufficiently large to ensure that (6.7) is definite. On the other hand, the V-statistic based estimate (6.7) is definite for all non-zero λ since W_V is positive semi-definite. Thus, the optimal $\widehat{\boldsymbol{\alpha}}$ can be obtained by solving the first-order stationary condition and if L is positive definite, the solution has a closed form expression, $\widehat{\boldsymbol{\alpha}} = (LW_V L + \lambda L)^{-1} LW_V \boldsymbol{y}$. Thus, we will focus on the V-statistic version in our experiments. In the following, we refer to this algorithm as **MMR-IV (RKHS)**.

Nyström approximation. The MMR-IV (RKHS) algorithm is computationally costly for large datasets as it requires a matrix inversion. To improve the scalability, we resort to Nyström approximation [Williams and Seeger, 2001] to accelerate the matrix inversion in $\widehat{\boldsymbol{\alpha}} = (LW_V L + \lambda L)^{-1} LW_V \boldsymbol{y}$. First, we randomly select a subset of $m (\ll n)$ samples from the original dataset and construct the corresponding sub-matrices of W_V, namely, W_{mm} and W_{nm} based on this subset. Second, let V and U be the eigenvalue vector and the eigenvector matrix of W_{mm}. Then, the Nyström approximation is obtained as $W_V \approx \widetilde{U} \widetilde{V} \widetilde{U}^\top$ where $\widetilde{U} := \sqrt{\frac{m}{n}} W_{nm} U V^{-1}$ and $\widetilde{V} := \frac{n}{m} V$. We finally apply the Woodbury formula [Flannery et al., 1992, p. 75] to obtain

$$\begin{aligned}(LW_V L + \lambda L)^{-1} LW_V &= L^{-1} (W_V + \lambda L^{-1})^{-1} W_V \\ &\approx \lambda^{-1} [I - \widetilde{U} (\lambda^{-1} \widetilde{U}^\top L \widetilde{U} + \widetilde{V}^{-1})^{-1} \widetilde{U}^\top \lambda^{-1} L] \widetilde{U} \widetilde{V} \widetilde{U}^\top.\end{aligned}$$

We will refer to this algorithm as **MMR-IV (Nyström)**; Algorithm 3 in Appendix D.5.1. The runtime complexity of this algorithm is $O(nm^2 + n^2)$.

6.4 Hyper-parameter Selection

We develop a convenient hyperparameter selection method for MMR-IV (Nyström) with our objective (6.7) and the V-statistic. This method benefits from the fact that our ERM form can be interpreted as a stochastic model with a Gaussian process, inspired

by Vehtari et al. [2016]. This approach is an analog of parameter selection in the ordinary kernel regression. However, we need to apply additional analysis due to the weight matrix W_V in (6.7) with V-statistics. Full details are provided in Appendix D.3.

Gaussian process (GP) interpretation. We start the GP interpretation by defining the energy-based likelihood $p(D|f(\boldsymbol{x})) \propto \exp(-\hat{R}_V(f)/2) \propto N(f(\boldsymbol{x})|\boldsymbol{y}, K_z^{-1})$, where $D \coloneqq \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$ denotes the dataset. Then we assign a GP prior to f, i.e., $f(\boldsymbol{x}) \sim \mathrm{GP}(\boldsymbol{0}, \delta l(\boldsymbol{x}, \boldsymbol{x}))$ with a real constant $\delta > 0$,s and the posterior distribution of $f(\boldsymbol{x})$ is straight-forwardly derived based on the Gaussian-like prior and likelihood,

$$p(f(\boldsymbol{x})|D) = N(f(\boldsymbol{x})|\boldsymbol{c}, C) \propto p(f(\boldsymbol{x}))p(D|f(\boldsymbol{x})),$$

where $\boldsymbol{c} = CK_{\boldsymbol{z}}\boldsymbol{y}$ and $C = (K_{\boldsymbol{z}} + (\delta L)^{-1})^{-1} \approx \delta[L - L\widetilde{U}(n^2\delta\widetilde{U}^\top L\widetilde{U} + \widetilde{V}^{-1})^{-1}\widetilde{U}(n^2\delta)L]$ by Nyström approximation . The connections between the GP model and the regularized $\widehat{R}_V(f)$ are elaborated in Appendix D.3 and summarized as Theorem 7. In a word, maximization of $p(f(\boldsymbol{x})|D)$ is alternative to minimization of (6.7) and the Bayesian inference is a substitute for the frequentist prediction.

Theorem 7. *Given* $\delta = (\lambda n^2)^{-1}$ *and* $\hat{f} = \mathrm{argmin}_{f \in \mathcal{H}_l}$ *Equation* (6.7)*, there are*
(i) $\mathrm{argmax}_{f(\boldsymbol{x})}\, p(f(\boldsymbol{x})|D) = \hat{f}(\boldsymbol{x})$*;*
(ii) prediction at $\boldsymbol{x}^*$*:* $\mathrm{argmax}_{f(\boldsymbol{x}_*)}\, p(f(\boldsymbol{x}_*)|D) = \hat{f}(\boldsymbol{x}_*)$*.*

Analytical cross-validation error. Now we derive the analytical error of the leave-M-out cross validation (LMOCV) from the perspective of the GP. We split the whole dataset D into training and development datasets, denoted as D_{tr} and $D_{de} \coloneqq \{\boldsymbol{x}_{de}, \boldsymbol{y}_{de}, \boldsymbol{z}_{de}\}$ respectively, where D_{de} has M triplets of data points. Given D_{tr}, the predictive probability on D_{de} can be obtained by Bayes' rules $p(f(\boldsymbol{x}_{de})|D_{tr}) \propto \frac{p(f(\boldsymbol{x}_{de})|D)}{p(D_{de}|f(\boldsymbol{x}_{de}))}$, where $p(f(\boldsymbol{x}_{de})|D) = N(\boldsymbol{c}_{de}, C_{de})$ with $\boldsymbol{c}_{de}$, C_{de} the mean and covariance of $f(\boldsymbol{x}_{de})$ in $p(f(\boldsymbol{x})|D)$ and $p(D_{de}|f(\boldsymbol{x}_{de})) \propto N(\boldsymbol{y}_{de}, K_{de}^{-1})$, $K_{de} = k(\boldsymbol{z}_{de}, \boldsymbol{z}_{de})$. By this result, the function estimated on the training set can be represented w.r.t. that on the whole dataset. It turns out that $p(f(\boldsymbol{x}_{de})|D_{tr}) = N(\boldsymbol{b}, B)$ where $B^{-1} = C_{de}^{-1} - K_{de}$ and $\boldsymbol{b} = B(C_{de}^{-1}\boldsymbol{c}_{de} - K_{de}\boldsymbol{y}_{de})$, and by (ii) in Theorem 7, the error of m repeated LMOCV is

$$\text{LMOCV Error} \coloneqq \textstyle\sum_{i=1}^{m} (r^{(i)})^\top K_{de}^{(i)} r^{(i)}, \tag{6.8}$$

where (i) denotes the experiment index and the residual $r^{(i)} \coloneqq \boldsymbol{b}^{(i)} - \boldsymbol{y}_M^{(i)} = (I - C_{de}^{(i)}K_{de}^{(i)})^{-1}(\boldsymbol{c}_{de}^{(i)} - \boldsymbol{y}_{de}^{(i)})$. The analytical error enables an efficient parameter selection with the CV procedure.

6.5 Consistency and Asymptotic Normality

We provide the consistency and asymptotic normality of $\widehat{f}_V$. We also develop the same results for $\widehat{f}_U$, but we defer them to the appendix due to space limitation. All proofs are presented in the appendix.

6.5.1 Consistency

We first show the consistency of $\hat{f}_V$, which depends on the uniform convergence of the risk functions. The result holds regardless of the shape of $\Omega(f)$, so we can utilize the regularization $\|\theta\|_2^2$ which is common for NN but non-convex in terms of f.

Theorem 8 (Consistency of $\hat{f}_V$)**.** *Assume that* $\mathbb{E}[|Y|^2] < \infty$, $\mathbb{E}[\sup_{f\in\mathcal{F}} |f(X)|^2] < \infty$, $\mathcal{F}$ *is compact, Assumption 1, 2 hold,* $\Omega(f)$ *is a bounded function and* $\lambda \xrightarrow{\mathrm{p}} 0$*. Then* $\hat{f}_V \xrightarrow{\mathrm{p}} f^*$*.*

If the $\Omega(f)$ is convex in f, the consistency can be obtained more easily by Newey and McFadden [1994, Theorem 2.7]. In this case, we can avoid several conditions. We provide an additional result with this setting in Appendix D.2.6.

6.5.2 Asymptotic Normality

We analyze asymptotic normality of the estimator $\hat{f}_V$, which is important to advanced statistical analysis such as tests. Here, we investigate two different cases: the estimator has finite- and infinite-dimension.

Finite-dimension case. We consider the $\hat{f}_V$ is characterized by a finite-dimensional parameter from a parameter space Θ. We rewrite the regularized V-statistic risk as a compact form $\widehat{R}_{V,\lambda}(f_\theta) := \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n h_\theta(u_i,u_j) + \lambda\Omega(\theta)$, $h_\theta(u_i,u_j) := (y_i - f(x_i))(y_j - f(x_j))k(z_i,z_j)$, and consider $R_k(f_\theta)$ is uniquely minimized at $\theta^* \in \Theta$.

Theorem 9 (Asymptotic normality of $\hat{\theta}_V$)**.** *Suppose that* f_θ *and* $\Omega(\theta)$ *are twice continuously differentiable about* θ*,* Θ *is compact,* $H = \mathbb{E}[\nabla_\theta^2 h_{\theta^*}(U,U')]$ *is non-singular,* $\mathbb{E}[|Y|^2] < \infty$*,* $\mathbb{E}\left[\sup_{\theta\in\Theta}|f_\theta(X)|^2\right] < \infty$*,* $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla_\theta f_\theta(X)\|_2^2\right] < \infty$*,* $\mathbb{E}\left[\sup_{\theta\in\Theta}\|\nabla_\theta^2 f_\theta(X)\|_F^2\right] < \infty$*,* $\sqrt{n}\lambda \xrightarrow{\mathrm{p}} 0$*,* $R_k(f_\theta)$ *is uniquely minimized at* θ^* *which is an interior point of* Θ*, and Assumption 1 holds. Then,* $\sqrt{n}(\hat{\theta}_V - \theta^*) \rightsquigarrow N(\mathbf{0}, \Sigma_V)$ *holds, where*

$$\Sigma_V = 4H^{-1}\mathrm{diag}(\mathbb{E}_U[\mathbb{E}^2_{U'}[h_{\theta^*}(U,U')]])H^{-1}$$

and $\rightsquigarrow$ *denotes a convergence in law.*

Infinite-dimension case. We show the asymptotic normality of an infinite-dimensional estimator $\hat{f}_V$. That is, we show that an error of $\hat{f}_V$ weakly converges to a Gaussian process that takes values in a function space $\mathcal{H}_l$. We set $\Omega(f) = \|f\|_{\mathcal{H}_l}^2$ and consider a minimizer: $f^*_{\lambda_0} \in \mathrm{argmin}_{f\in\mathcal{F}}\ R_k(f) + \lambda_0\|f\|_{\mathcal{H}_l}^2$ with arbitrary $\lambda_0 > 0$. We also define $\mathcal{N}(\varepsilon, \mathcal{H}, \|\cdot\|)$ as an ε-covering number of a set $\mathcal{H}$ in terms of $\|\cdot\|$. Then, we obtain the following:

Theorem 10. *Suppose Assumption 1 holds, l is a bounded kernel, k is a uniformly bounded function, and* $\lambda - \lambda_0 = o(n^{-1/2})$ *holds. Also, suppose that* $\mathcal{X}$*,* $\mathcal{Z}$*, and* $\mathcal{Y}$ *are compact spaces, and there exists* $s \in (0,2)$ *and a constant* $C_H > 0$ *such that* $\log\mathcal{N}(\varepsilon, \mathcal{H}_l, \|\cdot\|_{L^\infty}) \le C_H\varepsilon^{-s}$ *for any* $\varepsilon \in (0,1)$*. Then, there exists a Gaussian process* $\mathbb{G}_P^*$ *such that* $\sqrt{n}(\hat{f}_V - f^*_{\lambda_0}) \rightsquigarrow \mathbb{G}_P^*$ *in* $\mathcal{H}_l$*.*

This result allows statistical inference on functional estimators such as kernel machines. Although there are many conditions, all of them are valid for many well-known kernels; see Appendix D.2.11 for examples.

6.6 Experimental Results

We present the experimental results in a wide range of settings for IV estimation. Following Lewis and Syrgkanis [2018] and Bennett et al. [2019], we consider both low and high-dimensional scenarios. In our experiments, we compare our algorithms to the following baselines: (i) `DirectNN` (ii) `2SLS` (iii) `Poly2SLS` (iv) `DeepIV` [Hartford et al., 2017] (v) `KernelIV` [Singh et al., 2019] (vi) `GMM+NN` (vii) `AGMM` [Lewis and Syrgkanis, 2018] (viii) `DeepGMM` [Bennett et al., 2019]. (ix) `AGMM-K` [Dikkala et al., 2020]. We refer readers to Appendix D.5 for more experimental details.

6.6.1 Low-dimensional Scenarios

Following Bennett et al. [2019], we employ the following data generation process:

$$Y = f^*(X) + e + \delta, \quad X = Z_1 + e + \gamma,$$

where $Z := (Z_1, Z_2) \sim \text{Uniform}([-3,3]^2)$, $e \sim \mathcal{N}(0,1)$, and $\gamma, \delta \sim \mathcal{N}(0, 0.1^2)$. In words, Z is a two-dimensional IV, but only the first instrument Z_1 has an effect on X. The variable e is the confounding variable that creates the correlation between X and the residual $Y - f^*(X)$. We vary the true function f^* between the following cases to enrich the datasets: (i) `sin`: $f^*(x) = \sin(x)$. (ii) `step`: $f^*(x) = \mathbb{1}_{\{x \geq 0\}}$. (iii) `abs`: $f^*(x) = |x|$. (iv) `linear`: $f^*(x) = x$. We consider both small-sample ($n = 200$) and large-sample ($n = 2000$) regimes.

Table 6.1 reports the results for the large-sample regime (and Table D.1 in Appendix D.5 for the small-sample regime). Our findings are as follows: (i) `MMR-IV`s perform reasonably well in most cases. (ii) The linear methods (`2SLS` and `Poly2SLS`) perform best when the linearity assumption is satisfied. (iii) Some nonlinear but com-

Table 6.1: The mean square error (MSE) $\pm$ one standard deviation in the large-sample regime ($n = 2000$).

Algorithm	**True Function f^***			
	abs	linear	sin	step
`DirectNN`	.116 $\pm$ .000	.035 $\pm$ .000	.189 $\pm$ .000	.199 $\pm$ .000
`2SLS`	.522 $\pm$ .000	**.000 $\pm$ .000**	.254 $\pm$ .000	.050 $\pm$ .000
`Poly2SLS`	.083 $\pm$ .000	**.000 $\pm$ .000**	.133 $\pm$ .000	.039 $\pm$ .000
`GMM+NN`	.318 $\pm$ .000	.044 $\pm$ .000	.694 $\pm$ .000	.500 $\pm$ .000
`AGMM`	.600 $\pm$ .001	.025 $\pm$ .000	.274 $\pm$ .000	.047 $\pm$ .000
`DeepIV`	.247 $\pm$ .004	.056 $\pm$ .003	.165 $\pm$ .003	.038 $\pm$ .001
`DeepGMM`	.027 $\pm$ .009	.005 $\pm$ .001	.160 $\pm$ .025	*.025 $\pm$ .006*
`KernelIV`	*.019 $\pm$ .000*	.009 $\pm$ .000	*.046 $\pm$ .000*	.026 $\pm$ .000
`AGMM-K`	181 $\pm$ .000	2.34 $\pm$ .000	19.4 $\pm$ .000	4.13 $\pm$ .000
`MMR-IV (NN)`	**.011 $\pm$ .002**	.005 $\pm$ .000	.153 $\pm$ .019	.040 $\pm$ .004
`MMR-IV (Nys)`	**.011 $\pm$ .001**	*.001 $\pm$ .000*	**.006 $\pm$ .002**	**.020 $\pm$ .002**

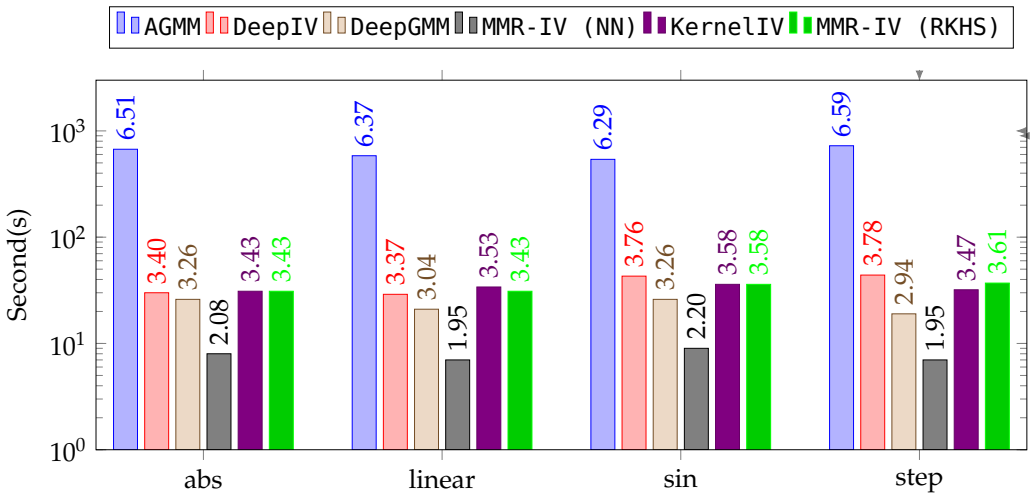


Figure 6.1: Runtime comparison in the large-sample regime ($n = 2000$). The computational time of parameter selection is excluded from the comparison. `AGMM-K (No Nyström)` and `MMR-IV (RKHS)` overlap due to the same runtime.

plicated methods (`GMM+NN`, `DeepIV` and `DeepGMM`) are unstable due to the sensitivity to hyper-parameters. (iv) We suspect that the `AGMM-K` performance is unsatisfactory because the hyper-parameter selection is not flexible enough. On the other hand, `MMR-IV (Nyström)` has the advantage of adaptive hyper-parameter selection (cf. Section 6.4). Appendix D.5.2 provides more details.

We also record the runtimes of all methods on the large-sample regime and report them in Figure 6.1. Compared to the NN-based methods, i.e., `AGMM`, `DeepIV`, `DeepGMM`, our `MMR-IV (NN)` is the most computationally efficient method, which is clearly a result of a simpler objective. Using a minimax optimization between two NNs, `AGMM` is the least efficient method. `DeepGMM` and `DeepIV` are more efficient than `AGMM`, but are less efficient than `MMR-IV (NN)`. Lastly, all three RKHS-based methods, namely, `KernelIV`, `AGMM-K` and `MMR-IV (RKHS)`, have similar computational time. All three methods are observed to scale poorly on large datasets.

6.6.2 High-dimensional Structured Scenarios

In high-dimensional setting, we employ the same data generating process as in Section 6.6.1. We consider only the absolute function for f^*, but map Z, X, or both X and Z to MNIST images (784-dim) [LeCun et al., 1998]. Let us denote the original outputs in Section 6.6.1 by X^{low}, Z^{low} and let $\pi(u) := \text{round}(\min(\max(1.5u + 5, 0), 9))$ be a transformation function mapping inputs to an integer between 0 and 9, and let $\text{RI}(d)$ be a function that selects a random MNIST image from the digit class d. We set $n = 10,000$. Then, the scenarios we consider are

(i) **MNIST$_Z$**: $Z \leftarrow \text{RI}(\pi(Z_1^{\text{low}}))$,
(ii) **MNIST$_X$**: $X \leftarrow \text{RI}(\pi(X^{\text{low}}))$,

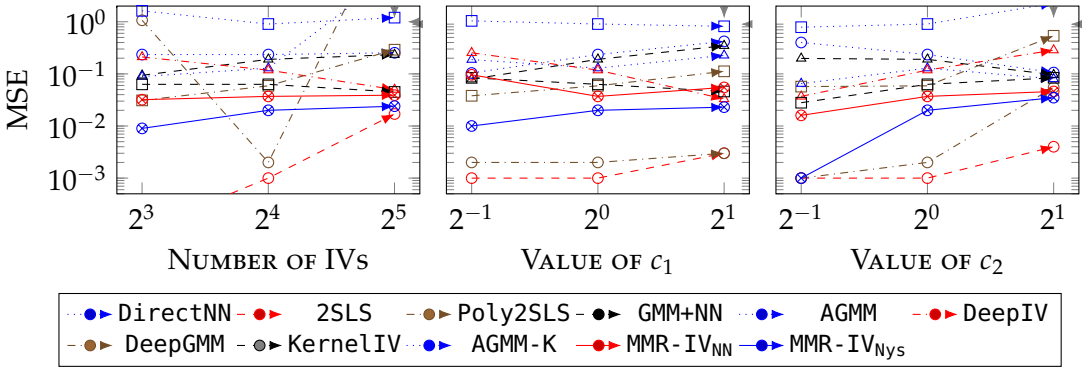


Figure 6.2: The MSE of different methods on Mendelian randomization experiments as we vary the numbers of instruments (left), the strength of confounders to exposures c_1 (middle), and the strength of confounders to instruments c_2 (right). The MSE is obtained from 10 repetitions of the experiment.

(iii) **MNIST$_{XZ}$**: $X \leftarrow \mathrm{RI}(\pi(X^{\mathrm{low}}))$, $Z \leftarrow \mathrm{RI}(\pi(Z_1^{\mathrm{low}}))$.

We report the results in Table 6.2. We summarize our findings: (i) `MMR-IV (NN)` performs well across scenarios. (ii) `MMR-IV (Nyström)` fails except MNIST$_Z$, because the ARD kernel with PCA are not representative enough for MNIST$_X$. (iii) `DeepIV` does not work when X is high-dimensional, similar to Bennett et al. [2019]. (iv) The two-step methods (`Poly2SLS` and `Ridge2SLS`) has large errors because the first-stage regression from Z to X is ill-posed. Appendix D.5.3 provides more details.

6.6.3 Mendelian Randomization

We demonstrate our method in the setting of Mendelian randomization which relies on genetic variants that satisfy the IV assumptions. The "exposure" X and outcome Y are univariate and generated from the simulation process by Hartwig et al. [2017]:

$$Y = \beta X + c_1 e + \delta, \quad X = \textstyle\sum_{i=1}^{m} \alpha_i Z_i + c_2 e + \gamma,$$

where $Z \in \mathbb{R}^{d'}$ with each entry $Z_i \sim B(2, p_i)$, $p_i \sim \mathrm{unif}(0.1, 0.9)$, $e \sim \mathcal{N}(0,1)$, $\alpha_i \sim \mathrm{unif}([0.8/d', 1.2/d'])$, and $\gamma, \delta \sim \mathcal{N}(0, 0.1^2)$. $Z_i \sim B(2, p_i)$ mimics the frequency of an individual getting one or more genetic variants. The parameters β, c_1 control the strength of exposures and confounders to outcomes, while c_2, α_i control the strength of instruments and confounders to exposures. We set $\alpha_i \sim \mathrm{unif}([0.8/d', 1.2/d'])$ so that as the number of IVs increases, each instrument becomes weaker while the overall strength of instruments ($\sum_i^{d'} \alpha_i$) remains constant.

In Mendelian randomization, it is known that genetic variants may act as *weak* IVs [Kuang et al., 2020; Hartford et al., 2020; Burgess et al., 2020], so this experiment aims to evaluate the sensitivity of different methods to the number of instruments (d') and confounder strengths (c_1, c_2). We consider three experiments: (i) $d' = 8, 16, 32$; (ii) $c_1 = 0.5, 1, 2$; (iii) $c_2 = 0.5, 1, 2$; unmentioned parameters use default values: $\beta = 1$,

Table 6.2: The mean square error (MSE) ± one standard deviation on high-dimensional structured data. We run each method 10 times.

Algorithm	**Setting**		
	MNIST_z	MNIST_x	MNIST_xz
`DirectNN`	.134 ± .000	*.229 ± .000*	*.196 ± .011*
`2SLS`	.563 ± .001	>1000	>1000
`Ridge2SLS`	.567 ± .000	.431 ± .000	.705 ± .000
`GMM+NN`	.121 ± .004	.235 ± .002	.240 ± .016
`AGMM`	*.017 ± .007*	.732 ± .107	.529 ±.163
`DeepIV`	.114 ± .005	n/a	n/a
`DeepGMM`	.038 ± .004	.315 ± .130	.333 ± .168
`AGMM-K+NN`	.021 ± .007	1.05 ± .366	.327 ± .192
`MMR-IV (NN)`	.024 ± .006	**.124 ± .021**	**.130 ± .009**
`MMR-IV (Nys)`	**.015 ± .002**	.442 ± .000	.425 ± .002

$d' = 16$, $c_1 = 1$, $c_2 = 1$. We draw $10,000$ samples for the training, validation and test sets, respectively, and train `MMR-IV (Nyström)` only on the training set. Other settings are the same as those of the low-dimensional scenario.

Figure 6.2 depicts the experimental results. Overall, `2SLS` performs well on all settings due to the linearity assumption, except particular sensitivity to the number of (weak) instruments, which is a well-known property of `2SLS` [Angrist and Pischke, 2008]. Although imposing no such assumption, `MMR-IV`s perform competitively and even more stably, since the information of instruments is effectively captured by the kernel k and we only need to deal with a simple objective, and also the analytical CV error plays an essential role. Section D.5.5 contains additional findings.

6.6.4 Application on the Vitamin D data

Lastly, we apply our algorithm to the Vitamin D data [Sjolander and Martinussen, 2019, Sec. 5.1]. The data were collected from a 10-year study on 2571 individuals aged 40–71 and 4 variables are employed: *age* (at baseline), *filaggrin* (binary indicator of filaggrin mutations), *VitD* (Vitamin D level at baseline) and *death* (binary indicator of death during study). The goal is to evaluate the potential effect of VitD on death. We follow Sjolander and Martinussen [2019] by controlling age in the analyses, using filaggrin as instrument, and then applying the MMR-IV (Nyström) algorithm. Sjolander and Martinussen [2019] modeled the effect of VitD on death by a generalized linear model and found the effect is insignificant by 2SLS (p-value on the estimated coefficient is 0.13 with the threshold of 0.05). More details can be found in Appendix D.5.6. The estimated effect is illustrated in Figure D.1 in the appendix. We observe that: (i) By using instruments, both our method and Sjolander and Martinussen [2019] output more reasonable results compared with those without instruments: a low VitD level at a young age has a slight effect on death, but a more adverse effect at an old age [Meehan and Penckofer, 2014]; (ii) Unlike Sjolander and Martinussen [2019], our

method allows more flexible non-linearity for causal effect.

6.7 Conclusion

Learning causal relations when hidden confounders are present is a cornerstone of decision making. IV regression is a standard tool to tackle this task, but currently faces challenges in nonlinear settings. The present work presents a simple framework that overcomes some of these challenges. We employ RKHS theory in the reformulation of conditional moment restriction (CMR) as a maximum moment restriction (MMR) based on which we can approach the problem from the empirical risk minimization (ERM) perspective. As we demonstrate, this framework not only facilitates theoretical analysis, but also results in easy-to-use algorithms that perform well in practice compared to existing methods. The paper also shows a way of combining the elegant theoretical approach of kernel methods with practical merits of deep neural networks. Despite these advantages, the optimal choice of the kernel k remains an open question which we hope to address in future work.

In the next chapter, we present a conclusion summarizing the thesis and provide future directions.

Chapter 7

Conclusions and Future Work

In this chapter, we first summarize the contributions we made in the thesis. We then present some potential research directions for future work.

7.1 Conclusions

Given a set of data and the model hypothesis space, accurate estimation of the model parameters is crucial for e.g. data modeling and model-based inference. The thesis is motivated by the non-parametric Bayesian estimation of the Hawkes process. The non-parametric form provides powerful modeling ability and the approximate Bayesian inference approach is a natural choice for less sensitivity to randomness of finite samples and better scalability on large-scale datasets. For this end, the thesis studies two different kinds of efficient approximate inference frameworks, namely, the Laplace Bayesian Hawkes process (LBHP) as Chapter 3 and the Variational Bayesian Hawkes process (VBHP) as Chapter 4. Both frameworks employ the branching structure of the Hawkes process to simplify the Bayesian inference and the finite support assumption for acceleration. Notably, they rely on different approximate principles: LBHP exploits Gibbs sampling as the high-level procedure with Laplace approximation applied in each iteration; VBHP only exploits the variational inference method. As a result, VBHP has an advantage of efficiently selecting the hyper-parameters without grid search, while LBHP is a more general framework and is able to estimate the distribution of other quantities related to the Hawkes process, such as the branching factor.

Beyond the studies on the Hawkes process, the thesis further explores a new approximate Bayesian inference approach in Chapter 5, which is applicable to more general Gaussian process models. The approach builds on the expectation propagation algorithm and substitutes the L_2 Wasserstein distance to the Kullback-Leibler (KL) divergence. Interestingly, although equipped with a more complicated distance, the approach preserves some desirable properties, such as the locality property of the EP update.

Finally, the thesis studies a simple and robust frequentist estimation framework as Chapter 6, which is outside the Bayesian field. The framework is designed for estimation of conditional moment restriction models, including the Hawkes process, and depends on the kernel maximum moment restriction. The framework has a

simple risk function and eases the proofs of consistency and asymptotic normality properties. It can also result in efficient hyper-parameter selection for some specific model classes.

7.2 Future Research Directions

According to recent frequentist estimation methods of the Hawkes process as reviewed in Section 2.5, it would be interesting to develop efficient Bayesian versions for these methods to gain more robustness and better modeling performance. Some of these methods are based on normalizing flows, which can be applied for approximate Bayesian inference. Therefore, developing more practical normalizing flows is an interesting direction.

The kernel maximum moment restriction based method is general and we expect it to be applied in different subfields of stochastic point processes.

Appendix: Laplace Bayesian Hawkes Process

A.1 Computing the Integral Term of the Log-likelihood

We consider $\Omega = [0, T]$, the background intensity μ, the triggering kernel $\phi(\cdot) = 1/2 f(\cdot)^2$, $f(\cdot) = \boldsymbol{\omega}^T \boldsymbol{e}(\cdot)$, and data $\{x_i\}_{i=1}^N$, and the integral term in the log-likelihood is calculated as below

$$\begin{aligned}
\text{Integral Term} &= -\frac{1}{2}\sum_{i=1}^{N}\int_0^T f^2(t-x_i)dt \\
&= -\frac{1}{2}\sum_{i=1}^{N}\int_0^{T-x_i}[\sum_{k=1}^{K}\omega_k e_k(t)]^2 dt \\
&= -\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{k'=1}^{K}\omega_k\omega_{k'}\underbrace{\int_0^{T-x_i} e_k(t)e_{k'}(t)dt}_{U_{kk'}^{(i)}} \\
&= -\frac{1}{2}\sum_{i=1}^{N}\boldsymbol{\omega}^T U^{(i)}\boldsymbol{\omega}.
\end{aligned}$$

In our case, Equation (3.9) has $d = 1$, i.e., $\phi_k(x) = (2/\pi)^{1/2}\sqrt{1/2}^{[k-1=0]}\cos[(k-1)x]$, $k = 1, 2, \cdots$. The matrix $U^{(i)}$ is calculated as below:

$$\begin{aligned}
U_{1,1}^{(i)} &= \int_0^{T-t_1}\frac{1}{\pi}dt = \frac{T-x_i}{\pi}, \\
U_{k>1,1}^{(i)} &= U_{1,k>1}^{(i)} = \frac{\sqrt{2}}{\pi}\frac{\sin[(k-1)(t_m-x_i)]}{k-1}, \\
U_{k,k(k>1)} &= \frac{1}{\pi}\Big\{T-x_i+\frac{\sin[2(k-1)(T-x_i)]}{2(k-1)}\Big\}, \\
U_{k,k'(k\neq k')} &= \frac{1}{\pi}\Big\{\frac{\sin[(k-k')(T-x_i)]}{k-k'}+\frac{\sin[(k+k'-2)(T-x_i)]}{k+k'-2}\Big\}.
\end{aligned}$$

A.2 M.A.P. μ and ϕ Given Infinite Branching Structures

M.A.P. μ and ϕ given Infinite branching structures is written as:

$$\begin{aligned}
&\operatorname*{argmax}_{\boldsymbol{\omega},\mu} \mathbb{E}_B[\log p(\boldsymbol{\omega},\mu|B,\{x_i\}_{i=1}^N,\Omega,k)] \\
&= \operatorname*{argmax}_{\boldsymbol{\omega},\mu} \underbrace{\mathbb{E}_B[\log p(\{x_i\}_{i=1}^N|\boldsymbol{\omega},\mu,B,\Omega,k)]}_{\text{Expected Log-likelihood}} + \underbrace{\log p(\boldsymbol{\omega}) + \log p(\mu)}_{\text{Constraints}} \\
&= \operatorname*{argmax}_{\boldsymbol{\omega},\mu} \sum_{i=1} \{ \sum_{x_j<x_i} p_{ij} \log \frac{1}{2}[\boldsymbol{\omega}^T \boldsymbol{e}(x_i-x_j)]^2 - p_{i0}\log\mu - \frac{1}{2}\int_0^{T-x_i} [\boldsymbol{\omega}^T\boldsymbol{e}(t)]^2 dt\} \\
&\qquad - (\beta+1)\mu T - \frac{1}{2}\boldsymbol{\omega}^T\Lambda^{-1}\boldsymbol{\omega} - (\alpha-1)\log\mu,
\end{aligned}$$

where B represents the branching structure, p_{ij} the probabilities of triggering relationships shown as Equation (2.2), and α, β are parameters of the Gamma prior of μT. The second line is obtained using Bayes' rule, which shows M.A.P. μ and ϕ given infinite branching structures is equivalent to maximizing the constrained expected log-likelihood, i.e., the objective function for the M-step of the EM algorithm and the third line is an explicit expression of the second line.

A.3 Mode-Finding the Triggering Kernel

Here we demonstrate in detail the computational challenges involved in finding the posterior mode with respect to the value of the triggering kernel at multiple point locations. Consider the triggering kernel $\phi(\cdot) = \frac{1}{2}f^2(\cdot)$ where $f(\cdot)$ is Gaussian process distributed. For a dataset $\{x_i\}_{i=1}^N$, $\boldsymbol{X} \equiv \{f(x_i)\}_{i=1}^N = \{X_i\}_{i=1}^N$ has a normal distribution, i.e., $\{f(x_i)\}_{i=1}^N \sim \mathcal{N}(\boldsymbol{m},\boldsymbol{\Sigma})$ where $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ are the mean and the covariance matrix. The distribution of $\boldsymbol{Y} \equiv \{\phi(x_i)\}_{i=1}^N = \{Y_i\}_{i=1}^N$ is derived as below where F is the cumulative density function and f the probabilistic density function.

$$\begin{aligned}
&F_{\boldsymbol{Y}}(\boldsymbol{y}) \\
&= P(-\sqrt{2y_i} < X_i < \sqrt{2y_i}, i = 1,\cdots,N) \\
&= \int_{-\sqrt{2y_1}}^{\sqrt{2y_1}} \cdots \int_{-\sqrt{2y_N}}^{\sqrt{2y_N}} \frac{1}{\sqrt{(2\pi)^N|\boldsymbol{\Sigma}^{-1}|}} \exp[-\frac{(\boldsymbol{X}-\boldsymbol{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{m})}{2}]\, \mathrm{d}X_1 \cdots\, \mathrm{d}X_N, \\
&f_{\boldsymbol{Y}}(\boldsymbol{y}) \\
&= \frac{\partial^N}{\partial y_1 \cdots \partial y_N} F_{\boldsymbol{Y}}(\boldsymbol{y}) \\
&= \frac{1}{\sqrt{(2\pi)^N|\boldsymbol{\Sigma}^{-1}|}} (\Pi_{i=1}^N \frac{1}{2\sqrt{2y_1}}) \sum_{\boldsymbol{X}\in\times_{i=1}^N\{\sqrt{2y_1},-\sqrt{2y_1}\}} \exp[-\frac{(\boldsymbol{X}-\boldsymbol{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{m})}{2}],
\end{aligned}$$

where $\times$ is the Cartesian product. There are 2^N summations of exponential functions, which is intractable.

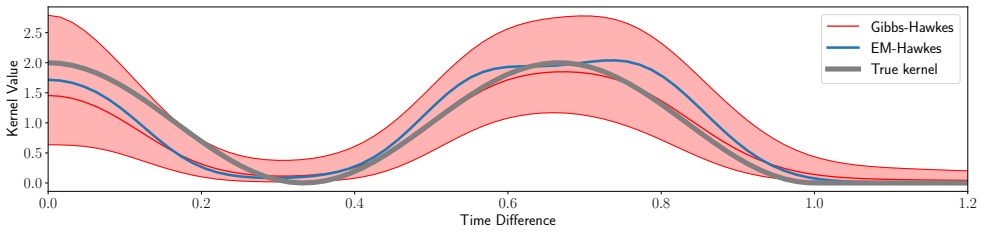


Figure A.1: Triggering kernels estimated by the Gibbs-Hawkes method (Section 3.2) and the EM-Hawkes method (Section 3.3.2). The true kernel is plotted as the bold gray curve. We plot the median (red) and [0.1, 0.9] interval (filled red) of the approximate predictive distribution, along with the triggering kernel inferred by the EM Hawkes method (blue). The hyper-parameters a and b of the Gaussian process kernel are set to 0.002.

Appendix: Variational Bayesian Hawkes Process

B.1 Deriving Equation (4.5)

As per Equation (2.7), there is

$$\begin{aligned}&\mathrm{CELBO}(q(B,\mu,f,\boldsymbol{u}),p(\mathcal{D}|B,\mu,f,\boldsymbol{u}),p(B,\mu,f,\boldsymbol{u}))\\&=\mathbb{E}_{q(B,\mu,f)}[\log p(\mathcal{D}|B,\mu,f)]-\mathrm{KL}(q(B,\mu,f,\boldsymbol{u})||p(B,\mu,f,\boldsymbol{u}))\end{aligned}$$

where the KL term can be simplified as

$$\begin{aligned}&\mathrm{KL}(q(B,\mu,f,\boldsymbol{u})||p(B,\mu,f,\boldsymbol{u}))\\&=\sum_B\int\int\int q(B,\mu,f,\boldsymbol{u})\log\frac{q(B)q(\mu)p(f|\boldsymbol{u})q(\boldsymbol{u})}{p(B)p(\mu)p(f|\boldsymbol{u})p(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u}\,\mathrm{d}f\,\mathrm{d}\mu\quad\text{(Equation (4.4) and Bayes' rule)}\\&=\sum_B\int\int\int q(B,\mu,f,\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}\,\mathrm{d}f\,\mathrm{d}\mu\log\frac{q(B)}{p(B)}\\&\quad+\sum_B\int\int\int q(B,\mu,f,\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}\,\mathrm{d}f\log\frac{q(\mu)}{p(\mu)}\,\mathrm{d}\mu\\&\quad+\sum_B\int\int\int q(B,\mu,f,\boldsymbol{u})\,\mathrm{d}f\,\mathrm{d}\mu\log\frac{q(\boldsymbol{u})}{p(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u}\quad\text{(simplification)}\\&=\sum_B q(B)\log\frac{q(B)}{p(B)}+\int q(\mu)\log\frac{q(\mu)}{p(\mu)}\,\mathrm{d}\mu+\int q(\boldsymbol{u})\log\frac{q(\boldsymbol{u})}{p(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u}\quad\text{(simplification)}\\&=\mathrm{KL}(q(B)||p(B))+\mathrm{KL}(q(\mu)||p(\mu))+\mathrm{KL}(q(\boldsymbol{u})||p(\boldsymbol{u})).\end{aligned}$$

We utilise the likelihood $p(\mathcal{D},B|\mu,f)$ by the reconstruction term and the KL term w.r.t. B

$$\begin{aligned}&\mathbb{E}_{q(B,\mu,f)}[\log p(\mathcal{D}|B,\mu,f)]-\mathrm{KL}(q(B)||p(B))\\&=\sum_B\int\int q(B,\mu,f)\log p(\mathcal{D}|B,\mu,f)\,\mathrm{d}f\,\mathrm{d}\mu-\sum_B q(B)\log\frac{q(B)}{p(B)}\quad\text{(definition)}\end{aligned}$$

$$= \sum_B \int\int q(B,\mu,f) \log p(\mathcal{D}|B,\mu,f)\ \mathrm{d}f\ \mathrm{d}\mu - \sum_B \int\int q(B,\mu,f) \log \frac{q(B)}{p(B)}\ \mathrm{d}f\ \mathrm{d}\mu$$

(align probabilities)

$$= \sum_B \int\int q(B,\mu,f) \log \frac{p(\mathcal{D}|B,\mu,f)p(B)}{q(B)}\ \mathrm{d}f\ \mathrm{d}\mu \quad \text{(merge)}$$

$$= \sum_B \int\int q(B,\mu,f) \log p(\mathcal{D},B|\mu,f)\ \mathrm{d}f\ \mathrm{d}\mu - \sum_B q(B)\log q(B) \quad \text{(merge)}$$

$$= \mathbb{E}_{q(B,\mu,f)}\Big[\log p(\mathcal{D},B|\mu,f)\Big] + H_B$$

where $H_B = -\sum_B q(B)\log q(B)$ is the entropy of B and further computed as follows. We adopt $q(B)$ from Equation (4.3) and the close form expression of H_B is derived as

$$H_B = -\sum_{\{\boldsymbol{b}_i\}_{i=1}^N} \Big(\prod_{i=1}^N\prod_{j=0}^{i-1} q_{ij}^{b_{ij}}\Big) \log\Big(\prod_{i=1}^N\prod_{j=0}^{i-1} q_{ij}^{b_{ij}}\Big)$$

$$= -\sum_{j=0}^{k-1}\sum_{\{\boldsymbol{b}_i\}_{i\neq k}} \Big(q_{kj}\prod_{i\neq k}\prod_{j=0}^{i-1} q_{ij}^{b_{ij}}\Big) \log\Big(q_{kj}\prod_{i\neq k}\prod_{j=0}^{i-1} q_{ij}^{b_{ij}}\Big) \quad \text{(split the summation)}$$

$$= -\sum_{j=0}^{k-1}\sum_{\{\boldsymbol{b}_i\}_{i\neq k}} \Big(q_{kj}\prod_{i\neq k}\prod_{j=0}^{i-1} q_{ij}^{b_{ij}}\Big) \Big[\log q_{kj} + \log\Big(\prod_{i\neq k}\prod_{j=0}^{i-1} q_{ij}^{b_{ij}}\Big)\Big] \quad \text{(split the log)}$$

$$= -\sum_{j=0}^{k-1} q_{kj}\log q_{kj} \underbrace{\sum_{\{\boldsymbol{b}_i\}_{i\neq k}} \Big(\prod_{i\neq k}\prod_{j=0}^{i-1} q_{ij}^{b_{ij}}\Big)}_{=1} - \underbrace{\sum_{j=0}^{k-1} q_{kj}}_{=1} \sum_{\{\boldsymbol{b}_i\}_{i\neq k}} \Big(\prod_{i\neq k}\prod_{j=0}^{i-1} q_{ij}^{b_{ij}}\Big) \log\Big(\prod_{i\neq k}\prod_{j=0}^{i-1} q_{ij}^{b_{ij}}\Big)$$

(distributive law of multiplication)

$$= -\sum_{j=0}^{k-1} q_{kj}\log q_{kj} - \sum_{\{\boldsymbol{b}_i\}_{i\neq k}} \Big(\prod_{i\neq k}\prod_{j=0}^{i-1} q_{ij}^{b_{ij}}\Big) \log\Big(\prod_{i\neq k}\prod_{j=0}^{i-1} q_{ij}^{b_{ij}}\Big) \quad \text{(simplification)}$$

$$= \cdots \quad \text{(do the same operations as above)}$$

$$= -\sum_{i=1}^N\sum_{j=0}^{i-1} q_{ij}\log q_{ij}.$$

B.2 Closed Forms of KL Terms in the ELBO

$$\mathrm{KL}(q(\mu)||p(\mu)) = (k-k_0)\psi(k) - k_0\log(c/c_0) - k - \log[\Gamma(k)/\Gamma(k_0)] + ck/c_0$$

$$\mathrm{KL}(q(\boldsymbol{u})||p(\boldsymbol{u})) = [\mathrm{Tr}(\boldsymbol{K}_{zz'}^{-1}\boldsymbol{S}) + \log|\boldsymbol{K}_{zz'}|/|\boldsymbol{S}| - M + \boldsymbol{m}^T\boldsymbol{K}_{zz'}^{-1}\boldsymbol{m}]/2$$

B.3 Extra Closed Form Expressions for Equation (4.2.2)

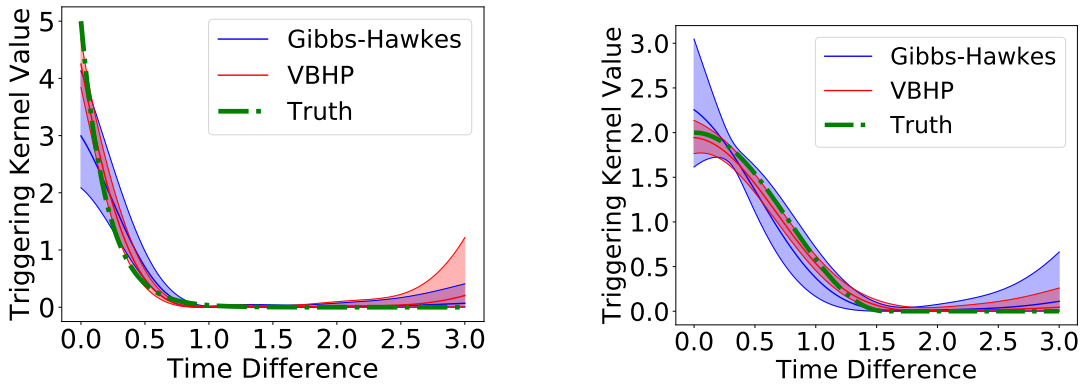
$$\begin{aligned}\int_{\mathcal{T}_i} \mathbb{E}^2_{q(f)}(f) &= \int_{\mathcal{T}_i} \boldsymbol{m}^T \boldsymbol{K}_{zz'}^{-1} \boldsymbol{K}_{zx} \boldsymbol{K}_{xz} \boldsymbol{K}_{zz'}^{-1} \boldsymbol{m} \, \mathrm{d}\boldsymbol{x} \\ &= \boldsymbol{m}^T \boldsymbol{K}_{zz'}^{-1} \Big(\int_{\mathcal{T}_i} \boldsymbol{K}_{zx} \boldsymbol{K}_{xz} \, \mathrm{d}\boldsymbol{x} \Big) \boldsymbol{K}_{zz'}^{-1} \boldsymbol{m} \\ &\equiv \boldsymbol{m}^T \boldsymbol{K}_{zz'}^{-1} \boldsymbol{\Psi}_i \boldsymbol{K}_{zz'}^{-1} \boldsymbol{m}.\end{aligned}$$

$$\begin{aligned}\int_{\mathcal{T}_i} \mathrm{Var}_{q(f)}[f] &= \int_{\mathcal{T}_i} \boldsymbol{K}_{xx} - \boldsymbol{K}_{xz} \boldsymbol{K}_{zz'}^{-1} \boldsymbol{K}_{zx} + \boldsymbol{K}_{xz} \boldsymbol{K}_{zz'}^{-1} \boldsymbol{S} \boldsymbol{K}_{zz'}^{-1} \boldsymbol{K}_{zx} \, \mathrm{d}\boldsymbol{x} \\ &= \int_{\mathcal{T}_i} \gamma - \mathrm{Tr}(\boldsymbol{K}_{zz'}^{-1} \boldsymbol{K}_{zx} \boldsymbol{K}_{xz}) + \mathrm{Tr}(\boldsymbol{K}_{zz'}^{-1} \boldsymbol{S} \boldsymbol{K}_{zz'}^{-1} \boldsymbol{K}_{zx} \boldsymbol{K}_{xz}) \, \mathrm{d}\boldsymbol{x} \\ &= \int_{\mathcal{T}_i} \gamma \, \mathrm{d}\boldsymbol{x} - \mathrm{Tr}(\boldsymbol{K}_{zz'}^{-1} \int_{\mathcal{T}_i} \boldsymbol{K}_{zx} \boldsymbol{K}_{xz} \, \mathrm{d}\boldsymbol{x}) + \mathrm{Tr}(\boldsymbol{K}_{zz'}^{-1} \boldsymbol{S} \boldsymbol{K}_{zz'}^{-1} \int_{\mathcal{T}_i} \boldsymbol{K}_{zx} \boldsymbol{K}_{xz} \, \mathrm{d}\boldsymbol{x}) \\ &\equiv \gamma |\mathcal{T}_i| - \mathrm{Tr}(\boldsymbol{K}_{zz'}^{-1} \boldsymbol{\Psi}_i) + \mathrm{Tr}(\boldsymbol{K}_{zz'}^{-1} \boldsymbol{S} \boldsymbol{K}_{zz'}^{-1} \boldsymbol{\Psi}_i).\end{aligned}$$

$$\begin{aligned}\boldsymbol{\Psi}_i(\boldsymbol{z}, \boldsymbol{z}') &= \int_{\mathcal{T}_i} \boldsymbol{K}_{zx} \boldsymbol{K}_{xz'} \, \mathrm{d}\boldsymbol{x} \\ &= \int_{\mathcal{T}_i} \gamma^2 \prod_{r=1}^{R} \exp\Big(-\frac{(x_r - z_r)^2}{2\alpha_r} \Big) \exp\Big(-\frac{(z'_r - x_r)^2}{2\alpha_r} \Big) \, \mathrm{d}\boldsymbol{x} \\ &= \int_{\mathcal{T}_i} \gamma^2 \prod_{r=1}^{R} \exp\Big(-\frac{(z_r - z'_r)^2}{4\alpha_r} \Big) \exp\Big(-\frac{(\bar{z}_r - x_r)^2}{\alpha_r} \Big) \, \mathrm{d}\boldsymbol{x} \\ &= \gamma^2 \prod_{r=1}^{R} \exp\Big(-\frac{(z_r - z'_r)^2}{4\alpha_r} \Big) \int_{\mathcal{T}_{i,r}} \exp\Big(-\frac{(\bar{z}_r - x_r)^2}{\alpha_r} \Big) \, \mathrm{d}x_r \quad (y_r = (\bar{z}_r - x_r)/\alpha_r) \\ &= \gamma^2 \prod_{r=1}^{R} -\sqrt{\alpha_r} \exp\Big(-\frac{(z_r - z'_r)^2}{4\alpha_r} \Big) \int_{(\bar{z}_r - \mathcal{T}_{i,r}^{\min})/\sqrt{\alpha_r}}^{(\bar{z}_r - \mathcal{T}_{i,r}^{\max})/\sqrt{\alpha_r}} \exp(-y_r^2) \, \mathrm{d}y_r \\ &= \gamma^2 \prod_{r=1}^{R} -\frac{\sqrt{\pi \alpha_r}}{2} \exp\Big(-\frac{(z_r - z'_r)^2}{4\alpha_r} \Big) \Big[\mathrm{erf}\Big(\frac{\bar{z}_r - \mathcal{T}_{i,r}^{\max}}{\sqrt{\alpha_r}} \Big) - \mathrm{erf}\Big(\frac{\bar{z}_r - \mathcal{T}_{i,r}^{\min}}{\sqrt{\alpha_r}} \Big) \Big].\end{aligned}$$

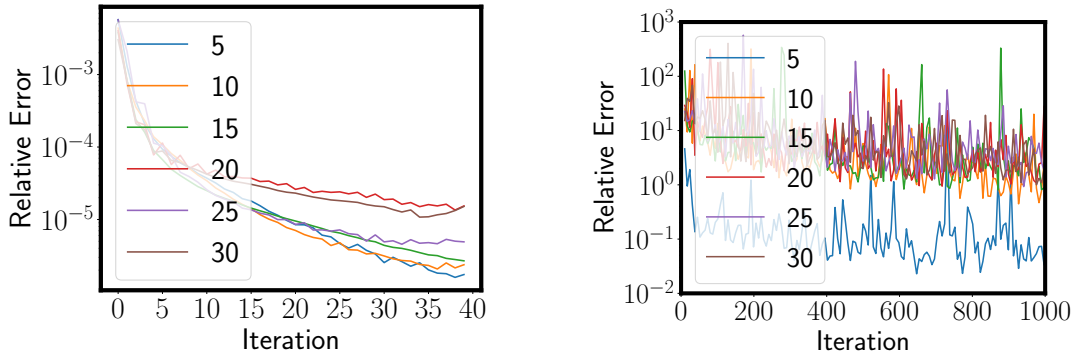
where we explicitly express $\mathcal{T}_i$ as a Cartesian product $\mathcal{T}_i \equiv \bigtimes_{r=1}^{R} [\mathcal{T}_{i,r}^{\min}, \mathcal{T}_{i,r}^{\max}]$ for R-dimensional data.

B.4 Additional Experiment Results



(a) Exp (b) Cos

Figure B.1: Posterior Triggering Kernels Inferred By VBHP and Gibbs Hawkes. Results of Gibbs Hawkes are obtained in 2000 iterations.



(a) VBHP (b) Gibbs Hawkes

Figure B.2: Convergence Rate of VBHP and Gibbs Hawkes with Different Numbers of Inducing Points. VBHP and Gibbs Hawkes measure respectively the relative error of the approximate marginal likelihood and of the posterior distribution of the Gaussian process.

Appendix: Quantile Propagation

C.1 Minimization of L_2 WD between Univariate Gaussian and Non-Gaussian Distributions

In this section, we derive the formulas of the optimal μ^* and σ^* for the L_2 WD, i.e., , Equation (5.1). Recall the optimization problem: we use a univariate Gaussian distribution $\mathcal{N}(f|\mu,\sigma^2)$ to approximate a univariate non-Gaussian distribution $q(f)$ by minimizing the L_2 WD between them:

$$\min_{\mu,\sigma} \mathrm{W}_2^2(q,\mathcal{N}) = \min_{\mu,\sigma} \int_0^1 \left| F_q^{-1}(y) - \mu - \sqrt{2}\sigma \mathrm{erf}^{-1}(2y-1) \right|^2 \mathrm{d}y,$$

where F_q^{-1} is the quantile function of the non-Gaussian distribution q, namely the pseudo-inverse function of the corresponding cumulative distribution function F_q defined in Proposition 1.

To solve this problem, we first calculate derivatives about μ and σ:

$$\begin{aligned}
\frac{\partial \mathrm{W}_2^2}{\partial \mu} &= -2 \int_0^1 F_q^{-1}(y) - \mu - \sqrt{2}\sigma \mathrm{erf}^{-1}(2y-1)\ \mathrm{d}y, \\
\frac{\partial \mathrm{W}_2^2}{\partial \sigma} &= -2 \int_0^1 (F_q^{-1}(y) - \mu - \sqrt{2}\sigma \mathrm{erf}^{-1}(2y-1))\sqrt{2}\mathrm{erf}^{-1}(2y-1)\ \mathrm{d}y.
\end{aligned}$$

Then, by zeroing derivatives, we obtain the optimal parameters:

$$\begin{aligned}
\mu^* &= \int_0^1 F_q^{-1}(y) - \sqrt{2}\sigma \mathrm{erf}^{-1}(2y-1)\ \mathrm{d}y \\
&= \int_{-\infty}^{\infty} x q(x)\ \mathrm{d}x - \frac{\sqrt{2}}{2}\sigma \int_{-1}^{1} \mathrm{erf}^{-1}(y)\ \mathrm{d}y \\
&= \mu_q - \sqrt{2}\sigma \int_{-\infty}^{\infty} x \mathcal{N}(x|0,1/2)\ \mathrm{d}x \\
&= \mu_q, \\
\sigma^* &= \sqrt{2} \int_0^1 (F_q^{-1}(y) - \mu)\mathrm{erf}^{-1}(2y-1)\ \mathrm{d}y \Big/ \int_0^1 2(\mathrm{erf}^{-1})^2(2y-1)\ \mathrm{d}y
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{2}\int_0^1 F_q^{-1}(y)\mathrm{erf}^{-1}(2y-1)\ \mathrm{d}y \Big/ \underbrace{\int_{-\infty}^{\infty} 2x^2\mathcal{N}(x|0,1/2)\ \mathrm{d}x}_{=1} \\
&= \sqrt{2}\int_0^1 F_q^{-1}(y)\mathrm{erf}^{-1}(2y-1)\ \mathrm{d}y \\
&= \sqrt{2}\int_{-\infty}^{\infty} f\mathrm{erf}^{-1}(2F_q(f)-1)\ \mathrm{d}F_q(f) \\
&= -\sqrt{\frac{1}{2\pi}}\int_{-\infty}^{\infty} f\ \mathrm{d}\, e^{-[\mathrm{erf}^{-1}(2F_{\widetilde{q}}(f)-1)]^2} \\
&= 0 + \sqrt{\frac{1}{2\pi}}\int_{-\infty}^{\infty} e^{-[\mathrm{erf}^{-1}(2F_{\widetilde{q}}(f)-1)]^2}\ \mathrm{d}f. \qquad \text{(C.1)}
\end{aligned}
$$

C.2 Minimization of L_p WD between Univariate Gaussian and Non-Gaussian Distributions

In this section, we describe a gradient descent approach to minimizing an L_p WD, for $p \neq 2$, in order to handle cases with no analytical expressions for the optimal parameters. Our goal is to use a univariate Gaussian distribution $\mathcal{N}(f|\mu,\sigma^2)$ to approximate a univariate non-Gaussian distribution $q(f)$. Specifically, we seek the minimizer in μ and σ of $\mathrm{W}_p^p(q,\mathcal{N})$; the derivatives of the objective function about μ and σ are:

$$
\begin{aligned}
\partial_\mu \mathrm{W}_p^p &= -p\int_0^1 |\varepsilon(y)|^{p-1}\mathrm{sgn}(\varepsilon(y))\ \mathrm{d}y = -p\int_{-\infty}^{\infty} |\eta(x)|^{p-1}\mathrm{sgn}(\eta(x))q(x)\ \mathrm{d}x, \\
\partial_\sigma \mathrm{W}_p^p &= -p\int_0^1 |\varepsilon(y)|^{p-1}\mathrm{sgn}(\varepsilon(y))\mathrm{erf}^{-1}(2y-1)\ \mathrm{d}y \\
&= -p\int_{-\infty}^{\infty} |\eta(x)|^{p-1}\mathrm{sgn}(\eta(x))\mathrm{erf}^{-1}(2F_q(x)-1)q(x)\ \mathrm{d}x.
\end{aligned}
$$

where for simplification, we define $\varepsilon(y) = F_q^{-1}(y) - \mu - \sqrt{2}\sigma\mathrm{erf}^{-1}(2y-1)$ and $\eta(x) = x - \mu - \sqrt{2}\sigma\mathrm{erf}^{-1}(2F_q(x)-1)$, with F_q and F_q^{-1} being the CDF and the quantile function of q. Note the derivatives have no analytical expressions. However, if the CDF F_q is available, we can use the standard numerical integration routines; otherwise, we resort to Monte Carlo sampling. In the framework of EP or QP, $q(x) \propto q^{\backslash i}(x)p(y_i|x)$ and $q^{\backslash i}$ is Gaussian, so we may draw samples from a Gaussian proposal distribution to obtain a simple Monte Carlo method.

C.3 Computations for Different Likelihoods

Given the likelihood $p(y|f)$ and the cavity distribution $q^{\backslash i}(f) = \mathcal{N}(f|\mu,\sigma^2)$, a stable way to compute the mean and the variance of the tilted distribution $\widetilde{q}(f) = p(y|f)q^{\backslash i}(f)/Z$ where the normalizer $Z = \int_{-\infty}^{\infty} p(y|f)q^{\backslash i}(f)\ \mathrm{d}f$, can be found in the software manual of Rasmussen and Williams [2005]. We present the key formulae

below, for use in subsequent derivations:

$$
\begin{aligned}
\partial_\mu Z &= \int_{-\infty}^{\infty} \frac{f-\mu}{\sigma^2} p(y|f)\mathcal{N}(f|\mu,\sigma^2)\,\mathrm{d}f \\
\frac{\partial_\mu Z}{Z} &= \frac{1}{\sigma^2}\int_{-\infty}^{\infty} f\frac{p(y|f)\mathcal{N}(f|\mu,\sigma^2)}{Z}\,\mathrm{d}f - \frac{\mu}{\sigma^2}\int_{-\infty}^{\infty}\frac{p(y|f)\mathcal{N}(f|\mu,\sigma^2)}{Z}\,\mathrm{d}y \\
\frac{\partial_\mu Z}{Z} &= \frac{1}{\sigma^2}\mu_{\widetilde{q}} - \frac{\mu}{\sigma^2} \\
\implies \mu_{\widetilde{q}} &= \frac{\sigma^2\partial_\mu Z}{Z} + \mu = \sigma^2\partial_\mu \log Z + \mu, \\
\partial_\mu^2 Z &= \int_{-\infty}^{\infty} -\frac{1}{\sigma^2}p(y|f)\mathcal{N}(f|\mu,\sigma^2) + \left(\frac{f-\mu}{\sigma^2}\right)^2 p(y|f)\mathcal{N}(f|\mu,\sigma^2)\,\mathrm{d}f \\
\frac{\partial_\mu^2 Z}{Z} &= \int_{-\infty}^{\infty}\left(-\frac{1}{\sigma^2} + \frac{\mu^2}{\sigma^4} + \frac{f^2}{\sigma^4} - \frac{2\mu f}{\sigma^4}\right)\frac{p(y|f)\mathcal{N}(f|\mu,\sigma^2)}{Z}\,\mathrm{d}f \\
\frac{\partial_\mu^2 Z}{Z} &= -\frac{1}{\sigma^2} + \frac{\mu^2}{\sigma^4} + \frac{1}{\sigma^4}(\sigma_{\widetilde{q}}^2 + \mu_{\widetilde{q}}^2) - \frac{2\mu}{\sigma^4}\mu_{\widetilde{q}} \\
\frac{\partial_\mu^2 Z}{Z} &= -\frac{1}{\sigma^2} + \frac{\sigma_{\widetilde{q}}^2}{\sigma^4} + \frac{(\mu-\mu_{\widetilde{q}})^2}{\sigma^4} = -\frac{1}{\sigma^2} + \frac{\sigma_{\widetilde{q}}^2}{\sigma^4} + \left(\frac{\partial_\mu Z}{Z}\right)^2 \\
\implies \sigma_{\widetilde{q}}^2 &= \sigma^4\left[\frac{\partial_\mu^2 Z}{Z} - \left(\frac{\partial_\mu Z}{Z}\right)^2\right] + \sigma^2 = \sigma^4\partial_\mu^2 \log Z + \sigma^2.
\end{aligned}
$$

C.3.1 Probit Likelihood for Binary Classification

For the binary classification with labels $y \in \{-1, 1\}$, the PDF of the tilted distribution $\widetilde{q}(f)$ with the probit likelihood is provided by Rasmussen and Williams [2005]:

$$
\widetilde{q}(f) = Z^{-1}\Phi(fy)\mathcal{N}(f|\mu,\sigma^2), \quad Z = \Phi(z), \quad z = \frac{\mu}{y\sqrt{1+\sigma^2}},
$$

and the mean estimate also has a closed form expression:

$$
\mu^\star = \mu_{\widetilde{q}} = \mu + \frac{\sigma^2\mathcal{N}(z)}{\Phi(z)y\sqrt{1+\sigma^2}}.
$$

As per Equation (5.1), the computation of the optimal $\sigma^\star$ requires the CDF of $\widetilde{q}$, denoted as $F_{\widetilde{q}}$. For positive $y > 0$, the CDF is derived as

$$
\begin{aligned}
F_{\widetilde{q},y>0}(x) &= Z^{-1}\int_{-\infty}^{x}\Phi\left(fy\right)\mathcal{N}\left(f|\mu,\sigma^2\right)\,\mathrm{d}f \\
&= \frac{Z^{-1}}{2\pi\sigma y}\int_{-\infty}^{\mu}\int_{-\infty}^{x-\mu}\exp\left(-\frac{1}{2}\begin{bmatrix}w\\f\end{bmatrix}^T\begin{bmatrix}v^{-2}+\sigma^{-2} & v^{-2}\\ v^{-2} & v^{-2}\end{bmatrix}\begin{bmatrix}w\\f\end{bmatrix}\right)\,\mathrm{d}w\,\mathrm{d}f \\
&= Z^{-1}\int_{-\infty}^{k}\int_{-\infty}^{h}\mathcal{N}\left(\begin{bmatrix}w\\f\end{bmatrix}\middle|\mathbf{0},\begin{bmatrix}1 & -\rho\\ -\rho & 1\end{bmatrix}\right)\,\mathrm{d}w\,\mathrm{d}f
\end{aligned}
$$

$$\overset{\text{(a)}}{=} Z^{-1}\left[\frac{1}{2}\Phi(h) - T\left(h, \frac{k+\rho h}{h\sqrt{1-\rho^2}}\right) + \frac{1}{2}\Phi(k) - T\left(k, \frac{h+\rho k}{k\sqrt{1-\rho^2}}\right) + \eta\right]$$
$$k = \frac{\mu}{\sqrt{\sigma^2+1}},\ \ h = \frac{x-\mu}{\sigma},\ \ \rho = \frac{1}{\sqrt{1+1/\sigma^2}},\ \ x \neq \mu,\ \ \mu \neq 0,$$

where the step (a) is obtained by exploiting the work of Owen [1956] and $T(\cdot,\cdot)$ is the Owen's T function:

$$T(h,a) = \frac{1}{2\pi}\int_0^a \frac{\exp\left[-(1+x^2)h^2/2\right]}{1+x^2}\,\mathrm{d}x,$$

and η is defined as

$$\eta = \begin{cases} 0 & hk > 0 \text{ or } (hk = 0 \text{ and } h + k \geq 0), \\ -0.5 & \text{otherwise.} \end{cases}$$

Similarly, for $y < 0$, the CDF is

$$F_{\widetilde{q},y<0}(x) = Z^{-1}\left[\frac{1}{2}\Phi(h) + T\left(h, \frac{k+\rho h}{h\sqrt{1-\rho^2}}\right) - \frac{1}{2}\Phi(k) + T\left(k, \frac{h+\rho k}{k\sqrt{1-\rho^2}}\right) - \eta\right].$$

Summarizing the two cases, we get the closed form expression of $F_{\widetilde{q}}$:

$$\begin{aligned} &F_{\widetilde{q}}(x) \\ &= Z^{-1}\left[\frac{1}{2}\Phi(h) - yT\left(h, \frac{k+\rho h}{h\sqrt{1-\rho^2}}\right) + \frac{y}{2}\Phi(k) - yT\left(k, \frac{h+\rho k}{k\sqrt{1-\rho^2}}\right) + y\eta\right] \\ &= Z^{-1}\left[\frac{1}{2}\Phi(h) - yT\left(h, \frac{k}{h\sqrt{1-\rho^2}} + \sigma\right) + \frac{y}{2}\Phi(k) - yT\left(k, \frac{h}{k\sqrt{1-\rho^2}} + \sigma\right) + y\eta\right]. \end{aligned}$$

Provided the above, the optimal $\sigma^\star$ can be computed by numerical integration of Eqn (C.1). For special cases, we provide additional formulas:

$$(1)\ x = \mu,\ \mu \neq 0 : F_{\widetilde{q}}(x) = Z^{-1}\left[\frac{1}{4} - \frac{y\mathrm{sign}(k)}{4} + \frac{y}{2}\Phi(k) - yT(k,\sigma) + y\eta\right];$$
$$(2)\ x \neq \mu,\ \mu = 0 : F_{\widetilde{q}}(x) = 2\left[\frac{1}{2}\Phi(h) - yT(h,\sigma) + \frac{y}{4} - \frac{y\mathrm{sign}(h)}{4} + y\eta\right];$$
$$(3)\ x = \mu,\ \mu = 0 : F_{\widetilde{q}}(x) = \frac{1}{2} - \frac{y}{\pi}\arctan(\sigma).$$

C.3.2 Square Link Function for Poisson Regression

Consider Poisson regression, which uses the Poisson likelihood $p(y|g) = g^y \exp(-g)/y!$ to model count data $y \in \mathbb{N}$, with the square link function $g(f) = f^2$ [Walder and

Bishop, 2017; Flaxman et al., 2017]. We use the square link function because it is more mathematically convenient than the exponential function. Given the cavity distribution $q^{\setminus i}(f) = \mathcal{N}(f|\mu, \sigma^2)$, we want the tilted distribution $\widetilde{q}(f) = q^{\setminus i}(f)p(y|g(f))/Z$ where the normalizer Z is derived as:

$$
\begin{aligned}
Z &= \int_{-\infty}^{\infty} q^{\setminus i}(f)p(y|g)\ \mathrm{d}f \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(f-\mu)^2}{2\sigma^2}\right) f^{2y} \exp(-f^2)/y!\ \mathrm{d}f \\
&\overset{\text{(a)}}{=} \frac{1}{\sqrt{2\pi\sigma^2}y! \exp(\mu^2/(1+2\sigma^2))} \int_{-\infty}^{\infty} f^{2y} \exp\left(-\frac{(f-\mu/(1+2\sigma^2))^2}{2\sigma^2/(1+2\sigma^2)}\right)\ \mathrm{d}f \\
&\overset{\text{(b)}}{=} \frac{\left(\frac{2\sigma^2}{1+2\sigma^2}\right)^{y+\frac{1}{2}}}{\sqrt{2\pi\sigma^2}y! \exp(\mu^2/(1+2\sigma^2))} \Gamma\left(y+\frac{1}{2}\right)\ {}_1F_1\left(-y; \frac{1}{2}; -\frac{\mu^2}{2\sigma^2(1+2\sigma^2)}\right) \\
&= \frac{\alpha^{y+\frac{1}{2}}}{\sqrt{2\pi\sigma^2}y! \exp(h)} \Gamma\left(y+\frac{1}{2}\right)\ {}_1F_1\left(-y; \frac{1}{2}; -\frac{h}{2\sigma^2}\right), \\
\alpha &= \frac{2\sigma^2}{1+2\sigma^2}, \quad h = \frac{\mu^2}{1+2\sigma^2}
\end{aligned}
\tag{C.2}
$$

where the step (a) rewrites the product of two exponential functions into the form of the Gaussian distribution, (b) is achieved through Mathematica [Wolfram, 2019], $\Gamma(\cdot)$ is the Gamma function and ${}_1F_1\left(-y; \frac{1}{2}; -\frac{h^2}{2\sigma^2}\right)$ is the confluent hypergeometric function of the first kind. Furthermore, we compute the first derivative of $\log Z$ w.r.t. μ and then the mean of the tilted distribution:

$$
\begin{aligned}
\partial_\mu \log Z &= \left(\frac{y\ {}_1F_1\left(-y+1; \frac{3}{2}; -\frac{h}{2\sigma^2}\right)}{\sigma^2\ {}_1F_1\left(-y; \frac{1}{2}; -\frac{h}{2\sigma^2}\right)} - 1\right) \frac{2\mu}{1+2\sigma^2} \\
\implies \mu_{\widetilde{q}} &= \sigma^2 \partial_\mu \log Z + \mu. \\
\partial_\mu^2 \log Z &= \left(\frac{y\ {}_1F_1\left(-y+1; \frac{3}{2}; -\frac{h}{2\sigma^2}\right)}{\sigma^2\ {}_1F_1\left(-y; \frac{1}{2}; -\frac{h}{2\sigma^2}\right)} - 1\right) \frac{2}{1+2\sigma^2} - \\
&\qquad \left(\frac{2(1-y)\ {}_1F_1\left(-y+2; \frac{5}{2}; -\frac{h}{2\sigma^2}\right)}{3\ {}_1F_1\left(-y; \frac{1}{2}; -\frac{h}{2\sigma^2}\right)} + \frac{2y\ {}_1F_1\left(-y+1; \frac{3}{2}; -\frac{h}{2\sigma^2}\right)^2}{{}_1F_1\left(-y; \frac{1}{2}; -\frac{h}{2\sigma^2}\right)^2}\right) \frac{2\mu^2 y}{\sigma^4(1+2\sigma^2)^2} \\
\implies \sigma^2_{\widetilde{q}} &= \sigma^4 \partial_\mu^2 \log Z + \sigma^2
\end{aligned}
$$

Finally, we derive the CDF of the tilted distribution $\widetilde{q}$ by using the binomial theorem:

$$
F_{\widetilde{q}}(x) = Z^{-1} \int_{-\infty}^{x} p(y|g)\mathcal{N}(f|\mu, \sigma^2)\ \mathrm{d}f
$$

$$
\begin{aligned}
&\overset{(a)}{=} A \int_{-\infty}^{x} f^{2y} \exp\left(-\frac{(f-\mu/(1+2\sigma^2))^2}{2\sigma^2/(1+2\sigma^2)}\right) \mathrm{d}f \\
&= A \int_{-\infty}^{x-\frac{\mu}{1+2\sigma^2}} \left(f+\frac{\mu}{1+2\sigma^2}\right)^{2y} \exp\left(-\frac{f^2}{2\sigma^2/(1+2\sigma^2)}\right) \mathrm{d}f \\
&\overset{(b)}{=} A \int_{-\infty}^{x-\beta} \left[\sum_{k=0}^{2y} \binom{2y}{k} f^k \beta^{2y-k}\right] \exp\left(-\frac{f^2}{\alpha}\right) \mathrm{d}f \\
&= A \sum_{k=0}^{2y} \binom{2y}{k} \beta^{2y-k} \left[\int_{-\infty}^{0} f^k \exp\left(-\frac{f^2}{\alpha}\right) \mathrm{d}f + \int_{0}^{x-\beta} f^k \exp\left(-\frac{f^2}{\alpha}\right) \mathrm{d}f\right] \\
&\overset{(c)}{=} \frac{A}{2} \sum_{k=0}^{2y} \binom{2y}{k} \beta^{2y-k} \alpha^{\frac{k+1}{2}} \left[(-1)^k \Gamma\left(\frac{k+1}{2}\right) + \right. \\
&\qquad \left. \operatorname{sgn}(x-\beta)^{k+1} \left(\Gamma\left(\frac{k+1}{2}\right) - \Gamma\left(\frac{k+1}{2}, \frac{(x-\beta)^2}{\alpha}\right)\right)\right] \\
A &= \frac{Z^{-1}}{\sqrt{2\pi\sigma^2 y!} \exp(\mu^2/(1+2\sigma^2))} = \left[\alpha^{y+\frac{1}{2}} \Gamma\left(y+\frac{1}{2}\right) {}_1F_1\left(-y; \frac{1}{2}; -\frac{h}{2\sigma^2}\right)\right]^{-1}, \\
\beta &= \frac{\mu}{1+2\sigma^2},
\end{aligned}
$$

where the step (a) has been derived in (a) of Equation (C.2), (b) applies the binomial theorem and (c) is obtained through Mathematica [Wolfram, 2019]. And, the function $\Gamma(a,z) = \int_z^\infty t^{a-1}e^{-t}\,\mathrm{d}t$ is the upper incomplete gamma function and $\operatorname{sgn}(x)$ is the sign function, equaling 1 when $x > 0$, 0 when $x = 0$ and -1 when $x < 0$.

C.4 Proof of Convexity

Theorem. Given two probability measures in $\mathcal{M}_+^1(\mathbb{R})$: a Gaussian $\mathcal{N}(\mu, \sigma^2)$ with mean μ and standard deviation $\sigma > 0$, and an arbitrary measure $\widetilde{q}$, the L_p WD $\mathrm{W}_p^p(\widetilde{q}, \mathcal{N})$ is strictly convex about μ and σ.

Proof. Let $F_{\widetilde{q}}^{-1}(y)$ and $F_{\mathcal{N}}^{-1}(y) = \mu + \sqrt{2}\sigma\mathrm{erf}^{-1}(2y-1)$, $y \in [0,1]$, be the quantile functions of $\widetilde{q}$ and the Gaussian $\mathcal{N}$, where erf is the error function. Then, we consider two distinct Gaussian measures $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ and a convex combination w.r.t. their parameters $\mathcal{N}(a_1\mu_1 + a_2\mu_2, (a_1\sigma_1 + a_2\sigma_2)^2)$ with $a_1, a_2 \in \mathbb{R}_+$ and $a_1 + a_2 = 1$. Given the above, we further define $\varepsilon_k(y) = F_{\widetilde{q}}^{-1}(y) - \mu_k - \sigma_k\sqrt{2}\mathrm{erf}^{-1}(2y-1)$, $k = 1, 2$, for notational simplification, and derive the convexity as:

$$
\mathrm{W}_p^p(\widetilde{q}, \mathcal{N}(a_1\mu_1 + a_2\mu_2, (a_1\sigma_1 + a_2\sigma_2)^2)) \overset{(a)}{=} \int_0^1 |a_1\varepsilon_1(y) + a_2\varepsilon_2(y)|^p \,\mathrm{d}y \overset{(b)}{\leq} \int_0^1 (a_1|\varepsilon_1(y)| +
$$

$$
a_2|\varepsilon_2(y)|)^p \,\mathrm{d}y \overset{(c)}{\leq} a_1\mathrm{W}_p^p(\widetilde{q}, \mathcal{N}(\mu_1, \sigma_1^2)) + a_2\mathrm{W}_p^p(\widetilde{q}, \mathcal{N}(\mu_2, \sigma_2^2)),
$$

where steps (a), (b) and (c) are obtained by applying Proposition 1, non-negativity

of the absolute value, and the convexity of $f(x) = x^p$, $p \geq 1$, over $\mathbb{R}_+$ respectively. The equality at (b) holds iff $\varepsilon_k(y) \geq 0, k = 1, 2, \forall y \in [0, 1]$, and (c)'s equality holds iff $|\varepsilon_1(y)| = |\varepsilon_2(y)|$, $\forall y \in [0, 1]$. These two conditions for equality can't be attained simultaneously as otherwise it would contradict that $\mathcal{N}(\mu_1, \sigma_1^2)$ is different from $\mathcal{N}(\mu_2, \sigma_2^2)$. So, $\mathrm{W}_p^p(\widetilde{q}, \mathcal{N})$, $p \geq 1$, is strictly convex about μ and σ. □

C.5 Proof of Variance Difference

Theorem 11. *The variance of the Gaussian approximation to a univariate tilted distribution $\widetilde{q}(f)$ as estimated by QP and EP satisfy $\sigma_{\mathrm{QP}}^2 \leq \sigma_{EP}^2$.*

Proof. Let $\mathcal{N}(\mu_{\mathrm{QP}}, \sigma_{\mathrm{QP}}^2)$ be the optimal Gaussian in QP. As per Proposition 1, we reformulate the L_2 WD based projection $\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu_{\mathrm{QP}}, \sigma_{\mathrm{QP}}^2))$ w.r.t. quantile functions:

$$\begin{aligned}
\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}(\mu_{\mathrm{QP}}, \sigma_{\mathrm{QP}}^2)) &= \int_0^1 |F_{\widetilde{q}}^{-1}(y) - \mu_{\mathrm{QP}} - \sqrt{2}\sigma_{\mathrm{QP}}\mathrm{erf}^{-1}(2y-1)|^2 \, \mathrm{d}y \\
&= \int_0^1 \underbrace{(F_{\widetilde{q}}^{-1}(y) - \mu_{\mathrm{QP}})^2}_{\sigma_{\mathrm{EP}}^2} + \underbrace{(\sqrt{2}\sigma_{\mathrm{QP}}\mathrm{erf}^{-1}(2y-1))^2}_{\sigma_{\mathrm{QP}}^2} \\
&\quad \underbrace{- 2(F_{\widetilde{q}}^{-1}(y) - \mu_{\mathrm{QP}})\sqrt{2}\sigma_{\mathrm{QP}}\mathrm{erf}^{-1}(2y-1)}_{\text{(A)}} \, \mathrm{d}y \\
&= \sigma_{\mathrm{EP}}^2 - \sigma_{\mathrm{QP}}^2,
\end{aligned}$$

where for (A), we used $\int \mu_{\mathrm{QP}}\sigma_{\mathrm{QP}}\mathrm{erf}^{-1}(2y-1) \, \mathrm{d}y = 0$ and the remaining factor can be easily shown to be equal to $2\sigma_{\mathrm{QP}}^2$. Furthermore, due to the non-negativity of the WD, we have $\sigma_{\mathrm{EP}}^2 \geq \sigma_{\mathrm{QP}}^2$, and the equality holds if and only if $\widetilde{q}$ is Gaussian. □

C.6 Proof of Locality Property

Theorem. Minimization of $\mathrm{W}_2^2(\widetilde{q}(\boldsymbol{f}), \mathcal{N}(\boldsymbol{f}))$ w.r.t. $\mathcal{N}(\boldsymbol{f})$ results in $q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i) = \mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)$.

Proof. We first apply the decomposition of the L_2 norm to rewriting the $\mathrm{W}_2^2(\widetilde{q}(\boldsymbol{f}), \mathcal{N}(\boldsymbol{f}))$ as below (see detailed derivations in Appendix C.6.2):

$$\mathrm{W}_2^2\left(\widetilde{q}, \mathcal{N}\right) = \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i - f_i'\|_2^2 + \mathrm{W}_2^2(q_{\setminus i|i}^{\setminus i}, \mathcal{N}_{\setminus i|i})\Big], \tag{C.3}$$

where the prime indicates that the variable is from the Gaussian $\mathcal{N}$, and for simplification, we use the notation π_i for the joint distribution $\pi(f_i, f_i')$ which belongs to a set of measures $U(\widetilde{q}_i, \mathcal{N}_i)$. Since $q^{\setminus i}(\boldsymbol{f})$ is known to be Gaussian, we define it in a

partitioned form:

$$q^{\backslash i}(\boldsymbol{f}) \equiv \mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}_{\backslash i}\\ f_i\end{bmatrix} \;\middle|\; \begin{bmatrix}\boldsymbol{m}_{\backslash i}\\ m_i\end{bmatrix}, \begin{bmatrix}\boldsymbol{S}_{\backslash i} & \boldsymbol{S}_{\backslash ii}\\ \boldsymbol{S}_{\backslash ii}^{\mathrm{T}} & S_i\end{bmatrix}\right), \tag{C.4}$$

and the conditional $q^{\backslash i}(\boldsymbol{f}_{\backslash i}|f_i)$ is expressed as:

$$q^{\backslash i}(\boldsymbol{f}_{\backslash i}|f_i) = \mathcal{N}(\boldsymbol{f}_{\backslash i}|\boldsymbol{m}_{\backslash i|i}, \boldsymbol{S}_{\backslash i|i}), \;\; \boldsymbol{m}_{\backslash i|i} = \boldsymbol{m}_{\backslash i} + \boldsymbol{S}_{\backslash ii}S_i^{-1}(f_i - m_i) \equiv \boldsymbol{a}f_i + \boldsymbol{b}, \tag{C.5}$$
$$\boldsymbol{S}_{\backslash i|i} = \boldsymbol{S}_{\backslash i} - \boldsymbol{S}_{\backslash ii}S_i^{-1}\boldsymbol{S}_{\backslash ii}^{\mathrm{T}}.$$

We define a similar partitioned expression for the Gaussian $\mathcal{N}(\boldsymbol{f}')$ by adding primes to variables and parameters on the r.h.s. of Equation (C.4), and as a result, the conditional $\mathcal{N}(\boldsymbol{f}'_{\backslash i}|f'_i)$ is written as:

$$\mathcal{N}(\boldsymbol{f}'_{\backslash i}|f'_i) = \mathcal{N}(\boldsymbol{m}'_{\backslash i|i}, \boldsymbol{S}'_{\backslash i|i}), \;\; \boldsymbol{m}'_{\backslash i|i} = \boldsymbol{m}'_{\backslash i} + \boldsymbol{S}'_{\backslash ii}S_i'^{-1}(f'_i - m'_i) \equiv \boldsymbol{a}'f'_i + \boldsymbol{b}', \tag{C.6}$$
$$\boldsymbol{S}'_{\backslash i|i} = \boldsymbol{S}'_{\backslash i} - \boldsymbol{S}'_{\backslash ii}S_i'^{-1}\boldsymbol{S}'^{\,\mathrm{T}}_{\backslash ii}. \tag{C.7}$$

Given the above definitions, we exploit Proposition 2 to take the means out of the L_2 WD on the r.h.s. of Equation (C.3):

$$\mathrm{W}_2^2(\widetilde{q}, \mathcal{N}) = \inf_{\pi_i} \mathbb{E}_{\pi_i}\left[\|f_i - f'_i\|_2^2 + \|\boldsymbol{m}_{\backslash i|i} - \boldsymbol{m}'_{\backslash i|i}\|_2^2\right] + \underbrace{\mathrm{W}_2^2\left(\mathcal{N}(\boldsymbol{0}, \boldsymbol{S}_{\backslash i|i}), \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}'_{\backslash i|i})\right)}_{(\mathrm{A})} \tag{C.8}$$

Minimizing this function requires optimizing m'_i, $\boldsymbol{m}'_{\backslash i}$, S'_i, $\boldsymbol{S}'_{\backslash i}$ and $\boldsymbol{S}'_{\backslash ii}$. As $\boldsymbol{S}'_{\backslash i}$ is only contained in $\boldsymbol{S}_{\backslash i|i}$ and isolated into the term (A), it can be optimized by simply setting

$$\boldsymbol{S}'_{\backslash i|i} = \boldsymbol{S}_{\backslash i|i} \overset{\text{Equation (C.7)}}{\Longrightarrow} \boldsymbol{S}_{\backslash i}^{(n)*} = \boldsymbol{S}_{\backslash i|i} + \boldsymbol{S}'_{\backslash ii}S_i'^{-1}\boldsymbol{S}'^{\,\mathrm{T}}_{\backslash ii}. \tag{C.9}$$

As a result, (A) is minimized to zero. Next, we plug in expressions of $\boldsymbol{m}_{\backslash i|i}$ and $\boldsymbol{m}'_{\backslash i|i}$ (Equation (C.5) and Equation (C.6)) into optimized Equation (C.8):

$$\min_{\boldsymbol{S}'_{\backslash i}} (\mathrm{C.8}) = \inf_{\pi_i} \mathbb{E}_{\pi_i}\left[\|f_i - f'_i\|_2^2 + \|\boldsymbol{a}f_i - \boldsymbol{a}'f'_i + \boldsymbol{b} - \boldsymbol{b}'\|_2^2\right], \tag{C.10}$$

where $\boldsymbol{m}'_{\backslash i}$ is only contained by $\boldsymbol{b}'$. Thus, we can optimize it by zeroing the derivative of the above function about $\boldsymbol{m}'_{\backslash i}$, which results in:

$$\boldsymbol{b}' = \boldsymbol{b} + \boldsymbol{a}\mu_{\widetilde{q}_i} - \boldsymbol{a}'m'_i \overset{\text{Equation (C.6)}}{\Longrightarrow} \boldsymbol{m}_{\backslash i}^{(n)*} = \boldsymbol{S}'_{\backslash ii}S_i'^{-1}m'_i + \boldsymbol{b} + \boldsymbol{a}\mu_{\widetilde{q}_i} - \boldsymbol{a}'m'_i, \tag{C.11}$$

where $\mu_{\widetilde{q}_i}$ is the mean of $\widetilde{q}(f_i)$. The minimum value of Equation (C.10) thereby is (see

details in subsection C.6.3):

$$\min_{\boldsymbol{m}'_{\setminus i}} (\text{C.10}) = (1 + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}')\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}'\|_2^2 S_i' - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'\left[\sigma_{\widetilde{q}_i}^2 + S_i' + (\mu_{\widetilde{q}_i} - m_i')^2\right] \tag{C.12}$$

where $\sigma_{\widetilde{q}_i}^2$ is the variance of $\widetilde{q}(f_i)$. This function can be further simplified using the quantile based reformulation of $\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i)$ (see details in Appendix C.6.4) which results in:

$$(\text{C.12}) = \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 \underbrace{-2^{\frac{3}{2}}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}' c_{\widetilde{q}_i} {S_i'}^{\frac{1}{2}} + \|\boldsymbol{a}'\|_2^2 S_i'}_{(\mathrm{B})}. \tag{C.13}$$

Now, we are left with optimizing m_i', S_i' and $\boldsymbol{S}'_{\setminus ii}$. To optimize $\boldsymbol{S}'_{\setminus ii}$, which only exists in the above term (B), we zero the derivative of (B) w.r.t. $\boldsymbol{S}'_{\setminus ii}$ and this yields:

$$\boldsymbol{a}'^{*} = 2^{\frac{1}{2}}(S_i')^{-\frac{1}{2}} c_{\widetilde{q}_i}\boldsymbol{a} \overset{\text{Equation (C.6)}}{\Longrightarrow} \boldsymbol{S}'^{*}_{\setminus ii} = (2S_i')^{\frac{1}{2}} c_{\widetilde{q}_i}\boldsymbol{a}, \tag{C.14}$$

and the minimum value of Equation (C.13) is

$$\min_{\boldsymbol{S}'_{\setminus ii}} (\text{C.13}) = \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2(\sigma_{\widetilde{q}_i}^2 - 2c_{\widetilde{q}_i}^2). \tag{C.15}$$

The results of optimizing m_i' and S_i' in the above equation have already been provided in Equation (5.1): $m_i'^{*} = \mu_{\widetilde{q}_i}$ and $S_i'^{*} = 2c_{\widetilde{q}_i}^2$. By plugging them into Equation (C.14) and Equation (C.11), we have $\boldsymbol{a}'^{*} = \boldsymbol{a}$ and $\boldsymbol{b}'^{*} = \boldsymbol{b}$. Finally, using Equation (C.9), we obtain $q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i) = \mathcal{N}(\boldsymbol{f}_{\setminus i}|\boldsymbol{a}f_i + \boldsymbol{b}, \boldsymbol{S}_{\setminus i|i}) = \mathcal{N}(\boldsymbol{f}_{\setminus i}|\boldsymbol{a}'f_i + \boldsymbol{b}', \boldsymbol{S}'_{\setminus i|i}) = \mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)$, which concludes the proof. □

C.6.1 Details of Equation 5.2

$$\begin{aligned}
\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|\mathcal{N}(\boldsymbol{f})) &= \int \widetilde{q}(\boldsymbol{f}) \log \frac{\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)\widetilde{q}(f_i)}{\mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)\mathcal{N}(f_i)}\,\mathrm{d}\boldsymbol{f} \\
&= \int \widetilde{q}(f_i) \log \frac{\widetilde{q}(f_i)}{\mathcal{N}(f_i)}\,\mathrm{d}f_i + \int \widetilde{q}(f_i) \int \widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i) \log \frac{\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)}{\mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)}\,\mathrm{d}\boldsymbol{f}_{\setminus i}\,\mathrm{d}f_i \\
&= \mathrm{KL}\big(\widetilde{q}(f_i)\|\mathcal{N}(f_i)\big) + \mathbb{E}_{\widetilde{q}(f_i)}\Big[\mathrm{KL}\big(\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)\|\mathcal{N}(\boldsymbol{f}_{\setminus i}|f_i)\big)\Big] \\
\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i) = \frac{\widetilde{q}(\boldsymbol{f})}{\widetilde{q}(f_i)} &\propto \frac{p(\boldsymbol{f})\cancel{p(y_i|f_i)}\prod_{j\neq i} t_j(\boldsymbol{f})}{q^{\setminus i}(f_i)\cancel{p(y_i|f_i)}} \\
&= q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i).
\end{aligned} \tag{C.16}$$

C.6.2 Details of Equation C.3

$$
\begin{aligned}
\mathrm{W}_2^2\left(\widetilde{q}(\boldsymbol{f}), \mathcal{N}(\boldsymbol{f})\right) &\equiv \inf_{\pi \in U(\widetilde{q}, \mathcal{N})} \mathbb{E}_{\pi}\left(\|\boldsymbol{f}-\boldsymbol{f}'\|_2^2\right) \\
&= \inf_{\pi \in U(\widetilde{q}, \mathcal{N})} \mathbb{E}_{\pi}\left(\|f_i-f_i'\|_2^2\right) + \mathbb{E}_{\pi}\left(\|\boldsymbol{f}_{\setminus i}-\boldsymbol{f}'_{\setminus i}\|_2^2\right) \\
&\overset{\text{(a)}}{=} \inf_{\pi \in U(\widetilde{q}, \mathcal{N})} \mathbb{E}_{\pi_i}\Big[\|f_i-f_i'\|_2^2 + \mathbb{E}_{\pi_{\setminus i|i}}\left(\|\boldsymbol{f}_{\setminus i}-\boldsymbol{f}'_{\setminus i}\|_2^2\right)\Big] \\
&\overset{\text{(b)}}{=} \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i-f_i'\|_2^2 + \inf_{\pi_{\setminus i|i}} \mathbb{E}_{\pi_{\setminus i|i}}\left(\|\boldsymbol{f}_{\setminus i}-\boldsymbol{f}'_{\setminus i}\|_2^2\right)\Big] \\
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i-f_i'\|_2^2 + \mathrm{W}_2^2(\widetilde{q}_{\setminus i|i}, \mathcal{N}_{\setminus i|i})\Big] \\
&\overset{\text{(c)}}{=} \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i-f_i'\|_2^2 + \mathrm{W}_2^2(q^{\setminus i}_{\setminus i|i}, \mathcal{N}_{\setminus i|i})\Big],
\end{aligned}
$$

where the superscript prime indicates that the variable is from the Gaussian $\mathcal{N}$. In (a), $\pi_i = \pi(f_i, f_i')$ and $\pi_{\setminus i|i} = \pi(\boldsymbol{f}_{\setminus i}, \boldsymbol{f}'_{\setminus i}|f_i, f_i')$. In (b), the first and the second inf are over $U(\widetilde{q}_i, \mathcal{N}_i)$ and $U(\widetilde{q}_{\setminus i|i}, \mathcal{N}_{\setminus i|i})$ respectively. (c) is due to $\widetilde{q}(\boldsymbol{f}_{\setminus i}|f_i)$ being equal to $q^{\setminus i}(\boldsymbol{f}_{\setminus i}|f_i)$ (refer to Equation (C.16)).

C.6.3 Details of Equation C.12

$$
\begin{aligned}
&\min_{\boldsymbol{m}'_{\setminus i}} \text{Equation (C.10)} \\
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i-f_i'\|_2^2 + \|\boldsymbol{a}(f_i-\mu_{\widetilde{q}_i}) - \boldsymbol{a}'(f_i'-m_i')\|_2^2\Big] \\
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i-f_i'\|_2^2\Big] + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}'\|_2^2 S_i' - 2\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'\mathbb{E}_{\pi_i}\left(f_i f_i' - \mu_{\widetilde{q}_i} m_i'\right) \\
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i-f_i'\|_2^2\Big] + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}'\|_2^2 S_i' + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'\mathbb{E}_{\pi_i}\left(\|f_i-f_i'\|_2^2 - f_i^2 - (f_i')^2 + 2\mu_{\widetilde{q}_i} m_i'\right) \\
&= \inf_{\pi_i} \mathbb{E}_{\pi_i}\Big[\|f_i-f_i'\|_2^2\Big] + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}'\|_2^2 S_i' + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'\mathbb{E}_{\pi_i}\Big(\|f_i-f_i'\|_2^2 - (f_i-\mu_{\widetilde{q}_i})^2 - \\
&\quad 2f_i\mu_{\widetilde{q}_i} + \mu_{\widetilde{q}_i}^2 - (f_i'-m_i')^2 - 2f_i'm_i' + (m_i')^2 + 2\mu_{\widetilde{q}_i}m_i'\Big) \\
&= (1+\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}')\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}'\|_2^2 S_i' - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'\left(\sigma_{\widetilde{q}_i}^2 + \mu_{\widetilde{q}_i}^2 + S_i' + (m_i')^2 - 2\mu_{\widetilde{q}_i}m_i'\right) \\
&= (1+\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}')\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}'\|_2^2 S_i' - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'\Big[\sigma_{\widetilde{q}_i}^2 + S_i' + (\mu_{\widetilde{q}_i} - m_i')^2\Big]
\end{aligned}
$$

C.6.4 Details of Equation C.12

We first use Proposition 1 to reformulate the L_2 WD $\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i)$ as:

$$\begin{aligned}
\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) &= \int_0^1 \left(F_{\widetilde{q}_i}^{-1}(y) - m_i' - \sqrt{2S_i'}\mathrm{erf}^{-1}(2y-1)\right)^2 \mathrm{d}y, \\
&= \int_0^1 (F_{\widetilde{q}_i}^{-1}(y) - m_i')^2 + 2S_i'\mathrm{erf}^{-1}(2y-1)^2 - 2\sqrt{2S_i'}\mathrm{erf}^{-1}(2y-1)(F_{\widetilde{q}_i}^{-1}(y) - m_i') \, \mathrm{d}y, \\
&= \int_0^1 (F_{\widetilde{q}_i}^{-1}(y) - \mu_{\widetilde{q}_i} + \mu_{\widetilde{q}_i} - m_i')^2 \, \mathrm{d}y + S_i' - 2\sqrt{2S_i'}c_{\widetilde{q}_i}, \\
&= \sigma_{\widetilde{q}_i}^2 + (\mu_{\widetilde{q}_i} - m_i')^2 + S_i' - 2c_{\widetilde{q}_i}\sqrt{2S_i'},
\end{aligned}$$

where $F_{\widetilde{q}_i}^{-1}(y)$ is the quantile function of $\widetilde{q}(f_i)$ and $c_{\widetilde{q}_i} \equiv \int_0^1 F_{\widetilde{q}_i}^{-1}(y)\mathrm{erf}^{-1}(2y-1) \, \mathrm{d}y$. Next, we plug this reformulation into Equation (C.12):

$$\begin{aligned}
\text{Equation (C.12)} &= \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'\mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}'\|_2^2 S_i' - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'\left[\sigma_{\widetilde{q}_i}^2 + S_i' + (\mu_{\widetilde{q}_i} - m_i')^2\right] \\
&= \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'\left[\cancel{\sigma_{\widetilde{q}_i}^2 + (\mu_{\widetilde{q}_i} - m_i')^2} + S_i' - 2c_{\widetilde{q}_i}\sqrt{2S_i'}\right] + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}'\|_2^2 S_i' \\
&\quad - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}'\left[\cancel{\sigma_{\widetilde{q}_i}^2 + S_i' + (\mu_{\widetilde{q}_i} - m_i')^2}\right] \\
&= \mathrm{W}_2^2(\widetilde{q}_i, \mathcal{N}_i) - 2c_{\widetilde{q}_i}\sqrt{2S_i'}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}' + \|\boldsymbol{a}\|_2^2\sigma_{\widetilde{q}_i}^2 + \|\boldsymbol{a}'\|_2^2 S_i'
\end{aligned}$$

C.7 More Details of EP

We use the expressions $\widetilde{q}(\boldsymbol{f}) = q^{\backslash i}(\boldsymbol{f})p(y_i|f_i)/Z_{\widetilde{q}}$ and $q^{\backslash i}(\boldsymbol{f}) = q(\boldsymbol{f})/(t_i(f_i)Z_{q^{\backslash i}})$, and the derivation of $\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|q(\boldsymbol{f})) = \mathrm{KL}(\widetilde{q}(f_i)\|q(f_i))$ is shown as below:

$$\begin{aligned}
\mathrm{KL}(\widetilde{q}(\boldsymbol{f})\|q(\boldsymbol{f})) &= \int \widetilde{q}(\boldsymbol{f}) \log \frac{q^{\backslash i}(\boldsymbol{f})p(y_i|f_i)}{Z_{\widetilde{q}}q(\boldsymbol{f})} \, \mathrm{d}\boldsymbol{f} \\
&= \int \widetilde{q}(\boldsymbol{f}) \log \frac{\cancel{q(\boldsymbol{f})}p(y_i|f_i)}{Z_{q^{\backslash i}}Z_{\widetilde{q}}\cancel{q(\boldsymbol{f})}t_i(f_i)} \, \mathrm{d}\boldsymbol{f} \\
&= \int \widetilde{q}(f_i) \log \frac{p(y_i|f_i)}{Z_{q^{\backslash i}}Z_{\widetilde{q}}t_i(f_i)} \, \mathrm{d}f_i \\
&= \int \widetilde{q}(f_i) \log \frac{q^{\backslash i}(f_i)p(y_i|f_i)}{Z_{q^{\backslash i}}Z_{\widetilde{q}}q^{\backslash i}(f_i)t_i(f_i)} \, \mathrm{d}f_i \\
&= \int \widetilde{q}(f_i) \log \frac{\widetilde{q}(f_i)}{q(f_i)} \, \mathrm{d}f_i \\
&= \mathrm{KL}(\widetilde{q}(f_i)\|q(f_i))
\end{aligned}$$

C.8 Predictive Distributions of Poisson Regression

Given the approximate predictive distribution $f(\boldsymbol{x}_*) = \mathcal{N}(\mu_*, \sigma_*^2)$ and the relation $g(f) = f^2$, it is straightforward to derive the corresponding $g(\boldsymbol{x}_*) \sim \text{Gamma}(k_*, c_*)$[1] where the shape k_* and the scale c_* are expressed as [Walder and Bishop, 2017; **Zhang** et al., 2020b]:

$$k_* = \frac{(\mu_*^2 + \sigma_*^2)^2}{2\sigma_*^2(2\mu_*^2 + \sigma_*^2)}, \quad c_* = \frac{2\sigma_*^2(2\mu_*^2 + \sigma_*^2)}{\mu_*^2 + \sigma_*^2}.$$

Furthermore, the predictive distribution of the count value $y \in \mathbb{N}$ can also be derived straightforwardly:

$$\begin{aligned} p(y) &= \int_0^\infty p(g_*)p(y|g_*)\ \mathrm{d}g_* \\ &= \int \text{Gamma}(g_*|k_*, c_*)\text{Poisson}(y|g_*)\ \mathrm{d}g_* \\ &= \frac{c_*^y(c_*+1)^{-k_*-y}\Gamma(k_*+y)}{y!\Gamma(k_*)} = \text{NB}(y|k_*, c_*/(1+c_*)), \end{aligned}$$

where $g_* = g(\boldsymbol{x}_*)$ and NB denotes the negative binomial distribution. The mode is obtained as $\lfloor c_*(k_*-1) \rfloor$ if $k_* > 1$ else 0.

C.9 Proof of Corollary 3.2

Since the site approximations of both EP and QP are Gaussian, we may analyse the predictive variances using results from the regression with Gaussian likelihood function case, namely the well known Equation (3.61) in [Rasmussen and Williams, 2005]:

$$\sigma^2(f_*) = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^{\mathrm{T}}(K + \widetilde{\Sigma})^{-1}\boldsymbol{k}_*, \tag{C.17}$$

where $f_* = f(\boldsymbol{x}_*)$ is the evaluation of the latent function at $\boldsymbol{x}_*$ and $\boldsymbol{k}_* = [k(\boldsymbol{x}_*, \boldsymbol{x}_i)]_{i=1}^{N\ \mathrm{T}}$ is the covariance vector between the test data $\boldsymbol{x}_*$ and the training data $\{\boldsymbol{x}_i\}_{i=1}^N$, K is the prior covariance matrix and $\widetilde{\Sigma}$ is the diagonal matrix with elements of site variances $\widetilde{\sigma}_i^2$.

After updating the parameters of a site function $t_i(f_i)$, the term $(K + \widetilde{\Sigma})^{-1}$ is updated to $(K + \widetilde{\Sigma} + (\widetilde{\sigma}_{i,\text{new}}^2 - \widetilde{\sigma}_i^2)\boldsymbol{e}_i\boldsymbol{e}_i^{\mathrm{T}})^{-1}$ where $\widetilde{\sigma}_{i,\text{new}}$ is the site variance estimated by EP or QP and $\boldsymbol{e}_i$ is a unit vector in direction i. Using the Woodbury, Sherman & Morrison formula [Rasmussen and Williams, 2005, A.9], we rewrite $(K + \widetilde{\Sigma} + (\widetilde{\sigma}_{i,\text{new}}^2 - \widetilde{\sigma}_i^2)\boldsymbol{e}_i\boldsymbol{e}_i^{\mathrm{T}})^{-1}$ as

$$\begin{aligned} (K + \widetilde{\Sigma} + (\widetilde{\sigma}_{i,\text{new}}^2 - \widetilde{\sigma}_i^2)\boldsymbol{e}_i\boldsymbol{e}_i^{\mathrm{T}})^{-1} &\equiv (A^{-1} + (\widetilde{\sigma}_{i,\text{new}}^2 - \widetilde{\sigma}_i^2)\boldsymbol{e}_i\boldsymbol{e}_i^{\mathrm{T}})^{-1} \\ &= A - A\boldsymbol{e}_i[(\widetilde{\sigma}_{i,\text{new}}^2 - \widetilde{\sigma}_i^2)^{-1} + \boldsymbol{e}_i^{\mathrm{T}}A\boldsymbol{e}_i]^{-1}\boldsymbol{e}_i^{\mathrm{T}}A \end{aligned}$$

[1] $\text{Gamma}(x|k, c) = \frac{1}{\Gamma(k)c^k}x^{k-1}e^{-x/c}$.

$$\equiv A - \boldsymbol{s}_i[(\widetilde{\sigma}^2_{i,\text{new}} - \widetilde{\sigma}^2_i)^{-1} + A_{ii}]^{-1}\boldsymbol{s}_i^{\mathrm{T}}$$
$$= A - \frac{1}{(\widetilde{\sigma}^2_{i,\text{new}} - \widetilde{\sigma}^2_i)^{-1} + A_{ii}}\boldsymbol{s}_i\boldsymbol{s}_i^{\mathrm{T}}$$

where $A = (K + \widetilde{\Sigma})^{-1}$ and $\boldsymbol{s}_i$ is the i'th column of A. Putting the above expression into Equation (C.17), we have that the predictive variance is updated according to:

$$\sigma^2_{\text{new}}(f_*) = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^{\mathrm{T}} A \boldsymbol{k}_* + \frac{1}{(\widetilde{\sigma}^2_{i,\text{new}} - \widetilde{\sigma}^2_i)^{-1} + A_{ii}}\boldsymbol{k}_*^{\mathrm{T}}\boldsymbol{s}_i\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{k}_*.$$

In EP and QP, the first two terms on the r.h.s. of the above equation are equivalent. As the site variance provided by QP is less or equal to that by EP, i.e., , $\widetilde{\sigma}^2_{i,\text{QP}} \leq \widetilde{\sigma}^2_{i,\text{EP}}$, the third term on the r.h.s. for QP is less or equal to that for EP. Therefore, the predictive variance of QP is less or equal to that of EP: $\sigma^2_{\text{QP}}(f_*) \leq \sigma^2_{\text{EP}}(f_*)$.

C.10 Lookup Tables

To speed up updating variances σ^2_{QP} in QP, we pre-compute the integration in Equation (5.1) over a grid of cavity parameters μ and σ, and store the results into lookup tables. Consequently, each update step obtains σ^2_{QP} simply based on the lookup tables. Concretely, for the GP binary classification, we compute Equation (5.1) with μ, σ and y varying from -10 to 10, 0.1 to 10 and $\{-1, 1\}$ respectively. μ and σ vary in a linear scale and a log10 scale respectively, and both have a step size of 0.001. The resulting lookup tables has a size of 20001×2001. In a similar way, we make the lookup table for the Poisson regression. In the experiments, we exploit the linear interpolation to fit σ^2_{QP} given $\mu \in [-10, 10]$ and $\sigma \in [0.1, 10]$, and if μ and σ lie out of the lookup table, σ^2_{QP} is approximately computed by the EP update formula, i.e., $\sigma^2_{\text{QP}} \approx \sigma^2_{\text{EP}}$. On Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz, we observe the running time of EP and QP is almost the same.

Algorithm 2 Expectation (Quantile) Propagation

Input: $p(\boldsymbol{f})$, $p(y_i|f_i)$, $t_i(f_i)$, $i = 1, \cdots, N$, $\boldsymbol{\theta}$
Output: $q(\boldsymbol{f})$ approximate posterior

repeat
  compute $q(\boldsymbol{f}) \propto p(\boldsymbol{f}) \prod_i t_i(f_i)$ by (2.8)
  repeat
    for $i = 1$ to N **do**
      compute $q^{\backslash i}(f_i) \propto q(f_i)/t_i(f_i)$ cavity
      compute $\widetilde{q}(f_i) \propto q^{\backslash i}(f_i) p(y_i|f_i)$ tilted
      if EP **then**
        $t_i(f_i) \propto \text{proj}_{\text{KL}}[\widetilde{q}(f_i)]/q^{\backslash i}(f_i)$ by (2.10)(2.11)
      else if QP **then**
        $t_i(f_i) \propto \text{proj}_{\text{W}}[\widetilde{q}(f_i)]/q^{\backslash i}(f_i)$ by (5.1)(2.11)
      end if
      update $q(\boldsymbol{f}) \propto p(\boldsymbol{f}) \prod_i t_i(f_i)$ by (2.8)
    end for
  until convergence
  $\boldsymbol{\theta} = \text{argmax}_{\boldsymbol{\theta}} \log q(\mathcal{D})$ by (2.9)
until convergence
return $q(\boldsymbol{f})$

Appendix: Kernel Maximum Moment Restriction

D.1 Integrally Strictly Positive Definite (ISPD) Kernels

Popular kernel functions that satisfy Assumption 1 are the Gaussian RBF kernel and Laplacian kernel

$$k(z, z') = \exp\left(-\frac{\|z - z'\|_2^2}{2\sigma^2}\right), \qquad k(z, z') = \exp\left(-\frac{\|z - z'\|_1}{\sigma}\right),$$

where σ is a positive bandwidth parameter. Another important kernel is an inverse multiquadric (IMQ) kernel

$$k(z, z') = (c^2 + \|z - z'\|_2^2)^{-\gamma}$$

where c and γ are positive parameters [Steinwart and Christmann, 2008, Ch. 4]. This class of kernel functions is closely related to the notions of universal kernels [Steinwart, 2002] and characteristic kernels [Fukumizu et al., 2004]. The former ensures that kernel-based classification/regression algorithms can achieve the Bayes risk, whereas the latter ensures that the kernel mean embeddings can distinguish different probability measures. In principle, they guarantee that the corresponding RKHSs induced by these kernels are sufficiently rich for the tasks at hand. We refer the readers to Sriperumbudur et al. [2011] and Simon-Gabriel and Schölkopf [2018] for more details.

D.2 Detailed Proofs

This section contains detailed proofs of the results that are missing in the main paper. Most of the proofs on consistency and asymptotic normality take advantages of the useful resource by Newey and McFadden [1994]. Readers are referred to it for more detailed discussions on e.g. assumptions.

D.2.1 Proof of Lemma 1

Proof. Since $\mathcal{H}_k$ is the RKHS, we can rewrite (6.2) as

$$\begin{aligned} R_k(f) &= \sup_{h \in \mathcal{H}_k, \|h\| \leq 1} \left(\mathbb{E}[(Y - f(X))\langle h, k(Z, \cdot)\rangle_{\mathcal{H}}]\right)^2 \\ &= \sup_{h \in \mathcal{H}_k, \|h\| \leq 1} \left(\langle h, \mathbb{E}[(Y - f(X))k(Z, \cdot)]\rangle_{\mathcal{H}}\right)^2 \\ &= \|\mathbb{E}[(Y - f(X))k(Z, \cdot)]\|^2_{\mathcal{H}_k}, \end{aligned} \tag{D.1}$$

where we used the reproducing property of $\mathcal{H}_k$ in the first equality and the fact that $\mathcal{H}_k$ is a vector space in the last equality. By assumption, $\mathbb{E}[(Y - f(X))k(Z, \cdot)]$ is Bochner integrable [Steinwart and Christmann, 2008, Def. A.5.20]. Hence, we can write (D.1) as

$$\begin{aligned} \|\mathbb{E}[(Y - f(X))k(Z, \cdot)]\|^2_{\mathcal{H}_k} &= \langle \mathbb{E}[(Y - f(X))k(Z, \cdot)], \mathbb{E}[(Y - f(X))k(Z, \cdot)]\rangle_{\mathcal{H}_k} \\ &= \mathbb{E}\left[\langle (Y - f(X))k(Z, \cdot), \mathbb{E}[(Y - f(X))k(Z, \cdot)]\rangle_{\mathcal{H}_k}\right] \\ &= \mathbb{E}\left[\langle (Y - f(X))k(Z, \cdot), (Y' - f(X'))k(Z', \cdot)\rangle_{\mathcal{H}_k}\right] \\ &= \mathbb{E}\left[(Y - f(X))(Y' - f(X'))k(Z, Z')\right], \end{aligned}$$

as required. □

D.2.2 Proof of Theorem 6

Proof. First, the law of iterated expectation implies that

$$\mathbb{E}[(Y - f(X))k(Z, \cdot)] = \mathbb{E}_Z[\mathbb{E}_{XY}[(Y - f(X))k(Z, \cdot)|Z]] = \mathbb{E}_Z[\mathbb{E}_{XY}[Y - f(X)|Z]k(Z, \cdot)].$$

By Lemma 1, we know that $R_k(f) = \|\mathbb{E}[(Y - f(X))k(Z, \cdot)]\|^2_{\mathcal{H}_k}$. As a result, $R_k(f) = 0$ if $\mathbb{E}[Y - f(X) \,|\, z] = 0$ for P_Z-almost all z. To show the converse, we assume that $R_k(f) = 0$ and rewrite it as

$$R_k(f) = \iint_{\mathcal{Z}} g(z)k(z, z')g(z') \, \mathrm{d}z \, \mathrm{d}z' = 0,$$

where we define $g(z) := \mathbb{E}_{XY}[Y - f(X) \,|\, z]p(z)$. Since k is ISPD by assumption, this implies that g is a zero function with respect to P_Z, i.e., $\mathbb{E}[Y - f(X) \,|\, z] = 0$ for P_Z-almost all z. □

D.2.3 Convexity Result

Theorem 12. *If $\mathcal{F}$ is a convex set and Assumptions 1, 2 hold, then the risk R_k given in* (6.3) *is strictly convex on $\mathcal{F}$.*

Proof. Given $\alpha \in (0, 1)$ and any functions $f, g : \mathcal{X} \to \mathbb{R}$, we will show that

$$R_k(\alpha f + (1 - \alpha)g) - \alpha R_k(f) - (1 - \alpha)R_k(g) < 0.$$

By Lemma 1, we know that $R_k(f) = \|\mathbb{E}[(Y - f(X))k(Z,\cdot)]\|^2_{\mathcal{H}_k}$. Hence, we can rewrite the above function as

$$
\begin{aligned}
&R_k(\alpha f + (1-\alpha)g) - \alpha R_k(f) - (1-\alpha)R_k(g) \\
&= \|\mathbb{E}[(Y - \alpha f(X) - (1-\alpha)g(X))k(Z,\cdot)]\|^2_{\mathcal{H}_k} - \alpha\|\mathbb{E}[(Y - f(X))k(Z,\cdot)]\|^2_{\mathcal{H}_k} \\
&\quad - (1-\alpha)\|\mathbb{E}[(Y - g(X))k(Z,\cdot)]\|^2_{\mathcal{H}_k} \\
&\overset{(a)}{=} \alpha(\alpha-1)\|\mathbb{E}[(Y - f(X))k(Z,\cdot)]\|^2_{\mathcal{H}_k} + \alpha(\alpha-1)\|\mathbb{E}[(Y - g(X))k(Z,\cdot)]\|^2_{\mathcal{H}_k} \\
&\quad - 2\alpha(\alpha-1)\langle\mathbb{E}[(Y - f(X))k(Z,\cdot)], \mathbb{E}[(Y - g(X))k(Z,\cdot)]\rangle_{\mathcal{H}_k} \\
&\overset{(b)}{=} \alpha(\alpha-1)\underbrace{\|\mathbb{E}[(f(X) - g(X))k(Z,\cdot)]\|^2_{\mathcal{H}_k}}_{>0} < 0
\end{aligned}
$$

The equality (a) is obtained by considering $Y = \alpha Y + (1-\alpha)Y$ in $\|\mathbb{E}[(Y - \alpha f(X) - (1-\alpha)g(X))k(Z,\cdot)]\|^2_{\mathcal{H}_k}$ on the left hand side of (a). We note that the right hand side of (a) is quadratic in $\|\mathbb{E}[(Y - f(X))k(Z,\cdot)]\|_{\mathcal{H}_k}$ and $\|\mathbb{E}[(Y - g(X))k(Z,\cdot)]\|_{\mathcal{H}_k}$, and can be further expressed as a square binomial as the right hand side of (b). Therefore, the convexity follows from the fact that k is the ISPD kernel, $\|\mathbb{E}[(f(X) - g(X))|Z]\|_2 \neq 0$ and $\alpha(\alpha-1) < 0$. □

D.2.4 Uniform Convergence of Risk Functionals

The results presented in this section are used to prove the consistency of $\hat{f}_V$ and $\hat{f}_U$.

Lemma 2 (Uniform consistency of $\widehat{R}_V(f)$)**.** *Assume that* $\mathbb{E}[|Y|^2] < \infty$*,* $\mathcal{F}$ *is compact,* $\mathbb{E}[\sup_{f\in\mathcal{F}}|f(X)|^2] < \infty$*, and Assumption 1 holds. Then, the risk* $R_k(f)$ *is continuous about* $f \in \mathcal{F}$ *and* $\sup_{f\in\mathcal{F}}|\widehat{R}_V(f) - R_k(f)| \xrightarrow{\text{p}} 0$*.*

Proof. First, let $u := (x,y,z)$, $u' := (x',y',z')$, and $h_f(u,u') := (y - f(x))(y' - f(x'))k(z,z')$ for some $(x,y,z),(x',y',z') \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$. To prove that $\widehat{R}_V$ converges uniformly to R_k, we need to show that (i) $h_f(u,u')$ is continuous at each f with probability one; (ii) $\mathbb{E}_{U,U'}\left[\sup_{f\in\mathcal{F}}|h_f(U,U')|\right] < \infty$, and $\mathbb{E}_{U,U}\left[\sup_{f\in\mathcal{F}}|h_f(U,U)|\right] < \infty$ Newey and McFadden [1994, Lemma 8.5]. To this end, it is easy to see that

$$
\begin{aligned}
|h_f(u,u')| &= |(y - f(x))(y' - f(x'))k(z,z')| \\
&\leq |y - f(x)||y' - f(x')||k(z,z')| \\
&\leq |y - f(x)||y' - f(x')|\sqrt{k(z,z)k(z',z')} \\
&\leq (|y| + |f(x)|)(|y'| + |f(x')|)\sqrt{k(z,z)k(z',z')}.
\end{aligned}
$$

The third inequality follows from the Cauchy-Schwarz inequality. Since $\mathcal{F}$ is compact, every $f \in \mathcal{F}$ has $f(x)$ bounded for $\|x\| < \infty$. In term of $k(\cdot,\cdot)$ is bounded as per Assumption 1, we have $h_f(u,u') < \infty$ and thus $h_f(u,u')$ continuous at each f with

probability one. Furthermore, we obtain the following inequalities

$$\begin{aligned}
\mathbb{E}_{U,U'}\left[\sup_{f\in\mathcal{F}}|h_f(U,U')|\right] &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}(|Y|+|f(X)|)(|Y'|+|f(X')|)\sqrt{k(Z,Z)k(Z',Z')}\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}(|Y|+|f(X)|)\sup_{f\in\mathcal{F}}(|Y'|+|f(X')|)\right]\sup_z k(z,z) \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}(|Y|+|f(X)|)\right]^2\sup_z k(z,z) \\
&= \mathbb{E}\left[|Y|+\sup_{f\in\mathcal{F}}|f(X)|\right]^2\sup_z k(z,z) < \infty \\
\mathbb{E}_{U,U}\left[\sup_{f\in\mathcal{F}}|h_f(U,U)|\right] &\leq \mathbb{E}_u\left[\sup_{f\in\mathcal{F}}(|Y|+|f(X)|)^2\right]\sup_z k(z,z) \\
&= \mathbb{E}\left[(|Y|+\sup_{f\in\mathcal{F}}|f(X)|)^2\right]\sup_z k(z,z) \\
&\leq 2\left(\mathbb{E}\left[|Y|^2\right]+\mathbb{E}\left[\sup_{f\in\mathcal{F}}|f(X)|^2\right]\right)\sup_z k(z,z) < \infty
\end{aligned}$$

Hence, our assertion follows from Newey and McFadden [1994, Lemma 8.5]. □

Lemma 3 (Uniform consistency of $\widehat{R}_U(f)$)**.** *Assume that* $\mathbb{E}[|Y|] < \infty$, $\mathcal{F}$ *is compact,* $\mathbb{E}[|f(X)|] < \infty$ *and Assumption 1 holds. Then,* $R_k(f)$ *is continuous about* f *and* $\sup_{f\in\mathcal{F}}|\widehat{R}_U(f) - R_k(f)| \xrightarrow{\mathrm{P}} 0$.

Proof. First, let $u := (x,y,z)$, $u' := (x',y',z')$, and $h_f(u,u') := (y-f(x))(y'-f(x'))k(z,z')$ for some $(x,y,z),(x',y',z') \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$. To prove the uniform consistency of $\widehat{R}_U$, we need to show that (i) $h_f(u,u')$ is continuous at each f with probability one; (ii) there is $d(u,u')$ with $|h_f(u,u')| \leq d(u,u')$ for all $f \in \mathcal{F}$ and $\mathbb{E}_{U,U'}[d(U,U')] < \infty$ [Newey and McFadden, 1994, Lemma 2.4]; (iii) $(u_i,u_j)_{i\neq j}^{n,n}$ has strict stationarity and ergodicity in the sense of Newey and McFadden [1994, Footnote 18 in P.2129]. To this end, it is easy to see that

$$\begin{aligned}
|h_f(u,u')| &= |(y-f(x))(y'-f(x'))k(z,z')| \\
&\leq |y-f(x)||y'-f(x')||k(z,z')| \\
&\leq |y-f(x)||y'-f(x')|\sqrt{k(z,z)k(z',z')} \\
&\leq (|y|+|f(x)|)(|y'|+|f(x')|)\sqrt{k(z,z)k(z',z')} \equiv d(u,u').
\end{aligned}$$

The third inequality follows from the Cauchy-Schwarz inequality. Since $\mathcal{F}$ is compact, every $f \in \mathcal{F}$ has $f(x)$ bounded for $\|x\| < \infty$. In terms of $k(\cdot,\cdot)$ is bounded as per Assumption 1, we have $h_f(u,u') < \infty$ and thus it proves that (i) $h_f(u,u')$ is continuous

at each f with probability one. To prove (ii) $\mathbb{E}_{U,U'}[d(U,U')] < \infty$, we show that

$$\mathbb{E}_{U,U'}[d(U,U')] \leq \mathbb{E}\left[|Y| + |f(X)|\right]^2 \sup_z k(z,z) < \infty$$

Furthermore, we show that $(u_i, u_j)_{i\neq j}^{n,n}$ has strict stationarity and ergodicity. Strict stationarity means that the distribution of a set of data $(u_i, u_{j\neq i})_{i=1,j=1}^{i+m,j+m'}$ does not depend on the starting indices i, j for any m and m', which is easy to check. Ergodicity means that $\widehat{R}_U(f) \xrightarrow{\text{p}} R_k(f)$ for all $f \in \mathcal{F}$ and $\mathbb{E}[|h_f(U,U')|] < \infty$. We have already shown that $h_f(u,u')$ is bounded and $\mathbb{E}[|h_f(U,U')|] < \infty$, so $\widehat{R}_U(f) \xrightarrow{\text{p}} R_k(f)$ follows by Hoeffding [1963, P.25]. Therefore, ergodicity holds, and we have shown all conditions required by extended results of Newey and McFadden [1994, Lemma 2.4]. Then, it follows that $\sup_{f\in\mathcal{F}} |\widehat{R}_U(f) - R_k(f)| \xrightarrow{\text{p}} 0$ and $R_k(f)$ is continuous. □

D.2.5 Indefiniteness of Weight Matrix W_U

Theorem 13. *If Assumption 1 holds, W_U is indefinite.*

Proof. By definition, we have

$$W_U = \frac{1}{n(n-1)}[K(\boldsymbol{z},\boldsymbol{z}) - \text{diag}(k(z_1,z_1),\ldots,k(z_n,z_n))] = \frac{1}{n(n-1)}K_U,$$

where $\text{diag}(a_1,\ldots,a_n)$ denotes an $n \times n$ diagonal matrix whose diagonal elements are $a_1,\ldots,a_n$. We can see that the diagonal elements of K_U are zeros and therefore $\text{trace}(W_U) = 0$. Let us denote the eigenvalues of W_U by $\{\lambda_i\}_{i=1}^n$. Since $\sum_{i=1}^n \lambda_i = \text{trace}(W_U)$, we conclude that there exist both positive and negative eigenvalues (all eigenvalues being zeros yields trivial $W_U = \mathbf{0}$). As a result, W_U is indefinite. □

D.2.6 Consistency of $\hat{f}_V$ with Convex $\Omega(f)$

Theorem 14 (Consistency of $\hat{f}_V$ with convex $\Omega(f)$)**.** *Assume that $\mathcal{F}$ is a convex set, f^* is an interior point of $\mathcal{F}$, $\Omega(f)$ is convex about f, $\lambda \xrightarrow{\text{p}} 0$ and Assumptions 1, 2 holds. Then, $\hat{f}_V$ exists with probability approaching one and $\hat{f}_V \xrightarrow{\text{p}} f^*$.*

Proof. Given $\Omega(f)$ is convex about f, we prove the consistency based on Newey and McFadden [1994, Theorem 2.7] which requires (i) $R_k(f)$ is uniquely maximized at f^*; (ii) $\widehat{R}_V(f) + \lambda\Omega(f)$ is convex; (iii) $\widehat{R}_V(f) + \lambda\Omega(f) \xrightarrow{\text{p}} R_k(f)$ for all $f \in \mathcal{F}$.

Recall that $\widehat{R}_V(f) = \|\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))k(z_i,\cdot)\|_{\mathcal{H}_k}^2$, and by the law of large number, we have that $\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))k(z_i,\cdot) \xrightarrow{\text{p}} \mathbb{E}[(Y - f(X))k(Z,\cdot)]$. Then $\widehat{R}_V(f) \xrightarrow{\text{p}} R_k(f)$ follows from the Continuous Mapping Theorem [Mann and Wald, 1943] based on the fact that the function $g(\cdot) = \|\cdot\|_{\mathcal{H}_k}^2$ is continuous. As $\lambda \xrightarrow{\text{p}} 0$, we obtain (iii) $\widehat{R}_V(f) + \lambda\Omega(f) \xrightarrow{\text{p}} R_k(f)$ by Slutsky's theorem [Van der Vaart, 2000, Lemma 2.8]. Besides, it is easy to see that $\widehat{R}_V(f)$ is convex because the weight matrix W_V is positive definite, and (ii) $\widehat{R}_V(f) + \lambda\Omega(f)$ is convex due to convex $\Omega(f)$. Further, the condition

(i) directly follows from Theorem 12, and given that f^* is an interior point of the convex set $\mathcal{F}$, our assertion follows from Newey and McFadden [1994, Theorem 2.7]. □

D.2.7 Proof of Theorem 8

Proof. From the conditions of Lemma 2, we know that $\mathcal{F}$ is compact, $R_k(f)$ is continuous about f and $\sup_{f\in\mathcal{F}} |\widehat{R}_V(f) - R_k(f)| \xrightarrow{\text{p}} 0$. As Assumptions 1, 2 hold, $R_k(f)$ is uniquely minimized at f^*. Based on the conditions that $\Omega(f)$ is bounded and $\lambda \xrightarrow{\text{p}} 0$, we obtain by Slutsky's theorem that

$$\sup_{f\in\mathcal{F}} \left|\widehat{R}_V(f) + \lambda\Omega(f) - R_k(f)\right| \leq \sup_{f\in\mathcal{F}} \left|\widehat{R}_V(f) - R_k(f)\right| + \lambda \sup_{f\in\mathcal{F}} \Omega(f) \xrightarrow{\text{p}} 0.$$

Consequently, we assert the conclusion by Newey and McFadden [1994, Theorem 2.1]. □

D.2.8 Consistency of $\hat{f}_U$

Theorem 15 (Consistency of $\hat{f}_U$)**.** *Assume that conditions of Lemma 3 and Assumption 2 hold, $\Omega(f)$ is a bounded function and $\lambda \xrightarrow{\text{p}} 0$. Then $\hat{f}_U \xrightarrow{\text{p}} f^*$.*

Proof. By the conditions of Lemma 3, we know that $\mathcal{F}$ is compact, $R_k(f)$ is continuous about f and $\sup_{f\in\mathcal{F}} |\widehat{R}_U(f) - R_k(f)| \xrightarrow{\text{p}} 0$. As Assumptions 1, 2 hold, $R_k(f)$ is uniquely minimized at f^*. Based on the conditions that $\Omega(f)$ is bounded and $\lambda \xrightarrow{\text{p}} 0$, we obtain by Slutsky's theorem that

$$\sup_{f\in\mathcal{F}} \left|\widehat{R}_U(f) + \lambda\Omega(f) - R_k(f)\right| \leq \sup_{f\in\mathcal{F}} \left|\widehat{R}_U(f) - R_k(f)\right| + \lambda \sup_{f\in\mathcal{F}} \Omega(f) \xrightarrow{\text{p}} 0.$$

Consequently, we assert the conclusion by Newey and McFadden [1994, Theorem 2.1]. □

D.2.9 Asymptotic Normality of $\hat{\theta}_U$

In this section, we consider the regularized U-statistic risk $\widehat{R}_{U,\lambda}(f_\theta)$. For $u_i := (x_i, y_i, z_i)$ and $u_j := (x_j, y_j, z_j)$, we express it in a compact form

$$\begin{aligned} \widehat{R}_{U,\lambda}(f_\theta) &:= \underbrace{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i}^{n} h_\theta(u_i, u_j)}_{\widehat{R}_U(f_\theta)} + \lambda\Omega(\theta) \\ h_\theta(u_i, u_j) &:= (y_i - f_\theta(x_i))k(z_i, z_j)(y_j - f_\theta(x_j)). \end{aligned}$$

We will assume that f_θ and $\Omega(\theta)$ are twice continuously differentiable about θ. The first-order derivative $\nabla_\theta \widehat{R}_U(f_\theta)$ can also be written as

$$\begin{aligned}\nabla_\theta \widehat{R}_{U,\lambda}(f_\theta) &= \underbrace{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\nabla_\theta h_\theta(u_i,u_j)}_{\nabla_\theta \widehat{R}_U(f_\theta)} + \lambda\nabla_\theta\Omega(\theta)\\ \nabla_\theta h_\theta(u_i,u_j) &= -\left[(y_i - f_\theta(x_i))\nabla_\theta f_\theta(x_j) + (y_j - f_\theta(x_j))\nabla_\theta f_\theta(x_i)\right]k(z_i,z_j).\end{aligned}$$

Asymptotic normality of $\nabla_\theta \widehat{R}_{U,\lambda}(f_{\theta^*})$. We first show the asymptotic normality of $\nabla_\theta \widehat{R}_{U,\lambda}(f_{\theta^*})$. We assume that there exists $z \in \mathcal{Z}$ such that $\mathbb{E}_X[\nabla_\theta f_{\theta^*}(X) \,|\, z]p(z) \neq 0$ or $\mathbb{E}_{XY}[Y - f_{\theta^*}(X) \,|\, z]p(z) \neq 0$. Both terms being equal to zeros for all $z \in \mathcal{Z}$ leads to a singular $\nabla^2_\theta \widehat{R}_U(f_{\theta^*})$ and the asymptotic distribution therefore becomes much more complicated to analyze.

Lemma 4. *Suppose that f_θ and $\Omega(\theta)$ are first continuously differentiable about θ, $\mathbb{E}[\|\nabla_\theta h_{\theta^*}(U,U')\|_2^2] < \infty$, there exists $z \in \mathcal{Z}$ such that $\mathbb{E}_X[\nabla_\theta f_{\theta^*}(X) \,|\, z]p(z) \neq 0$ or $\mathbb{E}_{XY}[Y - f_{\theta^*}(X) \,|\, z]p(z) \neq 0$, and $\sqrt{n}\lambda \xrightarrow{\mathrm{p}} 0$. Then,*

$$\sqrt{n}\nabla_\theta \widehat{R}_{U,\lambda}(f_{\theta^*}) \xrightarrow{\mathrm{p}} N(\mathbf{0}, 4\mathrm{diag}(\mathbb{E}_U[\mathbb{E}^2_{U'}[\nabla_\theta h_{\theta^*}(U,U')]])).$$

Proof. The proof follows from Serfling [1980, Section 5.5.1 and Section 5.5.2] and we need to show that (i) $\nabla_\theta \widehat{R}_U(f_{\theta^*}) \xrightarrow{\mathrm{p}} \mathbf{0}$ and (ii) whether $\mathrm{Var}_U[\mathbb{E}_{U'}[\nabla_\theta h_{\theta^*}(U,U')]] > 0$ or not. (i) can be obtained by the law of large numbers because $\nabla_\theta \widehat{R}_U(f_{\theta^*})$ is a sample average of $\nabla_\theta R_k(f_{\theta^*}) = \mathbf{0}$.

To prove (ii), we first note that $\mathrm{Var}_U[\mathbb{E}_{U'}[\nabla_\theta h_{\theta^*}(U,U')]] = \mathbb{E}_U[\mathbb{E}^2_{U'}[\nabla_\theta h_{\theta^*}(U,U')]] - \underbrace{\mathbb{E}^2_{UU'}[\nabla_\theta h_{\theta^*}(U,U')]}_{=\mathbf{0}} \geq \mathbf{0}$, where equality holds if for any U, there is $\mathbb{E}_{U'}[\nabla_\theta h_{\theta^*}(U,U')] = 0$, i.e.,

$$\begin{aligned}&\mathbb{E}_{U'}[\nabla_\theta h_{\theta^*}(U,U')]\\ &= -\mathbb{E}_{X'Z'}[\nabla_\theta f_{\theta^*}(X')k(Z',Z)](Y - f_{\theta^*}(X)) - \mathbb{E}_{X'Y'Z'}[(Y' - f_{\theta^*}(X'))k(Z',Z)]\nabla_\theta f_{\theta^*}(X)\\ &= \mathbf{0}.\end{aligned}$$

As the above equation holds for any Y, the coefficient of Y must be 0:

$$\mathbb{E}_{X'Z'}[\nabla_\theta f_{\theta^*}(X')k(Z',Z)] = \mathbb{E}_{Z'}[\mathbb{E}_{X'}[\nabla_\theta f_{\theta^*}(X')|Z']k(Z',Z)] = 0,$$

where we note that $\mathbb{E}[\nabla_\theta f_{\theta^*}(X') \,|\, Z']p(Z') = 0$ for any Z' implied by the second function above. Similarly, the coefficient of $\nabla_\theta f_{\theta^*}(X)$ must be zero, which implies that $\mathbb{E}_{X'Y'}[(Y' - f_{\theta^*}(X')) \,|\, Z']p(Z') = 0$ for any Z'. The two coefficients cannot be zero at the same time (otherwise against the given conditions), so $\mathrm{Var}_U[\mathbb{E}_{U'}[\nabla_\theta h_{\theta^*}(U,U')]] > 0$. Further due to the given condition $\mathbb{E}[\|\nabla_{\theta^*} h(U,U')\|_2^2] < \infty$, we obtain $\sqrt{n}\nabla_\theta \widehat{R}_U(f_{\theta^*}) \xrightarrow{\mathrm{p}} N(\mathbf{0}, 4\mathbb{E}_U[\mathbb{E}^2_{U'}[\nabla_\theta h_{\theta^*}(U,U')]])$ as per Serfling [1980, Section 5.5.1]. Finally, as $\sqrt{n}\lambda \xrightarrow{\mathrm{p}} 0$

and $\nabla_\theta \Omega(\theta^*) < \infty$ by the condition that $\Omega(\theta)$ is first continuously differentiable, we assert the conclusion by Slutsky's theorem,

$$\sqrt{n}\nabla_\theta \widehat{R}_{U,\lambda}(f_{\theta^*}) = \sqrt{n}\nabla_\theta \widehat{R}_U(f_{\theta^*}) + \sqrt{n}\lambda\nabla_\theta\Omega(\theta^*) \xrightarrow{\mathrm{p}} N(\mathbf{0}, 4\mathrm{diag}(\mathbb{E}_U[\mathbb{E}^2_{U'}[\nabla_\theta h_{\theta^*}(U,U')]])).$$

This concludes the proof. □

Uniform consistency of $\nabla^2_\theta \widehat{R}_{U,\lambda}(f_\theta)$. Next, we consider the second derivative $\nabla^2_\theta \widehat{R}_{U,\lambda}(f_\theta)$ and show its uniform consistency. In what follows, we denote by $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm. We can express $\nabla^2_\theta \widehat{R}_{U,\lambda}(f_\theta)$ as

$$\begin{aligned}
\nabla^2_\theta \widehat{R}_{U,\lambda}(f_\theta) &= \underbrace{\frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j\neq i}^n \nabla^2_\theta h_\theta(u_i,u_j)}_{\nabla^2_\theta \widehat{R}_U(f_\theta)} + \lambda\nabla^2_\theta\Omega(\theta) \\
\nabla^2_\theta h_\theta(u_i,u_j) &= [\nabla_\theta f_\theta(x_i)\nabla_\theta f_\theta^\top(x_j) - (y_i - f_\theta(x_i))\nabla^2_\theta f_\theta(x_j) \\
&\quad + \nabla_\theta f_\theta(x_j)\nabla_\theta f_\theta^\top(x_i) - (y_j - f_\theta(x_j))\nabla^2_\theta f_\theta(x_i)]k(z_i,z_j).
\end{aligned}$$

Lemma 5. *Suppose that f_θ and $\Omega(\theta)$ are twice continuously differentiable about θ, Θ is compact, $\mathbb{E}\left[|f_\theta(X)|\right] < \infty$, $\mathbb{E}\left[\|\nabla_\theta f_\theta(X)\|_2\right] < \infty$, $\mathbb{E}\left[\|\nabla^2_\theta f_\theta(X)\|_F\right] < \infty$, $\mathbb{E}[|Y|] < \infty$, $\lambda \xrightarrow{\mathrm{p}} 0$ and Assumption 1 holds. Then, $\mathbb{E}[\nabla^2_\theta h_\theta(U,U')]$ is continuous about θ and*

$$\sup_{\theta\in\Theta} \left\| \nabla^2_\theta \widehat{R}_{U,\lambda}(f_\theta) - \mathbb{E}[\nabla^2_\theta h_\theta(U,U')] \right\|_{\mathrm{F}} \xrightarrow{\mathrm{p}} 0.$$

Proof. The proof is similar to that of Lemma 3 and both applies extended results of Newey and McFadden [1994, Lemma 2.4]. As $(u_i,u_j)_{i\neq j}$ being strictly stationary in the sense of Newey and McFadden [1994, Footnote 18 in P.2129] has been shown in Lemma 3, we only need to show that (i) $\nabla^2_\theta h_\theta(u,u')$ is continuous at each $\theta\in\Theta$ with probability one and (ii) there exists $d(u,u') \geq \|\nabla^2_\theta h_\theta(u,u')\|_F$ for all $\theta\in\Theta$ and $\mathbb{E}[d(U,U')] < \infty$. We exploit the triangle inequality of the Frobenius norm and obtain

$$\begin{aligned}
&\left\|\nabla^2_\theta h_\theta(u,u')\right\|_{\mathrm{F}} \\
&\leq \left[2\|\nabla_\theta f_\theta(x)\nabla_\theta f_\theta^\top(x')\|_F + (|y| + |f_\theta(x)|)\|\nabla^2_\theta f_\theta(x')\|_{\mathrm{F}} + (|y'| + |f_\theta(x')|)\|\nabla^2_\theta f_\theta(x)\|_{\mathrm{F}}\right] k(z,z') \\
&\equiv d(u,u'),
\end{aligned}$$

We first show $d(u,u')$ is bounded for bounded u,u'. As f_θ is twice continuously differentiable about θ and Θ is compact, we have $f_\theta(x)$ bounded as well as each entry of $\nabla_\theta f_\theta(x)$ and $\nabla^2_\theta f_\theta(x)$ for $\|x\| < \infty$. Further taking into account that $k(\cdot,\cdot)$ is bounded as per Assumption 1, we know that $d(u,u') < \infty$ if u,u' are bounded, and it follows that (i) $\nabla^2_\theta h_\theta(u,u')$ is continuous at each $\theta\in\Theta$ with probability one as f_θ is twice continuously differentiable.

We then show that (ii) $\mathbb{E}_{U,U'}[d(U,U')] < \infty$ by the following inequalities:

$$\begin{aligned}
&\mathbb{E}_{U,U'}[d(U,U')] \\
&\leq 2\mathbb{E}\left[\|\nabla_\theta f_\theta(X)\nabla_\theta f_\theta^\top(X')\|_F + (|Y| + |f_\theta(X)|)\|\nabla_\theta^2 f_\theta(X')\|_{\mathrm{F}}\right] \sup_z k(z,z) \\
&= 2\left(\underbrace{\mathbb{E}\left[\|\nabla_\theta f_\theta(X)\|_F\right]}_{<\infty}\mathbb{E}\left[\|\nabla_\theta f_\theta^\top(X')\|_F\right] + \underbrace{\mathbb{E}\left[|Y| + |f_\theta(X)|\right]}_{<\infty}\underbrace{\mathbb{E}\left[\|\nabla_\theta^2 f_\theta(X')\|_{\mathrm{F}}\right]}_{<\infty}\right)\sup_z k(z,z) \\
&< \infty.
\end{aligned}$$

Therefore, we obtain $\sup_{\theta\in\Theta}\|\nabla_\theta^2\widehat{R}_U(f_\theta) - \mathbb{E}[\nabla_\theta^2 h_\theta(U,U')]\|_{\mathrm{F}} \xrightarrow{\mathrm{p}} 0$ following from the extended results in the remarks of Newey and McFadden [1994, Lemma 2.4]. Furthermore, from the conditions that $\Omega(\theta)$ is twice continuously differentiable and the parameter space Θ is compact, we obtain that $\|\nabla_\theta^2\Omega(\theta)\|_{\mathrm{F}} < \infty$ for any $\theta \in \Theta$. Finally, it follows from the Slutsky's theorem that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left\|\nabla_\theta^2\widehat{R}_{U,\lambda}(f_\theta) - \mathbb{E}[\nabla_\theta^2 h_\theta(U,U')]\right\|_{\mathrm{F}} \\
&\leq \sup_{\theta\in\Theta}\left\|\nabla_\theta^2\widehat{R}_U(f_\theta) - \mathbb{E}[\nabla_\theta^2 h_\theta(U,U')]\right\|_{\mathrm{F}} + \lambda\sup_{\theta\in\Theta}\left\|\nabla_\theta^2\Omega(\theta)\right\|_{\mathrm{F}} \xrightarrow{\mathrm{p}} 0.
\end{aligned}$$

This concludes the proof. □

Theorem 16 (Asymptotic normality of $\hat{\theta}_U$)**.** *Suppose that $H = \mathbb{E}[\nabla_\theta^2 h_{\theta^*}(U,U')]$ is non-singular, Θ compact, $\mathbb{E}\left[|f_\theta(X)|\right] < \infty$, $\mathbb{E}[|Y|] < \infty$, f_θ and $\Omega(\theta)$ are twice continuously differentiable about θ, $\mathbb{E}\left[\|\nabla_\theta f_\theta(X)\|_2\right] < \infty$, $\mathbb{E}\left[\|\nabla_\theta^2 f_\theta(X)\|_F\right] < \infty$, $\sqrt{n}\lambda \xrightarrow{\mathrm{p}} 0$, $R_k(f_\theta)$ is uniquely minimized at θ^* which is an interior point of Θ, $\mathbb{E}[\|\nabla_\theta h_{\theta^*}(U,U')\|_2^2] < \infty$ and Assumptions 1 hold. Then*

$$\sqrt{n}(\hat{\theta}_U - \theta^*) \xrightarrow{\mathrm{p}} N(\mathbf{0}, 4H^{-1}\mathrm{diag}(\mathbb{E}_U[\mathbb{E}_{U'}^2[h_{\theta^*}(U,U')]])H^{-1}).$$

Proof. The proof follows by Newey and McFadden [1994, Theorem 3.1] and we need to show that (i) $\widehat{\theta}_U \xrightarrow{\mathrm{p}} \theta^*$; (ii) $\widehat{R}_{U,\lambda}(\theta)$ is twice continuously differentiable; (iii) $\sqrt{n}\nabla_\theta\widehat{R}_{U,\lambda}(f_{\theta^*}) \xrightarrow{\mathrm{p}} N(\mathbf{0}, 4\mathbb{E}_U[\mathbb{E}_{U'}^2[h_{\theta^*}(U,U')]])$; (iv) there is $H(\theta)$ that is continuous at θ^* and $\sup_{\theta\in\Theta}\|\nabla_\theta^2\widehat{R}_{U,\lambda}(f_\theta) - H(\theta)\|_F \xrightarrow{\mathrm{p}} 0$; (v) $H(\theta^*)$ is non-singular.

The proof of (i) is very similar to Theorem 15 except that we consider finite dimensional parameter space instead of functional space. For a neat proof, we would like to omit the detailed proof here. We can first show the uniform consistency $\sup_{\theta\in\Theta}|\widehat{R}_{U,\lambda}(f_\theta) - R_k(f_\theta)| \xrightarrow{\mathrm{p}} 0$ and $R_k(f_\theta)$ is continuous about θ similarly to Lemma 3. Here, the proof is based on the conditions $\mathbb{E}[|Y|] < \infty$, Θ is compact, $\mathbb{E}[|f_\theta(X)|] < \infty$ and f_θ is twice continuously differentiable about θ, and Assumption 1 holds. Then, $\widehat{\theta}_U \xrightarrow{\mathrm{p}} \theta^*$ similarly to Theorem 15, because of the extra condition $R_k(f_\theta)$ is uniquely minimized at θ^*.

Furthermore, from the conditions that Θ is compact, f_θ is twice continuously differentiable about θ, $\mathbb{E}\left[|f_\theta(X)|\right] < \infty$, $\mathbb{E}\left[\|\nabla_\theta f_\theta(X)\|_2\right] < \infty$, $\mathbb{E}\left[\|\nabla^2_\theta f_\theta(X)\|_F\right] < \infty$, $\mathbb{E}[|Y|] < \infty$ and $k(z, z')$ is bounded as implied by Assumption 1, we can obtain (ii) $\widehat{R}_{V,\lambda}(\theta)$ is twice continuously differentiable about θ. Given $H = \mathbb{E}[\nabla^2_\theta h_{\theta^*}(U, U')] = \nabla^2_\theta R_k(\theta)$ is non-singular and $R_k(f_\theta)$ is uniquely minimized at θ^*, we can obtain that the Hessian matrix H is positive definite,

$$H = 2\mathbb{E}_{(XYZ),(X'Y'Z')}\left[\left(\nabla_\theta f_{\theta^*}(X)\nabla_\theta f^\top_{\theta^*}(X') - (Y - f_{\theta^*}(X))\nabla^2_\theta f_{\theta^*}(X')\right)k(Z, Z')\right] \succ \mathbf{0}.$$

If for all $z \in \mathcal{Z}$, there is $\mathbb{E}_X[\nabla_\theta f_{\theta^*}(X) \,|\, z]p(z) = \mathbf{0}$ and $\mathbb{E}_{XY}[Y - f_{\theta^*}(X) \,|\, z]p(z) = 0$, then we can see that the above function $H = \mathbf{0}$ which contradicts $H \succ \mathbf{0}$. Therefore, there must exist z s.t. $\mathbb{E}_X[\nabla_\theta f_{\theta^*}(X) \,|\, z]p(z) \neq \mathbf{0}$ or $\mathbb{E}_{XY}[Y - f_{\theta^*}(X) \,|\, z]p(z) \neq 0$. Then, it follows by Lemma 4 that (iii) $\sqrt{n}\nabla_\theta \widehat{R}_{U,\lambda}(f_{\theta^*}) \xrightarrow{\mathrm{p}} N(\mathbf{0}, 4\mathbb{E}_U[\mathbb{E}^2_{U'}[h_{\theta^*}(U, U')]])$.

Finally by Lemma 5, we know that $H(\theta) = \mathbb{E}[\nabla^2_\theta h_\theta(U, U')]$ and $H(\theta^*) = H$, so (iv) and (v) are satisfied. Now, conditions of Newey and McFadden [1994, Theorem 3.1] are all satisfied, so we assert the conclusion. □

D.2.10 Proof of Theorem 9

We restate the notations

$$\begin{aligned}
\widehat{R}_{V,\lambda}(f_\theta) &:= \underbrace{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n h_\theta(u_i, u_j)}_{\widehat{R}_V(f_\theta)} + \lambda\Omega(\theta) \\
h_\theta(u_i, u_j) &:= (y_i - f_\theta(x_i))k(z_i, z_j)(y_j - f_\theta(x_j)),
\end{aligned}$$

Lemma 6. *Suppose that conditions of Lemma 4 hold. Then*

$$\sqrt{n}\nabla_\theta \widehat{R}_{V,\lambda}(f_{\theta^*}) \xrightarrow{\mathrm{p}} N(\mathbf{0}, 4\mathbb{E}_U[\mathbb{E}^2_{U'}[\nabla_\theta h_{\theta^*}(U, U')]]).$$

Proof. As $\mathbb{E}[\|\nabla_\theta h_{\theta^*}(U, U')\|_2^2] < \infty$, $\sqrt{n}\nabla_\theta \widehat{R}_V(f_{\theta^*})$ has the same limit distribution as that of $\sqrt{n}\nabla_\theta \widehat{R}_U(f_{\theta^*})$ by Serfling [1980, Section 5.7.3]. Furthermore, by $\sqrt{n}\lambda \xrightarrow{\mathrm{p}} 0$ and $\nabla_\theta\Omega(\theta^*) < \infty$ from that $\Omega(\theta)$ is first continuously differentiable, we assert the conclusion by Slutsky's theorem

$$\sqrt{n}\nabla_\theta \widehat{R}_{V,\lambda}(f_{\theta^*}) = \sqrt{n}\nabla_\theta \widehat{R}_V + \sqrt{n}\lambda\nabla_\theta\Omega(\theta^*) \xrightarrow{\mathrm{p}} N(\mathbf{0}, 4\mathbb{E}_U[\mathbb{E}^2_{U'}[\nabla_\theta h_{\theta^*}(U, U')]]).$$

□

Lemma 7. *Suppose that f_θ and $\Omega(\theta)$ are twice continuously differentiable about θ, Θ is compact, $\mathbb{E}[\sup_{\theta\in\Theta}|f_\theta(X)|^2] < \infty$, $\mathbb{E}[\sup_{\theta\in\Theta}\|\nabla_\theta f_\theta(X)\|_2^2] < \infty$, $\mathbb{E}[\sup_{\theta\in\Theta}\|\nabla^2_\theta f_\theta(X)\|_F^2] < \infty$, $\mathbb{E}[|Y|^2] < \infty$, $\lambda \xrightarrow{\mathrm{p}} 0$ and Assumption 1 holds. Then, $\mathbb{E}[\nabla^2_\theta h_\theta(U, U')]$ is continuous about θ and $\sup_{\theta\in\Theta}\|\nabla^2_\theta \widehat{R}_V(f_\theta) - \mathbb{E}[\nabla^2_\theta h_\theta(U, U')]\|_{\mathrm{F}} \xrightarrow{\mathrm{p}} 0$.*

Proof. We apply Newey and McFadden [1994, Lemma 8.5] for this proof and need to show (i) $\nabla_\theta^2 h_\theta(u,u')$ is continuous about each $\theta \in \Theta$ with probability one, and (ii) $\mathbb{E}[\sup_{\theta\in\Theta} \|\nabla_\theta^2 h_\theta(U,U')\|_{\mathrm{F}}] < \infty$ and $\mathbb{E}[\sup_{\theta\in\Theta} \|\nabla_\theta^2 h_\theta(U,U)\|_{\mathrm{F}}] < \infty$.

We first see that $\|\nabla_\theta^2 h_\theta(u,u')\|_{\mathrm{F}}$ and $\|\nabla_\theta^2 h_\theta(u,u)\|_{\mathrm{F}}$ are bounded for finite u, u' because f_θ is twice continuously differentiable about θ. It follows that (i) $\nabla_\theta^2 h_\theta(u,u')$ is continuous about θ with probability one. We then derive upper bounds for $\mathbb{E}[\sup_{\theta\in\Theta} \|\nabla_\theta^2 h_\theta(U,U')\|_{\mathrm{F}}]$ and $\mathbb{E}[\sup_{\theta\in\Theta} \|\nabla_\theta^2 h_\theta(U,U)\|_{\mathrm{F}}]$ so as to show their boundedness,

$$
\begin{aligned}
&\mathbb{E}[\sup_{\theta\in\Theta} \|\nabla_\theta^2 h_\theta(U,U')\|_{\mathrm{F}}] \\
&\le 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\nabla_\theta f_\theta^\top(X')\|_F + (|Y| + |f_\theta(X)|)\|\nabla_\theta^2 f_\theta(X')\|_{\mathrm{F}}\right] \sup_z k(z,z) \\
&\le 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\|_2\right]^2 + 2\mathbb{E}\left[|Y| + \sup_{\theta\in\Theta}|f_\theta(X)|\right] \mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(X')\|_{\mathrm{F}}\right] \sup_z k(z,z) \\
&< \infty,
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}[\sup_{\theta\in\Theta} \|\nabla_\theta^2 h_\theta(U,U)\|_{\mathrm{F}}] \\
&\le 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\nabla_\theta f_\theta^\top(X)\|_F + (|Y| + |f_\theta(X)|)\|\nabla_\theta^2 f_\theta(X)\|_{\mathrm{F}}\right] \sup_z k(z,z) \\
&\le 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\|_2^2\right] + 2\mathbb{E}\left[\left(|Y| + \sup_{\theta\in\Theta}|f_\theta(X)|\right)^2\right] \mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(X')\|_{\mathrm{F}}^2\right] \sup_z k(z,z) \\
&\le 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\|_2^2\right] + \mathbb{E}\left[(|Y| + \sup_{\theta\in\Theta}|f_\theta(X)|)^2\right] + \mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(X')\|_{\mathrm{F}}^2\right] \sup_z k(z,z) \\
&\le 2\mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(X)\|_2^2\right] + 2\mathbb{E}\left[|Y|^2 + \sup_{\theta\in\Theta}|f_\theta(X)|^2\right] + \mathbb{E}\left[\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(X')\|_{\mathrm{F}}^2\right] \sup_z k(z,z) \\
&< \infty.
\end{aligned}
$$

Thus, we assert the conclusion by Newey and McFadden [1994, Lemma 8.5]. □

Proof of Theorem 9. The proof is the same as that of Theorem 16 except that $\widehat{R}_U$ is replaced by $\widehat{R}_V$. □

D.2.11 Asymptotic Normality in the Infinite-dimension Case

We firstly state the asymptotic normality theorem for $\hat{f}_U$ and its proof. Afterwards, we provide the proof of Theorem 10 whose proof is a slightly modified version of that of $\hat{f}_U$.

Theorem 17. *Suppose Assumption 1 holds, l is a bounded kernel, k is a uniformly bounded function, and $\lambda \geq \lambda_0$ holds. Also, suppose that $\mathcal{X}$, $\mathcal{Z}$, and $\mathcal{Y}$ are compact spaces, and there exists $s \in (0,2)$ and a constant $C_H > 0$ such that $\log \mathcal{N}(\varepsilon, \mathcal{H}_l, \|\cdot\|_{L^\infty}) \leq C_H \varepsilon^{-s}$ for any $\varepsilon \in (0,1)$. If $\lambda - \lambda_0 = o(n^{-1/2})$ holds, then there exists a Gaussian process $\mathbb{G}_P^*$ such that*

$$\sqrt{n}(\hat{f}_U - f_{\lambda_0}^*) \rightsquigarrow \mathbb{G}_P^* \text{ in } \mathcal{H}_l.$$

An exact covariance of $\mathbb{G}_P^*$ is described in the proof. The proof is based on the uniform convergence of U-processes on the function space [Arcones and Gine, 1993] and the functional delta method using the asymptotic expansion of the loss function [Hable, 2012]. This asymptotic normality allows us to perform statistical inference, such as tests, even in the non-parametric case.

We discuss the boundedness and covering number assumptions in Theorem 10 and Theorem 17. For the boundedness assumption, many common kernels, such as the Gaussian RBF kernel, the Laplacian kernel, and the Mercer kernel, satisfy it. For the covering number, the common kernels above also satisfy it with a certain parameter configuration. For example, the Gaussian kernel (see Section 4 in Steinwart and Christmann [2008]) and the Mercer kernel (explained in Zhou [2002]).

To prove the theorem, we provide some notation. Let P be a probability measure which generates $u = (x, y, z)$ and $\mathcal{W} = \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Also, we define a function $h_f(u, u') = (y - f(x))(y' - f(x'))k(z, z')$. Let $\mathbb{H} := \{h_f : \mathcal{W} \times \mathcal{W} \to \mathbb{R} \mid f \in \mathcal{H}_l\}$. For preparation, we define $P^1 h_f : \mathcal{W} \to \mathbb{R}$ as $P^1 h_f(\cdot) = (\int h_f(u, \cdot) + h_f(\cdot, u) dP(u))/2$ for $h_f \in \mathbb{H}$. For a signed measure Q on $\mathcal{W}$, we define a measure $Q^2 := Q \otimes Q$ on $\mathcal{W} \times \mathcal{W}$. Then, we can rewrite the U-statistic risk as

$$\widehat{R}_U(f) = \frac{(n-2)!}{n!} \sum_{i=1}^{n} \sum_{j \neq i}^{n} h_f(u_i, u_j) =: U_n^2 h_f,$$

where U_n is an empirical measure for the U-statistics. Similarly, we can rewrite the V-statistic risk as

$$\widehat{R}_V(f) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_f(u_i, u_j) =: P_n^2 h_f,$$

where P_n is an empirical measure of u.

Further, we define a functional associated with measure. We consider functional spaces $\mathcal{G}_1 := \{g : \mathcal{W} \times \mathcal{W} \to [0,1] \mid \text{a convex set } \omega \text{ s.t. } g = \mathbf{1}\{\cdot \leq \omega\}\}$ and $\mathcal{G}_2 := \{g : \mathcal{W} \times \mathcal{W} \to [0,1] \mid \exists f, f' \in \mathcal{H}_l, g(u, u') = h_f(u, u')(f'(x) + f(x'))\}$. Note that $\mathcal{G}_1$ contains a functional which corresponds to the U_n^2. Then, we consider a set of

functionals

$$B_S := \left\{ F : \mathcal{G}_1 \cup \mathcal{G}_2 \to \mathbb{R} \mid \exists \text{non-zero finite } Q^2 \text{ s.t. } F(g) = \int g \, dQ^2, \quad \forall g \in \mathcal{G}_1 \cup \mathcal{G}_2 \right\},$$

and also let B_0 be the closed linear span of B_S. For a functional $F \in B_S$, let $\iota(F)$ be a measure which satisfies

$$F(g) = \int g \, d\iota(F).$$

Uniform Central Limit Theorem: We firstly achieve the uniform convergence. Note that a measure P^2 satisfies $P^2 h_f := \mathbb{E}_{(U,U')}[h_f(U,U')]$. The convergence theorem for U-processes is as follows:

Theorem 18 (Theorem 4.4 in Arcones and Gine [1993])**.** *Suppose $\mathbb{H}$ is a set of uniformly bounded class and symmetric functions, such that $\{P^1 h_f \mid h_f \in \mathbb{H}\}$ is a Donsker class and*

$$\lim_{n\to\infty} \mathbb{E}[n^{-1/2} \log \mathcal{N}(n^{-1/2}\varepsilon, \mathbb{H}, \|\cdot\|_{L^1(U_n^2)})] = 0$$

holds, for all $\varepsilon > 0$. Then, we obtain

$$\sqrt{n}(U_n^2 - P^2) \rightsquigarrow 2\mathbb{G}_{P^1}, \text{ in } \ell^\infty(\mathbb{H}).$$

Here, $\mathbb{G}_{P^1}$ denotes a Brownian bridge, which is a Gaussian process on $\mathbb{H}$ with zero mean and a covariance

$$\mathbb{E}_U[P^1 h_f(U) P^1 h'_f(U)] - \mathbb{E}_U[P^1 h_f(U)]\mathbb{E}_U[P^1 h'_f(U)],$$

with $h_f, h'_f \in \mathbb{H}$.

To apply the theorem, we have to show that $\mathbb{H}$ satisfies the condition in Theorem 18. We firstly provide the following bound:

Lemma 8. *For any $f, f' \in \mathcal{H}_l$ such that $\|f\|_{L^\infty} \vee \|f'\|_{L^\infty} \leq B$ holds, $y, y' \in [-B, B]$ and $u, u' \in \mathcal{W}$, we have*

$$|h_f(u,u') - h_{f'}(u,u')| \leq 4B|k(z,z')|\|f - f'\|_{L^\infty}.$$

Proof of Lemma 8. We simply obtain the following:

$$\begin{aligned}
&|h_f(u,u') - h_{f'}(u,u')| \\
&= |(y - f(x))k(z,z')(y' - f(x')) - (y - f'(x))k(z,z')(y' - f'(x'))| \\
&= |k(z,z')||(y - f(x))(y' - f(x')) - (y - f'(x))(y' - f'(x'))| \\
&= |k(z,z')||y'(f'(x) - f(x)) + y(f'(x') - f(x')) + f(x)f(x') - f'(x)f'(x')| \\
&= |k(z,z')||y'(f'(x) - f(x)) + y(f'(x') - f(x')) + f(x')(f(x) - f'(x)) - f'(x)(f(x') - f'(x'))| \\
&\leq |k(z,z')|\{|y'||f'(x) - f(x)| + |y||f'(x') - f(x')| + |f(x')||f(x) - f'(x)| - |f'(x)||f(x') - f'(x')|\} \\
&\leq 4B|k(z,z')|\|f - f'\|_{L^\infty},
\end{aligned}$$

as required. □

Lemma 9. *Suppose the assumptions of Theorem 10 hold. Then, the followings hold:*

1. $\{P^1 h_f \mid h_f \in \mathbb{H}\}$ *is a Donsker class.*
2. *For any* $\varepsilon > 0$*, the following holds:*

$$\lim_{n\to\infty} E[n^{-1/2} \log \mathcal{N}(n^{-1/2}\varepsilon, \mathbb{H}, \|\cdot\|_{L^1(U_n^2)})] = 0.$$

Proof of Lemma 9. For preparation, fix $\varepsilon > 0$ and set $N = \mathcal{N}(\varepsilon, \mathcal{H}_l, \|\cdot\|_{L^\infty})$. Also, let Q be an arbitrary finite discrete measure. Then, by the definition of a bracketing number, there exist N functions $\{f_i \in \mathcal{H}_l\}_{i=1}^N$ such that for any $f \in \mathcal{H}_l$ there exists $i \in \{1, 2, ..., N\}$ such as $\|f - f_i\|_{L^\infty} \leq \varepsilon$.

For the first condition, as shown in Equation (2.1.7) in Van der Vaart and Wellner [1996], it is sufficient to show

$$\sup_Q \log \mathcal{N}(\varepsilon, \{P^1 h_f \mid h_f \in \mathbb{H}\}, \|\cdot\|_{L^2(Q)}) \leq c\varepsilon^{-2\delta} \ (\varepsilon \to 0),$$

for arbitrary $\delta \in (0, 1)$. Here, $c > 0$ is some constant, and Q is taken from all possible finite discrete measure. To this end, it is sufficient to show that $\log \mathcal{N}(\varepsilon, \{P^1 h_f \mid h_f \in \mathbb{H}\}, \|\cdot\|_{L^2(Q)}) \leq c' \log \mathcal{N}(\varepsilon, \mathcal{H}_l, \|\cdot\|_{L^\infty})$ with a constant $c' > 0$. Fix $P^1 h_f \in \{P^1 h_f \mid h_f \in \mathbb{H}\}$ arbitrary, and set f_i which satisfies $\|f - f_i\|_{L^2(Q)} \leq \varepsilon$. Then, we have

$$\begin{aligned}
&\left\| P^1 h_f - P^1 h_{f_i} \right\|_{L^2(Q)}^2 \\
&= \int \left\{ \int (h_f(u,u') + h_f(u',u))/2 - (h_{f_i}(u,u') + h_{f_i}(u',u))/2 \, dP(u') \right\}^2 dQ(u) \\
&= \int \left(\int h_f(u,u') - h_{f_i}(u,u') \, dP(u') \right)^2 dQ(u) \\
&\leq C \int \left(\int |k(z,z')| \, dP(u') \right)^2 dQ(u) \|f - f_i\|_{L^\infty}^2 \\
&\leq C' \|f - f_i\|_{L^\infty}^2 \\
&\leq C' \varepsilon^2,
\end{aligned}$$

with constants $C, C' > 0$. The first inequality follows Lemma 8 with the bounded property of f, f' and $\mathcal{Y}$. The second inequality follows the bounded condition of k in Theorem 17. Hence, the entropy condition shows the first statement.

For the second condition, we have the similar strategy. For any $h_f \in \mathbb{H}$, we consider $i \in \{1, 2, \ldots, N\}$ such that $\|f - f_i\|_{L^\infty} \leq \varepsilon$. Then, we measure the following value

$$\|h_f - h_{f_i}\|_{L^1(U_n^2)} = \int |h_f(u,u') - h_{f_i}(u,u')| \, dU_n^2(u,u') \leq C'' \|f - f_i\|_{L^\infty} \leq C'' \varepsilon,$$

with a constant $C'' > 0$. Hence, we have

$$\mathbb{E}[n^{-1/2}\log\mathcal{N}(n^{-1/2}\varepsilon, \mathbb{H}, \|\cdot\|_{L^1(U_n^2)})] \leq n^{-1/2}\log\mathcal{N}(n^{-1/2}\varepsilon, \mathcal{H}_l, \|\cdot\|_{L^\infty})$$
$$\leq Cn^{-1/2}\left(\frac{n^{1/2}}{\varepsilon}\right)^s = Cn^{(s-1)/2} \to 0, \ (n \to \infty),$$

since $s \in (0,1)$. □

From Theorem 18 and Lemma 9, we rewrite the central limit theorem utilizing terms of functionals. Note that $\iota^{-1}(U_n^2), \iota^{-1}(P^2) \in B_S$ holds. Then, we can obtain

$$\sqrt{n}(\iota^{-1}(U_n^2) - \iota^{-1}(P^2)) \rightsquigarrow 2\mathbb{G}_{P^1} \text{ in } \ell^\infty(\mathbb{H}).$$

Learning Map and Functional Delta Method: We consider a learning map $S : B_S \to \mathcal{H}_l$. For a functional $F \in B_S$, we define

$$S_\lambda(F) := \operatorname*{argmin}_{f\in\mathcal{H}_l} \iota(F)h_f + \lambda\|f\|_{\mathcal{H}_l}^2.$$

Obviously, we have

$$\hat{f} = S_\lambda(\iota^{-1}(U_n^2)), \text{ and } f_{\lambda_0}^* = S_{\lambda_0}(\iota^{-1}(P^2)).$$

We consider a derivative of S_λ in the sense of the Gateau differentiation by the following steps.

Firstly, we define a partial derivative of the map $R_{Q^2}(f)$. To investigate the optimality of the minimizer of

$$R_{Q^2,\lambda}(f) := \int h_f(u,u')\,dQ^2(u,u') + \lambda\|f\|_{\mathcal{H}_l}^2.$$

To this end, we consider the following derivative $\nabla R_{Q^2,\lambda}[f] : \mathcal{H}_l \to \mathcal{H}_l$ with a direction f as

$$\nabla R_{Q^2,\lambda}[f](f') := 2\lambda f + \int \partial_{f,1}h_f(u,u')f'(x) + \partial_{f,2}h_f(u,u')f'(x')\,dQ^2(u,u').$$

Here, $\partial_{f,1}h_f$ is a partial derivative of h_f in terms of the input $f(x)$ as

$$\partial_{f,1}h_f(u,u') = \partial_{t|t=f(x)}(y-t)k(z,z')(y'-f(x')) = -(y'-f(x'))k(z,z'),$$

and $\partial_{f,2}h_f$ follows it respectively. The following lemma validates the derivative:

Lemma 10. *If the assumptions in Theorem 10 hold, then $\nabla R_{Q^2,\lambda}[f]$ is a Gateau-derivative of $R_{Q^2,\lambda}$ with the direction $f \in \mathcal{H}_l$.*

Proof of Lemma 10. We consider a sequence of functions $h_n \in \mathcal{H}_l$ for $n \in \mathbb{N}$, such that $h_n(x) \neq 0, \forall x \in \mathcal{X}$ and $\|h_n\|_{L^\infty} \to 0$ as $n \to \infty$. Then, for $f \in \mathcal{H}_l$, a simple calculation

yields

$$\begin{aligned}
&\left| \frac{R_{Q^2,\lambda}(f+h_n) - R_{Q^2,\lambda}(f) - \nabla R_{Q^2,\lambda}[f](h_n)}{\|h_n\|_{L^\infty}} \right| \\
&\quad \leq \int \|h_n\|_{L^\infty}^{-1} |k(z,z')((y-f(x))h_n(x) \\
&\qquad + (y'-f(x'))h_n(x') + h_n(x)h_n(x')) - \nabla R_{Q^2,\lambda}[f](h_n)| \, dQ^2(u,u') \\
&\quad \leq \int \|h_n\|_{L^\infty}^{-1} |k(z,z')h_n(x)h_n(x')| \, dQ^2(u,u') \\
&\quad \leq \int \|h_n\|_{L^\infty}^{-1} |k(z,z')\|h_n\|_{L^\infty}^2| \, dQ^2(u,u') \\
&\quad \leq \|h_n\|_{L^\infty} \int |k(z,z')| \, dQ^2(u,u') \to 0, \ (n \to \infty).
\end{aligned}$$

The convergence follows the definition of h_n and the absolute integrability of k, which follows the bounded property of k and compactness of $\mathcal{Z}$. Then, we obtain the statement. □

Here, we consider its RKHS-type formulation of $\nabla R_{Q^2,\lambda}$, which is convenient to describe a minimizer. Let $\Phi_l : \mathcal{X} \to \mathcal{H}_l$ be the feature map associated with the RKHS $\mathcal{H}_l$, such that $\langle \Phi_l[x], f \rangle_{\mathcal{H}_l} = f(x)$ for any $x \in \mathcal{X}$ and $f \in \mathcal{H}_l$. Let $\nabla \tilde{R}_{Q^2,\lambda} : \mathcal{H}_l \to \mathcal{H}_l$ be an operator such that

$$\nabla \tilde{R}_{Q^2,\lambda}(f) := 2\lambda f + \int \partial_{f,1} h_f(u,u') \Phi_l[x](\cdot) + \partial_{f,2} h_f(u,u') \Phi_l[x'](\cdot) \, dQ^2(u,u').$$

Obviously, $\nabla R_{Q^2,\lambda}[f](\cdot) = \langle \nabla \tilde{R}_{Q^2,\lambda}(f), \cdot \rangle_{\mathcal{H}_l}$. Now, we can describe the first-order condition of the minimizer of the risk. Namely, we can state that

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{H}_l} R_{Q^2,\lambda}(f) \Leftrightarrow \nabla \tilde{R}_{Q^2,\lambda}(\hat{f}) = 0.$$

This equivalence follows Theorem 7.4.1 and Lemma 8.7.1 in Luenberger [1997].

Next, we apply the implicit function theorem to obtain an explicit formula of the derivative of S. To this end, we consider a second-order derivative $\nabla^2 \tilde{R}_{Q^2,\lambda} : \mathcal{H}_l \to \mathcal{H}_l$ as

$$\nabla^2 \tilde{R}_{Q^2,\lambda}(f) := 2\lambda f + \int k(z,z')(f(x)\Phi_l[x](\cdot) + f(x')\Phi_l[x'](\cdot)) \, dQ^2(u,u'),$$

which follows (b) in Lemma A.2 in Hable [2012]. Its basic properties are provided in the following result:

Lemma 11. *If Assumption 1 and the assumptions in Theorem 10 hold, then $\nabla^2 \tilde{R}_{Q^2,\lambda}$ is a continuous linear operator and it is invertible.*

Proof of Lemma 11. By (b) in Lemma A.2 in Hable [2012], $\nabla^2 \tilde{R}_{Q^2,\lambda}$ is a continuous linear operator. In the following, we define $A : \mathcal{H}_l \to \mathcal{H}_l$ as $A(f) = \int k(z,z') f(x) \Phi_l[x](\cdot) +$

$f(x')\Phi_l[x'](\cdot))\, dQ^2(u,u')$. To show $\nabla^2 \tilde{R}_{Q^2,\lambda}$ is invertible, it is sufficient to show that (i) $\nabla^2 \tilde{R}_{Q^2,\lambda}$ is injective, and (ii) A is a compact operator.

For the injectivity, we fix non-zero $f \in \mathcal{H}_l$ and obtain

$$
\begin{aligned}
&\|\nabla^2 \tilde{R}_{Q^2,\lambda}(f)\|^2_{\mathcal{H}_l} \\
&= \langle 2\lambda f + A(f), 2\lambda f + A(f)\rangle_{\mathcal{H}_l} \\
&= 4\lambda^2 \|f\|^2_{\mathcal{H}_l} + 4\lambda \langle f, A(f)\rangle_{\mathcal{H}_l} + \|A(f)\|^2_{\mathcal{H}_l} \\
&> 4\lambda \langle f, A(f)\rangle_{\mathcal{H}_l} \\
&= 4\lambda \left\langle f, \int k(z,z') f(x) \Phi[x]\, dQ^2(u,u') \right\rangle_{\mathcal{H}_l} + 4\lambda \left\langle f, \int k(z,z') f(x') \Phi[x']\, dQ^2(u,u') \right\rangle_{\mathcal{H}_l} \\
&= 4\lambda \int k(z,z') f(x)^2\, dQ^2(u,u') + 4\lambda \int k(z,z') f(x')^2\, dQ^2(u,u') \\
&\geq 0.
\end{aligned}
$$

The last equality follows the property of Φ_l and the last inequality follows the ISPD property in Assumption 1.

For the compactness, we follow Lemma A.5 in Hable [2012] and obtain that operators $(f \mapsto \int k(z,z') f(x) \Phi_l[x](\cdot)\, dQ^2(u,u'))$ and $(f \mapsto \int k(z,z') f(x') \Phi_l[x'](\cdot)\, dQ^2(u,u'))$ are compact. □

We define the Gateau derivative of S. For a functional $F' \in \ell^\infty(\mathcal{G}_1 \cup \mathcal{G}_2)$, we define the following function

$$
\nabla S_{Q^2,\lambda}(F') := -\nabla^2 \tilde{R}^{-1}_{Q^2,\lambda} \left(\int \partial_{f,1} h_{f_{Q^2}}(u,u') \Phi_l[x](\cdot) + \partial_{f,2} h_{f_{Q^2}}(u,u') \Phi_l[x'](\cdot)\, d\iota(F')(u,u') \right),
$$

where $f_{Q^2} = S_\lambda(\iota^{-1}(Q^2))$ and Q^2 is a signed measure on $\mathcal{W} \times \mathcal{W}$. Then, we provide the following derivative theorem:

Proposition 3. *Suppose the assumptions in Theorem 10 hold. For $F \in B_S$, $F' \in \ell^\infty(\mathcal{G}_1 \cup \mathcal{G}_2)$, and $s \in \mathbb{R}$ such that $F + sF' \in B_S$, $\nabla S_{\iota(F),\lambda}(F')$ is a Gateau-derivative of S_λ, namely,*

$$
\lim_{s \to 0} \left\| \frac{S_\lambda(F + sF') - S_\lambda(F)}{s} - \nabla S_{\iota(F),\lambda}(F') \right\|_{\mathcal{H}_l} = 0.
$$

Proof of Proposition 3. This proof has the following two steps, (i) define a proxy operator Γ, then (ii) prove the statement by the implicit function theorem.

(i) Define Γ: Note that $\iota(F')$ exists since $F + sF' \in B_S$ implies $F' \in B_0$. We define the following operator $\Gamma(s,f,\lambda) : \mathcal{H}_l \to \mathcal{H}_l$ for $f \in \mathcal{H}_l$:

$$
\begin{aligned}
\Gamma(s,f,\lambda) &:= \nabla \tilde{R}_{\iota(F)+s\iota(F'),\lambda} \\
&= 2\lambda f + \int \partial_{f,1} h_f(u,u') \Phi_l[x](\cdot) + \partial_{f,2} h_f(u,u') \Phi_l[x'](\cdot)\, d\iota(F)(u,u') \\
&\quad + s \int \partial_{f,1} h_f(u,u') \Phi_l[x](\cdot) + \partial_{f,2} h_f(u,u') \Phi_l[x'](\cdot)\, d\iota(F')(u,u').
\end{aligned}
$$

For simple derivatives, Lemma A.2 in Hable [2012] provides

$$\nabla_s \Gamma(s, f, \lambda) = \int \partial_{f,1} h_f(u, u') \Phi_l[x](\cdot) + \partial_{f,2} h_f(u, u') \Phi_l[x'](\cdot) \, d\iota(F')(u, u'),$$

and

$$\nabla_f \Gamma(s, f, \lambda) = \nabla^2 \tilde{R}_{\iota(F)+s\iota(F'),\lambda}.$$

(ii) Apply the implicit function theorem: By its definition and the optimal conditions, we have

$$\Gamma(s, f, \lambda) = 0 \Leftrightarrow f = S_\lambda(\iota(F) + s\iota(F')).$$

Also, we obtain

$$\nabla_f \Gamma(0, S_\lambda(F), \lambda) = \nabla^2 \tilde{R}_{\iota(F),\lambda}.$$

Then, for each $\lambda > 0$, by the implicit function theorem, there exists a smooth map $\varphi_\lambda : \mathbb{R} \to \mathcal{H}_l$ such that

$$\Gamma(s, \varphi_\lambda(s), \lambda) = 0, \ \forall s,$$

and it satisfies

$$\nabla_s \varphi_\lambda(0) = -\left(\nabla_f \Gamma(0, \varphi_\lambda(0), \lambda)\right)^{-1} \left(\nabla_s \Gamma(0, \varphi_\lambda(0), \lambda)\right) = \nabla S_{\iota(F),\lambda}(F').$$

Also, we have $\varphi_\lambda(s) = S(Q^2 + s\mu^2)$. Then, we have

$$\begin{aligned} &\lim_{s \to 0} \left\| \frac{S_\lambda(F + sF') - S_\lambda(F')}{s} - \nabla S_{\iota(F),\lambda}(F') \right\|_{\mathcal{H}_l} \\ &= \lim_{s \to 0} \left\| \frac{\varphi_\lambda(s) - \varphi_\lambda(0)}{s} - \nabla_s \varphi_\lambda(0) \right\|_{\mathcal{H}_l} = 0. \end{aligned}$$

Then, we obtain the statement. □

Now, we are ready to prove Theorem 10 and Theorem 17.

Proof of Theorem 17. As a preparation, we mention that S_λ is differentiable in the Hadamard sense, which is Gateau differentiable by Proposition 3. Lemma A.7 and A.8 in Hable [2012] show that $\nabla S_{\iota(F),\lambda,\iota(G)}$ is Hadamard-differentiable for any λ, G and F.

Then, we apply the functional delta method. As shown in Theorem 18 and Lemma 9, we have

$$\sqrt{n}(\iota^{-1}(U_n^2) - \iota^{-1}(P^2)) \rightsquigarrow 2\mathbb{G}_{P^1}.$$

Hence, we obtain

$$\sqrt{n}((\lambda_0/\lambda)\iota^{-1}(U_n^2) - \iota^{-1}(P^2)) = \frac{\lambda_0}{\lambda}\sqrt{n}(\iota^{-1}(U_n^2) - \iota^{-1}(P^2)) + \frac{\sqrt{n}(\lambda - \lambda_0)}{\lambda} \rightsquigarrow 2\mathbb{G}_{P^1},$$

since $\lambda - \lambda_0 = o(n^{-1/2})$. Utilizing the result, we can obtain

$$\sqrt{n}(\hat{f} - f^*_{\lambda_0}) = \sqrt{n}(S_{\lambda_0}((\lambda_0/\lambda)\iota^{-1}(U_n^2)) - S_{\lambda_0}(\iota^{-1}(P^2))) + o_P(1) \rightsquigarrow \nabla S_{P^2,\lambda_0}(2\mathbb{G}_{P^1}),$$

in $\ell^\infty(\mathbb{H})$. The convergence follows the functional delta method. □

Proof of Theorem 10. This proof is completed by substituting the Central Limit Theorem part (Theorem 18) of the proof of Theorem 17. From Section 3 in Akritas et al. [1986], the V- and U- processes have the same limit distribution asymptotically, so the same result holds. □

D.3 Gaussian Process (GP) Interpretation

We present the close connection between the non-parametric model and the Gaussian process (GP) in this section. We will show that the RKHS solution of our objective function (6.6) is equivalent to the maximum function of the posterior distribution of the GP. The relationship is inspired by the similarity of our objective function to that of GLS, and it will be used to derive an efficient cross validation error.

By Mercer theorem, the kernel $l(x, x')$ can be expanded as $l(x, x') = \sum_{i=1}^{N} \lambda_i \phi_i(x)\phi_i(x')$, where $\phi_i(x)$ are orthonormal in $L^2(\mathbb{R}^d)$, a space of square-integrable real-valued functions, and $N < \infty$ for degenerate kernels and $N = \infty$ otherwise. To satisfy for arbitrary $f(x) = \sum_{i=1}^{N} \alpha_i \phi_i(x)$ the reproducing property

$$\left\langle \sum_i \alpha_i \phi_i, l(x, \cdot) \right\rangle_{\mathcal{H}_l} = f(x),$$

we choose $\phi_i(x)$ such that ϕ_i is orthogonal in $\mathcal{H}_l$ as the Mercer expansion is not unique. Based on the reproducing property, we can obtain $\langle \phi_i, \phi_j \rangle_{\mathcal{H}_l} = \delta_{ij}\lambda_i^{-1}$ where $\delta_{ij} = 1$, if $i = j$, and zero otherwise. Similar settings can be found in Walder and Bishop [2017], for example.

Let us consider a GP over a space of functions

$$f(x) = \Phi(x)\boldsymbol{w},$$

where $\Phi(x) = [\phi_1(x), \phi_2(x), \ldots, \phi_N(x)]$ is the feature vector and $\boldsymbol{w}$ is the parameter vector. We assume a prior distribution of $\boldsymbol{w} \sim N(\mathbf{0}, \delta\Lambda)$[1] where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_N$, and $\delta > 0$ is a hyper-parameter, which plays the same role as the regularization hyper-parameter (as we show later).

[1] Throughout the paper, we denote by $N(\mu, \sigma^2)$ a Gaussian distribution with the mean μ and variance σ^2.

This definition is equivalent to assuming that on a set of input $\boldsymbol{x}$, $f(\boldsymbol{x}) \sim \mathrm{GP}(\mathbf{0}, \delta l(\boldsymbol{x}, \boldsymbol{x}))$ because

$$\begin{aligned} \mathbb{E}[f(\boldsymbol{x})] &= \Phi(\boldsymbol{x})\mathbb{E}[\boldsymbol{w}] = \mathbf{0}, \\ \mathrm{Var}[f(\boldsymbol{x})] &= \delta\Phi(\boldsymbol{x})\Lambda\Phi(\boldsymbol{x})^\top = \delta l(\boldsymbol{x}, \boldsymbol{x}). \end{aligned}$$

We refer the interested readers to Rasmussen and Williams [2005] for further details on Gaussian process models.

D.3.1 Likelihood Function

Given the prior distribution on $\boldsymbol{w}$, we aim to characterize the posterior distribution $p(\boldsymbol{w} \mid D) \propto p(D \mid \boldsymbol{w})p(\boldsymbol{w})$ where $D = \{(x_i, y_i, z_i)\}_{i=1}^n$ is an i.i.d. sample of size n, and $f(\boldsymbol{x}) = \Phi(\boldsymbol{x})\boldsymbol{w}$ where the i's row of $\Phi(\boldsymbol{x})$ contains the feature vector of x_i, namely, $\Phi(x_i)^\top$. To define the likelihood $p(D \mid \boldsymbol{w})$, we recall from Lemma 1 that the risk $R_k(f)$ can be expressed in terms of two independent copies of random variables (X, Y, Z) and (X', Y', Z'). To this end, let $D' := \{(x_i', y_i', z_i')\}_{i=1}^n$ be an independent sample of size n with an identical distribution to D. Given a pair of samples (x, y, z) and (x', y', z') from D and D', respectively, we then define the likelihood as

$$\begin{aligned} p(\{(x, y, z), (x', y', z')\} \mid \boldsymbol{w}) &\propto \exp\left[-\frac{1}{2}(y - \Phi(x)\boldsymbol{w})k(z, z')(y' - \Phi(x')\boldsymbol{w})\right] \\ &= \exp\left[-\frac{1}{2}(y - f(x))k(z, z')(y' - f(x'))\right]. \end{aligned}$$

Hence, the likelihood on both D and D' can be expressed as

$$\begin{aligned} p(\{D, D'\} \mid \boldsymbol{w}) &\propto \prod_{i=1}^n \prod_{j=1}^n p(\{(x_i, y_i, z_i), (x_j', y_j', z_j')\} \mid \boldsymbol{w}) \\ &= \exp\left[-\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n (y_i - f(x_i))k(z_i, z_j')(y_j' - f(x_j'))\right]. \end{aligned} \quad \text{(D.17)}$$

In practice, however, we only have access to a single copy of sample, i.e., D, but not D'. One way of constructing D' is through data splitting: the original dataset is split into two halves of equal size where the former is used to construct D and the latter is used to form D'. Unfortunately, this approach reduces the effective sample size that can be used for learning. Alternatively, we propose to estimate the original likelihood (D.17) by using M-estimators: given the full dataset $D = \{(x_i, y_i, z_i)\}_{i=1}^n$, our approximated likelihood can be defined as

$$\begin{aligned} p(D \mid \boldsymbol{w}) &\propto \prod_{i=1}^n \prod_{j=1}^n p(\{(x_i, y_i, z_i), (x_j, y_j, z_j)\} \mid \boldsymbol{w}) \\ &= \exp\left[-\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n (y_i - f(x_i))k(z_i, z_j)(y_j - f(x_j))\right] \end{aligned}$$

$$
\begin{aligned}
&= \exp\left[-\frac{1}{2}(\boldsymbol{y} - f(\boldsymbol{x}))^\top K_z(\boldsymbol{y} - f(\boldsymbol{x}))\right] \\
&= (2\pi)^{n/2}|K_z|^{-1/2}N(f(\boldsymbol{x}) \mid \boldsymbol{y}, K_z^{-1}).
\end{aligned}
$$

The matrix K_z, as a kernel matrix defined above, plays a role of correlating the residuals $y_i - f(x_i)$. Besides, the standard GP regression has a similar likelihood, which is Gaussian but the covariance matrix is computed on $\boldsymbol{x}$ rather than $\boldsymbol{z}$. Based on the likelihood $p(D \mid \boldsymbol{w})$, the maximum likelihood (ML) estimator hence coincides with the minimizer of the non-regularized version of our objective (6.6).

D.3.2 Maximum A Posteriori (MAP)

Combining the above likelihood function with the prior on $\boldsymbol{w}$, the maximum a posteriori (MAP) estimate of $\boldsymbol{w}$ can be obtained by solving

$$
\begin{aligned}
&\underset{w}{\text{argmax}}\ \log p(\boldsymbol{w} \mid D) \\
&= \underset{w}{\text{argmax}}\ \Big[\log p(D \mid \boldsymbol{w}) + \log p(\boldsymbol{w}) - \log p(D)\Big] \\
&= \underset{w}{\text{argmax}}\ \Big[-\frac{1}{2}(\boldsymbol{y} - \Phi(\boldsymbol{x})\boldsymbol{w})^\top K_z(\boldsymbol{y} - \Phi(\boldsymbol{x})\boldsymbol{w}) \\
&\qquad\qquad - \frac{n}{2}\log 2\pi - \frac{1}{2}\log|\delta\Lambda| - \frac{1}{2}\boldsymbol{w}^\top(\delta\Lambda)^{-1}\boldsymbol{w} - \log p(D)\Big] \\
&= \underset{w}{\text{argmax}}\ \Big[-\frac{1}{2}(\boldsymbol{y} - \Phi(\boldsymbol{x})\boldsymbol{w})^\top K_z(\boldsymbol{y} - \Phi(\boldsymbol{x})\boldsymbol{w}) - \frac{1}{2}\boldsymbol{w}^\top(\delta\Lambda)^{-1}\boldsymbol{w}\Big]. \qquad \text{(D.18)}
\end{aligned}
$$

Specifically, setting the first derivative of (D.18) to zero yields the first-order condition

$$
-(\delta\Lambda)^{-1}\hat{\boldsymbol{w}} - \Phi(\boldsymbol{x})^\top K_z\Phi(\boldsymbol{x})\hat{\boldsymbol{w}} + \Phi(\boldsymbol{x})^\top K_z\boldsymbol{y} = 0
$$

and the MAP estimate

$$
\hat{\boldsymbol{w}} = \delta\Lambda\Phi(\boldsymbol{x})^\top K_z(\boldsymbol{y} - \Phi(\boldsymbol{x})^\top\hat{\boldsymbol{w}}). \qquad \text{(D.19)}
$$

Note that we express $\hat{\boldsymbol{w}}$ using the implicit expression. Given a test data point $\boldsymbol{x}^*$, the corresponding prediction can be obtained by computing the posterior expectation, i.e.,

$$
\begin{aligned}
\hat{f}(\boldsymbol{x}^*) &\equiv \mathbb{E}_{f|D}[f(\boldsymbol{x}^*)] \\
&= \Phi(\boldsymbol{x}^*)\hat{\boldsymbol{w}} \\
&= \Phi(\boldsymbol{x}^*)(\delta\Lambda)\Phi(\boldsymbol{x})^\top K_z(\boldsymbol{y} - \Phi(\boldsymbol{x})^\top\hat{\boldsymbol{w}}) \\
&\equiv l(\boldsymbol{x}^*, \boldsymbol{x})\hat{\boldsymbol{\alpha}},
\end{aligned}
$$

where we define $\hat{\boldsymbol{\alpha}} \equiv \delta K_z(\boldsymbol{y} - \Phi(\boldsymbol{x})^\top\hat{\boldsymbol{w}})$. The second equality holds because $\hat{\boldsymbol{w}}$ is the posterior mean. The third equality is obtained by substituting (D.19) into the second equation. Finally, the last equation follows from the Mercer theorem:

$\Phi(\boldsymbol{x}')\Lambda\Phi(\boldsymbol{x})^\top = l(\boldsymbol{x}', \boldsymbol{x})$.

We can see that without using the representer theorem here, we get the same expression of the optimal solution $\hat{f}$ as the one we obtain by invoking the representer theorem in Section 6.3.2. By substituting $\hat{\boldsymbol{\alpha}} = \delta K_z(\boldsymbol{y} - \Phi(\boldsymbol{x})^\top \hat{\boldsymbol{w}})$ back into (D.19), we obtain

$$\hat{\boldsymbol{w}} = \Lambda\Phi(\boldsymbol{x})^\top \hat{\boldsymbol{\alpha}}. \tag{D.20}$$

Furthermore, putting (D.20) into (D.18) also yields the following optimization problem:

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left[\delta^{-1}\boldsymbol{\alpha}^\top L\boldsymbol{\alpha} + (\boldsymbol{y} - L\boldsymbol{\alpha})^\top K_z(\boldsymbol{y} - L\boldsymbol{\alpha})\right] \\ &= \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left[(n^2\delta)^{-1}\boldsymbol{\alpha}^\top L\boldsymbol{\alpha} + (\boldsymbol{y} - L\boldsymbol{\alpha})^\top W(\boldsymbol{y} - L\boldsymbol{\alpha})\right],\end{aligned}$$

where $L = l(\boldsymbol{x}, \boldsymbol{x})$ is the kernel matrix. This problem has the same form as that of the V-statistic objective (6.7) except that $(n^2\delta)^{-1}$ replaces λ. Therefore, we conclude that the predictive mean is equivalent to the optimal function of the V-statistic objective (6.7) given that $(n^2\delta)^{-1} = \lambda$, and δ plays the role of the regularization parameter. Such a GP interpretation is used for an elegant derivation of the efficient cross validation error in Section 6.4.

D.4 Related Works in Reinforcement Learning

Firstly, the idea of "kernel loss" was proposed in Feng et al. [2019] to estimate the value function in reinforcement learning. Similar idea has been used to estimate the importance ratio of two state or state-action distributions in Liu et al. [2018] and Uehara et al. [2020]. Kallus [2018] and Kallus [2020] considered the estimation of the average policy effect (APE) and policy learning. Secondly, in the area of causal inference, Wong and Chan [2018] employed similar technique to estimate an average treatment effect. Despite the methodological similarity to our work, these works did not consider the IV regression setting. To better understand the connection, we first introduce the objective functions of the aforementioned works and then highlight the differences, challenges, and novelties of our work.

Kernel loss in reinforcement learning. Although estimating different subjects f, Liu et al. [2018] and Feng et al. [2019] employ similar population objective functions in the following form:

$$\min_f \mathbb{E}_X \mathbb{E}_{X'}[((\mathcal{A}f)(X) - f(X))k(X, X')((\mathcal{A}f)(X') - f(X'))], \tag{D.21}$$

where $\mathcal{A}$ is a task-specific operator acting on f and X' is an independent copy of X. More specifically, Liu et al. [2018] aims to estimate the importance ratio of state distributions of two policies as $f(X)$. In the task of value function estimation by Feng et al. [2019], $f(X)$ is the value function. In both works, X represents the state variable

in the context of reinforcement learning.

Loss for APE and policy learning. As mentioned by Kallus [2018], the estimation of the APE is similar to that of the average treatment effect, so we introduce the former as an example. By Kallus [2018], the target is to estimate the average effect of a provided policy $\pi_T(X)$ for a treatment $T \in [1, \dots, m]$ conditioned on a covariate X. More specifically, given a set of observations of T, X and the outcome Y, $\{t_i, x_i, y_i\}_{i=1}^n$, the APE estimate is $f := \sum_{i=1}^n w_i y_i$ where $\{w_i\}_{i=1}^n$ are unknown weights. We use bold character $\boldsymbol{z} := [z_i]_{i=1}^n$ to denote a column vector of variables z_i. Then, the weights $\boldsymbol{w}$ are determined by the following objective:

$$\min_{\boldsymbol{w}} (\sum_T \gamma_T^q \mathcal{B}_T^q)^{2/q} + n^{-2} \boldsymbol{w}^\top \Lambda \boldsymbol{w} \tag{D.22}$$
$$\mathcal{B}_T(\boldsymbol{w}) := (\boldsymbol{w}^\top \delta_{\boldsymbol{t}T} - \pi_T(\boldsymbol{x}))^\top k_T(\boldsymbol{x}, \boldsymbol{x})(\boldsymbol{w}^\top \delta_{\boldsymbol{t}T} - \pi_T(\boldsymbol{x})),$$

where $q \in (0, 1)$ and $\gamma_T > 0$ are constants, $\delta_{s,t} = \mathbb{I}[s = t]$ is the Kronecker delta and Λ is a positive definite matrix dependent on $\boldsymbol{y}$. Furthermore, $\pi_T(\boldsymbol{x})$ and $\delta_{\boldsymbol{t}T}$ are (column) function values evaluated on $\boldsymbol{x}$ and $\boldsymbol{t}$, and $k_T(\boldsymbol{x}, \boldsymbol{x})$ is the kernel matrix on $\boldsymbol{x}$ conditioned on T. The policy learning task assumes that no policy is known and requires optimizing both $\boldsymbol{w}$ and π on a slight variant of the above objective.

Comparison between related losses (D.21)**,** (D.22) **and MMR loss.** Compared with our population and empirical risks (6.3) and (6.6), the kernel loss (D.21) and $\mathcal{B}_T(\boldsymbol{w})$ in the loss for APE (D.22) have a similar quadratic form. The difference is that the kernel loss and $\mathcal{B}_T(\boldsymbol{w})$ have variables, which appear in the kernel function, involved in the residual, whereas the kernel function in our objective depends on the instrument Z which does not appear in the residual.

Challenges and novelties of MMR. The difference in losses simplifies the estimation in the related work and requires us to perform new analyses. Specifically, the consistence of the estimators in the above related work holds under mild conditions. That is, in the estimation of the value function, minimizing the population objective function to 0 guarantees a solution for the Bellman equation, which is consistent to the value function according to the unique solution property of the Bellman equation [Feng et al., 2019]; for the importance ratio and the APE estimation, the consistency holds under mild conditions of data distributions [Liu et al., 2018; Kallus, 2018]. In contrast, the estimator by minimizing the MMR risk (6.3) to 0 is hardly consistent to the the true $f(X)$ under mild conditions, because there can be $\hat{f} \neq f$ satisfying $\mathbb{E}[Y - \hat{f}(X) \mid Z] = 0$ almost surely. Therefore, we first introduce the completeness condition in Assumption 2 for the estimation, which is classical in the IV regression field and is needed to identify $f(X)$ from the CMR. Second, a new theoretical analysis is necessary to clarify the nature of the estimation on $f(X)$, such as consistency and asymptotic normality. We develop a novel theory for this point in Section 6.5, which has not been studied by these related work.

Algorithm 3 MMR-IV (Nyström)

Input: Dataset $D = \{(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})\}_{i=1}^{n}$, kernel functions k and l with parameters θ_k and θ_l, regularization parameter λ, leave out data size M, Nyström approximation sample size M', LMOCV times m

Output: predictive value at $\boldsymbol{x}_*$

1: $K = k(\boldsymbol{z}, \boldsymbol{z}; \theta_k)$

2: $\hat{\delta}, \hat{\theta}_l = \mathrm{argmin}_{\delta, \theta_l} \sum_{i=1}^{m} (\boldsymbol{b}^{(i)} - \boldsymbol{y}_M^{(i)})^\top K_M^{(i)} (\boldsymbol{b}^{(i)} - \boldsymbol{y}_M^{(i)})$

3: Equation (6.8)

4: $K/n^2 \approx \widetilde{U}\widetilde{V}\widetilde{U}^\top$ Nyström Approx with M' samples

5: $\hat{\alpha} = \lambda^{-1}[I - \widetilde{U}(\lambda^{-1}\widetilde{U}^\top L \widetilde{U} + \widetilde{V}^{-1})^{-1}\widetilde{U}^\top \lambda^{-1} L]\widetilde{U}\widetilde{V}\widetilde{U}^\top \boldsymbol{y}$

6: **return** $\hat{f}(\boldsymbol{x}_*) = l(\boldsymbol{x}_*, \boldsymbol{x}; \hat{\theta}_l)\hat{\alpha}$

Algorithm 4 MMR-IV (NN)

Input: Dataset $D = \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$, kernel function k with parameters θ_k, NN f_{NN} with parameters θ_{NN}, regularization parameter λ

Output: predictive value at $\boldsymbol{x}_*$

1: Compute $K = k(\boldsymbol{z}, \boldsymbol{z}; \theta_k)$

2: $\hat{f}_{NN} = \mathrm{argmin}_{f_{\mathrm{NN}}} (\boldsymbol{y} - f_{\mathrm{NN}}(\boldsymbol{x}))^\top K (\boldsymbol{y} - f_{\mathrm{NN}}(\boldsymbol{x}))/n^2 + \lambda \|\theta_{\mathrm{NN}}\|_2^2$

3: **return** $\hat{f}_{NN}(\boldsymbol{x}_*)$

D.5 Additional Details of Experiments

In this section, we provide additional details about the experiments as well as more discussions on the experimental results presented in Section 6.6.

D.5.1 Baseline Algorithms

The details of MMR-IV (Nyström) and MMR-IV (NN) algorithms are given in Algorithm 3 and Algorithm 4, respectively. In the experiments, we compare our algorithms to the following baseline algorithms:

- `DirectNN`: A standard least square regression on X and Y using a neural network (NN).
- `2SLS`: A vanilla 2SLS on raw X and Z.
- `Poly2SLS`: The 2SLS that is performed on polynomial features of X and Z via ridge regressions.
- `DeepIV` [Hartford et al., 2017]: A nonlinear extension of 2SLS using deep NNs. We use the implementation available at https://github.com/microsoft/EconML.
- `KernelIV` [Singh et al., 2019]: A generalization of 2SLS by modeling relations among X, Y, and Z as nonlinear functions in RKHSs. We use the publicly available implementation at https://github.com/r4hu1-5in9h/KIV.

- `GMM+NN`: An optimally-weighted GMM [Hansen, 1982] is combined with a NN $f(X)$. The details of this algorithm can be found in Bennett et al. [2019, Section 5].

- `AGMM` [Lewis and Syrgkanis, 2018]: This algorithm models $h(Z)$ by a deep NN and employs a minimax optimization to solve for $f(X)$. The implementation we use is available at https://github.com/vsyrgkanis/adversarial_gmm.

- `DeepGMM` [Bennett et al., 2019]: This algorithm is a variant of `AGMM` with optimal inverse-covariance weighting matrix. The publicly available implementation at https://github.com/CausalML/DeepGMM is used and all of the above baselines are provided in the package except KernelIV.

- `AGMM-K` [Dikkala et al., 2020]: This algorithm extends AGMM by modeling $h(Z)$ and $f(X)$ as RKHSs. Nyström approximation is applied for fast computation. The publicly available implementation at https://github.com/microsoft/AdversarialGMM is used.

D.5.2 Experimental Settings on Low-dimensional Scenario

For the experiments in Section 6.6.1, we consider both small-sample ($n = 200$) and large-sample ($n = 2000$) regimes, in which n points are sampled for training, validation and test sets, respectively. In both regimes, we standardize the values of Y to have zero mean and unit variance for numerical stability. Hyper-parameters of NNs in Algorithm 4, including the learning rate and the regularization parameter, are chosen by 2-fold CV for fair comparisons with the baselines. These hyper-parameters are provided in Appendix D.5.4. Besides, we use the well-tuned hyper-parameter selections of baselines provided in their packages without changes. We fix the random seed to 527 for all data generation and model initialization.

In contrast to our NN-based method, the RKHS-based method in Algorithm 3 has the analytic form of CV error. We combine the training and validation sets to perform leave-2-out CV to select parameters of the kernel l and the regularization parameter λ. We choose the sum of Gaussian kernels for k and the Gaussian kernel for l. For the Nyström approximation, we subsample 300 points from the combined set. As a small subset of the Gram matrix is used as Nyström samples, which misses much information of data, we avoid outliers in test results by averaging 10 test errors with different Nyström approximations in each experiment. All methods are repeated 10 times on each dataset with random initialization.

Detailed comments on the experimental results. First, under the influence of confounders, `DirectNN` performs worst as it does not use instruments. Second, `MMR-IV`s perform reasonably well in both small-sample and large-sample regimes. For the linear function, `2SLS` and `Poly2SLS` tend to outperform other algorithms as the linearity assumption is satisfied in this case. For non-linear functions, some NN based methods show competitive performance in certain cases. Notably, `GMM+NN` has unstable performance as the function h in (6.1) is designed manually. `KernelIV` performs quite

Table D.1: The mean square error (MSE) ± one standard deviation in the small-sample regime ($n = 200$).

Algorithm	Function f^*			
	abs	linear	sin	step
`DirectNN`	.143 ± .000	.046 ± .000	.404 ± .006	.253 ± .000
`2SLS`	.564 ± .000	**.003 ± .000**	.304 ± .000	.076 ± .000
`Poly2SLS`	.125 ± .000	**.003 ± .000**	.164 ± .000	.077 ± .000
`GMM+NN`	.792 ± .000	.203 ± .000	1.56 ± .001	.550 ± .000
`AGMM`	.031 ± .000	.011 ± .000	.330 ± .000	.080 ± .000
`DeepIV`	.204 ± .008	.047 ± .004	.197 ± .004	**.039 ± .001**
`DeepGMM`	*.022 ± .003*	.032 ± .016	.143 ± .030	*.039 ± .002*
`KernelIV`	.063 ± .000	.024 ± .000	*.086 ± .000*	.055 ± .000
`AGMM-K`	12.3 ± .000	1.32 ± .000	1.57 ± .000	1.71 ± .000
`DualIV`	.202 ± .000	0.103 ± .000	.251 ± .000	.362 ± .000
`MMR-IV (NN)`	**.019 ± .003**	*.004 ± .001*	.292 ± .024	.075 ± .008
`MMR-IV (RKHS)`	.030 ± .000	.011 ± .000	**.075 ± .000**	.057 ± .000

well but not as well as our method. `AGMM-K` is similar in principle to our method while its errors are high. We suspect that this is due to the hyper-parameter selection is not flexible enough. Additionally, we observe that the `AGMM-K`'s performance becomes better as the numbers of CV folds and of hyper-parameter candidates increase. A similar observation is also obtained on DualIV. Thus, it show that it is desirable to have the analytical CV error, which is a advantage of our method against other RKHS baselines. Besides, the selection of the weight matrix remains an open question, and although DualIV, AGMM-K and MMRIV (RKHS) rely on the median heuristic to select the bandwidth, we believe that the selection has different effects on different methods and it is difficult to get fair selection. We will leave this problem to future work. Moreover, the performances of the complicated methods like `DeepIV` and `DeepGMM` deteriorate more in the large-sample regimes than the small-sample regimes. We suspect that this is because these methods rely on two NNs and are thus sensitive to different hyper-parameters. In contrast, `MMR-IV (Nyström)` has the advantage of adaptive hyper-parameter selection.

D.5.3 Experimental Settings on High-dimensional Scenario

For experiments in Section 6.6.2, we sample $n = 10,000$ points for the training, validation, and test sets, and run each method 10 times. For `MMR-IV (Nyström)`, we run it only on the training set to reduce computational workload, and use principal component analysis (PCA) to reduce dimensions of X from 728 to 8. We still use the sum of Gaussian kernels for $k(z, z')$, but use the automatic relevance determination (ARD) kernel for $l(x, x')$. We omit `KernelIV` and `DualIV` in this experiment since their kernel parameter selection is not suitable for image data.

Detailed comments on the experimental results. Like Bennett et al. [2019], we

observe that the `DeepIV` code returns NaN when X is high-dimensional. Besides, we run `Ridge2SLS`, which is `Poly2SLS` with fixed linear degree, in place of `Poly2SLS` to reduce computational workload. `AGMM-K+NN`, i.e., the KLayerTrained method in [Dikkala et al., 2020], is a variant of `AGMM-K`, which uses neural networks to model $f(X)$, and to extract features of Z as inputs of the kernel k. To reinforce the flexibility, `AGMM-K+NN` also trains the kernel hyper-parameter, i.e., the lengthscale of the RBF kernel k. `2SLS` has large errors for high-dimensional X because the first-stage regression from Z to X is ill-posed. The average performance of `GMM+NN` suggests that manually designed functions for instruments is insufficient to extract useful information. Furthermore, `MMR-IV (NN)` performs competitively across scenarios. On MNIST_Z, `MMR-IV`s perform better than other methods, which implies using the sum of Gaussian kernels for the kernel k is proper. `DeepGMM` has competitive performance as well. On MNIST_X and MNIST_{XZ}, `MMR-IV (NN)` outperforms all other methods. Compared with `MMR-IV (NN)`, `AGMM-K+NN` has a more flexible kernel k but is lack of good kernel hyper-parameter selection. So, it is vulnerable to local minimum and shows unstable performance. The ARD kernel with PCA of `MMR-IV (Nyström)` fails because the features from PCA are not representative enough. In addition, we observe that `DeepGMM` often produces results that are unreliable across all settings. We suspect that hard-to-optimize objective and complicated optimization procedure are the causes. Compared with `DirectNN`, most of baselines can hardly deal with high-dimensional structured instrumental regressions.

D.5.4 More Details of MMR-IVs

For the kernel function on instruments, we employ the sum of Gaussian kernels

$$k(z, z') = \frac{1}{3}\sum_{i=1}^{3} \exp\left(-\frac{\|z - z'\|_2^2}{2\sigma_{ki}^2}\right)$$

and the Gaussian kernel for $l(x, x') = \exp(-\|x - x'\|_2^2/(2\sigma_l^2))$, where σ_{k1} is chosen as the median interpoint distance of $\boldsymbol{z} = \{z_i\}_{i=1}^n$ and $\sigma_{k2} = 0.1\sigma_{k1}$, $\sigma_{k3} = 10\sigma_{k1}$. The motivation of such a kernel k is to optimize f on multiple kernels, and we leave parameter selection of k to the future work.

As per the dimensions of X, we parametrize f either as a fully connected neural network with leaky ReLU activations and 2 hidden layers, each of which has 100 cells, for non-image data, or a deep convolutional neural network (CNN) architecture for MNIST data. We denote the fully connected neural network as FCNN(100,100) and refer readers to our code release for exact details on our CNN construction. Learning rates and regularization parameters are summarized in Table D.2.

D.5.5 Additional Comments on Mendelian Randomization

For experiments in Section 6.6.3, `KernelIV` also achieves competitive and stable performance across all settings, but not as good as ours. `DirectNN` is always among

Table D.2: Hyper-parameters of neural networks used in the experiments.

Scenario	Model f	Learning Rates	λ
Low Dimensional	FCNN(100,100)	$(10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6})$	$(5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4})$
MNIST_Z	FCNN(100,100)	$(10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6})$	$(5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4})$
MNIST_X	CNN	$(10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6})$	$(5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4})$
MNIST_XZ	CNN	$(10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6})$	$(5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4})$
Mendelian	FCNN(100,100)	$(10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6})$	$(5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4})$

the worst approaches on all settings as no instrument is used. `Poly2SLS` performs accurately on the last two experiments, while presents significant instability with the number of instruments in Figure 6.2(left) because of failure of the hyper-parameter selection. In Figure 6.2(middle) and Figure 6.2(right), we can observe that the performance of most approaches deteriorates as the effect of confounders becomes stronger. `MMR-IV (Nyström)` has promising performance and shows a bit more sensitivity to c_2 than c_1, and the good performance take the advantage of the hyper-parameter selection compared with `AGMM-K` and `DualIV`.

D.5.6 Experimental Settings on Vitamin D Data

We normalize each variable to have a zero mean and unit variance to reduce the influence of different scales. We consider two cases: (i) without instrument and (ii) with instruments. By without instrument, we mean that the W_V matrix of `MMR-IV (Nyström)` becomes an identity matrix. Following Sjolander and Martinussen [2019], we assess the effect of Vitamin D (exposure) on mortality rate (outcome), control the age in the analyses, and use filaggrin as the instrument. We illustrate original Vitamin D, age and death in Figure D.2. We randomly pick (random seed is 527) 300 Nyström samples and use leave-2-out cross validation to select hyper-parameters. The generalized linear models in [Sjolander and Martinussen, 2019] are a linear function in the first step and a logistic regression model in the second step.

In this experiment, we use *age* as a control variable by considering a structural equation model, which is similar to the model (2.13) except the presence of the controlled (exogenous) variable C,

$$Y = f(X, C) + \varepsilon, \quad X = t(Z, C) + g(\varepsilon) + \nu$$

where $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\nu] = 0$. We further assume that the instrument Z satisfies the following three conditions:

(i) *Relevance:* Z has a causal influence on X;

(ii) *Exclusion restriction:* Z affects Y only through X, i.e., $Y \perp\!\!\!\perp Z | X, \varepsilon, C$;

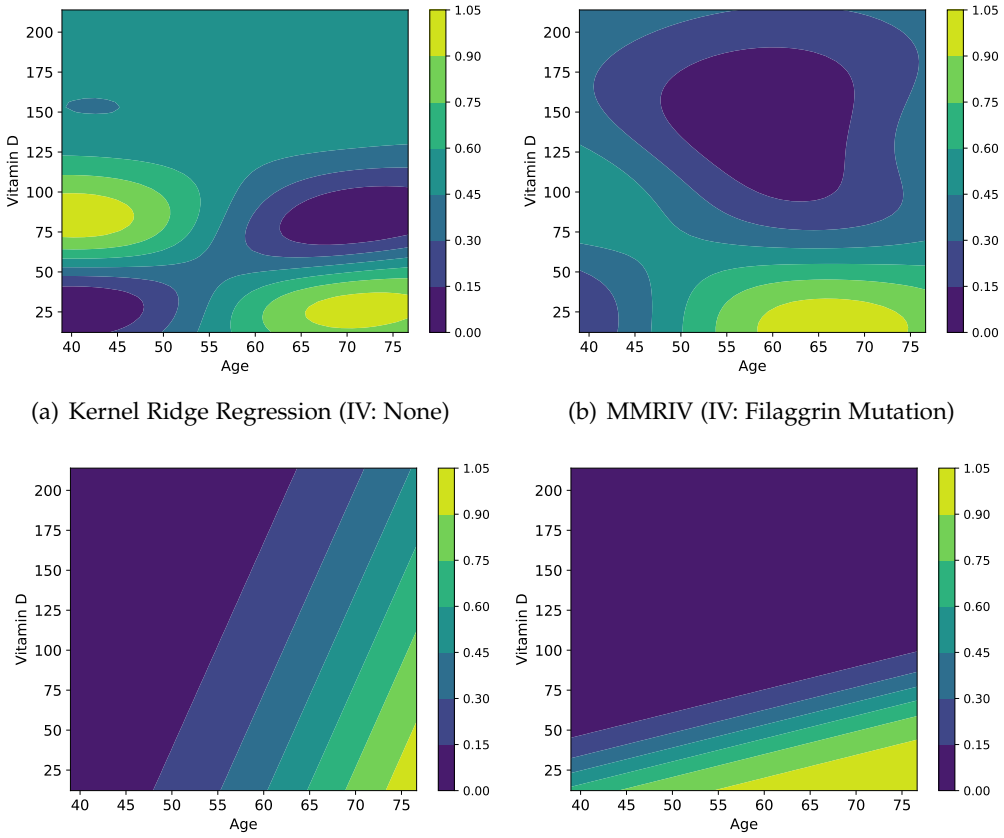
(iii) *Unconfounded instrument(s):* Z is conditionally independent of the error, i.e., $\varepsilon \perp\!\!\!\perp Z \,|\, C$.

Unlike the conditions specified in the main text, (ii) and (iii) also include the controlled variable C. A similar model is employed in Hartford et al. [2017]. From Assumption

(iii), we can see that $\mathbb{E}[\varepsilon \mid C, Z] = \mathbb{E}[\varepsilon \mid C]$, and based on this, we further obtain

$$\mathbb{E}[(Y - f(X, C) - \mathbb{E}[\varepsilon|C])h(Z, C)] = 0$$

for any measurable function h. Note that $\mathbb{E}[\varepsilon \mid C]$ is only conditioned on C, remains constant on arbitrary values of C, and is typically non-zero. To adapt our method to this model, we only need to use the kernel k with (Z, C) as inputs and be aware that the output of the method is an estimate of $f'(X, C) := f(X, C) + \mathbb{E}[\varepsilon|C]$ instead of just $f(X, C)$.



(a) Kernel Ridge Regression (IV: None)

(b) MMRIV (IV: Filaggrin Mutation)

(c) Generalized Linear Model (GLM, IV: None)

(d) 2SLS + GLM (IV: Filaggrin Mutation)

Figure D.1: Estimated effect of vitamin D level on mortality rate, controlled by age. All plots depict normalized contours of $\hat{f}'$ defined in Section D.5.6 where blue represents low mortality rate and yellow the opposite. We can divide each plot roughly into left (young age) and right (old age) parts. While the right parts reflect similar information (i.e., lower vitamin D level at an old age leads to higher mortality rate), the left parts are different. In (a), a high level of vitamin D at a young age can result in a high mortality rate, which is counter-intuitive. A plausible explanation is that it is caused by some unobserved confounders between vitamin D level and mortality rate. In (b), on the other hand, this spurious effect disappears when the filaggrin mutation is used as instrument, i.e., a low vitamin D level at a young age has only a slight effect on death, but a more adverse effect at an old age [Meehan and Penckofer, 2014]. This comparison demonstrates the benefit of an instrument variable. (c) and (d) correspond to the results obtained by using Sjolander and Martinussen [2019]' generalized linear model (GLM), from which we can draw similar conclusions. It is noteworthy that MMRIV allows more flexible non-linearity for causal effect.

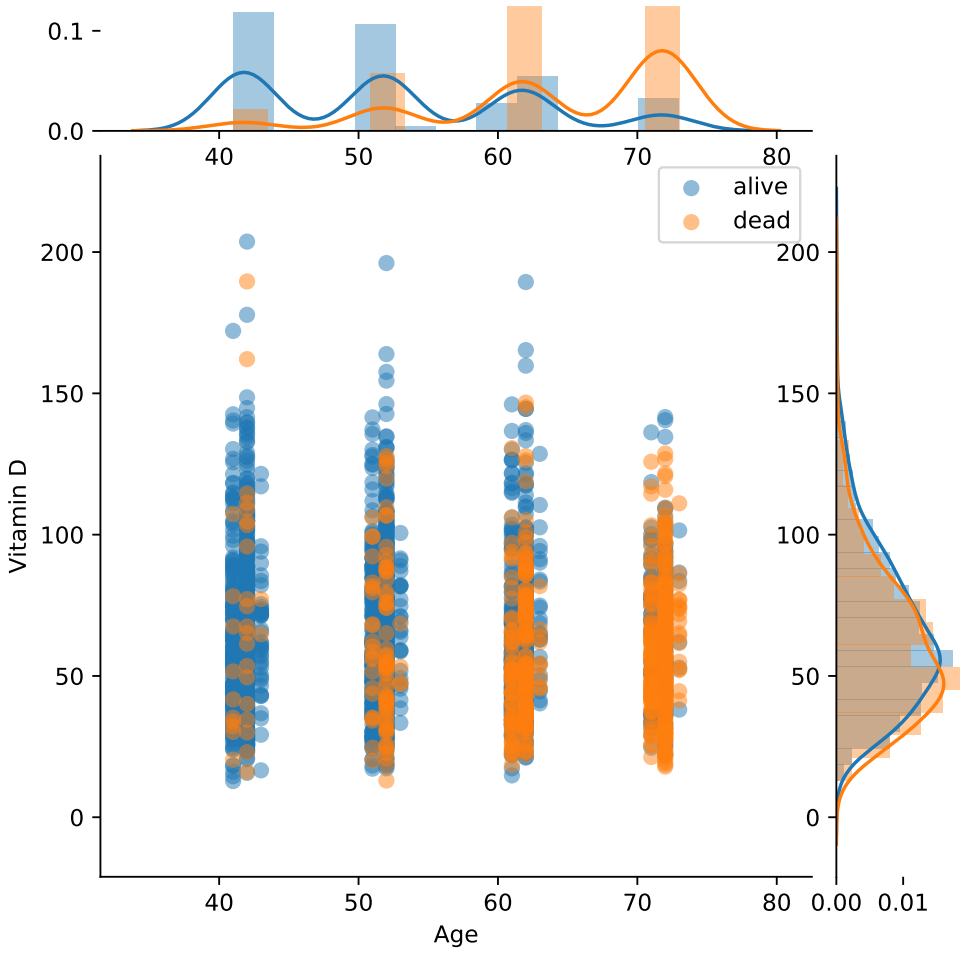


Figure D.2: Distribution of Vitamin D data. Data points are plotted in the middle, the solid curve and histogram on the right describe the kernel density estimation and histogram of Vitamin D, and those on the top are for Age.

Bibliography

Achab, M.; Bacry, E.; Gaïffas, S.; Mastromatteo, I.; and Muzy, J.-F., 2017. Uncovering causality from multivariate hawkes integrated cumulants. *J. Mach. Learn. Res.*, 18, 1 (Jan. 2017), 6998–7025. (cited on page 18)

Adams, R. P.; Murray, I.; and MacKay, D. J. C., 2009. Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. *International Conference on Machine Learning (ICML)*, (2009), 9–16. (cited on page 24)

Ai, C. and Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71, 6 (2003), 1795–1843. (cited on pages 15 and 20)

Akritas, M. G. et al., 1986. Empirical processes associated with V-statistics and a class of estimators under random censoring. *The Annals of Statistics*, 14, 2 (1986), 619–637. (cited on page 115)

Ambrogioni, L.; Güçlü, U.; Güçlütürk, Y.; Hinne, M.; van Gerven, M. A.; and Maris, E., 2018. Wasserstein variational inference. In *Advances in Neural Information Processing Systems*, 2473–2482. (cited on pages 19 and 51)

Ancarani, L. U. and Gasaneo, G., 2008. Derivatives of any order of the confluent hypergeometric function f 1 1 (a, b, z) with respect to the parameter a or b. *Journal of Mathematical Physics*, 49, 6 (2008), 063508. (cited on page 39)

Andersen, T. G. and Sorensen, B. E., 1996. GMM estimation of a stochastic volatility model: A monte carlo study. *Journal of Business & Economic Statistics*, 14, 3 (1996), 328–352. (cited on page 61)

Angrist, J. D.; Imbens, G. W.; and Rubin, D. B., 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 434 (1996), 444–455. (cited on pages 21 and 22)

Angrist, J. D. and Pischke, J.-S., 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. (cited on pages 60 and 71)

Arcones, M. A. and Gine, E., 1993. Limit theorems for U-processes. *The Annals of Probability*, (1993), 1494–1542. (cited on pages 108 and 109)

Arjovsky, M.; Chintala, S.; and Bottou, L., 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, (2017). (cited on pages 19 and 51)

Aronszajn, N., 1950. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68, 3 (1950), 337–404. (cited on page 61)

Athey, S.; Tibshirani, J.; Wager, S.; et al., 2019. Generalized random forests. *Annals of Statistics*, 47, 2 (2019), 1148–1178. (cited on page 20)

Attias, H., 1999. Inferring parameters and structure of latent variable models by variational bayes. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 21–30. Morgan Kaufmann Publishers Inc. (cited on page 40)

Bacry, E.; Bompaire, M.; Gaïffas, S.; and Poulsen, S., 2017. tick: a Python library for statistical learning, with a particular emphasis on time-dependent modeling. *ArXiv e-prints*, (2017). (cited on pages 29 and 43)

Bacry, E.; Mastromatteo, I.; and Muzy, J.-F., 2015. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1, 01 (2015), 1550005. (cited on page 1)

Bacry, E. and Muzy, J.-F., 2016. First- and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62, 4 (2016), 2184–2202. (cited on pages 2, 4, 17, 18, 29, and 45)

Bao, P.; Shen, H.-W.; Jin, X.; and Cheng, X.-Q., 2015. Modeling and predicting popularity dynamics of microblogs using self-excited hawkes processes. In *Proceedings of the 24th International Conference on World Wide Web*, 9–10. ACM. (cited on page 2)

Barabási, A.-L., 2005. The origin of bursts and heavy tails in human dynamics. *Nature*, 435, 7039 (may 2005), 207–11. doi:10.1038/nature03459. URL http://dx.doi.org/10.1038/nature03459. (cited on page 2)

Benamou, J.-D. and Brenier, Y., 2000. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84, 3 (Jan 2000), 375–393. doi:10.1007/s002110050002. (cited on page 20)

Bennett, A.; Kallus, N.; and Schnabel, T., 2019. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems 32*, 3564–3574. Curran Associates, Inc. (cited on pages 16, 20, 21, 22, 61, 68, 70, 121, and 122)

Berlinet, A. and Thomas-Agnan, C., 2004. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers. (cited on page 61)

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning*. Springer. (cited on pages 11 and 12)

Blundell, R.; Chen, X.; and Kristensen, D., 2007. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica*, 75, 6 (2007), 1613–1669. (cited on page 21)

Bonneel, N.; Rabin, J.; Peyré, G.; and Pfister, H., 2015. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51, 1 (2015), 22–45. (cited on page 51)

Bui, T. D.; Yan, J.; and Turner, R. E., 2017. A Unifying Framework for Gaussian Process Pseudo-point Approximations Using Power Expectation Propagation. *J. Mach. Learn. Res.*, 18, 1 (Jan. 2017), 3649–3720. (cited on page 19)

Burgess, S.; Foley, C. N.; Allara, E.; Staley, J. R.; and Howson, J. M. M., 2020. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications*, 11, 1 (2020), 376. (cited on pages 60 and 70)

Burgess, S.; Small, D. S.; and Thompson, S. G., 2017a. A review of instrumental variable estimators for mendelian randomization. *Statistical Methods in Medical Research*, 26, 5 (2017), 2333–2355. (cited on page 20)

Burgess, S.; Small, D. S.; and Thompson, S. G., 2017b. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, 26, 5 (2017), 2333–2355. (cited on page 60)

Byrd, R. H.; Lu, P.; Nocedal, J.; and Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16, 5 (1995), 1190–1208. (cited on page 54)

Card, D., 1999. The causal effect of education on earnings. In *Handbook of Labor Economics* (Eds. O. Ashenfelter and D. Card), vol. 3 of *Handbook of Labor Economics*, chap. 30, 1801–1863. Elsevier. (cited on page 60)

Carrasco, M., 2012. A regularization approach to the many instruments problem. *Journal of Econometrics*, 170, 2 (2012), 383–398. (cited on page 61)

Carrasco, M. and Florens, J.-P., 2000. Generalization of gmm to a continuum of moment conditions. *Econometric Theory*, 16, 6 (2000), 797–834. (cited on page 61)

Carrasco, M. and Florens, J.-P., 2014. On the asymptotic efficiency of gmm. *Econometric Theory*, 30, 2 (2014), 372–406. doi:10.1017/S0266466613000340. (cited on page 61)

Carrasco, M.; Florens, J.-P.; and Renault, E., 2007. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In *Handbook of Econometrics* (Eds. J. Heckman and E. Leamer), vol. 6B, chap. 77. Elsevier, 1 edn. (cited on pages 61 and 64)

Celeux, G. and Diebolt, J., 1985. The sem algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comp. Stat. Quarterly*, 2 (1985), 73–82. (cited on pages 4, 23, and 28)

Chen, X. and Christensen, T. M., 2018. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9, 1 (2018), 39–84. (cited on page 21)

Chen, X. and Pouzo, D., 2012. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80, 1 (2012), 277–321. (cited on page 21)

Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; and Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. (cited on page 20)

Courty, N.; Flamary, R.; Tuia, D.; and Corpetti, T., 2016. Optimal transport for data fusion in remote sensing. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 3571–3574. IEEE. (cited on page 50)

Cuturi, M., 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, 2292–2300. (cited on page 51)

Daley, D. J. and Vere-Jones, D., 2003. *An introduction to the theory of point processes: volume I: Elementary Theory and Methods*. Springer Science & Business Media. (cited on page 7)

Darolles, S.; Fan, Y.; Florens, J. P.; and Renault, E., 2011. Nonparametric instrumental regression. *Econometrica*, 79, 5 (2011), 1541–1565. (cited on page 21)

Daskalakis, C. and Panageas, I., 2018. The limit points of (optimistic) gradient descent in min-max optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 9256–9266. (cited on page 21)

Davey Smith, G. and Ebrahim, S., 2003. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32, 1 (2003), 1–22. (cited on page 20)

de Jong, R. M., 1996. The bierens test under data dependence. *Journal of Econometrics*, 72, 1 (1996), 1–32. (cited on page 61)

Dehaene, G. and Barthelmé, S., 2018. Expectation propagation in the large data limit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 1 (2018), 199–217. (cited on page 20)

Dempster, A. P.; Laird, N. M.; and Rubin, D. B., 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1 (1977), 1–38. (cited on pages 4, 23, and 28)

Dezfouli, A. and Bonilla, E. V., 2015. Scalable inference for gaussian process models with black-box likelihoods. In *Advances in Neural Information Processing Systems*, 1414–1422. (cited on page 19)

D'Haultfoeuille, X., 2011. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27, 3 (2011), 460–471. URL http://www.jstor.org/stable/27975488. (cited on page 62)

Dieng, A. B.; Tran, D.; Ranganath, R.; Paisley, J.; and Blei, D., 2017. Variational Inference via \chi Upper Bound Minimization. In *Advances in Neural Information Processing Systems 30* (Eds. I. Guyon; U. V. Luxburg; S. Bengio; H. Wallach; R. Fergus; S. Vishwanathan; and R. Garnett), 2732–2741. Curran Associates, Inc. (cited on page 50)

Diggle, P. J.; Moraga, P.; Rowlingson, B.; and Taylor, B. M., 2013. Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, 28, 4 (2013), 542–563. (cited on page 24)

Dikkala, N.; Lewis, G.; Mackey, L.; and Syrgkanis, V., 2020. Minimax estimation of conditional moment models. *CoRR*, abs/2006.07201 (2020). (cited on pages 20, 22, 68, 121, and 123)

Dominguez, M. and Lobato, I., 2004. Consistent estimation of models defined by conditional moment restrictions. *Econometrica*, 72, 5 (2004), 1601–1615. (cited on page 61)

Donald, S. G.; Imbens, G. W.; and Newey, W. K., 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117, 1 (2003), 55–93. doi:https://doi.org/10.1016/S0304-4076(03)00118-0. URL https://www.sciencedirect.com/science/article/pii/S0304407603001180. (cited on page 61)

Donald, S. G. and Newey, W. K., 2001. Choosing the number of instruments. *Econometrica*, 69, 5 (2001), 1161–1191. (cited on page 61)

Donnet, S.; Rivoirard, V.; and Rousseau, J., 2018. Nonparametric bayesian estimation of multivariate hawkes processes. *arXiv preprint arXiv:1802.05975*, (2018). (cited on pages 3, 18, and 19)

Dowson, D. and Landau, B., 1982. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12, 3 (1982), 450 – 455. (cited on page 20)

Du, N.; Dai, H.; Trivedi, R.; Upadhyay, U.; Gomez-Rodriguez, M.; and Song, L., 2016. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1555–1564. ACM. (cited on pages 17 and 18)

Dua, D. and Graff, C., 2017. UCI machine learning repository. URL http://archive.ics.uci.edu/ml. (cited on page 55)

Eichler, M.; Dahlhaus, R.; and Dueck, J., 2017. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38, 2 (2017), 225–242. (cited on pages 2, 4, 17, and 18)

El Moselhy, T. A. and Marzouk, Y. M., 2012. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231, 23 (2012), 7815–7850. (cited on page 20)

Feng, Y.; Li, L.; and Liu, Q., 2019. A kernel loss for solving the bellman equation. *Advances in neural information processing systems*, 32 (2019). (cited on pages 118 and 119)

Filimonov, V. and Sornette, D., 2015. Apparent criticality and calibration issues in the hawkes self-excited point process model: application to high-frequency financial data. *Quantitative Finance*, 15, 8 (2015), 1293–1314. (cited on page 2)

Flannery, B. P.; Press, W. H.; Teukolsky, S. A.; and Vetterling, W., 1992. Numerical recipes in C. *Press Syndicate of the University of Cambridge, New York*, 24 (1992), 78. (cited on page 65)

Flaxman, S.; Teh, Y. W.; and Sejdinovic, D., 2017. Poisson intensity estimation with reproducing kernels. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54 of *Proceedings of Machine Learning Research*, 270–279. PMLR, Fort Lauderdale, FL, USA. URL http://proceedings.mlr.press/v54/flaxman17a.html. (cited on pages 17, 24, 26, 56, and 87)

Fukumizu, K.; Bach, F. R.; and Jordan, M. I., 2004. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5 (Dec. 2004), 73–99. (cited on page 97)

Gelbrich, M., 1990. On a Formula for the L_2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces. *Mathematische Nachrichten*, 147, 1 (1990), 185–203. doi:10.1002/mana.19901470121. (cited on page 20)

Gelman, A.; Vehtari, A.; Jylänki, P.; Sivula, T.; Tran, D.; Sahai, S.; Blomstedt, P.; Cunningham, J. P.; Schiminovich, D.; and Robert, C., 2017. Expectation propagation as a way of life: A framework for bayesian inference on partitioned data. *arXiv preprint arXiv:1412.4869*, (2017). (cited on pages 3, 20, and 50)

GPy, since 2012. GPy: A Gaussian process framework in python. http://github.com/SheffieldML/GPy. (cited on page 54)

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C., 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*, 5767–5777. (cited on page 19)

Hable, R., 2012. Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis*, 106 (2012), 92–117. (cited on pages 108, 112, 113, and 114)

Hall, A., 2005. *Generalized Method of Moments*. Advanced texts in econometrics. Oxford University Press. (cited on pages 61 and 64)

Hall, A. R. et al., 2005. *Generalized method of moments*. Oxford university press. (cited on page 61)

Hall, P. and Horowitz, J. L., 2005. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33, 6 (12 2005), 2904–2929. (cited on page 21)

Halpin, P. F., 2012. An em algorithm for hawkes process. *Psychometrika*, 2 (2012). (cited on pages 27 and 42)

Hansen, L. P., 1982. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, (1982), 1029–1054. (cited on pages 21, 61, and 121)

Hartford, J.; Lewis, G.; Leyton-Brown, K.; and Taddy, M., 2017. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 1414–1423. PMLR. (cited on pages 16, 21, 22, 68, 120, and 124)

Hartford, J. S.; Veitch, V.; Sridhar, D.; and Leyton-Brown, K., 2020. Valid causal inference with (some) invalid instruments. *CoRR*, abs/2006.11386 (2020). (cited on page 70)

Hartwig, F. P.; Davey Smith, G.; and Bowden, J., 2017. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology*, 46, 6 (2017), 1985–1998. (cited on page 70)

Hawkes, A. G., 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58, 1 (1971), 83–90. (cited on pages 1 and 8)

Hawkes, A. G. and Oakes, D., 1974. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11, 3 (1974), 493–503. (cited on page 1)

Heess, N.; Tarlow, D.; and Winn, J., 2013. Learning to pass expectation propagation messages. In *Advances in Neural Information Processing Systems*, 3219–3227. (cited on pages 5 and 50)

Hensman, J.; Matthews, A.; and Ghahramani, Z., 2015. Scalable variational Gaussian process classification. *Journal of Machine Learning Research*, (2015). (cited on page 50)

Hensman, J.; Zwiessele, M.; and Lawrence, N., 2014. Tilted variational bayes. In *Artificial Intelligence and Statistics*, 356–364. (cited on pages 19 and 50)

Hernández-Lobato, J. M.; Li, Y.; Rowland, M.; Hernández-Lobato, D.; Bui, T.; and Turner, R., 2016. Black-box α-divergence minimization. *International Conference on Machine Learning*, (2016). (cited on pages 5 and 19)

Hoeffding, W., 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58, 301 (1963), 13–30. (cited on page 101)

Horowitz, J. L., 2011. Applied nonparametric instrumental variables estimation. *Econometrica*, 79, 2 (2011), 347–394. (cited on page 21)

Huang, G.; Guo, C.; Kusner, M. J.; Sun, Y.; Sha, F.; and Weinberger, K. Q., 2016. Supervised word mover's distance. In *Advances in Neural Information Processing Systems*, 4862–4870. (cited on page 50)

Ishwaran, H. and James, L. F., 2001. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96, 453 (2001), 161–173. (cited on pages 4, 23, and 25)

Ivanov, V. K.; Vasin, V. V.; and Tanana, V., 2002. *Theory of Linear Ill-posed Problems and Its Applications*. Inverse and ill-posed problems series. VSP. (cited on page 22)

Jarrett, R., 1979. A note on the intervals between coal-mining disasters. *Biometrika*, 66, 1 (1979), 191–193. (cited on page 56)

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K., 1999. An introduction to variational methods for graphical models. *Machine learning*, 37, 2 (1999), 183–233. (cited on pages 12, 19, and 50)

Jylänki, P.; Vanhatalo, J.; and Vehtari, A., 2011. Robust gaussian process regression with a student-t likelihood. *Journal of Machine Learning Research*, 12, Nov (2011), 3227–3257. (cited on pages 13 and 50)

Kallus, N., 2018. Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31 (2018). (cited on pages 118 and 119)

Kallus, N., 2020. Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21, 62 (2020), 1–54. (cited on page 118)

Kariya, T. and Kurata, H., 2004. *Generalized least squares*. John Wiley & Sons. (cited on page 63)

Klungel, O.; Jamal Uddin, M.; de Boer, A.; Belitser, S.; Groenwold, R.; and Roes, K., 2015. Instrumental variable analysis in epidemiologic studies: an overview of the estimation methods. *Pharm Anal Acta*, 6, 353 (2015), 2. (cited on page 60)

Kuang, Z.; Sala, F.; Sohoni, N.; Wu, S.; Córdova-Palomera, A.; Dunnmon, J.; Priest, J.; and Re, C., 2020. Ivy: Instrumental variable synthesis for causal inference. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, vol. 108 of *Proceedings of Machine Learning Research*, 398–410. PMLR. (cited on page 70)

Kurashima, T.; Althoff, T.; and Leskovec, J., 2018. Modeling interdependent and periodic real-world action sequences. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 803–812. International World Wide Web Conferences Steering Committee. (cited on page 17)

Kuss, M. and Rasmussen, C. E., 2005. Assessing approximate inference for binary gaussian process classification. *Journal of machine learning research*, 6, Oct (2005), 1679–1704. (cited on pages xix, 55, and 56)

Lallouache, M. and Challet, D., 2016. The limits of statistical significance of hawkes processes fitted to financial data. *Quantitative Finance*, 16, 1 (2016), 1–11. (cited on page 2)

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 11 (1998), 2278–2324. (cited on page 69)

Lewis, E. and Mohler, G., 2011. A nonparametric em algorithm for multiscale hawkes processes. *Journal of Nonparametric Statistics*, 1, 1 (2011), 1–20. (cited on pages 2, 4, 8, 17, and 19)

Lewis, G. and Syrgkanis, V., 2018. Adversarial generalized method of moments. (cited on pages 16, 21, 22, 60, 68, and 121)

Li, S.; Xiao, S.; Zhu, S.; Du, N.; Xie, Y.; and Song, L., 2018. Learning temporal point processes via reinforcement learning. In *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2018/file/5d50d22735a7469266aab23fd8aeb536-Paper.pdf. (cited on page 18)

Li, Y.; Hernández-Lobato, J. M.; and Turner, R. E., 2015. Stochastic expectation propagation. In *Advances in neural information processing systems*, 2323–2331. (cited on pages 3 and 20)

Li, Y. and Turner, R. E., 2016. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, 1073–1081. (cited on pages 19 and 50)

Liao, L.; Chen, Y.; Yang, Z.; Dai, B.; Wang, Z.; and Kolar, M., 2020. Provably efficient neural estimation of structural equation model: An adversarial approach. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc. (cited on page 21)

Liao, L.; Fu, Z.; Yang, Z.; Kolar, M.; and Wang, Z., 2021. Instrumental variable value iteration for causal offline reinforcement learning. *arXiv preprint arXiv:2102.09907*, (2021). (cited on page 20)

Linderman, S. and Adams, R., 2014. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, 1413–1421. (cited on pages 3 and 18)

Linderman, S. W. and Adams, R. P., 2015. Scalable bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*, (2015). (cited on pages 3, 18, and 19)

Liu, Q.; Li, L.; Tang, Z.; and Zhou, D., 2018. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc. (cited on pages 118 and 119)

Lloyd, C.; Gunter, T.; Osborne, M.; and Roberts, S., 2015. Variational inference for gaussian process modulated poisson processes. In *International Conference on Machine Learning*, 1814–1822. (cited on pages 4, 23, 24, 35, 36, 37, and 41)

Luenberger, D. G., 1997. *Optimization by vector space methods*. John Wiley & Sons. (cited on page 112)

Mallasto, A. and Feragen, A., 2017. Learning from uncertain curves: The 2-wasserstein metric for gaussian processes. In *Advances in Neural Information Processing Systems 30* (Eds. I. Guyon; U. V. Luxburg; S. Bengio; H. Wallach; R. Fergus; S. Vishwanathan; and R. Garnett), 5660–5670. Curran Associates, Inc. (cited on page 20)

Mann, H. B. and Wald, A., 1943. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14, 3 (09 1943), 217–226. (cited on page 101)

Matheron, G., 1963. Principles of geostatistics. *Economic Geology*, 58, 8 (12 1963), 1246–1266. (cited on pages 13 and 50)

Meehan, M. and Penckofer, S., 2014. The role of vitamin D in the aging adult. *Journal of aging and gerontology*, 2, 2 (2014), 60. (cited on pages xvii, 71, and 126)

Mei, H. and Eisner, J. M., 2017. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, 6757–6767. (cited on pages 17 and 18)

Mercer, J., 1909. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209 (1909), 415–446. URL http://www.jstor.org/sle/91043. (cited on pages 10 and 24)

Minka, T., 2004. Power ep. *Dep. Statistics, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep*, (2004). (cited on page 19)

Minka, T., 2005. Divergence measures and message passing. Technical report, Microsoft Research. (cited on pages 5, 19, and 50)

Minka, T. P., 2001a. *A family of algorithms for approximate Bayesian inference*. Ph.D. thesis, Massachusetts Institute of Technology. (cited on page 20)

Minka, T. P., 2001b. The EP energy function and minimization schemes. Technical report, Technical report. (cited on pages 13 and 19)

Minka, T. P., 2001c. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 362–369. Morgan Kaufmann Publishers Inc. (cited on pages 13, 19, and 50)

Mishra, S.; Rizoiu, M.-A.; and Xie, L., 2016. Feature driven and point process approaches for popularity prediction. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 1069–1078. ACM. (cited on pages 2, 17, and 32)

Møller, J.; Syversveen, A. R.; and Waagepetersen, R. P., 1998. Log gaussian cox processes. *Scandinavian journal of statistics*, 25, 3 (1998), 451–482. (cited on pages 24 and 50)

Montavon, G.; Müller, K.-R.; and Cuturi, M., 2016. Wasserstein training of restricted boltzmann machines. In *Advances in Neural Information Processing Systems*, 3718–3726. (cited on page 19)

Muandet, K.; Fukumizu, K.; Sriperumbudur, B.; and Schölkopf, B., 2017. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10, 1-2 (2017), 1–141. (cited on pages 21 and 61)

Muandet, K.; Jitkrittum, W.; and Kübler, J., 2020a. Kernel conditional moment test via maximum moment restriction. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, vol. 124 of *Proceedings of Machine Learning Research*, 41–50. PMLR. (cited on pages 22, 60, 61, and 62)

Muandet, K.; Mehrjou, A.; Lee, S. K.; and Raj, A., 2020b. Dual instrumental variable regression. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc. Forthcoming. (cited on pages 16, 20, 21, and 22)

Muth, J. F., 1961. Rational expectations and the theory of price movements. *Econometrica: Journal of the Econometric Society*, (1961), 315–335. (cited on page 20)

Muzellec, B. and Cuturi, M., 2018. Generalizing point embeddings using the wasserstein space of elliptical distributions. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18 (Montréal, Canada, 2018), 10258–10269. Curran Associates Inc., USA. (cited on page 20)

Neal, R. M., 1997. Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv preprint physics/9701026*, (1997). (cited on page 50)

Newey, W., 1993. Efficient estimation of models with conditional moment restrictions. In *Handbook of Statistics*, vol. 11, chap. 16, 419–454. Elsevier. (cited on pages 15, 20, and 60)

Newey, W. K. and McFadden, D., 1994. Chapter 36 large sample estimation and hypothesis testing. In *Handbook of Econometrics*, vol. 4, 2111 – 2245. Elsevier. (cited on pages 67, 97, 99, 100, 101, 102, 104, 105, 106, and 107)

Newey, W. K. and Powell, J. L., 2003. Instrumental variable estimation of non-parametric models. *Econometrica*, 71, 5 (2003), 1565–1578. (cited on pages 21 and 62)

Newey, W. K. and Smith, R. J., 2004. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72, 1 (2004), 219–255. (cited on page 61)

Omi, T.; ueda, n.; and Aihara, K., 2019. Fully neural network based model for general temporal point processes. In *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2019/file/39e4973ba3321b80f37d9b55f63ed8b8-Paper.pdf. (cited on pages 17 and 18)

Opper, M. and Archambeau, C., 2009. The variational Gaussian approximation revisited. *Neural computation*, 21, 3 (2009), 786–792. (cited on pages 50 and 54)

Opper, M. and Winther, O., 2000. Gaussian processes for classification: Mean-field algorithms. *Neural computation*, 12, 11 (2000), 2655–2684. (cited on pages 3, 13, 19, and 50)

Opper, M. and Winther, O., 2005. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6, Dec (2005), 2177–2204. (cited on page 20)

Owen, D. B., 1956. Tables for computing bivariate normal probabilities. *Ann. Math. Statist.*, 27, 4 (12 1956), 1075–1090. doi:10.1214/aoms/1177728074. (cited on page 86)

Ozaki, T., 1979. Maximum likelihood estimation of hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31, 1 (1979), 145–155. (cited on page 17)

Pascanu, R.; Mikolov, T.; and Bengio, Y., 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13 (Atlanta, GA, USA, 2013), III–1310–III–1318. JMLR.org. (cited on page 18)

Peyré, G.; Cuturi, M.; et al., 2019. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11, 5-6 (2019), 355–607. (cited on pages 14 and 15)

Quiñonero-Candela, J. and Rasmussen, C. E., 2005. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6, 65 (2005), 1939–1959. URL http://jmlr.org/papers/v6/quinonero-candela05a.html. (cited on page 10)

Rasmussen, C. E. and Williams, C. K. I., 2005. *Gaussian Processes for Machine Learning*. The MIT Press. ISBN 026218253X. (cited on pages 11, 24, 25, 52, 84, 85, 94, and 116)

Rasmussen, J. G., 2013. Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15, 3 (2013), 623–642. (cited on pages 3, 18, and 19)

Ripley, B. D., 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press. doi:10.1017/CBO9780511812651. (cited on page 55)

Rizoiu, M.-A.; Mishra, S.; Kong, Q.; Carman, M.; and Xie, L., 2017. SIR-Hawkes: on the Relationship Between Epidemic Models and Hawkes Point Processes. *arXiv preprint*, (nov 2017), 1–16. URL http://arxiv.org/abs/1711.01679. (cited on page 2)

Rizoiu, M.-A.; Mishra, S.; Kong, Q.; Carman, M.; and Xie, L., 2018. Sir-hawkes: Linking epidemic models and hawkes processes to model diffusions in finite populations. In *Proceedings of the 27th International Conference on World Wide Web*, 419–428. International World Wide Web Conferences Steering Committee. (cited on pages 17, 32, and 46)

Rubin, I., 1972. Regular point processes and their detection. *IEEE Transactions on Information Theory*, 18, 5 (1972), 547–557. (cited on page 7)

Rubner, Y.; Tomasi, C.; and Guibas, L. J., 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40, 2 (2000), 99–121. (cited on page 50)

Rue, H.; Martino, S.; and Chopin, N., 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 2 (2009), 319–392. doi:https://doi.org/10.1111/j.1467-9868.2008.00700.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2008.00700.x. (cited on pages 33 and 34)

Rue, H.; Riebler, A.; Sørbye, S. H.; Illian, J. B.; Simpson, D. P.; and Lindgren, F. K., 2017. Bayesian computing with inla: A review. *Annual Review of Statistics and Its Application*, 4, 1 (2017), 395–421. doi:10.1146/annurev-statistics-060116-054045. URL https://doi.org/10.1146/annurev-statistics-060116-054045. (cited on page 33)

Schölkopf, B.; Herbrich, R.; and Smola, A. J., 2001a. A generalized representer theorem. In *Computational Learning Theory*, 416–426. Springer Berlin Heidelberg, Berlin, Heidelberg. (cited on page 22)

Schölkopf, B.; Herbrich, R.; and Smola, A. J., 2001b. A generalized representer theorem. In *International conference on computational learning theory*, 416–426. Springer. (cited on page 64)

Schölkopf, B. and Smola, A., 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA. (cited on page 61)

Seeger, M., 2005. Expectation propagation for exponential families. Technical report, Department of EECS, University of California at Berkeley. (cited on pages 13, 19, and 53)

Serfling, R., 1980. *Approximation theorems of mathematical statistics*. John Wiley & Sons. (cited on pages 60, 63, 103, and 106)

Shafieezadeh-Abadeh, S.; Nguyen, V. A.; Kuhn, D.; and Esfahani, P. M., 2018. Wasserstein distributionally robust kalman filtering. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18 (Montréal, Canada, 2018), 8483–8492. Curran Associates Inc., USA. (cited on page 20)

Shchur, O.; Biloš, M.; and Günnemann, S., 2020a. Intensity-free learning of temporal point processes. *International Conference on Learning Representations (ICLR)*, (2020). (cited on page 18)

Shchur, O.; Gao, N.; Biloš, M.; and Günnemann, S., 2020b. Fast and flexible temporal point processes with triangular maps. In *Advances in Neural Information Processing Systems*, vol. 33, 73–84. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2020/file/00ac8ed3b4327bdd4ebbebcb2ba10a00-Paper.pdf. (cited on page 18)

Shirai, T. and Takahashi, Y., 2003. Random point fields associated with certain fredholm determinants ii: Fermion shifts and their ergodic and gibbs properties. *The Annals of Probability*, 31, 3 (2003), 1533–1564. (cited on page 24)

Simma, A. and Jordan, M. I., 2010. Modeling events with cascades of poisson processes. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10 (Catalina Island, CA, 2010), 546–555. AUAI Press, Arlington, Virginia, United States. URL http://dl.acm.org/citation.cfm?id=3023549.3023614. (cited on page 1)

Simon-Gabriel, C.-J. and Schölkopf, B., 2018. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19, 44 (2018), 1–29. (cited on pages 62 and 97)

Singh, R.; Sahani, M.; and Gretton, A., 2019. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems 32*, 4593–4605. Curran Associates, Inc. (cited on pages 16, 21, 22, 68, and 120)

Sjolander, A. and Martinussen, T., 2019. Instrumental variable estimation with the R package ivtools. *Epidemiologic Methods*, 8, 1 (2019). (cited on pages xvii, 71, 124, and 126)

Snelson, E. and Ghahramani, Z., 2006. Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, 1257–1264. (cited on page 41)

Snelson, E.; Ghahramani, Z.; and Rasmussen, C. E., 2004. Warped gaussian processes. In *Advances in Neural Information Processing Systems 16* (Eds. S. Thrun; L. K. Saul; and B. Scholkopf), 337–344. MIT Press. (cited on page 13)

Song, L.; Fukumizu, K.; and Gretton, A., 2013. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30, 4 (2013), 98–111. (cited on page 21)

Song, L.; Huang, J.; Smola, A.; and Fukumizu, K., 2009. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*. (cited on page 21)

Sriperumbudur, B. K.; Fukumizu, K.; and Lanckriet, G. R. G., 2011. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 (Jul. 2011), 2389–2410. (cited on page 97)

Steinwart, I., 2002. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2 (Mar. 2002), 67–93. (cited on page 97)

Steinwart, I. and Christmann, A., 2008. *Support vector machines*. Springer Science & Business Media. (cited on pages 97, 98, and 108)

Takatsu, A., 2011. Wasserstein geometry of gaussian measures. *Osaka J. Math.*, 48, 4 (12 2011), 1005–1026. (cited on page 20)

Zhang, R.; Imaizumi, M.; Schölkopf, B.; and Muandet, K., 2021a. Maximum moment restriction for instrumental variable regression. *arXiv preprint*, (2021). Submitted to NeurIPS 2021. (cited on pages ix and 3)

Zhang, R.; Muandet, K.; Schölkopf, B.; and Imaizumi, M., 2021b. Instrument space selection for kernel maximum moment restriction. *arXiv preprint*, (2021). Submitted to NeurIPS 2021. (cited on page ix)

Zhang, R.; Walder, C.; Bonilla, E. V.; Rizoiu, M.-A.; and Xie, L., 2020a. Quantile propagation for wasserstein-approximate gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, 21566–21578. Curran Associates, Inc. (cited on pages ix and 3)

Zhang, R.; Walder, C.; and Rizoiu, M.-A., 2020b. Variational inference for sparse gaussian process modulated hawkes proces. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*. New York, U.S.A. (cited on pages ix, 3, and 94)

Zhang, R.; Walder, C.; Rizoiu, M.-A.; and Xie, L., 2019. Efficient non-parametric bayesian hawkes processes. In *International Joint Conference on Artificial Intelligence (IJCAI 2019)*. Macao, China. (cited on pages ix, 3, 42, 43, and 50)

Titsias, M. K., 2009. Variational model selection for sparse gaussian process regression. Technical report, Technical report, School of Computer Science, University of Manchester. (cited on pages 10 and 11)

Tolstikhin, I.; Bousquet, O.; Gelly, S.; and Schoelkopf, B., 2017. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, (2017). (cited on page 51)

Tsybakov, A., 2008. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edn. (cited on page 21)

Uehara, M.; Huang, J.; and Jiang, N., 2020. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, 9659–9668. PMLR. (cited on page 118)

Van der Vaart, A., 2000. *Asymptotic Statistics*. Cambridge University Press. (cited on pages 64 and 101)

Van der Vaart, A. and Wellner, J., 1996. Weak convergence and empirical processes. (cited on page 110)

Vapnik, V. N., 1998. *Statistical Learning Theory*. Wiley-Interscience. (cited on page 63)

Vehtari, A.; Mononen, T.; Tolvanen, V.; Sivula, T.; and Winther, O., 2016. Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *The Journal of Machine Learning Research*, 17, 1 (2016), 3581–3618. (cited on page 66)

Villani, C., 2008. *Optimal transport: old and new*, vol. 338. Springer Science & Business Media. (cited on page 50)

Walder, C. J. and Bishop, A. N., 2017. Fast bayesian intensity estimation for the permanental process. In *International Conference on Machine Learning*, 3579–3588. (cited on pages 4, 18, 23, 24, 26, 27, 56, 86, 94, and 115)

Williams, C. K. I. and Barber, D., 1998. Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 12 (Dec 1998), 1342–1351. doi:10.1109/34.735807. (cited on pages 13 and 50)

Williams, C. K. I. and Seeger, M., 2001. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13* (Eds. T. K. Leen; T. G. Dietterich; and V. Tresp), 682–688. MIT Press. (cited on pages 43 and 65)

Wolfram, R. I., 2019. Mathematica, Version 12.0. Champaign, IL, 2019. (cited on pages 87 and 88)

Wong, R. K. and Chan, K. C. G., 2018. Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105, 1 (2018), 199–213. (cited on page 118)

Wu, F. and Huberman, B. A., 2007. Novelty and collective attention. *PNAS*, 104, 45 (nov 2007), 17599–601. doi:10.1073/pnas.0704916104. URL http://www.pnas.org/content/104/45/17599.abstract. (cited on page 2)

Xiao, S.; Farajtabar, M.; Ye, X.; Yan, J.; Song, L.; and Zha, H., 2017a. Wasserstein learning of deep generative point process models. In *Advances in Neural Information Processing Systems*, 3247–3257. (cited on pages 17 and 18)

Xiao, S.; Yan, J.; Yang, X.; Zha, H.; and Chu, S., 2017b. Modeling the intensity function of point process via recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31. (cited on pages 17 and 18)

Xu, H.; Farajtabar, M.; and Zha, H., 2016. Learning granger causality for hawkes processes. In *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48 of *Proceedings of Machine Learning Research*, 1717–1726. PMLR, New York, New York, USA. (cited on pages 17 and 19)

Xu, H.; Luo, D.; Chen, X.; and Carin, L., 2018. Benefits from superposed hawkes processes. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, (2018), 623–631. (cited on page 30)

Xu, M.; Lakshminarayanan, B.; Teh, Y. W.; Zhu, J.; and Zhang, B., 2014. Distributed bayesian posterior sampling via moment sharing. In *Advances in Neural Information Processing Systems*. (cited on page 20)

Zhang, Q.; Lipani, A.; Kirnap, O.; and Yilmaz, E., 2020. Self-attentive Hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 11183–11193. PMLR. URL http://proceedings.mlr.press/v119/zhang20q.html. (cited on pages 17 and 18)

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J., 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1513–1522. ACM. (cited on pages 1, 32, and 46)

Zhou, D.-X., 2002. The covering number in learning theory. *Journal of Complexity*, 18, 3 (2002), 739–767. (cited on page 108)

Zhou, K.; Zha, H.; and Song, L., 2013. Learning triggering kernels for multi-dimensional hawkes processes. In *International Conference on Machine Learning*, 1301–1309. (cited on pages 2, 4, 17, 19, 29, 43, and 45)

Zou, G., 2004. A modified poisson regression approach to prospective studies with binary data. *American journal of epidemiology*, 159, 7 (2004), 702–706. (cited on page 13)

Zuo, S.; Jiang, H.; Li, Z.; Zhao, T.; and Zha, H., 2020. Transformer hawkes process. In *International Conference on Machine Learning*, 11692–11702. PMLR. (cited on pages 17 and 18)