Isotonic Hawkes Processes

Yichen Wang, Bo Xie, Nan Du {YICHEN.WANG, BO.XIE, DUNAN}@GATECH.EDU
Le Song LSONG@CC.GATECH.EDU
College of Computing, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30332 USA

Abstract

Hawkes processes are powerful tools for modeling the mutual-excitation phenomena commonly observed in event data from a variety of domains, such as social networks, quantitative finance and healthcare records. The intensity function of a Hawkes process is typically assumed to be linear in the sum of triggering kernels, rendering it inadequate to capture nonlinear effects present in real-world data. To address this shortcoming, we propose an *Isotonic-Hawkes* process whose intensity function is modulated by an additional nonlinear link function. We also developed a novel iterative algorithm which learns both the nonlinear link function and other parameters provably. We showed that Isotonic-Hawkes processes can fit a variety of nonlinear patterns which cannot be captured by conventional Hawkes processes, and achieve superior empirical performance in real world applications.

1. Introduction

Temporal point processes are powerful tools for modeling the complex dynamics of events occurrences. In particular, Hawkes processes (Hawkes, 1971) are well-suited to capture the phenomenon of mutual excitation between the occurrence of events, and have been applied successfully in modeling criminal retaliations (Mohler et al., 2011), online users behaviors (Farajtabar et al., 2014; 2015; Du et al., 2015), and opinion dynamics (Wang et al., 2016).

A Hawkes process is characterized by a *linear* intensity function, *i.e.*, $\lambda(t) = w \cdot x_t$ where x_t denotes time-dependent features and w is the weight. The intensity function parametrizes the likelihood of observing an event in the time window $[t, t+dt)$ given that it has not occurred before t. Such linearity may be insufficient to model many real world scenarios. For example, after purchasing a new album, users may be initially highly engaged, and play the album over and over again. However, the engagement will saturate at some point as they become bored of the same album. Such plateau pattern may not be captured by a simple linear relation. In another scenario, a recent hospital visit may trigger more future visits due to the progression of a disease into a more severe stage. Such cumulative influence from recent events may grow faster than a linear model can explain.

Nonlinear Hawkes process (Brémaud & Massoulié, 1996) has been introduced to provide more flexibility in explaining the real-world phenomena. It applies a fixed nonlinear link function g to the linear combination, *i.e.*, $\lambda(t) = g(w \cdot x_t)$. For computational considerations, $g(\cdot)$ is often assumed to be in some simple parametric forms, such as $\exp(u)$ and $\max\{0, u\}$ (Carstensen et al., 2010; Hansen et al., 2015). Although these models are more flexible, they are still restricted to a few nonlinear patterns with a fixed parametrization, which may not be correct for real world data. Ideally, both $g(\cdot)$ and w should be learned from data. Unfortunately, such desideratum leads to a *non-convex* optimization problem, where efficient algorithms with provable guarantees do not exist.

To address these challenges, we propose a novel model, referred to as the *Isotonic-Hawkes* process, where both $g(\cdot)$ and w can be directly learned from data. Rather than committing to a fixed parametric form, we instead use a non-parametric, monotonic nonlinear link function. Therefore, it is extremely flexible to capture different temporal dynamics without the need to select a fixed form in advance.

To solve the non-convex learning problem with guarantees, we propose a different loss function than the typical log-likelihood for point processes. Moreover, by exploiting the problem structure, we are still able to provide theoretical guarantees on the computational and statistical performance. Our work makes the following contributions:

- We propose a novel method for nonlinear Hawkes process that can learn *both* the link function and other

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

parameters directly from data.

- Although the learning involves a *non-convex* problem, our algorithm can provably recover the true link function and the model parameters. This also requires a novel analysis for non *i.i.d.* observations.
- Our method achieves superior empirical performance, significantly outperforming alternatives on both synthetic and real-world datasets.

Related work. Prior work on nonlinear Hawkes process focuses on theoretical properties (Brémaud & Massoulié, 1996; Zhu, 2015; Hansen et al., 2015). The link function is usually given, and the discretization of time is needed in order to evaluate the integral of the intensity function. Hence, efficient algorithms are available only for specific link functions (Paninski, 2004; Truccolo et al., 2005; Carstensen et al., 2010). In contrast, our method is the first algorithm that can learn both the link function and the model parameters non-parametrically.

Our work is also closely related to Isotonic regression and Single Index Model (SIM). The Isotonic regression (Barlow et al., 1972; Robertson et al., 1988; Mair et al., 2009) is a well studied technique to fit an arbitrary monotonic 1-D function. SIM generalizes the linear regression and estimates both the nonlinear link function and the feature weights. However, earlier work are usually heuristics, which are not guaranteed to converge to a global optimum (Horowitz & Härdle, 1996; Hristache et al., 2001; Naik & Tsai, 2004; Delecroix et al., 2006). Only recently algorithms have been proposed with global convergence guarantees (Kalai & Sastry, 2009; Kakade et al., 2011; Acharyya & Ghosh, 2015) .

Unlike SIM, which only focuses on regression, our work is concerned with learning a temporal point process where the response variable is not directly observed. At the same time, the observations are non *i.i.d.* , a setting significantly different from previous works. The added complexity of temporal point processes requires us to develop a new efficient algorithm and its analysis.

2. Preliminaries

In this section, we will first provide some background on isotonic regression and point processes.

Isotonic Regression and Single Index Model. Given 1-D data points $\{(z_i, y_i)\}_{i=1}^n$, Isotonic regression solves a least-square problem with monotonicity constraints:

$$\min_{\hat{\boldsymbol{y}} \in \mathbb{R}^n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \text{ s.t. } \hat{y}_i \leq \hat{y}_j \text{ if } z_i \leq z_j. \quad (1)$$

The problem can be solved efficiently by a Pool Adjacent Violators (PAV) algorithm (Mair et al., 2009) in $O(n \log n)$, and the input and the solution $\{(z_i, \hat{y}_i)\}_{i=1}^n$ implicitly define a monotonic function with $g(z_i) = \hat{y}_i$.

The Single Index Model is a generalized linear model with the following assumption $\mathbb{E}[y|x] = g(w \cdot x)$, where g is the link function. The Isotron algorithm can provably recover w and g (Kalai & Sastry, 2009; Kakade et al., 2011) under the mild assumption that g is monotonic and Lipschitz continuous.

Temporal Point Processes A temporal point process is a random process of which the realization consists of a list of discrete temporal events $\{t_i\}_{i=1}^n$. It is equivalent to a counting process, $\{N(t), t \geq 0\}$, which records the cumulative number of events happening right before time t, and satisfies $N(t') \leq N(t)$ for $t' \leq t$ and $N(0) = 0$. A counting process is also a submartingale, *i.e.*, $\mathbb{E}[N(t)|\mathcal{H}_{t'}] \geq N(t')$ for all $t > t'$, where $\mathcal{H}_{t'} = \{t_i | t_i < t'\}$ denotes the history up to but not including time t'.

A useful characterization of temporal point processes is the intensity function. Specifically, according to Doob-Meyer theorem (Aalen et al., 2008), $N(t)$ has the unique decomposition: $N(t) = \Lambda(t) + M(t)$, where $\Lambda(t)$ is an increasing predictable process called the compensator (or cumulative intensity) and $M(t)$ is a zero-mean martingale. Alternatively, we have $\mathbb{E}[dM(t)|\mathcal{H}_t] = 0$, and

$$\mathbb{E}[dN(t)|\mathcal{H}_t] = d\Lambda(t) := \lambda(t)dt \quad (2)$$

where $\lambda(t)$ is the intensity function. Intuitively, the larger $\lambda(t)$, the greater the chance an event happens in time interval $[t, t + dt)$.

The functional form of the intensity function characterizes the temporal point process. A particular useful form is the intensity of a *Hawkes process*, which captures the mutual excitation phenomena between events:

$$\lambda(t) = \lambda_0 + \alpha \sum\nolimits_{t_i \in \mathcal{H}_t} \kappa(t - t_i) \quad (3)$$

where λ_0 captures the long-term incentive to generate events. $\kappa(t) \geqslant 0$ models temporal dependencies, and $\alpha \geqslant 0$ quantifies how the influence from each past event evolves over time, making the intensity function depend on the history $\mathcal{H}_t$.

In the Hawkes process, past events affect the occurrence of future events, which makes it particularly useful for modeling clustered event patterns. However, the linear link function of the intensity function may be insufficient to model many real world scenarios. In the next section, we propose Isotonic-Hawkes processes with a flexible link function and a provable learning algorithm.

3. Isotonic Hawkes Processes

We propose a new family of nonlinear Hawkes processes: Isotonic-Hawkes processes. We present the moment matching learning framework for the non-convex problem. To facilitate learning, we optimize the representation of the objective function by showing that the intensity integral in the objective function can be exactly computed. Then we present the overall algorithm, which applies an alternating minimization scheme to update the link function g and weights parameters w.

3.1. Model Formulation

In Isotonic-Hawkes processes, we model its intensity as the composition of a monotonic, non-parametric link function and a linear Hawkes intensity.

Definition 1. *A Isotonic-Hawkes process is a counting process $N(t)$, with associated history $\mathcal{H}_t = \{t_i | t_i < t\}$, such that the intensity function $\lambda(t)$ can be written as:*

$$\lambda(t) = g\left(\lambda_0 + \alpha \sum_{t_i \in \mathcal{H}_t} \kappa(t - t_i)\right) = g(w \cdot x_t) \qquad (4)$$

where $\kappa(t) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is a continuous monotonic decreasing triggering kernel capturing the influence of the history $\mathcal{H}_t$, $\lambda_0 \geqslant 0$ is a baseline intensity independent of the history, and $g \in \mathcal{G} : \mathbb{R} \rightarrow \mathbb{R}^+$, is a monotonic increasing and G-Lipschitz link function, i.e.,

$$0 \leqslant g(b) - g(a) \leqslant G(b - a) \textit{ for all } 0 \leqslant a \leqslant b \qquad (5)$$

We set $w = (\lambda_0, \alpha)^\top$, and $x_t = \left(1, \sum_{t_i \in \mathcal{H}_t} \kappa(t - t_i)\right)^\top$.

We require $\kappa(t)$ to be monotonically decreasing, such as the exponential kernel $\exp(-t)\mathbb{I}[t > 0]$, Gaussian kernel and heavy tailed log-logistic kernel. This property is useful for computing the integral of the intensity discussed later.

The linear term in Hawkes process alone is not sufficient to capture the general trend in real-world applications. For instance, linearity leads to unbounded intensity, which is at odds with the saturation phenomenon. The nonlinear link function g enables the model to adapt to such nonlinearities in the data, hence achieving better performance. We assume g is nonparametric and monotonic increasing, which covers a wide range of functions, and also maintains the properties of the composed intensity function.

3.2. Moment Matching Objective

Maximum Likelihood Estimation (MLE) is often used to learn Hawkes processes, yielding a convex problem w.r.t. w. The estimator has good statistical rates and is consistent and asymptotically efficient (Ozaki, 1979). However, if we want to learn $g(\cdot)$ and w jointly, MLE becomes non-convex, and we no longer have statistical guarantees. To solve this problem, we use a different learning criteria based on the moment matching idea (Aalen et al., 2008). We can establish global convergence guarantees despite the non-convexity by establishing the connections to Isotonic regression and SIM.

Let $N_i = N(t_i)$. Since $N(t)$ is a counting process, the count increases by one at each time t_i. Hence for a list of events $\{t_i\}_{i=1}^n$, we have $N_i = i$ for $i \in [n]$. We have

$$\mathbb{E}[N_i | \mathcal{H}_{t_i}] = \int_0^{t_i} \lambda(t) dt = \int_0^{t_i} g(w \cdot x_t) dt. \qquad (6)$$

Therefore, we can estimate the parameters g and w by matching the integral $\int_0^{t_i} \lambda(t) dt$ with observations N_i, which leads to the following objective function:

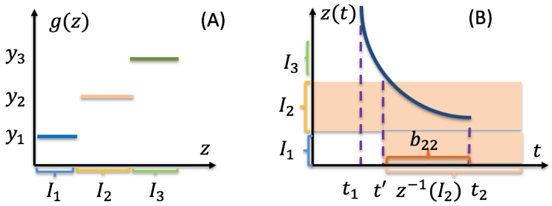


Figure 1. Illustration of integral computation. (A) the function g has 3 pieces and is constant on intervals I_1, I_2 and I_3. (B) The function $z(t) = w \cdot x_t$ is restricted on the interval $[t_1, t_2]$. It is continuous and monotonic decreasing due to the property of triggering kernel κ. The pre-image of I_2 is shown as the light yellow area on the t axis, and b_{22} is the intersection of $[t_1, t_2]$ and $z^{-1}(I_2)$. It is found by locating the pre-image of the endpoints, t' and another point outside the interval $[t_1, t_2]$ (not shown here).

$$\min_{g \in \mathcal{G}, w} \frac{1}{n} \sum_{i=1}^{n} \left(N_i - \int_0^{t_i} g(w \cdot x_t) dt\right)^2. \qquad (7)$$

Note that we need to optimize an integral w.r.t. a function g, which is challenging in representation and computation. Instead of optimizing over $\mathcal{G}$, we replace it with the family of *piecewise constant non-decreasing functions, $\mathcal{F}$, and the jumps of g is defined only at the intensity of each observed event*. As shown in Theorem 2, the integral of such functions can be computed *exactly* as weighted combinations of $g(w \cdot x_{t_i})$ defined on the observed time points t_i. For notation simplicity, we set $x_i = x_{t_i}$. The piecewise-constant function will provide a good approximation to the original function as we increase the number of training samples.

3.3. Integral Computation

We assume g is a piecewise constant function defined on each time t_i. Then we have the following result:

Theorem 2. *Assume g is piecewise constant, then the integral on $[0, t_i]$ is a weighted sum of $y_j = g(w \cdot x_j)$ with weights a_{ij}. That is $\int_0^{t_i} g(w \cdot x_t) dt = \sum_{j \in \mathcal{S}_i} a_{ij} y_j$.*

To efficiently compute the a_{ij}'s, we can first compute the integral on intervals $[t_{i-1}, t_i]$, then use cumulative sum to arrive at the final results.

Set $z(t) = w \cdot x_t$, since $g(\cdot)$ is a piecewise constant function, we have:

$$g(z(t)) = \sum_{j=1}^{n} y_j \mathbb{I}\left[z(t) \in I_j\right]$$

where $\mathbb{I}[\cdot]$ is the indicator function, and I_j denotes the j-th interval where $g(\cdot)$ is a constant. Therefore, we can write the integral on $[t_{i-1}, t_i]$ as:

$$\int_{t_{i-1}}^{t_i} g\big(z(t)\big) dt = \sum_{j=1}^{n} y_j \int_{t_{i-1}}^{t_i} \mathbb{I}\left[t \in z^{-1}(I_j)\right] dt$$

where $z^{-1}(I_j)$ denotes the pre-image of the interval I_j. Next, we need to compute $b_{ij} := \int_{t_{i-1}}^{t_i} \mathbb{I}\left[t \in z^{-1}(I_j)\right] dt$. Since it is the length of the intersection of two intervals,

Algorithm 1 COMPUTE-COEFFICIENT

```
Input: $t_i$, for $i = 1, \cdots, n$
Output: $a_{ij}$
for $j = 1, \cdots, n$ do
    Compute $t'_j$ that satisfies $x_{t'_j} = x_{t_j}$.
end for
for $j = 1, \cdots, n$ do
    Set $a_{0,j} = 0$
    for $i = 1, \cdots, n$ do
        Compute $b_{ij} = \min(t'_{j-1}, t_i) - \max(t'_j, t_{i-1})$
        Compute $a_{ij} = a_{(i-1),j} + b_{ij}$
    end for
end for
```

we can compute b_{ij} by finding all the endpoints of the pre-images $z^{-1}(I_j)$.

To do this, we first state a property of $z^{-1}(I_j)$. Restricted on $[t_{i-1}, t_i]$, $z(t)$ is a continuous and monotonic decreasing function due to the monotonic decreasing triggering kernel κ (Figure 1(B)). Combined with the fact that I_j are disjoint and share endpoints (Figure 1(A)), the pre-images $z^{-1}(I_j)$ are also *disjoint and share endpoints*.

With this property, we can compute b_{ij} easily. According to the definition of I_j, one endpoint of $z^{-1}(I_j)$ is $w \cdot x_{t_j}$, so we just need to find another endpoint as $t'_j = z^{-1}(w \cdot x_{t_j})$, which is equivalent to solving the equation $w \cdot x_{t'_j} = w \cdot x_{t_j}$.

Note x_t only has two dimensions, and the first dimension is a constant. Hence, the above equation does not depend on w, and it suffices to solve $x_{t'_j} = x_{t_j}$, where the left-hand side is a function of the unknown t'_j, and the right-hand side is a function of the observed data. It can be easily solved by root finding algorithms. We can then compute b_{ij} as:

$$b_{ij} = \min(t'_{j-1}, t_i) - \max(t'_j, t_{i-1})$$

The $\min$ and $\max$ operator implement the interval intersection. Since $z(t)$ is monotonic decreasing, we have $t'_{j-1} \geq t'_j$. Figure 1 illustrates the algorithm.

After we have computed b_{ij}, a_{ij} is readily available by $a_{ij} = \sum_{i'<=i} b_{i'j}$. The corresponding index sets $\mathcal{S}_i$ contain nonzero a_{ij}'s. The detailed procedures are presented in Algorithm 1.

3.4. Overall Algorithm

With Theorem 2, we can replace the integral of an unknown function by the weighted summation of its values defined at the intensity of each observed event. Hence we can represent the $g \in \mathcal{F}$ non-parametrically, and reformulate the objective function as:

$$\min_{g\in\mathcal{F}, w} \frac{1}{n}\sum_{i=1}^{n}\left(N_i - \sum_{j\in\mathcal{S}_i} a_{ij} g(w \cdot x_j)\right)^2. \tag{8}$$

We optimize g and w alternatively until convergence. The update rules for w and g are presented as follows.

Update $\hat{w}$. Given $\hat{g}^k$, the update rule for $\hat{w}^{k+1}$ is:

$$\hat{w}^{k+1} = \hat{w}^k + \frac{1}{n}\sum_{i=1}^{n}\left(N_i - \sum_{j\in\mathcal{S}_i} a_{ij}\hat{g}^k(\hat{w}^k \cdot x_j)\right)\sum_{j\in\mathcal{S}_i} a_{ij}x_j \tag{9}$$

Similar to the Isotron algorithm (Kalai & Sastry, 2009), this update rule is parameter free and Perceptron-like.

Update $\hat{g}$. Note that $\hat{g}$ is a non-parametric function which is represented by its values $\hat{y}_i^k$ at $\hat{w}^k \cdot x_i$. Therefore, we only need to determine its values on existing data points.

Given $\hat{w}^k$, we first sort $\left\{\hat{w}^k \cdot x_i\right\}_{i=1}^n$ such that it is an increasing sequence. That is, we re-label the data points according to the sorted order. Then we solve the following Quadratic Programming problem for $\left\{\hat{y}_i^{k+1}\right\}_{i=1}^n$:

$$\min \sum_{i=1}^{n}(N_i - \sum_{j\in\mathcal{S}_i} a_{ij}\hat{y}_j^{k+1})^2 \tag{10}$$

$$\text{s.t. } \hat{y}_i^{k+1} \leq \hat{y}_{i+1}^{k+1},\ 1 \leq i \leq n-1 \tag{11}$$

For simplicity we re-write the problem in matrix notations. Denote $\boldsymbol{N} = (N_1, \cdots, N_n)^\top$, $\hat{\boldsymbol{y}} = (\hat{y}_1, \cdots, \hat{y}_n)^\top$, $\boldsymbol{A}_{i,j} = a_{ij}$ if $j \in \mathcal{S}_i$ and $\boldsymbol{A}_{i,j} = 0$ otherwise. The monotonic constraint in (11) can be written as $\boldsymbol{B}\boldsymbol{y} \leq \boldsymbol{0}$ where $\boldsymbol{B}$ is the adjacent difference operator: $\boldsymbol{B}_{i,i} = 1$, $\boldsymbol{B}_{i,i+1} = -1$ and other entries are zero. Then we arrive at the following formulation:

$$\min_{\hat{\boldsymbol{y}}} \|\boldsymbol{N} - \boldsymbol{A}\hat{\boldsymbol{y}}\|^2,\ s.t.\ \boldsymbol{B}\hat{\boldsymbol{y}} \leq 0$$

This is a convex problem and can be computed efficiently using projected gradient descent:

$$\hat{\boldsymbol{y}}^{t+1} := \Pi\left[\hat{\boldsymbol{y}}^t + \eta\boldsymbol{A}^\top\left(\boldsymbol{N} - \boldsymbol{A}\hat{\boldsymbol{y}}^t\right)\right]$$

where $\Pi\left[\boldsymbol{u}\right]$ is an operator that projects $\boldsymbol{u}$ into the feasible set:

$$\Pi\left[\boldsymbol{u}\right] = \operatorname*{argmin}_{\boldsymbol{x}} \left\|\boldsymbol{x} - \boldsymbol{u}\right\|^2,\ s.t.\ \boldsymbol{B}\boldsymbol{x} \leq 0$$

The projection is exactly the Isotonic regression problem and can be solved by PAV (Mair et al., 2009) in $O(n\log n)$. In addition, the computation of the gradient is also efficient since $\boldsymbol{A}$ is a sparse matrix and it takes time $O(n + \text{nnz}(\boldsymbol{A}))$, where $\text{nnz}(\boldsymbol{A})$ is the number of nonzero elements. The number of iterations required to reach ϵ accuracy is $O(1/\epsilon)$, hence the overall complexity is $O((n\log n + \text{nnz}(\boldsymbol{A}))\,/\epsilon)$. This can also be accelerated to $O((n\log n + \text{nnz}(\boldsymbol{A}))\,/\sqrt{\epsilon})$ using Nesterov's acceleration scheme (Nesterov, 1983). The algorithm is illustrated in Algorithm 2 and the whole alternating minimization procedure is summarized in Algorithm 3. Such procedure will efficiently find the near-optimal $\hat{g}$ and $\hat{w}$.

4. Theoretical Guarantees

We now provide the theoretical analysis of convergence property. First we define the error as:

$$\varepsilon(\hat{g}^k, \hat{w}^k) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}^k(\hat{w}^k x_i) - g^*(w^* \cdot x_i)\right)^2 \tag{12}$$

Algorithm 2 LEARN-ISOTONIC-FUNC

```
Input: $\{N_i\}, \{a_{ij}\}, \eta$
Output: $\hat{\boldsymbol{y}}$
Initialize $\hat{\boldsymbol{y}}^0$ randomly
Construct matrices $\boldsymbol{N}$, $\boldsymbol{A}$ from input
$t = 0$
repeat
    $t = t + 1$
    $\hat{\boldsymbol{y}}^{t+1} = \Pi\left[\hat{\boldsymbol{y}}^t + \eta \boldsymbol{A}^\top \left(\boldsymbol{N} - \boldsymbol{A}\hat{\boldsymbol{y}}^t\right)\right]$
until convergence
```

Algorithm 3 ISOTONIC-HAWKES

```
Input: Sequences of events $\{t_i\}_{i=1}^n$
Output: $\hat{g}, \hat{w}$
Compute $x_i = (1, \sum_{t_j \in \mathcal{H}_{t_i}} \kappa(t_i - t_j))^\top$ for $i \in [n]$
$\{a_{ij}\}$ = COMPUTE-COEFFICIENT($\{t_i\}$)
Compute $N_i = i$ for $i \in [n]$
Initialize $w^0$, $g^0$ randomly
$k = 0$
repeat
    $k = k + 1$
    Sort the data according to $\hat{w}^k \cdot x_i$
    Update $\hat{g}^k$ = LEARN-ISOTONIC-FUNC($\{N_i\}$, $\{a_{ij}\}$)
    Update $\hat{w}^{k+1}$ using (9)
until loss$(\hat{g}, \hat{w}) \leqslant \epsilon$
```

where $g^*(\cdot)$ and w^* are the unknown true link function and model parameters, respectively. The goal is to analyze how quickly $\varepsilon(\hat{g}^k, \hat{w}^k)$ decreases with k.

Notations. Set $y_i^* = g^*(w^* \cdot x_i)$ to be the expected value of each y_i. Let $\bar{N}_i$ be the expected value of N_i. Then we have $\bar{N}_i = \sum_{j \in \mathcal{S}_i} a_{ij} y_j^*$. Clearly we do not have access to $\bar{N}_i$. However, consider a hypothetical call to the algorithm with input $\{(x_i, \bar{N}_i)\}_{i=1}^n$ and suppose it returns $\bar{g}^k$. In this case, we define $\bar{y}_i^k = \bar{g}^k(\bar{w}^k \cdot x_i)$.

We first bound the error using the squared distance $\|\hat{w}^k - w^*\|^2$ between two consecutive iterations:

Lemma 3. *Suppose that* $\|\hat{w}^k - w^*\| \leq W$, $\|x_i\| \leq 1$, $\sqrt{c} \leq \sum_{j \in \mathcal{S}_i} a_{ij} \leq \sqrt{C}$, $y_j \leq M$, *and*

$$\frac{1}{n}\sum_{i=1}^n \left|(N_i - \bar{N}_i)\right| \leq \eta_1, \quad \frac{1}{n}\sum_{i=1}^n \sum_{j \in \mathcal{S}_i} a_{ij} |\hat{y}_j^k - \bar{y}_j^k| \leq \eta_2$$

then we have:
$\|\hat{w}^k - w^*\|^2 - \|\hat{w}^{k+1} - w^*\|^2 \geq C_2 \varepsilon(\hat{g}^k, \hat{w}^k) - C_1(\eta_1 + \eta_2)$, *where* $C_1 = \max\{5CW, 4M\sqrt{c} + 2CW\}$, $C_2 = 2c - C$.

The squared distance decreases at a rate depending on $\varepsilon(\hat{g}^k, \hat{w}^k)$ and the upper bounds η_1 and η_2. The following two lemmas provide the concrete upper bounds.

Lemma 4 (Martingale Concentration Inequality). *Suppose* $dM(t) \leq K$, $V(t) \leq k$ *for all* $t > 0$ *and some* $K, k \geq 0$. *With probability at least* $1 - \delta$, *it holds that*

$$\frac{1}{n}\sum_{i=1}^n |N_i - \bar{N}_i| \leq O\left(\left(K + \sqrt{4K^2 + 8k^2}\right)\left(\log(1/\delta)\right)^{1/2}\right).$$

Note $N_i - \bar{N}_i = M_i$, which is the martingale at time t_i. A continuous martingale is a stochastic process such that $\mathbb{E}[M_t | \{M_\tau, \tau \leq s\}] = M_s$. It means the conditional expectation of an observation at time t is equal to the observation at time s, given all the observations up to time $s \leq t$. $V(t)$ is the variation process. The martingale serves as the noise term in point processes (similar to Gaussian noise in regression) and can be bounded using the Bernstein-type concentration inequality.

Lemma 5. *(Kakade et al., 2011) With probability at least* $1 - \delta$, *it holds for any* k *that*

$$\frac{1}{n}\sum_{j=1}^n |\hat{y}_j^k - \bar{y}_j^k| \leq O\left(\left(\frac{W^2 \log(Wn/\delta)}{n}\right)^{1/3}\right).$$

Lemma 5 relates $\hat{y}_j^k$ (the value we can actually compute) to $\bar{y}_j^k$ (the value we could compute if we had the conditional means of N_j). The proof of this lemma uses the covering number technique in (Kakade et al., 2011).

We now state the main theorem:

Theorem 6. *Suppose* $\mathbb{E}[N_i | \mathcal{H}_{t_i}] = \int_0^{t_i} g^*(w^* \cdot x_t) dt$, *where* g^* *is monotonic increasing, 1-Lipschitz and* $\|w^*\| \leq W$. *Then with probability at least* $1 - \delta$, *there exist some iteration* $k < O\left(\left(\frac{Wn}{\log(Wn/\delta)}\right)^{1/3}\right)$ *such that*

$$\varepsilon(\hat{g}^k, \hat{w}^k) \leq O\left(\left(\frac{W^2 \log(Wn/\delta)}{n}\right)^{1/3}\right).$$

Theorem 6 implies that some iteration has $\varepsilon(\hat{g}^k, \hat{w}^k) = O(1/\sqrt[3]{n})$. It is plausible the rate is sharp based on the information-theoretic lower bounds in (Zhang, 2002).

Proof sketch. We conduct a telescoping sum of Lemma 3 and show that there are at most $O\big(W/(\eta_1 + \eta_2)\big)$ iterations before the error $\varepsilon(\hat{g}^k, \hat{w}^k)$ is less than $O(\eta_1 + \eta_2)$. Set η_1, η_2 to be the right-hand sides in Lemma 4 and 5. Since η_2 is the dominant term compared with η_1, we replace η_1 by η_2 in the final results. This completes the proof.

5. Extensions

We provide several extensions of the Isotonic-Hawkes processes to more general cases.

General point processes. The idea and algorithm of Isotonic-Hawkes can be easily extended to other point processes. The time-dependent feature x_t in Isotonic-Hawkes is designed to capture the influence of history. However, one can also incorporate and extend other features in prior work (Perry & Wolfe, 2013; Li & Zha, 2014) or design it from users' experiences and the application domain.

Learning monotonic decreasing functions. Our model can be easily extended to learn a monotonically decreasing function. We just need to change the sign of the inequality in (11). Note that Theorem 6 still holds in this case.

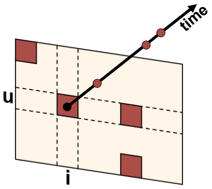


Figure 2. The sequence of events for each pair is modeled as an Isotonic-Hawkes process.

Low-rank Isotonic-Hawkes processes. We can also use our model to learn low rank parameters. For example, in the time-sensitive recommendations for online services (Du et al., 2015), we model user u's past consumption events on item i as an Isotonic-Hawkes process (Figure 2) and need to learn the parameters $\{\lambda_0^{ui}, \alpha^{ui}, g^{ui}\}$ for each user-item pair (u, i). That is, we Since group structure often exists within users' preferences and items' attributes, we assume that both matrices $\boldsymbol{\lambda}_0 = (\lambda_0^{ui})$ and $\boldsymbol{\alpha} = (\alpha^{ui})$ have low-rank structures. We can then factorize them as product of two rank r matrices: $\boldsymbol{\lambda}_0 = \boldsymbol{X}_0\boldsymbol{Y}_0$ and $\boldsymbol{\alpha}_0 = \boldsymbol{X}\boldsymbol{Y}$. Then we formulate the learning problem by applying our objective function in (7) for each observed pair (u, i):

$$\min_{\boldsymbol{X}_0, \boldsymbol{Y}_0, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{g}} \sum\nolimits_{\mathcal{H}^{ui} \in \mathcal{O}} \ell\left(\mathcal{H}^{ui}\right) \tag{13}$$

$$\ell\left(\mathcal{H}^{ui}\right) = \frac{1}{n^{ui}} \sum_{j=1}^{n^{ui}} \left(N_j^{ui} - \int_0^{t_i} g^{ui}(\boldsymbol{w}^{ui} \cdot x_t^{ui}) dt \right)^2$$

where $w^{ui} = (\lambda^{ui}, \alpha^{ui})$. n^{ui} is total number of events and $\mathcal{H}^{ui}$ is the set of history events for user-item pair (u, i). $\mathcal{O} = \{\mathcal{H}^{ui}\}$ is the collection of all observed sequences.

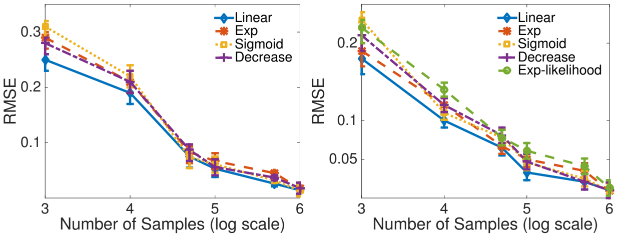
We use the alternating minimization technique to update $\boldsymbol{X}_0, \boldsymbol{Y}_0, \boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{g}$. First keep $\boldsymbol{g}^k$ fixed and update the parameters to $\boldsymbol{X}_0^{k+1}, \boldsymbol{Y}_0^{k+1}, \boldsymbol{X}^{k+1}, \boldsymbol{Y}^{k+1}$, then keep them fixed and update g^{k+1}. For the unobserved user-item pairs, after the algorithm stops, we obtain g^{ui} by taking average of the user's link functions learned from data.

Multi-dimensional Isotonic-Hawkes processes. We extend the Isotonic-Hawkes process to multi-dimension, which is particular useful to model information diffusion in social networks. It is defined by a U-dimensional point process $\boldsymbol{N}(t)$, with intensity for the u-th dimension as:

$$\lambda^u(t) = g^u \left(\lambda_0^u + \alpha^{uu_i} \sum_{i: t_i \in \mathcal{H}_t} \kappa(t - t_i) \right) = g^u(w^u \cdot x_t^u)$$

where $\alpha^{uu'}$ captures the degree of influence in the u'-th dimension to the u-th dimension. As for learning, the input data is a sequence of events observed in the form of $\{(t_i, u_i)\}$ and each pair represents an event occurring at the u_i-th dimension at time t_i. Hence, for each dimension u, set $N_i^u = N^u(t_i)$, and we solve the problem:

$$\min_{g^u, w^u} \frac{1}{n^u} \sum_{i=1}^{n^u} \left(N_i^u - \int_0^{t_i} g^u(w^u \cdot x_t^u) dt \right)^2 \tag{14}$$



(a) RMSE for function g (b) RMSE for parameter w

Figure 3. Convergence by number of samples.

where the i-th entry of w^u and x_t^u is $w^u(i) = (\lambda_0^u, \alpha^{uu_i})$ and $x_t^u(i) = (1, \sum_{t_i \in \mathcal{H}_t} \kappa(t, t_i))$ respectively. Our goal is to learn $\boldsymbol{w} = (w^u)$ and $\boldsymbol{g} = (g^u)$. From (14) we can see that learning in each dimension u is independent of others. Hence under this framework, w^u and g^u can be learned using Algorithm 3 in *parallel* efficiently.

6. Experiments

We evaluate the performance of Isotonic-Hawkes on both synthetic and real-world datasets with respect to the following tasks :

- **Convergence**: investigate how well Isotonic-Hawkes can learn the true parameters as the number of training samples increases.
- **Fitting capability**: study how well Isotonic-Hawkes can explain real-world data by comparing it with the classic Hawkes process.
- **Time-sensitive recommendation**: demonstrate that Isotonic Hawkes can improve the predictive performance in item recommendation and time prediction.
- **Diffusion network modeling**: evaluate how well Isotonic-Hawkes can model the information diffusion from cascades of temporal events.

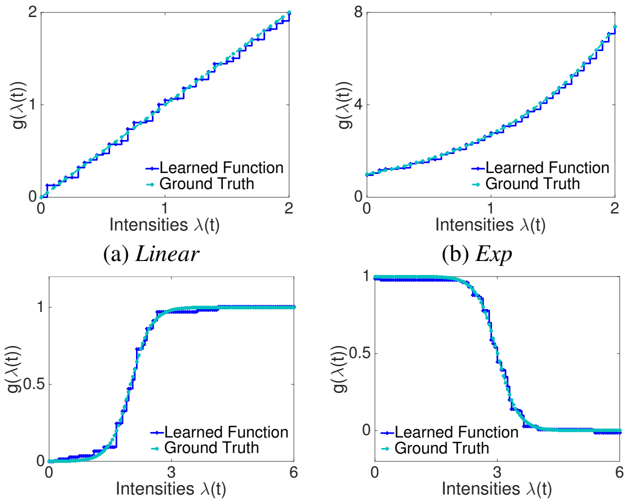
6.1. Experiments on Synthetic Data

Experimental setup. Table 1 lists the ground-truth setting with four typical link functions $g(\cdot)$ and the respective model parameters w. The first three link functions (Linear, Exp, Sigmoid) are monotonically increasing, while the last one is strictly decreasing. For the *Exp* link function, we explore the performance of learning self-inhibition by setting α to be negative. Without loss of generality, we use the unit-rate exponential decaying function as the triggering kernel. Then, based on the configuration of each row in Table 1, we simulate one million events using Ogata's Thinning algorithm (Ogata, 1981).

Table 1. Model configurations.

Name	link function g	Weights w
Linear	$g(x) = x$	$w = (1.2, 0.6)$
Exp	$g(x) = e^x$	$w = (0.5, -0.1)$
Sigmoid	$g(x) = 1/(1 + e^{-4(x-2)})$	$w = (0.5, 1.2)$
Decrease	$g(x) = 1 - 1/(1 + e^{-4(x-3)})$	$w = (0.5, 1.2)$

Convergence analysis. We first evaluate the convergence property of our learning algorithm by increasing the number of samples from 1,000 to 1,000,000. For each



(a) *Linear* (b) *Exp* (c) *Sigmoid* (d) *Decrease*

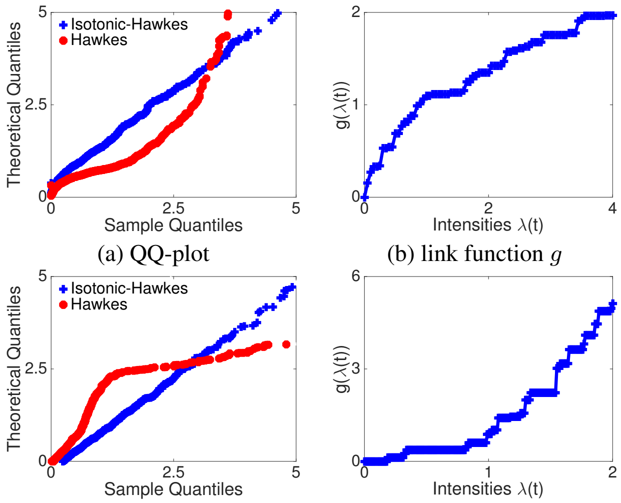
Figure 4. Comparison between learned link function and the ground truth on four synthetic datasets.

dataset, we repeat the simulations ten times and report the averaged results. Figure 3 (a) shows the Root Mean Squared Error (RMSE) between the values of the learned function and those given by the ground-truth link function as a function of training data size. Figure 3 (b) shows the RMSE of learning the model parameters. The x-axis is in log scale. Since in all cases, the RMSE decreases in a consistent way, it demonstrates that Isotonic-Hawkes is very robust regardless of the latent ground-truth link functions. Furthermore, for the *Exp* link function, we compare the RMSE between our method and the likelihood based approach, Exp-likelihood (Truccolo et al., 2005), which has access to the link function and discretizes the time interval to compute the integral in the likelihood. Our method works better at estimating w. Finally, the ability to recover the linear link function verifies that Isotonic-Hawkes naturally includes the classic Hawkes process as a special case and is much more expressive to explain the data.

Visualization of recovered link functions. We also plot each learned link function against the respective ground-truth in Figure 4 trained with 1,000,000 events. In all the cases, the algorithm can achieve the global optimal to precisely recover the true functions.

6.2. Experiments on Time-sensitive Recommendation

Experimental setup. For the task of time-sensitive recommendation, we fit a low-rank Isotonic-Hawkes process with the alternating minimization technique from (13) to solve the following two related problems proposed from (Du et al., 2015) : (1) how to recommend the most relevant item at the right moment; (2) how to accurately predict the next returning-time of users to existing services. We evaluate the predictive performance of our model on two real datasets. *last.fm*[1] consists of the music listening histories of around 1,000 users over 3,000 different albums. We use the events of the first three months for training and those of the next month for testing. There are around 20,000 user-album pairs with more than one million events. *tmall.com*[2] contains online shopping records. There are around 100K events between 26, 376 users and 2,563 stores. We use the events of the first four months for training and those of the last month for testing. The unit time is an hour.



(a) QQ-plot (b) link function g (c) QQ-plot (d) link function g

Figure 5. Experiment results on two randomly picked sequences from *last.fm* data. (a-b) and (c-d) correspond to two sequences.

Better data fitting capability. Since the true temporal dynamics governing the temporal point patterns are unknown, we first investigate whether our new model can better explain the data compared with the classic Hawkes process. According to the Time Changing Theorem (Daley & Vere-Jones, 2007), given a sequence $\mathcal{T} = \{t_i\}_{i=1}^n$ and a point process with intensity $\lambda(t)$, the set of samples $\{\int_{t_{i-1}}^{t_i} \lambda(t)dt\}_{i=1}^n$ should conform to a unit-rate exponential distribution if $\mathcal{T}$ is truly sampled from the process. As a consequence, we compare the theoretical quantiles from the unit-rate exponential distribution with the empirical quantiles of different models. The closer the slope of QQ-plot goes to one, the better a model matches the point patterns. (Du et al., 2015) has shown that Hawkes process fits the data better compared to other simple processes.

In Figure 5 (a) and (c), we show that Isotonic-Hawkes achieves much better fitting capability. Furthermore, (b) and (d) visualize the learned link functions. In Figure 5(b), the function captures the phenomenon that the user's

[1] http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-1K.html

[2] http://ijcai-15.org/index.php/repeat-buyers-prediction-competition

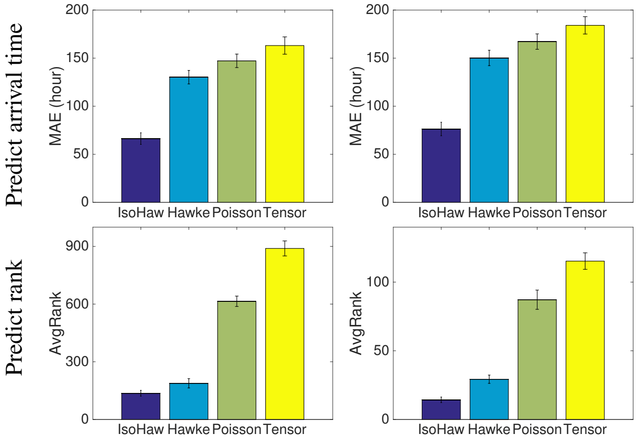
interests tend to saturate in the long-run despite that he may be excited about the item initially. Intuitively, we can also see this from (a), where Hawkes process has larger sample quantiles than the theoretical one, which means $\int_{t_{i-1}}^{t_i} \lambda(t)dt$ is larger than the value it should be. Hence using a saturating function in (b) helps adjusting the Hawkes intensity $\lambda(t)$ and make it smaller. In contrast, (d) presents the opposite trend where the user was not quite interested in the given item initially, but later became addicted to it. Since the Hawkes sample quantile is smaller than the theoretical one in (c), link function helps changing $\lambda(t)$ to be larger. Hence learning the link function is important.

Recommendation improvements. We evaluate the predictive performance on the two tasks following (Du et al., 2015) : (1) Rank prediction. At each testing moment, we record the predicted rank of the target item based on the respective intensity function. We report the average rank over all test events. Smaller value indicates better performance. (2) Arrival-time prediction. We predict the arrival time of the next testing event and report the mean absolute error (MAE) between the predicted time and the true value. In addition, besides Hawkes process, we also compare with the commonly used Poisson process, which is a relaxation of the Hawkes model by assuming that each user-item pair has constant base intensity independent of the history, as well as the state-of-the-art Tensor factorization method (Chi & Kolda, 2012) which applies Poisson factorization to fit the number of events in each discretized time slot and has better performance than methods based squared loss (Wang et al., 2015). We use the parameters averaged over all time intervals to make predictions. The latent rank of the low-rank Isotonic-Hawkes process and the tensor method are tuned to give the best performance.

We summarize the results in Figure 6. First, Hawkes outperforms the Poisson process, which means that considering the effects of history is helpful. Second, Isotonic-Hawkes outperforms Hawkes process for a significant margin thanks to the better data fitting capability shown in Figure 5. For time prediction, since the MAE's unit is hour, we can see that the error difference between Isotonic-Hawkes and Hawkes is about three days. The online shopping services can benefit a lot from this improvement and make better demand predictions.

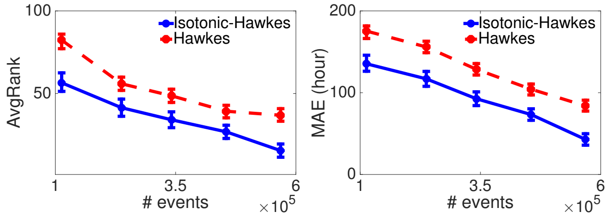
6.3. Experiments on Modeling Diffusion Networks

Finally, we apply the multi-dimension Isotonic-Hawkes process with the model estimation procedure in (14) to recover the latent information diffusion network reflected by the nonzero patterns of the mutual excitation matrix over the real *Network* dataset from (Farajtabar et al., 2014). This dataset comprises of all tweets posted by 2,241 users in eight month. The network has 4,901 edges. We split data into a training set (covering 85% of the total events) and a test set (covering the remaining 15%) according to time. Being similar to the time-sensitive recommendation task, we report the average rank of all testing events and MAE for the arrival time prediction with an increasing proportion of training events. Figure 7 verifies that Isotonic-Hawkes outperforms Hawkes process consistently.



last.fm *tmall.com*

Figure 6. Time-sensitive recommendation results.



(a) AvgRank (b) Arrival-time

Figure 7. Prediction results on the *Network* dataset.

7. Conclusion

We have proposed a novel nonlinear Hawkes process, the Isotonic-Hawkes process, with a flexible nonlinear link function. Along with the model, we have developed a computationally and statistically efficient algorithm to learn the link function and model parameters jointly, and rigorously show that under mild assumptions of the monotonicity, our algorithm is guaranteed to converge to the global optimal solution. Furthermore, our model is very general and can be extended to many different forms, including monotonically decreasing link functions, low-rank Isotonic-Hawkes processes model and multi-dimensional Isotonic-Hawkes processes. Experiments on both synthetic and real world datasets empirically verify the theoretical guarantees and demonstrate the superior predictive performance compared to other baselines.

Acknowledgements. This project was supported in part by NSF/NIH BIGDATA 1R01GM108341, ONR N00014-15-1-2340, NSF IIS-1218749, and NSF CAREER IIS-1350983.

References

Aalen, Odd, Borgan, Ornulf, and Gjessing, Hakon. *Survival and event history analysis: a process point of view*. Springer, 2008.

Acharyya, Sreangsu and Ghosh, Joydeep. Parameter estimation of generalized linear models without assuming their link function. In *AISTAT*, 2015.

Barlow, R.E., Bartholomew, D., Bremner, J. M., and Brunk, H. D. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. J. Wiley, 1972.

Brémaud, Pierre and Massoulié, Laurent. Stability of nonlinear hawkes processes. *The Annals of Probability*, pp. 1563–1588, 1996.

Carstensen, Lisbeth, Sandelin, Albin, Winther, Ole, and Hansen, Niels R. Multivariate hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics*, 11(1):456, 2010.

Chi, Eric C and Kolda, Tamara G. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, 2012.

Daley, D.J. and Vere-Jones, D. *An introduction to the theory of point processes: volume II: general theory and structure*, volume 2. Springer, 2007.

Delecroix, Michel, Hristache, Marian, and Patilea, Valentin. On semiparametric m-estimation in single-index regression. *Journal of Statistical Planning and Inference*, 2006.

Du, Nan, Wang, Yichen, He, Niao, and Song, Le. Time sensitive recommendation from recurrent user activities. In *NIPS*, 2015.

Farajtabar, Mehrdad, Du, Nan, Gomez-Rodriguez, Manuel, Valera, Isabel, Zha, Hongyuan, and Song, Le. Shaping social activity by incentivizing users. In *NIPS*, 2014.

Farajtabar, Mehrdad, Wang, Yichen, Gomez-Rodriguez, Manuel, Li, Shuang, Zha, Hongyuan, and Song, Le. Coevolve: A joint point process model for information diffusion and network co-evolution. In *NIPS*, 2015.

Hansen, Niels Richard, Reynaud-Bouret, Patricia, Rivoirard, Vincent, et al. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

Hawkes, Alan G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

Horowitz, Joel L and Härdle, Wolfgang. Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436): 1632–1640, 1996.

Hristache, Marian, Juditsky, Anatoli, and Spokoiny, Vladimir. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, 2001.

Kakade, Sham M, Kanade, Varun, Shamir, Ohad, and Kalai, Adam. Efficient learning of generalized linear and single index models with isotonic regression. In *NIPS*, 2011.

Kalai, Adam Tauman and Sastry, Ravi. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

Li, Liangda and Zha, Hongyuan. Learning parametric models for social infectivity in multi-dimensional hawkes processes. In *AAAI*, 2014.

Liptser, Robert and Shiryayev, Albert Nikolaevich. *Theory of martingales*, volume 49. Springer Science & Business Media, 2012.

Mair, Patrick, Hornik, Kurt, and de Leeuw, Jan. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32(5):1–24, 2009.

Mohler, George O, Short, Martin B, Brantingham, P Jeffrey, Schoenberg, Frederic Paik, and Tita, George E. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493), 2011.

Naik, Prasad A and Tsai, Chih-Ling. Isotonic single-index model for high-dimensional database marketing. *Computational statistics & data analysis*, 2004.

Nesterov, Yurii. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Soviet Math. Docl.*, 269:543–547, 1983.

Ogata, Yosihiko. On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1): 23–31, 1981.

Ozaki, Tohru. Maximum likelihood estimation of hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

Paninski, L. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50: 2200–2203, 2004.

Perry, Patrick O and Wolfe, Patrick J. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society*, 75(5):821–849, 2013.

Robertson, Tim, Wright, FT, Dykstra, Richard L, and Robertson, T. *Order restricted statistical inference*, volume 229. Wiley New York, 1988.

Truccolo, Wilson, Eden, Uri T, Fellows, Matthew R, Donoghue, John P, and Brown, Emery N. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.

Wang, Yichen, Chen, Robert, Ghosh, Joydeep, Denny, Joshua C, Kho, Abel, Chen, You, Malin, Bradley A, and Sun, Jimeng. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *KDD*, 2015.

Wang, Yichen, Theodorou, Evangelos, Verma, Apurv, and Song, Le. A stochastic differential equation framework for guiding information diffusion. *arXiv preprint arXiv:1603.09021*, 2016.

Zhang, Cun-Hui. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555, 2002.

Zhu, Lingjiong. Large deviations for markovian nonlinear hawkes processes. *The Annals of Applied Probability*, 2015.