

MULTIVARIATE ARRIVAL RATE ESTIMATION USING SEMIDEFINITE PROGRAMMING

Dávid Papp
Farid Alizadeh

Rutgers Center for Operations Research
640 Bartholomew Rd
Piscataway, NJ 08854, USA

ABSTRACT

An efficient method for the smooth estimation of the arrival rate of non-homogeneous, multi-dimensional Poisson processes from inexact arrivals is presented. The method provides a piecewise polynomial spline estimator. It is easily parallelized, and it exploits the sparsity of the neighborhood structure of the underlying spline space; as a result, it is very efficient and scalable. Numerical illustration is included.

1 INTRODUCTION

This paper presents an efficient method for the smooth estimation of arrival rates of non-homogeneous, multi-dimensional Poisson processes from inexact arrivals, using polynomial splines. The approach combines efficient methods of convex optimization with linear and semidefinite (more specifically, sum of squares) constraints, and decomposition methods of convex optimization. Its main advantages include the following: (1) it provides a smooth (not piecewise linear) estimator, (2) it exploits the sparsity of the neighborhood structure of the spline, leading to a very efficient method, (3) it is fully parallelizable, which further improves its scalability, and (4) it can easily be modified to incorporate various arbitrary constraints, including periodicity in one or several variables, and bound constraints.

While in this paper we concentrate only on arrival rate estimation, many of the ideas presented are also applicable for other function approximation and estimation problems, too. For further applications of similar techniques in monotone and concave regression, density estimation, and binary classification, see the thesis (Papp 2011).

The structure of this paper is as follows. The mathematical formulation of the problem is introduced in the next section, where the objective function of our optimization model is also derived, along with an observation of possibly independent interest: the expected number of arrivals of the maximum likelihood estimator (under very general conditions) agrees with the observed number of arrivals. In Section 3 we turn our attention to multivariate nonnegative splines. This functional cone is NP-hard to optimize over, hence, we consider two families of inner approximations, in which a maximum likelihood estimator can be found in polynomial time. The first approximation leads to optimization models with semidefinite constraints via “weighted-sum-of-squares” polynomials, the second one uses polynomials with nonnegative coefficients in a nonnegative basis, and results in a linearly constrained optimization model. We prove that (similarly to the cone of nonnegative splines) these spline cones are dense in the cone of nonnegative continuous functions. In Section 4 we adapt a decomposition algorithm of Ruszczyński designed for sparse convex optimization problems, which, as we demonstrate, particularly favors spline estimation problems with their sparse neighborhood structure. Combining the inner approximation ideas with this decomposition algorithm leads to an efficient, parallel algorithm for the maximum likelihood arrival rate estimation problem. An illustrative numerical example is provided in which the arrival rate of accidents on the New Jersey Turnpike is estimated.

2 MAXIMUM LIKELIHOOD ESTIMATION WITH INEXACT ARRIVALS

We consider the following problem. We observe (inexact) arrivals $\mathbf{x}_1, \dots, \mathbf{x}_N \in \Delta \subseteq [0, 1]^d$ assumed to have been generated by a non-homogeneous Poisson process with arrival rate $\lambda : \Delta \rightarrow \mathbb{R}_+$. We assume that λ is continuous, and we seek a sufficiently differentiable estimator $\hat{\lambda}$ for it, hence we use piecewise polynomial splines of sufficiently high order of differentiability to approximate the arrival rate. Furthermore, we may require λ to be periodic in any or all of its variables.

In most applications the arrivals are inexact (rounded). With sufficiently many arrivals, this means that a coordinate of multiple arrivals may appear to coincide, invalidating the Poisson assumption. Hence, the effect of rounding cannot be neglected, but a model for rounded arrival times is necessary. Such an approach is outlined next. The range of possible arrivals Δ is divided into small regions (representing the arrivals that would be rounded to the same point), and the data are the number of arrivals within each region. Suppose the arrivals are aggregated in regions $\Delta_i, i \in \mathcal{I}$, and let the number of arrivals recorded in region i be n_i . (Naturally, $\Delta = \cup_{i \in \mathcal{I}} \Delta_i$ and $N = \sum_{i \in \mathcal{I}} n_i$.) Then in the non-homogeneous Poisson model the likelihood associated with the arrival rate λ is

$$\begin{aligned} L_{\mathbf{n}}(\lambda) &= \Pr(\text{number of arrivals} = N \mid \lambda) \cdot \Pr(\text{distribution of exactly } N \text{ arrivals} = \mathbf{n} \mid \lambda) \\ &= \frac{I^N}{N!} e^{-I} \cdot \frac{N!}{\prod_{i \in \mathcal{I}} n_i!} \prod_{i \in \mathcal{I}} \left(\frac{\int_{\Delta_i} \lambda}{I} \right)^{n_i} = \frac{e^{-I}}{\prod_{i \in \mathcal{I}} n_i!} \prod_{i \in \mathcal{I}} \left(\int_{\Delta_i} \lambda \right)^{n_i}, \quad \text{where } I = \int_{\Delta} \lambda, \end{aligned}$$

whenever all the integrals in the above formulae exist (which they do, of course, if λ is a polynomial spline). Hence, maximizing the likelihood function is equivalent to maximizing

$$f(\lambda) \stackrel{\text{def}}{=} \ln \left(\left(\prod_{i \in \mathcal{I}} n_i! \right) L_{\mathbf{n}}(\lambda) \right) = - \int_{\Delta} \lambda + \sum_{i \in \mathcal{I}} n_i \ln \int_{\Delta_i} \lambda. \tag{1}$$

This is the objective function of our optimization model. Extending the above arguments to obtain maximum *penalized* likelihood models is straightforward.

Some optimization methods may benefit from a further simplification that is made possible by our next observation.

Lemma 1 Let \mathcal{K} be a cone of nonnegative functions over Δ whose restrictions to each Δ_i are integrable. Define $f : \mathcal{K} \rightarrow \mathbb{R}$ as in (1), and assume that there exists a $\lambda_0 \in \mathcal{K}$ satisfying $\int_{\Delta} \lambda_0 > 0$. Then every function $\hat{\lambda} \in \arg \max_{\lambda \in \mathcal{K}} f(\lambda)$ satisfies $\int_{\Delta} \hat{\lambda} = N$. Thus, the expected number of arrivals corresponding to the maximum likelihood estimator from \mathcal{K} equals the observed number of arrivals.

Proof. Suppose $\hat{\lambda}$ is an optimal solution (implying $\int_{\Delta} \hat{\lambda} > 0$), and consider feasible solutions $c\hat{\lambda}$ with $c > 0$. We have $f(c\hat{\lambda}) = -c \int_{\Delta} \hat{\lambda} + N \ln c + \sum_{i \in \mathcal{I}} n_i \ln \int_{\Delta_i} \lambda$, and by assumption $\frac{d}{dc} f(c\hat{\lambda})|_{c=1} = 0$. The last equation gives $\int_{\Delta} \hat{\lambda} = N$, in which case $c = 1$ indeed maximizes $f(c\hat{\lambda})$. \square

Lemma 1 allows us to remove the first term of the objective function f in (1), if we add the equation $\int_{\Delta} \lambda = N$ to our constraints. This, along with the constraint $\lambda \geq 0$ over Δ , renders the set of candidate arrival rate functions bounded. In this paper we concentrate on splines, that is, the optimal λ is chosen from a subset of piecewise polynomial functions nonnegative over Δ .

3 OPTIMIZATION OVER MULTIVARIATE NONNEGATIVE SPLINES

The maximum log-likelihood function f in (1) is a concave function with easily computable derivatives, whose optimization over reasonably “well-behaved” closed convex sets is straightforward via a number of different convex optimization methods. The same holds for maximum penalized likelihood models that can

be derived analogously. However, the set of multivariate polynomials nonnegative over a given domain, even though it is convex, is not an easy set to optimize over – in fact, as it is rather well known, even the *recognition* of nonnegative polynomials is difficult.

Proposition 2 (Boros and Hammer 2002) Deciding whether a k -variate polynomial is nonnegative over $[0, 1]^d$ (equivalently, minimizing a d -variate polynomial over the unit cube) is NP-hard, even for multilinear polynomials of degree two.

Similar statements can be made for polynomials nonnegative over other polyhedral sets (including simplexes), as well as for everywhere nonnegative polynomials. This shows that optimization over piecewise polynomial splines is difficult even if the shape of pieces is simple, and the the degree of the spline pieces is low. The overcome this difficulty, we shall consider inner approximations of the set of nonnegative polynomials: *weighted-sum-of-squares polynomials*.

3.1 Weighted-sum-of-squares Polynomials

We say that a polynomial is a *sum-of-squares* (or SOS) polynomial, if it is expressible as a sum of perfect squares. Obviously, d -variate SOS polynomials are nonnegative over the entire \mathbb{R}^d . If the domain is a semi-algebraic set $\Delta = \{\mathbf{x} \mid w_i(\mathbf{x}) \geq 0, i = 1, \dots, m\}$, where w_1, \dots, w_m are polynomials, then a sufficient (but not necessary!) condition for a polynomial p to be nonnegative over Δ is that p is expressible as

$$p(\mathbf{x}) = \sum_{I \subseteq \{1, \dots, m\}} \left(\prod_{i \in I} w_i(\mathbf{x}) \right) s_I(\mathbf{x}), \tag{2}$$

or simply as

$$p(\mathbf{x}) = \sum_{i=1}^m w_i(\mathbf{x}) s_i(\mathbf{x}), \tag{3}$$

where the polynomials s_i and s_I are SOS polynomials. Generally, polynomials expressible in the form $\sum_{i \in I} w_i s_i$, (I finite), where w_i are fixed polynomials (“weights”) and s_i are SOS polynomials, are called *weighted-sum-of-squares polynomials*, or WSOS polynomials for short.

An implication of a theorem of Nesterov (2000) is that WSOS polynomials admit a good characterization, also suitable for optimization, as long as the set of weights $\{w_i, i \in I\}$ and the spaces of the underlying SOS polynomials are both finite.

Theorem 3 (Nesterov 2000) Let $W = \{w_i \mid i \in I\}$ be a finite set of polynomials nonnegative over $\Delta \subset \mathbb{R}^d$, and consider finite-dimensional linear spaces of polynomials $V_i, i \in I$. Then the set of WSOS polynomials

$$\Sigma = \left\{ \sum_{i \in I} w_i \sum_j p_{i,j}^2 \mid p_{i,j} \in V_i \right\}$$

is a closed convex cone that is representable as the Minkowski sum of $|I|$ linear images of the cone of positive semidefinite matrices of order $\max_i(\dim(V_i))$.

Therefore, the maximization of the log-likelihood function f in (1) over sets defined by linear equations and constraints of the form $A_k(\lambda) \in \Sigma_k$, where each A_k is a linear operator and each Σ_k is a WSOS cone satisfying the conditions of Theorem 3 is a *semidefinite programming problem* (Wolkowicz, Saigal, and Vandenberghe 2000) with a convex objective function – a tractable optimization problem (Renegar 2001).

The approximability of nonnegative polynomials over semi-algebraic sets by WSOS polynomials is a well-studied problem. Unfortunately, to achieve good approximation of p , the SOS polynomials s_I and s_i in (2) and (3) need to have considerably higher degree than the degree of p itself, which is not practical for the large-scale optimization models of our interest. Since we are only interested in the approximation power of the piecewise polynomial spline estimator, we shall choose a different approach, and investigate

the approximation power of *piecewise WSOS polynomial splines*. In what follows, to keep the discussion simple, we assume $\Delta = [0, 1]^d$, and confine our discussion to *tensor product splines*, defined as the tensor product of (linear) spaces of univariate splines with free knots. (For an introduction to these basic concepts in splines, see (Schumaker 1981).)

As with every nonparametric method, we must address the issue of overfitting that may result from using splines with too many pieces. Due to space limitations we shall avoid a lengthy discussion on knot point selection, but we remark that all standard methods can be directly applied in connection with the approach proposed in this paper. In particular, the subdivision of the domain can be iteratively refined, and cross-validation, such as k -folding, or information criteria, such as AIC, AIC_c , and BIC, can be used to find the optimal subdivision. Additionally, a maximum penalized likelihood approach may be developed completely analogously, by adding a penalty term for non-smoothness to the objective function (1). If f is a piecewise polynomial function, standard penalties such as $\int_{\Delta} (f')^2$ and $\int_{\Delta} (f'')^2$ are convex quadratic functions of the coefficients, and the computational effort required to optimize the resulting penalized likelihood function is the same as required by the maximum likelihood approach without smoothness penalty. The penalized likelihood approach introduces an additional parameter (the coefficient of the penalty term), which can also be set using cross-validation. For the numerical illustration in Section 5 we used the likelihood function (1), without any penalty terms.

3.2 Piecewise Weighted-sum-of-squares Splines

3.2.1 Terminology and Notation

Let us assume that Δ is the d -dimensional interval $[0, 1]^d$, and consider splines over a rectilinear subdivision of Δ . Such a subdivision can be given by a list of vectors (of possibly different dimensions) $\mathbf{A} = (a_{i,j}), i = 1, \dots, d, j = 1, \dots, \ell_i$ such that the knot point sequence $(a_{i,1}, \dots, a_{i,\ell_i})$ defines the subdivision of Δ along the i th axis. Then each region of the subdivision is affinely similar to Δ , and we can represent a spline by the coefficients of its polynomial pieces scaled to Δ .

Formally, for each dimension $i = 1, \dots, d$ we fix a basis $\{u_0^{(i)}, \dots, u_{m_i}^{(i)}\}$ of univariate polynomials of degree m_i , with domain $[0, 1]$. The spline s is then given piecewise; for each multi-index $\mathbf{j} = (j_1, \dots, j_d)$, s over the \mathbf{j} -th piece is given by

$$s(\mathbf{x}) = q^{(\mathbf{j})}(\mathbf{x}) \quad \forall \mathbf{x} = [a_{1,j_1}, a_{1,j_1+1}] \times \dots \times [a_{d,j_d}, a_{d,j_d+1}].$$

Each polynomial piece $q^{(\mathbf{j})}$ is represented by the coefficients $p_{\mathbf{k}}^{(\mathbf{j})}$, $\mathbf{k} \in \{0, \dots, m_1\} \times \dots \times \{0, \dots, m_d\}$, of its affinely scaled counterpart $p^{(\mathbf{j})}: [0, 1]^d \rightarrow \mathbb{R}$ satisfying

$$q^{(\mathbf{j})}(\mathbf{x}) = \sum_{\mathbf{k}} p_{\mathbf{k}}^{(\mathbf{j})} \prod_{i=1}^d u_{k_i}^{(i)} \left(\frac{x_i - a_{i,j_i}}{a_{i,j_i+1} - a_{i,j_i}} \right). \tag{4}$$

Note that each $p^{(\mathbf{j})}$ has the same domain, $[0, 1]^d$. It is clear that $s(\mathbf{x}) \geq 0$ for every $\mathbf{x} \in \Delta$ if and only if $p^{(\mathbf{j})}(\mathbf{x}) \geq 0$ for every \mathbf{j} and $\mathbf{x} \in [0, 1]^d$. We refer to this representation of s by the coefficients $p_{\mathbf{k}}^{(\mathbf{j})}$ as the *scaled representation* of s .

3.2.2 The Approximation Power of Piecewise Weighted-sum-of-squares Splines

The nonnegativity of a spline s over Δ reduces to the nonnegativity of each polynomial $p^{(\mathbf{j})}$ over $\Delta = [0, 1]^d$, and our goal now is to identify proper subsets of polynomials nonnegative over Δ that give rise to piecewise polynomial splines with good approximation power. First, we need to introduce some more notation.

Let the subdivision of Δ be defined by a list of vectors \mathbf{A} , as above, and the *mesh size* of such a subdivision be defined as $\|\mathbf{A}\| = \max_{i,j}(a_{i,j+1} - a_{i,j})$. We say that a sequence of subdivision is *nested* if

each subdivision in the sequence refines the previous one. A sequence of subdivisions is *asymptotically nested* if each of its elements is included in an infinite nested subsequence.

Let us denote by Σ a fixed cone of WSOS polynomials with weights nonnegative over Δ . Finally, let $\mathcal{P}(\Sigma, \mathbf{A})$ denote the set of piecewise WSOS polynomial splines over the subdivision \mathbf{A} whose pieces (in their scaled representation) all belong to Σ . We have the following theorem.

Theorem 4 Assume that $1 \in \text{int}\Sigma$, where 1 denotes the constant one polynomial. Furthermore, let $\mathbf{A}_1, \mathbf{A}_2, \dots$ be an asymptotically nested sequence of subdivisions of $\Delta = [0, 1]^d$ with mesh sizes approaching zero. Then the set $\bigcup_i \mathcal{P}(\Sigma, \mathbf{A}_i)$ is a dense subcone of the cone of nonnegative continuous functions over Δ .

We shall not prove this theorem; instead, we prove a stronger assertion below.

A special case of the above approach is the following. A sufficient (but obviously not necessary) condition for a polynomial p to be nonnegative over Δ is that p has nonnegative coefficients in a basis $U = \{u_0, \dots, u_m\}$ that consists of polynomials nonnegative over Δ , that is, $p \in \text{cone}(U)$ for a nonnegative polynomial basis U . Similarly to the piecewise WSOS polynomial splines above, we can define a *piecewise U -spline* as a piecewise polynomial spline whose pieces, in their scaled representation, belong to $\text{cone}(U)$. The set of piecewise U -splines with subdivision \mathbf{A} is denoted by $\mathcal{P}(U, \mathbf{A})$.

Theorem 5 Consider a basis $U = \{u_0, \dots, u_m\}$ of d -variate polynomials of multi-degree $\mathbf{m} = (m_1, \dots, m_d)$ such that each u_i is nonnegative over $\Delta = [0, 1]^d$, and assume that $1 \in \text{int}\text{cone}(U)$, where 1 denotes the constant one polynomial. Furthermore, let $\mathbf{A}_1, \mathbf{A}_2, \dots$ be an asymptotically nested sequence of subdivisions with mesh sizes approaching zero. Then the set $\bigcup_i \mathcal{P}(U, \mathbf{A}_i)$ is a dense subcone of the cone of nonnegative functions over Δ .

Proof. First we show that for every polynomial p of degree \mathbf{m} , strictly positive over $[0, 1]$, there exist nonnegative constants C_i such that $p + C_i \in \mathcal{P}(U, \mathbf{A}_i)$ for every i , and $\lim C_i = 0$.

Fix i , and consider a piece in the subdivision from the knot point sequence \mathbf{A}_i :

$$[a_{1,j_1}, a_{1,j_1+1}] \times \dots \times [a_{d,j_d}, a_{d,j_d+1}].$$

The polynomial p can be represented as a piecewise polynomial spline of degree \mathbf{m} with knot point sequence \mathbf{A}_i ; its scaled representation is

$$p^{(\mathbf{j})}(x_1, \dots, x_d) = p((a_{1,j_1+1} - a_{1,j_1})x_1 + a_{1,j_1}, \dots, (a_{d,j_d+1} - a_{d,j_d})x_d + a_{d,j_d}).$$

Collecting terms in the standard basis, we see that every coefficient in the above expression is of order $\mathcal{O}(\|\mathbf{A}_i\|)$, except for the constant term, which is $p(a_{1,j_1}, \dots, a_{d,j_d})$. By assumption, this constant term is positive, because p is strictly positive on $[0, 1]$. By the assumption on U , $\sum_{k=0}^m \alpha_k u_k \equiv p(a_{1,j_1}, \dots, a_{d,j_d})$ for some positive $\alpha_0, \dots, \alpha_m$. Now, if we express $p^{(\mathbf{j})}$ in the basis U : $p^{(\mathbf{j})} = \sum p_k^{(\mathbf{j})} u_k$, we have that $p_k^{(\mathbf{j})} = \alpha_k - \delta_k^{(\mathbf{j})}$ with $|\delta_k^{(\mathbf{j})}| = \mathcal{O}(\|\mathbf{A}_i\|)$, consequently $p^{(\mathbf{j})} + p(a_{1,j_1}, \dots, a_{d,j_d}) \max_k (|\delta_k^{(\mathbf{j})}| / \alpha_k)$ has positive coefficients in the basis U . Applying the same argument for every \mathbf{j} , we obtain that $p + C_i \in \mathcal{P}(U, \mathbf{A}_i)$ for

$$C_i = \max_{\mathbf{j}} (p(a_{1,j_1}, \dots, a_{d,j_d}) \max_k (|\delta_k^{(\mathbf{j})}| / \alpha_k)).$$

Finally, as $|\delta_k^{(\mathbf{j})}| = \mathcal{O}(\|\mathbf{A}_i\|)$ and p is bounded, $C_i \rightarrow 0$ as $\|\mathbf{A}_i\| \rightarrow 0$.

The same argument also proves that for every strictly positive spline over $[0, 1]$, with knot point sequence \mathbf{A} , and for every sequence $\{\mathbf{A}_i\}$ consisting of subdivisions of \mathbf{A} satisfying $\lim \|\mathbf{A}_i\| = 0$, there exist nonnegative constants C_i such that $s + C_i \in \mathcal{P}(U, \mathbf{A}_i)$ for every i , and $\lim C_i = 0$.

Consequently, $\bigcup_i \mathcal{P}(U, \mathbf{A}_i)$ is a dense subset of nonnegative splines of multi-degree \mathbf{m} .

Now our assertion follows from the fact that tensor product splines (of every given order of differentiability) in $\Delta = [0, 1]^d$ are dense in the space of continuous functions over Δ ; see (Schumaker 1981, Theorem 13.21). □

Note that the conditions $1 \in \text{int}\Sigma$ and $1 \in \text{intcone}(U)$ are sufficient *and necessary* for the desired conclusion. For example, the cone of polynomials with nonnegative coefficients in the standard monomial basis is a WSOS cone with weights nonnegative over $[0, 1]^d$. It does not satisfy the condition, as the constant 1 is on the boundary of this polynomial cone. Incidentally, the corresponding spline cones $\mathcal{P}(U, \mathbf{A})$ consist of functions that are monotone nondecreasing and convex in every variable, hence the union of such cones cannot be not dense in the cone of nonnegative continuous functions over $[0, 1]^d$.

We shall clarify in what sense the second, polyhedral approximation, approach is a special case of the WSOS approach. Using the notation of Theorem 3 the cone $\text{cone}(U)$ for a nonnegative polynomial basis U can be considered a WSOS cone Σ with weights in U , whose spaces V_i are the one-dimensional linear spaces consisting only of constant polynomials. Furthermore, for every WSOS cone satisfying the conditions of Theorem 4 one can find a basis U such that the corresponding polyhedral spline cone satisfies the conditions of Theorem 5. Hence, Theorem 5 implies Theorem 4.

As a final remark, we must mention that, unlike the use of piecewise WSOS splines, the above polyhedral approach to the arrival rate estimation problem (and other shape-constrained estimation problems) is not completely new in the statistical and function estimation literature, even though the most common approach is somewhat different: typically a nonnegative basis of spline functions is chosen, such as *B-splines*, and the optimization is carried out over the nonnegative linear combination of this basis. However, as shown in (Papp 2011), cones generated by (univariate) B-splines of degree $m \geq 2$ and given subdivision of $\Delta = [0, 1]$ form a proper subcone of piecewise U -splines with nonnegative coefficients in their scaled representation (4) corresponding to the *Bernstein polynomial basis* given by $U = \{u_0, \dots, u_m\}$, $u_i(x) = \binom{m}{i} x^i (1-x)^{m-i}$. This inclusion extends to the corresponding tensor product splines. Hence, we do not consider B-splines in this paper.

4 A DECOMPOSITION METHOD FOR MULTIVARIATE SPLINE ESTIMATION

The size of the optimization models involving multivariate splines prohibits the solution of models of high dimension or small mesh size. On the other hand, these problems have a very regular and sparse structure that makes them potentially amenable to decomposition methods. In this section we outline an augmented Lagrangian decomposition method with particularly good convergence properties for spline estimation problems. We illustrate it, in the next section, by estimating the (two-dimensional) weekly arrival rate of car accidents on the New Jersey Turnpike.

There is a vast literature on decomposition methods for both linear and nonlinear convex optimization problems, an area initiated by Dantzig and Wolfe (Dantzig and Wolfe 1960) and Benders (Benders 1962), which we can hardly summarize in the space available.

The method we propose is a simplified version of the augmented Lagrangian-based method from (Ruszczynski 1995) specifically designed for sparse problems. Since some of the details of our specific estimation problem might obscure the main ideas of the algorithm, we shall discuss the method in a slightly more abstract form than necessary for our purposes.

4.1 Augmented Lagrangian Decomposition for Sparse Problems

Let $L \geq 2$, and let \mathcal{X}_i ($i = 1, \dots, L$) be a nonempty compact subset of \mathbb{R}^{n_i} . Finally, let $f_i: \mathcal{X}_i \rightarrow \mathbb{R}$ be convex. With these given, we consider the convex optimization problem

$$\text{minimize } f(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{i=1}^L f_i(\mathbf{x}_i) \tag{5a}$$

$$\text{subject to } \sum_{i=1}^L \mathbf{A}_i \mathbf{x}_i = \mathbf{b} \tag{5b}$$

$$\mathbf{x}_i \in \mathcal{X}_i \quad i = 1, \dots, L. \tag{5c}$$

As a specific example, we can model arrival rate estimation problems in this framework: the optimality of the estimator is defined piecewise, and so are the nonnegativity constraints, which are replaced by constraints that the scaled representation of each pieces belongs to a WSOS cone. (In the rest of the paper we will refer to these constraints as “the WSOS constraints” for short. The WSOS constraints are translated to semidefinite constraints via Theorem 3 or linear constraints if the polyhedral approach and Theorem 5 are used.)

Thus, we can set L to be the number of pieces, (5c) are the WSOS constraints, and (5b) includes the continuity of the estimator and its derivatives, as well as periodicity constraints, as needed for the problem. See (Papp 2011) for models of various shape-constrained estimation problems in the same framework.

Alternatively, if the estimator is a polynomial spline over a rectilinear grid we can set $L = 2$, since all constraints connect only pieces of two disjoint class, following a chessboard-like pattern. The same is true for some other regular subdivisions as well, including a regular simplicial subdivision. This property can be exploited by some methods. We shall focus on a direct consequence of this observation: that in our arrival rate estimation model *all of the coupling constraints in (5b) involve variables corresponding to only two different \mathcal{X}_i* . (Note that this is specific to the arrival rate estimation problem. The apparently similar problem of estimation of probability density functions can also be posed as an optimization problem over nonnegative functions, but these also include the constraint that the estimator integrates to one, which is a single linear equation that involves every variable.)

The condition that the sets \mathcal{X}_i be bounded is a rather technical condition, as one can always find reasonable bounds on the spline coefficients based on the data. In our arrival rate estimation model we do not need such assumptions, since the optimal estimator is a piecewise nonnegative polynomial function whose integral is given by Lemma 1 – the set of such polynomials is obviously bounded.

The method proposed in (Ruszczynski 1995) associates multipliers π to the linear coupling constraints, and considers a separable approximation Λ_{apx} of the augmented Lagrangian of (5),

$$\Lambda(\mathbf{x}, \pi) = f(\mathbf{x}) + \langle \pi, \mathbf{b} - \mathbf{A}\mathbf{x} \rangle + \frac{\rho}{2} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2,$$

in which the bilinear terms in the quadratic penalties are linearized around a point $\tilde{\mathbf{x}} \in \mathbb{R}^{\sum_{i=1}^L n_i}$:

$$\Lambda_{\text{apx}}(\mathbf{x}, \tilde{\mathbf{x}}, \pi) \stackrel{\text{def}}{=} \sum_{i=1}^L \Lambda_i(\mathbf{x}_i, \tilde{\mathbf{x}}_i, \pi) \stackrel{\text{def}}{=} \sum_{i=1}^L f_i(\mathbf{x}_i) - \langle \mathbf{A}_i^T \pi, \mathbf{x}_i \rangle + \frac{\rho}{2} \left\| \mathbf{b} - \mathbf{A}_i \mathbf{x}_i - \sum_{j \neq i} \mathbf{A}_j \tilde{\mathbf{x}}_j \right\|^2. \tag{6}$$

The approach then is to fix a set of multipliers, and find an approximate minimizer of the augmented Lagrangian $\Lambda(\cdot, \pi)$ by iteratively minimizing $\Lambda_{\text{apx}}(\cdot, \tilde{\mathbf{x}}, \pi)$ and updating $\tilde{\mathbf{x}}$ so as to approximate the optimal solution better. The components Λ_i of the approximate Lagrangian can be optimized separately, and even in parallel (in a Jacobi, rather than a Gauss–Seidel, fashion), which is a very desirable property when infrastructure for massively parallel computations is available. Once an approximately optimal solution to the augmented Lagrangian is found, the multipliers are updated as in the classic multiplier method

Algorithm 1: Simplified augmented Lagrangian decomposition

```

parameters:  $\rho > 0, \tau > 0$ 
1 initialize  $\pi$                                      /* arbitrary initial value */
2 repeat
3    $\pi \leftarrow \pi + \rho(\mathbf{b} - \mathbf{A}\mathbf{x})$ 
4   foreach  $i = 1, \dots, L$  do solve  $\min_{\mathbf{x}_i \in \mathcal{X}_i} \Lambda_i(\mathbf{x}_i, \tilde{\mathbf{x}}, \pi)$            /* parallel */
5   if  $\mathbf{A}_i \mathbf{x}_i \neq \mathbf{A}_i \tilde{\mathbf{x}}_i$  for any  $i = 1, \dots, L$  then
6      $\tilde{\mathbf{x}} \leftarrow \tilde{\mathbf{x}} + \tau(\mathbf{x} - \tilde{\mathbf{x}})$ 
7     go to step 4
8   end if
9 until  $\mathbf{A}\mathbf{x} = \mathbf{b}$ 
10 return  $\mathbf{x}$ 

```

(Rockafellar 1976). The formal definition of the method is given in Algorithm 1, which requires two parameters: the augmented Lagrangian coefficient ρ , and a step size parameter τ .

It can be shown that for every $\rho > 0$ and $0 < \tau < 1/(N - 1)$ Algorithm 1 is convergent. For the inner loop this is a special case of Theorem 1 in (Ruszczynski 1995), and for the outer loop this follows from the convergence of the method of multipliers (Rockafellar 1976). Perhaps the most attractive feature of Algorithm 1 is that its speed of convergence depends very highly on the largest number N of variable blocks \mathbf{x}_i linked by a coupling equality constraint, favoring problems with a sparse neighborhood structure. Let us measure the progress of the algorithm by the difference

$$\Delta^{(k)} \stackrel{\text{def}}{=} \Lambda(\tilde{\mathbf{x}}^{(k)}, \boldsymbol{\pi}^{(k)}) - \min_{\mathbf{x} \in \mathcal{X}} \Lambda(\mathbf{x}, \boldsymbol{\pi}^{(k)}),$$

where the superscript k refers to the number of times the outer loop of Algorithm 1 has been executed. The following theorem establishes the rate of convergence under a technical assumption.

Theorem 6 Assume that there is a $\gamma > 0$ such that for every $\boldsymbol{\pi}$ and every $\mathbf{x} \in \mathcal{X}$,

$$\Lambda(\mathbf{x}, \boldsymbol{\pi}) - \min_{\mathbf{x} \in \mathcal{X}} \Lambda(\mathbf{x}, \boldsymbol{\pi}) \geq \gamma \text{dist}(\mathbf{x}, \arg \min_{\mathbf{x} \in \mathcal{X}} \Lambda(\mathbf{x}, \boldsymbol{\pi}))^2,$$

and let $\alpha = \max_i \|\mathbf{A}_i\|_2$. Then for every $\rho > 0$ and $0 < \tau < 1/(N - 1)$,

$$\Delta^{(k+1)} \leq \left(1 - \frac{\tau(1 - \tau(N - 1))}{1 + 2\rho\alpha^2(N - 1)^2\gamma^{-1}} \right) \Delta^{(k)}. \tag{7}$$

Proof. This is a special case of Theorem 2 in (Ruszczynski 1995). □

We remark again that in our arrival rate estimation model we have $N = 2$, implying that the decomposition method is particularly suited for our problem.

The first condition of the theorem is the *quadratic growth condition* of (Ruszczynski 1995), and is satisfied by the log-likelihood objective. Theorem 6 also suggests a way to set the parameter τ : minimizing the coefficient on the right-hand side of (7) we obtain that $\tau = (2(N - 1))^{-1}$ is the recommended choice, regardless of ρ . (The theory does not provide guidance in the selection of ρ .)

5 NUMERICAL ILLUSTRATION – NJ TURNPIKE ACCIDENTS

We consider the two-dimensional point process of car accidents on the New Jersey Turnpike (NJTP). The two dimensions are time (with an assumed weekly periodicity) and location along the road. It is not entirely clear whether this is indeed a Poisson process, as accidents may change the traffic pattern, which in turn

affects the distribution of the accidents. Furthermore, coincidences in the location (a number of drivers hitting the same tree, or pedestrians or animals at the same crossing) have likelihood zero in every Poisson model. However, as the accidents are relatively rare (serious accidents that change the traffic pattern for a long period of time are even more so), and major highways are assumed to have no easy-to-hit objects, a Poisson model may be a reasonable approximation. The fact that accidents may occur more frequently close to exits does not contradict the non-homogeneous Poisson model.

The data. We obtained car accident data from the New Jersey Department of Transportation. The raw data contained information on every car accident in 2009 recorded at the accident locations by police officers. The time of the accident is rounded to the nearest minute, but it is not clear whether the recorded time is the approximate time of the accident, the time the police were notified of the accident, or the time the officers attended to the accident. Hence, we can consider this as noisy data, despite the apparent precision of the time data. The location is given by the Standard Route Identifier of the road segment and an approximate milepost reading (variably rounded, apparently to the nearest 0.05 mile or to the nearest mile).

We removed all entries from the data that corresponded to accidents in roads other than the NJTP segment marked I-95. This is an approximately 78-mile-long segment stretching between two state borders (with Pennsylvania and New York, respectively) with no forks or joins. We also removed entries with missing milepost information. (Date and time were present for every entry.) While we could take these accidents into account directly in a maximum-likelihood approach (and their time and date information we shall not discard), such incomplete entries were few, and it is reasonable to assume that accidents whose milepost is not recorded follow the same milepost and time distribution as the entries with complete information, and hence, the removal of this data should not introduce biases. We can then simplify our model, and estimate the arrival rate based only on the entries with specified milepost, which we can finally multiply with the appropriate constant to account for the discarded accidents. Finally, we removed all accidents that happened on ramps while entering or leaving the highway, as they are confounding in multiple ways. (Unfortunately, this binary flag was missing for most accidents; we assumed that the flag was false every time it was missing.) Finally, this left us with 4138 accidents.

Numerical results. In our example $\Delta = [0, T] \times [0, X]$; $T = 1$ week, $X = 77.96$ miles. Taking into consideration how the data are rounded, the regions \mathcal{S} in the objective function (1) can be rectangles no smaller than 1 minute by 0.1 miles, but even considerably larger rectangles are reasonable.

Figure 1 shows a biquadratic spline estimator, obtained using the polyhedral inner approximation, with 28×13 pieces (so each piece corresponds to 6 hours and roughly 6 miles), the regions \mathcal{S} were 1 minute by 1 mile rectangles. The estimator was obtained using an AMPL (Fourer, Gay, and Kernighan 2002) implementation of Algorithm 1, in which the subproblems were solved by the nonlinear solver KNITRO (Ziena). In this problem there was no clear difference between the biquadratic polyhedral and non-polyhedral WSOS approximations.

6 CONCLUSION

We have presented an efficient approach for the spline estimation of non-homogeneous, multi-dimensional Poisson processes from inexact (rounded) arrivals. The key idea behind the approach was to consider piecewise polynomial splines whose pieces are from a polynomial cone possessing two properties: that it admits a good characterization suitable for optimization, and that the splines built from it can uniformly approximate every continuous function. Two variants of this approach were considered: one that uses weighted-sum-of-squares polynomials and leads to optimization models with semidefinite constraints, and one that uses polynomials with nonnegative coefficients in a suitable fixed basis, and results in linearly constrained optimization models. A characterization of polynomial cones possessing the desired density property was provided for both the polyhedral and the non-polyhedral variant.

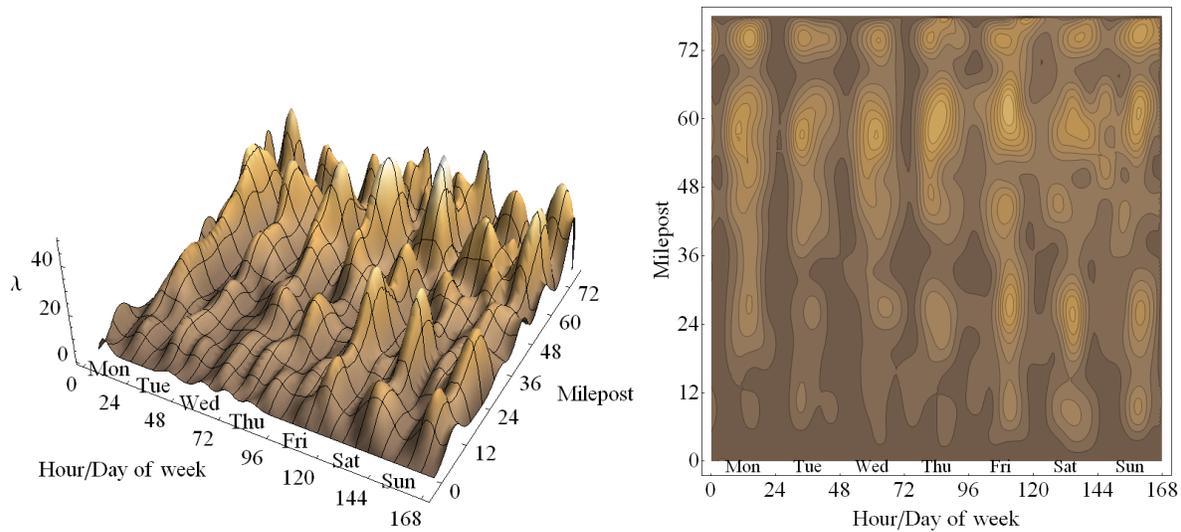


Figure 1: A piecewise biquadratic sum-of-squares spline estimator of the NJTP accident rate obtained using Algorithm 1. Left: three-dimensional plot. Right: contour plot. The horizontal axis shows the time (Monday–Sunday), the vertical axis is location.

The approach was then combined with a decomposition method whose worst-case running time explicitly depends on the neighborhood structure defined by the coupling constraints between the subproblems. This neighborhood structure is the sparsest possible for the spline estimation problems of our interest.

Several questions and future research directions emerge. While in this paper we concentrate only on arrival rate estimation, many of the ideas presented are also applicable for other estimation problems as well. Many of these ideas appear in the thesis of the first author (Papp 2011), but research in this area is far from complete. Another observation worthy of further investigation is the fact that most splines have a bipartite structure: their pieces can be partitioned into two disjoint classes such that only pieces from different classes are adjacent. This makes the arrival estimation model considered in this paper amenable to a number of other iterative decomposition methods, including alternating direction methods. Comparing these approaches and the one presented, especially on a parallel architecture, would be interesting.

ACKNOWLEDGMENTS

This work was partly supported by the National Science Foundation grant number CMMI-0935305.

REFERENCES

- Benders, J. F. 1962, December. “Partitioning procedures for solving mixed-variables programming problems”. *Numerische Mathematik* 4:238–252. doi:[10.1007/BF01386316](https://doi.org/10.1007/BF01386316).
- Boros, E., and P. L. Hammer. 2002, November. “Pseudo-Boolean optimization”. *Discrete Applied Mathematics* 123 (1–3): 155–225. doi:[10.1016/S0166-218X\(01\)00341-9](https://doi.org/10.1016/S0166-218X(01)00341-9).
- Dantzig, G. B., and P. Wolfe. 1960, January. “Decomposition Principle for Linear Programs”. *Operations Research* 8 (1): 101–111. doi:[10.1287/opre.8.1.101](https://doi.org/10.1287/opre.8.1.101).
- Fourer, R., D. M. Gay, and B. W. Kernighan. 2002. *AMPL: A Modeling Language for Mathematical Programming*. second ed. Pacific Grove, CA: Brooks/Cole–Thomson Learning.
- Nesterov, Y. 2000. “Squared Functional Systems and Optimization Problems”. In *High Performance Optimization*, edited by H. Frenk, K. Roos, T. Terlaky, and S. Zhang, Volume 33 of *Applied Optimization*, 405–440. Dordrecht, The Netherlands: Kluwer Academic Publishers.

- Papp, D. 2011, May. *Optimization models for shape-constrained function estimation problems involving nonnegative polynomials and their restrictions*. Ph. D. thesis, Rutgers University, New Brunswick, NJ, USA.
- Renegar, J. 2001. *A Mathematical View of Interior-Point Methods in Convex Optimization*. Number 03 in MPS-SIAM Series on Optimization. Philadelphia, PA: MPS-SIAM.
- Rockafellar, R. T. 1976, May. “Augmented Lagrangians and applications of the proximal point algorithm in convex programming”. *Mathematics of Operations Research* 1:97–116. doi:10.1287/moor.1.2.97.
- Ruszczynski, A. 1995, August. “On Convergence of an Augmented Lagrangian Decomposition Method for Sparse Convex Optimization”. *Mathematics of Operations Research* 20:634–656. doi:10.1287/moor.20.3.634.
- Schumaker, L. L. 1981. *Spline Functions: Basic Theory*. Wiley-Interscience.
- Wolkowicz, H., R. Saigal, and L. Vandenberghe. (Eds.) 2000. *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, Volume 27 of *International series in operations research & management science*. Norwell, MA: Kluwer.
- Ziena. “KNITRO Documentation”. <http://www.ziena.com/documentation.htm>.

AUTHOR BIOGRAPHIES

DÁVID PAPP is a postdoctoral fellow at the Northwestern University Department of Industrial Engineering and Management Sciences. He earned his Ph.D in operations research at Rutgers University. His research interests include mathematical programming, with an emphasis on stochastic optimization and semidefinite programming, with applications in statistics, machine learning, and chemical informatics. He is a member of SIAM, INFORMS, and MAA. His email address is dpapp@iems.northwestern.edu.

FARID ALIZADEH is an Associate Professor at the Rutgers Business School and the Rutgers Center for Operations Research. His e-mail is alizadeh@rutcor.rutgers.edu.