Improved Linear Programs for Discrete Barycenters

Steffen Borgwardt\textsuperscript{1} and Stephan Patterson\textsuperscript{2}

\textsuperscript{1}steffen.borgwardt@ucdenver.edu; University of Colorado Denver
\textsuperscript{2}stephan.patterson@ucdenver.edu; University of Colorado Denver

Abstract. Discrete barycenters are the optimal solutions to mass transport problems for a set of discrete measures. They arise in applications of operations research and statistics. The best known algorithms are based on linear programming, but these programs scale exponentially in the number of measures, making them prohibitive for practical purposes.

In this paper, we improve on these algorithms. First, by using the optimality conditions to restrict the search space, we provide a better linear program that reduces the number of variables dramatically. Second, we recall a proof method from the literature, which lends itself to a linear program that has not been considered for computations. We exhibit that this second formulation is a viable, and arguably the go-to approach, for data in general position. Third, we then combine the two programs into a single hybrid model that retains the best properties of both formulations for partially structured data.

We then study the models through both a theoretical analysis and computational experiments. We consider both the hardness of constructing the models and their actual solution. In doing so, we exhibit that each of the improved linear programs becomes the best, go-to approach for data of different underlying structure.

Keywords: discrete barycenter, optimal transport, linear programming

MSC 2010: 90B80, 90C05, 90C46, 90C90

1 Introduction

Optimal mass transport problems for several marginals arise in many applications of operations research and statistics \cite{3,7,9,10,12,13,18,22,25}. The \textit{weighted Wasserstein barycenters} are optimal solutions to these problems and have seen a lot of activity \cite{4,5,8,11,21,23,29,30}:

Given probability measures $P_1, \ldots, P_n$ on $\mathbb{R}^d$ and a weight vector $\lambda \in \mathbb{R}^n_>$ with $\sum_{i=1}^n \lambda_i = 1$, a Wasserstein barycenter is a probability measure $\bar{P}$ on $\mathbb{R}^d$ satisfying

$$\varphi(\bar{P}) := \sum_{i=1}^n \lambda_i W_2(\bar{P}, P_i)^2 = \inf_{P \in \mathcal{P}^2(\mathbb{R}^d)} \sum_{i=1}^n \lambda_i W_2(P, P_i)^2,$$

(1)

where $W_2$ is the quadratic Wasserstein distance and $\mathcal{P}^2(\mathbb{R}^d)$ is the set of all probability measures on $\mathbb{R}^d$ with finite second moments. Informally, a barycenter $\bar{P}$ is a measure such that the summed-up transport from $\bar{P}$ to all $P_i$ with respect to the quadratic Wasserstein distance is minimal. See the monographs \cite{27,28} and \cite{1} for the current state of the art for barycenters for continuous measures.

We study the frequent setting with data given as a set of \textit{discrete probability measures} $P_1, \ldots, P_n$, i.e. with finite support in $\text{supp}(P_i) \subset \mathbb{R}^d$. We denote the support set of $P_i$ as
supp\((P_i) = \{x_{ik} | k = 1, \ldots, |P_i|\}\), where \(|P_i|\) is the number of support points in \(P_i\). Each \(x_{ik} \in \text{supp}(P_i)\) has a corresponding mass \(d_{ik} > 0\), and \(\sum_{k=1}^{|P_i|} d_{ik} = 1\) for each \(P_i\).

We formally state the problem of computing a discrete barycenter, i.e. a barycenter for a given set of discrete measures.

**Discrete Barycenter Problem**

**Input:** Discrete probability measures \(P_1, \ldots, P_n\), weight vector \(\lambda \in \mathbb{R}^n_{> 0}\)

**Output:** Discrete barycenter \(\bar{P}\) for \(P_1, \ldots, P_n\) and \(\lambda\).

Discrete barycenters satisfy a number of interesting properties [2,17]. In contrast to the continuous case, it is possible to have several barycenters (i.e. optimizers of (1)). However, each barycenter is a discrete measure itself, supported on a subset of the set

\[
S = \{\sum_{i=1}^n \lambda_i x_{ik} | x_{ik} \in \text{supp}(P_i)\} := \{x_1, \ldots, x_{|S|}\}. \tag{2}
\]

This is the set of all convex combinations of support points, one from each measure \(P_i\), given by the fixed \(\lambda_i\). We call its elements \(x_j\) the weighted means. We say the measures \(P_1, \ldots, P_n\) are in general position if different combinations of \(x_{ik} \in \text{supp}(P_i)\) always induce different weighted means \(x_j\).

When representing a barycenter \(\bar{P}\), we use values \(z_j, j = 1, \ldots, |S|\), to denote the mass on support point \(x_j \in S\). Further, the values \(y_{ijk}\) denote mass transported from \(x_j \in S\) to \(x_{ik} \in \text{supp}(P_i)\) (for all \(i = 1, \ldots, n\) and \(k = 1, \ldots, |P_i|\)).

A proof that all barycenters are supported on a subset of \(S\) follows from a combination of two properties. First, the quadratic Wasserstein distance for the cost of transport between two points \(x_j\) and \(x_{ik}\) is simply the squared Euclidean distance \(\|x_j - x_{ik}\|^2\). With this notation, the transportation cost (1) can be written as

\[
\varphi(\bar{P}) := \inf_{P \in \mathcal{P}^+(\mathbb{R}^d)} \sum_{i=1}^n \lambda_i W_2(P, P_i)^2 = \min \sum_{i=1}^n \lambda_i \sum_{j=1}^{|S|} \sum_{k=1}^{|P_i|} \|x_j - x_{ik}\|^2 y_{ijk} \tag{3}
\]

in the discrete setting.

Second, for each barycenter, an optimal transport to the measures is non-mass splitting: The mass of each barycenter support point is transported only to a single support point in each measure. Algebraically, this can be represented as follows: Let \(x_j \in S\) have mass \(z_j\). Then for all \(i\), there is exactly one \(k\) with \(y_{ijk} = z_j\), while \(y_{ijk'} = 0\) for all \(k' \neq k\).

Combining the two properties, it is not hard to see why \(S\) has the aforementioned shape. The point \(x_j = \sum_{i=1}^n \lambda_i x_{ik}\) is the unique minimizer of the cost \(\sum_{i=1}^n \lambda_i \|x^* - x_{ik}\|^2\) of sending a unit of mass from \(x^* \in \mathbb{R}^d\) to a single, fixed \(x_{ik}\) in each measure \(P_i, i = 1, \ldots, n\).

Further, there always exists a barycenter with provably sparse support; we call this a sparse barycenter. More precisely, there is a barycenter \(\bar{P}\) with

\[
|\text{supp}(\bar{P})| \leq \left(\sum_{i=1}^n |P_i|\right) - n + 1. \tag{4}
\]

In particular, the number of support points of \(\bar{P}\) is bounded above by the sum of the number of support points in the measures \(P_1, \ldots, P_n\). This is generally just a tiny fraction of the
number of possible support points in $S$, whose size is bounded by the product, rather than the sum, of the sizes of the original measures.

It is open whether the Discrete Barycenter Problem can be solved in polynomial time. The best known algorithms are based on linear programming [2, 8]: Let the $z_j$ correspond to variables measuring the (unknown) masses of a barycenter supported on $S$. The mass $z_j$ is transported to each measure $P_i$, the amount of which is measured by the variables $y_{ijk}$. This yields constraints $\sum_{k=1}^{\vert P_i \vert} y_{ijk} = z_j$ (for all $i = 1, \ldots, n$ and $j = 1, \ldots, \vert S \vert$). Further, each support point $x_{ik}$ in each measure $P_i$ receives its mass $d_{ik}$ from barycenter support points $x_j$, which can be stated as $\sum_{j=1}^{\vert S \vert} y_{ijk} = d_{ik}$ (for all $i = 1, \ldots, n$ and $k = 1, \ldots, \vert P_i \vert$). Finally, note that (3) is a linear objective function.

Thus using $z_j$ and $y_{ijk}$ as variables, a linear program for the computation of a barycenter may be formulated as

$$\min \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{\vert S \vert} \sum_{k=1}^{\vert P_i \vert} \|x_j - x_{ik}\|^2 y_{ijk}$$

(original)

$$\sum_{k=1}^{\vert P_i \vert} y_{ijk} = z_j, \; \forall i = 1, \ldots, n, \; \forall j = 1, \ldots, \vert S \vert,$$

$$\sum_{j=1}^{\vert S \vert} y_{ijk} = d_{ik}, \; \forall i = 1, \ldots, n, \; \forall k = 1, \ldots, \vert P_i \vert,$$

$$y_{ijk} \geq 0, \; \forall i = 1, \ldots, n, \; \forall j = 1, \ldots, \vert S \vert, \; \forall k = 1, \ldots, \vert P_i \vert$$

$$z_j \geq 0, \; \forall j = 1, \ldots, \vert S \vert.$$

Moreover, any optimal vertex of the LP has sparse support, i.e. it satisfies (4). See [2, 8] for more details.

**Proposition 1.** The Discrete Barycenter Problem can be solved using LP (original). Any optimal vertex corresponds to a sparse barycenter.

One of the most important observations about LP (original) is that its size may scale exponentially in the number $n$ of measures [6]. For simplicity in denoting the worst case scenario, assume $\vert P_i \vert = p_{\text{max}}$ for all $i = 1, \ldots, n$. For measures $P_1, \ldots, P_n$ with support points in general position, one obtains $\vert S \vert = \prod_{i=1}^{n} \vert P_i \vert = (p_{\text{max}})^n$. Then LP (original) has $n(p_{\text{max}})^{n+1} + (p_{\text{max}})^n$ variables and $n(p_{\text{max}})^n + np_{\text{max}}$ equality constraints.

The exponential scaling of LP (original) means that, even for reasonably sized problems, computations are challenging. This is why it is of great interest to reduce the computational effort as much as possible. A first approach in the literature is a strongly polynomial 2-approximation algorithm based on the restriction of the set of support points for an approximate barycenter to the union of supports of the measures $P_1, \ldots, P_n$ [6].

**Contributions and Outline**

In this paper, we improve on LP (original) while retaining all optimal solutions to the Discrete Barycenter Problem. By using the optimality conditions of discrete barycenters in a better way, we present several ways to reduce the problem size.

In Section 2, we formally develop these improvements. First, we explain why just a subset of the variables $y_{ijk}$ is required for a better formulation LP (reduced) of the original LP.
We show that LP (reduced) always is an improved on LP (original). For data in general position, i.e. where each combination \( \sum_{i=1}^{n} \lambda_i x_{ik} \) gives a different point in \( S \), the reduction in the number of variables is dramatic (even for moderate problem sizes, one typically retains less than 1% compared to the original). For structured data, we still obtain a significant improvement (even though these are the cases where LP (original) already worked well). This makes LP (reduced) a go-to approach for structured data.

Second, we turn to an alternative LP (general) that finds a discrete barycenter. It has been used as a proof method in [2,17] to show the existence of a sparse barycenter for all discrete barycenter problems. Due to an inherent, unavoidable exponential scaling, independently of what the underlying data looks like, it has not been considered for computational purposes before. However, for data in general position, we show that it, in fact, is the best approach. For a set of just two measures, one obtains a transportation problem.

Third, we combine the two above approaches to a hybrid model (LP (hybrid)) that retains the best properties of both formulations for partially structured data. The key idea informally is that the decision which model to use can be split into independent decisions for each point in the set \( S \), respectively each combination of support points \( x_{ik} \in \text{supp}(P_i) \) with one from each measure. In partially structured data, it is best to use mix the two strategies.

Sections 3 and 4 are dedicated to the theoretical analysis and computational experiments for different types of data. We use three representative types of data: a geospatial data set in general position, the well-known digits data set, where handwritten digits are recorded in a 16 \times 16-grid (c.f. [15]), and a tailored data set that combines the properties of these two sets.

In Section 3, we discuss data in general position. We show that both LP (reduced) and LP (general) are vast improvements over LP (original) in both theory and practice. Further, we exhibit that a best implementation of LP (hybrid) becomes identical to LP (general).

In Section 4, we discuss data supported in a regular grid. In this highly structured setting, the size of \( S \) becomes polynomial. We begin the section with a discussion of an important distinction: A priori knowledge about this structure versus the lack thereof. We show that the construction of LPs (reduced) and LP (hybrid) is hard (exponential unless \( P = NP \)) if we lack this information, even though the resulting LPs are of polynomial size. We then devise an efficient preprocessing routine for LP (hybrid) (which also works for LP (reduced)) to achieve a significant, but not necessarily optimal improvement over LP (original) while avoiding the inefficient preprocessing required to set up the model exactly.

In Sections 4.2 and 4.3, we show computational experiments without and with a priori knowledge. We see that LPs (reduced) and LP (hybrid) perform best in this setting. We also see that the efficient preprocessing routine leads to a better total running time, even though the constructed LPs are larger. Finally, in Section 4.4, we show computational experiments for a data set supported in the combination of a regular grid with some additional points in general position. Here, LP (hybrid) greatly outperforms the other models, as it is able to adapt its representation of \( S \) to the different parts of the data.

2 Improved Linear Programs

In the following, we describe three ways to improve on the formulation of LP (original). The first one is a strict improvement, the second one the best approach for data in general position, and the third one a hybrid approach of the former two.
2.1 Optimality Conditions for \( y \) Variables

For an initial reduction in size, we note that variables can be dropped from LP (original) while keeping all optimal solutions in the feasible set. Due to the non-mass splitting property, we have seen that any \( x_j = \sum_{i=1}^{n} \lambda_i x_{ik} \in S \) is an optimal support point for mass to be transported to all of the \( x_{ik} \) from which it is constructed. This implies that in an optimal solution, \( x_j \) never transports to any \( x_{ik} \) not in a weighted mean calculation producing \( x_j \).

Equivalently, in the optimal solution, \( y_{ijk} = 0 \) for all such pairs \( x_j, x_{ik} \). We can therefore eliminate those \( y_{ijk} \) from the formulation by only introducing variables \( y_{ijk} \) when \( x_{ik} \) appears in a computation of weighted mean \( x_j \).

We require some notation. Let \( S_{ik} \) be the set of indices \( j \) for which \( x_j \in S \) can be computed as a weighted mean of a set of support points that includes \( x_{ik} \). Formally,

\[
S_{ik} = \{ j : x_j = \lambda_i x_{ik} + \sum_{l=1,l\neq i}^{n} \lambda_l x_{lk'} \text{ for some } x_{lk'} \in \text{supp}(P_l) \}.
\]

Conversely, let \( S_j \) to be the set of index tuples \((i,k)\) of support points \( x_{ik} \) which contribute to a computation of \( x_j \), i.e.

\[
S_j = \{ (i,k) : x_j = \lambda_i x_{ik} + \sum_{l=1,l\neq i}^{n} \lambda_l x_{lk'} \text{ for some } x_{lk'} \in \text{supp}(P_l) \}.
\]

With this notation, LP (original) can be improved to a smaller formulation that restricts the use of variables \( y_{ijk} \) to only when \( x_{ik} \) appears in a construction of the weighted means \( x_j \) as follows.

\[
\begin{align*}
\min & \; \sum_{i=1}^{n} \lambda_i \sum_{k=1}^{\left| P_i \right|} \sum_{j \in S_{ik}} \| x_j - x_{ik} \|^2 y_{ijk} \\
\sum_{(i,k) \in S_j} y_{ijk} &= z_j, \; \forall i = 1, \ldots, n, \; \forall j = 1, \ldots, |S| \\
\sum_{j \in S_{ik}} y_{ijk} &= d_{ik}, \; \forall i = 1, \ldots, n, \; \forall k = 1, \ldots, |P_i| \\
y_{ijk} &\geq 0, \; \forall i = 1, \ldots, n, \; \forall k = 1, \ldots, |P_i|, \; \forall j \in S_{ik} \\
z_j &\geq 0, \; \forall j = 1, \ldots, |S|
\end{align*}
\]

We obtain Theorem 1.

**Theorem 1.** *The Discrete Barycenter Problem can be solved using LP (reduced). Any optimal vertex corresponds to a sparse barycenter.*

The optimal vertices of LPs (original) and (reduced) are in one-to-one correspondence. This immediately gives the second part of the statement.

Note that at least one pair \((x_j, x_{ik})\) where \( x_{ik} \) is not part of any weighted means construction of \( x_j \), exists for any dimension \( \mathbb{R}^d \) as long as \( n \geq 2 \) and \( |P_i| \geq 2 \) for at least one \( i = 1, \ldots, n \). In other words, for non-trivial input there is an \((i, k) \notin S_j\) for some \( j \). Thus LP (reduced) provides a strict reduction in the number of variables and the number of non-zero entries in the constraint matrix for all practical examples.
Lemma 1. For a set of at least two measures $P_1, \ldots, P_n$, and at least one with $|P_i| \geq 2$, LP (reduced) has strictly fewer variables than LP (original). Further, the nonzero entries of the constraint matrix are a strict subset.

Informally, LP (reduced) always is a strict improvement over LP (original). In Section 4, we turn to the dramatic reduction in size—often several orders of magnitude—in more detail.

2.2 Fixed Transport

An alternative linear program for the Discrete Barycenter Problem has been used as a proof method in [2,17] for the properties in Section 1, but not considered for computational purposes. However, as we will see in Section 3, there are inputs where this is the preferred approach, in particular for data in general position.

The key idea is to treat each combination of original support points that gives a weighted mean separately, even if they produce the same coordinates. Applying this idea to LP (reduced) means that a variable $z_j$ is used for each combination of original support points. Then $|S_j| = n$, it contains precisely one pair $(i,k)$ for each $i \leq n$. If mass is associated to variable $z_j$, it is also fully associated to the corresponding $y_{ijk}$ with $(i,k) \in S_j$. This implies that the variables $y_{ijk}$ can be eliminated through $y_{ijk} = z_j$ for all $(i,k) \in S_j$. Informally, assigning mass to $z_j$ gives rise to a fixed transport to the measures.

We now develop an LP formulation based on this idea in our own notation. First, we define the set of all combinations of original support points, one from each measure, as

$$S^* = \{(x_{1k}, \ldots, x_{nk}) : x_{ik} \in \text{supp}(P_i)\} := \{s_1^*, \ldots, s_{|S^*|}^*\}.$$ 

There is an intimate relation of $S^*$ and $S$: Each tuple $s_h^* = (x_{1k}, \ldots, x_{nk})$, $h = 1, \ldots, |S^*|$, corresponds to a set of original support points $x_{ik}$, $i = 1, \ldots, n$. These support points add up to a weighted mean $x_j = \sum_{i=1}^n \lambda_i x_{ik} \in S$. So each $s_h^* \in S^*$ is associated to an $x_j \in S$. Generally, it is possible that multiple $s_h^* \in S^*$ are associated with the same $x_j \in S$. In turn, each $x_j \in S$ is associated to at least one $s_h^* \in S^*$. We obtain $|S^*| \geq |S|$. For data in general position, $|S^*| = |S|$, and there is a bijection between the tuples $s_h^*$ and weighted means $x_j$.

Next, we introduce a variable $w_h$ for each $s_h^* = (x_{1k}, \ldots, x_{nk})$, $h = 1, \ldots, |S^*|$, representing mass associated with it. The corresponding cost $c_h$ of transporting a unit of mass to the measures is

$$c_h = \sum_{i=1}^n \lambda_i ||x_j - x_{ik}||^2.$$ 

Finally, we define sets $S^*_{ik}$ similarly to the sets $S_{ik}$. $S^*_{ik}$ is the set of indices $h$ in $1, \ldots, S^*$ where the $i$th component of $s_h^*$ is $x_{ik}$. Formally,

$$S^*_{ik} = \{h: s_h^* = (\ldots, x_{ik}, \ldots)\}.$$ 

We now have all the ingredients for the LP. Note that in LPs (original) and (reduced), the first type of main constraints was used to connect the $z_j$ and $y_{ijk}$. They are not needed in the new variant, where the $y_{ijk}$ are eliminated due to the fixed transport. We obtain the following LP as an alternative to solving the Discrete Barycenter Problem.
The Discrete Barycenter Problem can be solved using LP (general). Any optimal vertex corresponds to a sparse barycenter.

For \( n = 2 \), i.e. for a set of two discrete measures, the computation of a discrete barycenter can be modeled as a transportation problem; see [17] where this was explained through a transformation of LP (original). We believe the best way to prove this claim is to perform a reindexing of LP (general) for this special case, as follows:

As there are only two measures, we can denote the support elements and masses as \( x_k \in \text{supp}(P_1) \) with mass \( d_k \) and \( x_l \in \text{supp}(P_2) \) with mass \( d_l \). Further, the weights for the measures can be represented as \( \lambda_1 = \lambda \) and \( \lambda_2 = 1 - \lambda \).

For a given \( x_k \in \text{supp}(P_1) \) and \( x_l \in \text{supp}(P_2) \), an optimal barycenter support point is \( x = \lambda x_k + (1 - \lambda)x_l \). The corresponding cost of transport is

\[
  c_{kl} = \lambda \|x - x_k\|^2 + (1 - \lambda)\|x - x_l\|^2 = \lambda(1 - \lambda)\|x_k - x_l\|^2 + \lambda^2(1 - \lambda)\|x_k - x_l\|^2
\]

\[
= \lambda(1 - \lambda)\|x_k - x_l\|^2 = \alpha\|x_k - x_l\|^2,
\]

for some constant \( \alpha \in \mathbb{R} \). Thus, the Discrete Barycenter Problem for \( n = 2 \) can be denoted as the following transportation problem.

\[
\min \sum_{k=1}^{\mid P_1 \mid} \sum_{l=1}^{\mid P_2 \mid} c_{kl}w_{kl} \quad \text{(transportation)}
\]

\[
\sum_{k=1}^{\mid P_1 \mid} w_{kl} = d_k, \quad \forall k = 1, \ldots, \mid P_1 \mid
\]

\[
\sum_{l=1}^{\mid P_2 \mid} w_{kl} = d_l, \quad \forall l = 1, \ldots, \mid P_2 \mid
\]

\[
w_{kl} \geq 0, \quad \forall k = 1, \ldots, \mid P_1 \mid, \quad \forall l = 1, \ldots, \mid P_2 \mid
\]

It is well-known that transportation problems can be solved in strongly polynomial time [24,14]. This immediately gives Theorem 3.

Theorem 3. The Discrete Barycenter Problem for \( n = 2 \) can be solved in strongly polynomial time by solving LP (transportation), respectively LP (general).

In Section 3, we will see that LP (general) also is the go-to approach for a set of measures in general position.
2.3 Hybrid Approach to Variable Introduction

The strategies of variable introduction presented in LP (reduced) and LP (general) are not mutually exclusive. For each tuple \( s_h^* \) individually, we can decide whether to use a representation with \( y \)-variables as in LP (reduced) or to use a representation with \( w \)-variables as in LP (general). To this end, we partition the set \( \{1, \ldots, |S^*|\} \) into two index sets \((S^*)_w\) and \((S^*)_y\), to indicate for which \( s_h^* \) we will use the \( y \)-representation, respectively the \( w \)-representation.

The original support points \( x_{ik} \) can receive mass through a transport denoted by some \( y_{ijk} \) and through some fixed transport of a \( w_h \). First, we introduce \((S^*)_w\) to be the set of indices \( h \) such that the \( i^{th} \) component of \( s_h^* \) is \( x_{ik} \). Then \((S^*)_w \subset S^*_w\) corresponds precisely to those combinations \( s_h^* \) which imply a fixed transport to \( x_{ik} \) if assigned mass. For a formal definition of \((S^*)_w\), we only have to restrict \( S^*_w \) to indices \( h \in (S^*)_w \), i.e.

\[
(S^*)_w = \{ \ h : \ s_h^* = (\ldots, x_{ik}, \ldots), h \in (S^*)_w \}. 
\]

For a proper indexing of the transport corresponding to \((S^*)_y\) we need index sets that mirror \( S^*_w \) and \( S^*_y \) as defined in Section 2.1, restricted to \((S^*)_y\). \( S^*_y \) contains the indices \( j \) of all \( x_j \in S \) produced by the weighted means of tuples in \((S^*)_y\) that contain \( x_{ik} \). Formally,

\[
S^*_y = \{ \ j : \ x_j = \lambda_i x_{ik} + \sum_{l=1,l\neq i}^n \lambda_l x_{ki'} \text{ for some } s_h^* = (x_{i1}, \ldots, x_{ik}, \ldots, x_{nk'}) \in (S^*)_y \}. 
\]

Further, \( S^*_y \) contains those index pairs \((i,k)\) for which \( x_j \in S \) is the weighted mean of a tuple in \((S^*)_y\) that contains \( x_{ik} \). This gives

\[
S^*_y = \{ \ (i,k) : \ x_j = \lambda_i x_{ik} + \sum_{l=1,l\neq i}^n \lambda_l x_{ki'} \text{ for some } s_h^* = (x_{i1}, \ldots, x_{ik}, \ldots, x_{nk'}) \in (S^*)_y \}. 
\]

Now, we are ready to state an LP that allows for the split of \( \{1, \ldots, |S^*|\} \) into index sets \((S^*)_y\) and \((S^*)_w\):

\[
\min \sum_{h \in (S^*)_w} c_h w_h + \sum_{i=1}^n \lambda_i \sum_{k=1 \in S^*_y}^{|P_i|} \|x_j - x_{ik}\|^2 y_{ijk} \quad \text{(hybrid)}
\]

\[
\sum_{(i,k) \in S^*_y} y_{ijk} = z_j, \quad \forall i = 1, \ldots, n, \quad \forall j = 1, \ldots, |S^*_y|.
\]

\[
\sum_{h \in (S^*)_w} w_h + \sum_{j \in S^*_y} y_{ijk} = d_{ik}, \quad \forall i = 1, \ldots, n, \quad \forall k = 1, \ldots, |P_i|.
\]

\[
w_h \geq 0, \quad \forall h = 1, \ldots, |(S^*)_w|,
\]

\[
y_{ijk} \geq 0, \quad \forall i = 1, \ldots, n, \quad \forall k = 1, \ldots, |P_i|, \quad \forall j \in S^*_y
\]

\[
z_j \geq 0, \quad \forall j = 1, \ldots, |S^*_y|
\]

Correctness of this model is a direct consequence of Theorems 1 and 2.

**Theorem 4.** The Discrete Barycenter Problem can be solved using LP (hybrid). Any optimal vertex corresponds to a sparse barycenter.
In the next sections, we study the advantages of LPs (reduced), (general), and (hybrid) in model size and practical performance in some representative settings. We begin with data in general position, where LP (general) is the best model and LP (hybrid) mirrors it. Then we turn to highly structured data, more precisely a set of measures supported in regular $d$-dimensional grids, where LPs (reduced) and (hybrid) dramatically outperform LP (general). Finally, we exhibit a type of input for which LP (hybrid) vastly outperforms the other approaches.

### 3 Data in General Position

#### 3.1 Theoretical Analysis

We begin with data in general position. Recall that measures $P_1, \ldots, P_n$ are in general position if different combinations of $x_{ik} \in \text{supp}(P_i)$ always induce different weighted means $x_j$. For a simple notation, we assume that all $P_i$ have the same number of support points $|P_i| = p_{\text{max}}$. Further, recall that LP (original) used $n(p_{\text{max}})^n + 1 + (p_{\text{max}})^n$ variables and $n(p_{\text{max}})^n + n p_{\text{max}}$ constraints.

First, we analyze the size reduction achieved through LP (reduced). The general position implies that all $(p_{\text{max}})^n$ combinations of support points in the original measures produce $(p_{\text{max}})^n$ different $x_j$, which then is also the number of required $z_j$ variables. Since each $x_j$ in LP (reduced) transports to $n$ points, one for each $P_i$, we get a total number of variables $(p_{\text{max}})^n + n(p_{\text{max}})^n = (1 + n)(p_{\text{max}})^n$. This has reduced the exponent on the growth by one compared to before; the number of variables is reduced by $n(p_{\text{max}})^n(p_{\text{max}} - 1)$.

A representation of this reduction as a percentage highlights the dramatic improvement over LP (original). The percentage reduction in number of variables can be represented as a function that is dominated by $p_{\text{max}}$:

$$1 - \frac{(1 + n)(p_{\text{max}})^n}{(p_{\text{max}})^n + n(p_{\text{max}})^n + 1} = 1 - \frac{1 + n}{1 + n p_{\text{max}}} = \frac{n(p_{\text{max}} - 1)}{1 + n p_{\text{max}}} = \frac{(p_{\text{max}} - 1)}{n + p_{\text{max}}}.$$  

For all reasonable sizes of $n$, this fraction is close to $\frac{(p_{\text{max}} - 1)}{p_{\text{max}}}$. For example, for $p_{\text{max}} = 256$, as is the case for $P_i$ supported densely on a $16 \times 16$ grid, this is about a 99.61% reduction in the number of variables. The digits data set used for computational experiments in Section 4 is supported on such a grid (but not densely).

Next, we turn to LP (general). Here, the number of variables depends only on the sizes of the support sets $\text{supp}(P_i)$ of the original measures. There are always $\prod_{i=1}^n |P_i|$ variables regardless of whether or not the measures are in general position. Informally, unlike LP (reduced) or LP (original), LP (general) always assumes the worst case of having data in general position. For all $|P_i| = p_{\text{max}}$, it has $(p_{\text{max}})^n$ variables. In this case, LP (original) had $(p_{\text{max}})^n + n(p_{\text{max}})^{n+1}$ variables. Using LP (general) eliminates all of the $y$ variables from LP (original), reducing the total number of variables by $n(p_{\text{max}})^{n+1}$. It also eliminates all $y$ variables from LP (reduced), for a reduction in the number of variables of $n(p_{\text{max}})^n$.

Looking again at the percentage reduction, for LP (original) the reduction is given by

$$1 - \frac{(p_{\text{max}})^n}{(n p_{\text{max}})^{n+1} + (p_{\text{max}})^n} = \frac{n p_{\text{max}}}{n p_{\text{max}} + 1} = \frac{p_{\text{max}}}{p_{\text{max}} + \frac{1}{n}}.$$  

This can get arbitrarily close to a 100% reduction in variables; at $n = 4$ and $p_{\text{max}} = 256$ this is already 99.9%. Comparing to LP (reduced), the reduction is given by
\[
1 - \frac{(p_{\text{max}})^n}{(1 + n)(p_{\text{max}})^n} = \frac{n}{n + 1} = \frac{1}{1 + \frac{1}{n}}.
\]

Thus the percentage reduction between LP (reduced) and LP (general) depends only on \(n\) and can also get arbitrary close to 100%; for \(n = 4\), this is an 80% reduction in number of variables.

It is also worth noting that the reduction of \(y\) variables when improving LP (original) to LP (reduced) does not reduce the number \(n(p_{\text{max}})^n + p_{\text{max}}\) of constraints. In contrast, LP (general) does reduce the number of constraints to \(np_{\text{max}}\), which is no longer exponential. This reduction can be significant for the solving time, especially if the number of required variables is comparable between formulations. It is always the case that LP (reduced) has at least as many constraints as LP (general). For data in general position, LP (general) is clearly preferable, with both fewer variables and fewer constraints.

Finally, consider LP (hybrid). For each of the \((p_{\text{max}})^n\) different \(x_j\), respectively each of the \(|S| = |S^*|\) combinations \(s_h\) of original support points, we have to decide whether to introduce \(w\) variables as in LP (general) or \(y, z\) variables as in LP (reduced). With the same arguments as above, it is best to always introduce \(w\) variables, which produces exactly LP (general). Formally, we obtain Theorem 5.

**Theorem 5.** For data in general position, choosing \((S^*)^w = S^*\) and \((S^*)^y = \emptyset\) gives a minimal number of variables and constraints in LP (hybrid). Then LP (hybrid) and LP (general) are identical.

We summarize the formulas for the number of variables and constraints for the various LP variants in Table 1.

| LP formulation | Variables | Constraints |
|----------------|-----------|-------------|
| (original)     | \(n(p_{\text{max}})^n + 1\) | \(n(p_{\text{max}})^n + np_{\text{max}}\) |
| (reduced)      | \((1 + n)(p_{\text{max}})^n\) | \((p_{\text{max}})^n + np_{\text{max}}\) |
| (general)      | \((p_{\text{max}})^n\) | \(np_{\text{max}}\) |
| (hybrid)       | \((p_{\text{max}})^n\) | \(np_{\text{max}}\) |

**Table 1.** The number of variables and constraints of the LPs for data in general position.

### 3.2 Computational Results

We exhibit the practical advantage of LP (general) over LP (original) through some sample computations. These computations are on crime data for Denver County, which is openly available as part of the Denver Open Data Catalog. For a data set in general position, we use the locations of murders during 2016. Each month forms a measure \(P_i, i = 1, \ldots, 12\), by weighting each murder equally during the month.

A discrete barycenter for this data set can be interpreted to indicate locations for police presence, such that a fast response to (at least) one of the incidents in each month is achieved. An aggregate image of all murder locations and a corresponding discrete barycenter are displayed in Figure 1. The radii of the shapes in both parts of the figure are relative to their masses. Larger masses occur for fewer murders in that month.
The data for this application was processed in Python, and the LP was solved using AMPL. Using LP (original), the largest number of months for which we were able to compute a discrete barycenter was 9 months. This computation took a minute on a standard laptop. In contrast, using LP (general) the same computation finished in less than a second. The barycenter for a full year (12 months) of data, depicted in Figure 1, was computed in 14 minutes. Note that the number of variables for the 12-month set is about 30 times larger than for 9 months; recall the exponential scaling, highlighted in Table 1.

4 Data in Regular Grids

LPs (original) and (reduced) do have a significant advantage over LP (general) in many practical applications: They are able to take advantage of the structure of the supports of $P_i$, whereas LP (general) cannot. To see this advantage, first recall that LPs (original) and (reduced) have variables $y$ and $z$ both indexed on $j = 1, \ldots, |S|$, where $|S|$ is the number of distinct weighted means. In contrast, LP (general) introduces variables $w$ indexed on $h = 1, \ldots, |S^*|$. $S$ can be of much smaller size than $S^*$, where each combination $s_h^*$ of
support points in the original measures is counted, even if they result in the same support point \( x_j \in S \).

We will highlight the differences in the various model sizes for structured data by considering measures that are supported in a \( d \)-dimensional regular grid. This is one of the most frequent settings in optimal mass transport problems. The MNIST digits data set of handwritten digits is a prime example for such data. It has been used for benchmarking for many machine learning algorithms; see [15] for more information. In this database, each measure is a handwritten digit scanned into a \( 16 \times 16 \) grid. Each measure is supported on a subset of the grid. The different shades of grey indicate different masses at the support points; the darker, the more mass it holds. The masses add up to 1. See Figure 2. In this figure, we also include a sample barycenter computed for four digits from the set. Note that the barycenter is supported sparsely on a finer grid (here \( 61 \times 61 \)).

![Figure 2](image)

**Fig. 2.** Left, an example of digit 8 from the MNIST digits data set. Right, an example barycenter calculated for four digit 8’s.

For our implementation, we considered two approaches. In the first, we implemented LPs (original), (reduced), (general), (hybrid) in C++. In this implementation, during the data processing we do not use the a priori knowledge that the final structure of \( S \) is a subset of the \((nK - n + 1)^d\)-length grid. Instead, we generate \( S \) by processing each element of \( S^* \) and checking for duplication in those elements already produced. This requires exponential effort, but allows us to determine exactly those \( x_j \in S \) which can be generated by the support set of the digits of consideration, and produce the smallest possible LP’s for each formulation. These provide a baseline for other applications in which there may be structure, but the presence or type of which is not known prior to processing.

The exponential effort of the preprocessing for the first implementation is in strong contrast to the polynomial size of the support set \( S \) in this application. Because of this, in our second implementation we do use the a priori knowledge of the structure of \( S \) to generate the possible \( x_j \). This is done without checking if \( x_k \) exist to produce \( x_j \), so it requires effort that is now linear in the number of measures \( n \). This leads to larger LP’s, due to the presence of additional \( x_j \), but significantly less processing time. We have implemented LPs (original) and (reduced) in this manner. Both efforts produce LP’s containing variable and constraint numbers bounded by the formulas given in Table 2, which are given under the assumptions that the original grids have non-zero mass at all elements of the grid. All LPs were solved using the Gurobi Optimizer 7.0.
We first present the theoretical size comparisons for grid-structured data. Included in the theory is the additional necessary grid knowledge which we use in the second implementation where we avoid processing $S^*$; that is, we consider how we may reduce the $y$ variables without construction of $S_{ik}$ and $S_j$. We also discuss a formula for determining if the hybrid approach may introduce fewer variables without processing $S^*$, but we do not implement this.

4.1 Theoretical Analysis

LPS (original) and (reduced) versus LP (general)

We begin by comparing the model sizes of LPs (original) and (general). Suppose that all measures $P_i$ are supported on a $d$-dimensional regular grid of integer step sizes in each direction, each coordinate going from 1 to $K$, and that $\lambda_i = \frac{1}{n}$ for all $i = 1, \ldots, n$. Informally, then $S$ becomes a $d$-dimensional grid that is $n$ times finer than the original grids. We get $|S| = (nK - n + 1)^d$, and the number of variables in LP (original) is

$$(nK - n + 1)^d(1 + \sum_{i=1}^n |P_i|) = (nK - n + 1)^d(1 + nK^d).$$

This growth is polynomial in $n$, with degree $d + 1$. By contrast, LP (general) will have $(K^d)^n$ variables, an exponential growth in $n$. The number of constraints in both LPs does not significantly impact the relation between the two LPs: While LP (general) has the minimum number of $nK^d$ constraints, LP (original) only has an additional $n(nK - n + 1)^d$ constraints (which again is polynomial in $n$ with degree $d + 1$). So both formulations have a polynomial number of constraints for a regular grid.

| LP formulation | Variables | Constraints |
|----------------|-----------|-------------|
| (original)     | $(nK - n + 1)^d(1 + nK^d)$ | $nK^d + n(nK - n + 1)^d$ |
| (general)      | $(K^d)^n$  | $nK^d$      |

Table 2. The number of variables and constraints of the LP’s for $n$ measures supported fully on a grid of length $K$.

For grid-structured data, the previous result that LP (reduced) is smaller than LP (original) still holds. In the first implementation, while processing $S^*$ we can again construct $S_{ik}$ and eliminate any $y$ not in the combination in that manner. But in fact, we now show that we can also use the structure to determine, prior to processing, all original support points $x_{ik}$ which can never produce a particular $x_j$. This knowledge will be used directly in the second implementation of LP (reduced).

For the sake of simplicity, let $\lambda_i = \frac{1}{n}$. Then $x_j = \frac{1}{n} \sum_{i=1}^n x_{ik}$ for all $x_j \in S$ and we can use $s_l = \sum_{i=1}^n x_{ikl}$ to denote each coordinate $l = 1, \ldots, d$. Consider an $x_j$ which has a coordinate $s_l$ within $K - 1$ of the minimum or of the maximum among all points in $S$ for at least one $l = 1, \ldots, d$. Then there exist points in the original grid to which it cannot transport, in any optimal solution. This is depicted in Figure 3, which shows the set $S$ for 4 measures in $\mathbb{R}^2$ in a regular grid with $K = 4$, resulting in a $13 \times 13$ grid. Any point with a coordinate of 1, 1.25, 1.5, or a coordinate of 4, 3.75, or 3.5 does not transport to all points.
in the original grids. All other points – those in the ‘center’ of the grid – will require all \( nK^d + 1 \) variables when the original measures are supported on the entirety of the original grid.

In summary, LP (reduced) is of smaller size than LP (original), and both are significantly smaller than (general), in this setting. Without a priori knowledge, the construction of LP (reduced) requires the implicit setup of sets \( S_{ik} \) and \( S_j \) (recall the discussion in Section 2.1), which we will now discuss the difficulty in doing. Due to the structure of the grid data, it is possible to efficiently – so in particular without processing \( S^* \) explicitly – take advantage of the effect highlighted in Figure 3.

In contrast, for general data, this is a difficult task: We prove that the construction of the sets \( S_j \) is hard, unless \( P \neq \text{NP} \). Of course, if \( S \) is of exponential size, the total effort of constructing all the \( S_j \) trivially is be exponential, too. However, even deciding whether a given \( x \) lies in \( S \) or not is already NP-hard.

**Lemma 2.** Let \( P_1, \ldots, P_n \subset \mathbb{R}^d \) be discrete measures, let \( \lambda_1, \ldots, \lambda_n \), and let \( x \in \mathbb{R}^d \). Then it is NP-hard to decide whether \( x \in S \), even for \( d = 1 \).

**Proof.** We prove the claim through a reduction to the subset sum problem, which is known to be NP-complete. The subset sum problem can be stated as follows: given a set of integer \( p_1, \ldots, p_n \), is there a non-empty subset whose sum is a given \( s \in \mathbb{Z} \)?

Let \( P_1, \ldots, P_n \subset \mathbb{Z} \) be discrete measures, and let \( P_i \) consist of the support points \( p_i \) and 0 (both of mass \( \frac{1}{2} \), but this is not relevant). Further, let \( \lambda_1, \ldots, \lambda_n = \frac{1}{n} \) and \( x \in \mathbb{Z} \).

To decide whether \( x \in S \), note that \( x \) has to be represented as \( x = \sum_{i=1}^n \lambda_i x_{ik} \), where \( x_{ik} \) is either \( p_i \) or 0 for all \( i \leq n \). Thus, the decision is equivalent to the decision whether there is a subset of the \( p_1, \ldots, p_n \) adding up to \( n \cdot x \). This proves the claim. \( \Box \)

We obtain the following direct consequence on the hardness of constructing \( S_j \) for a fixed \( x_j \in S \).

**Corollary 1.** Let \( P_1, \ldots, P_n \subset \mathbb{R}^d \) be discrete measures, let \( \lambda_1, \ldots, \lambda_n \), and let \( x_j \in S \). Then it is NP-hard to construct the set \( S_j \).

Recall that, formally, NP-hardness is a statement on decision problems. The phrasing in Corollary 1 means that it is NP-hard to decide whether a given set \( S_j \) is correct.

![Fig. 3](image-url) Combining \( n = 4 \) grids of length \( K = 4 \) results in a 13x13 grid for \( S \). The outer \( K - 1 \) rows and columns have at least \( nK \ x_{ik} \) to which they will not transport.
On the construction of LP (hybrid)

The key idea to LP (hybrid) is that the two strategies of variable introduction can be used interchangeably; recall Section 2. Note that choosing between a \( y, z \) set of variables and a set of \( w \) variables is different from the construction of sets \( S_{ik} \) and \( S_{j} \); we only have to count the number of combinations \( s_{ik}^{*} \) that give the same support point \( x_{j} \). Lemma 2 tells us that this remains hard for general data. However, for data in regular grids, we have a lot of extra information. In the following, we analyze choosing between a \( y, z \) set of variables and a set of \( w \) variables for each support point \( x_{j} \) in \( S \), introducing all of one type. So the combinations \( s_{ik}^{*} \) that give the same \( x_{j} \) have to be either all in \((S^{*})^{w}\) or all in \((S^{*})^{y}\), and not split up. This simplification is a restriction for general data, where obtaining a lower number of variables for a variant of LP (hybrid) may be possible by mixing the two strategies for an individual \( x_{j} \). We do not attempt to achieve this true minimum. Our goal is a significant improvement over LP (reduced).

To derive a formula for the number of combinations producing a particular \( x_{j} \), we begin in dimension \( d = 1 \). We again let \( \lambda_{i} = \frac{1}{n} \) for all \( i = 1, \ldots, n \). Then \( x_{j} = \frac{1}{n} \sum_{i=1}^{n} x_{ik} \) for each \( x_{j} \in S \). We consider \( s = n \cdot x_{j} = \sum_{i=1}^{n} x_{ik} \). Note that the \( x_{ik} \) are integers between 1 and \( K \).

The problem of determining the number of combinations of \( x_{ik} \) that give \( x_{j} \) is equivalent to the number

\[
F(s, K, n) = \text{number of solutions to } \begin{cases} \ a_1 + a_2 + \ldots + a_n = s \\ 1 \leq a_i \leq K, a_i \in \mathbb{Z} \forall i \leq n \end{cases}
\]

This is a standard counting problem, where \( n K \)-sided dice are tossed and one is interested in the number of combinations that give sum \( s \). Thus, \( F(s, K, n) \) is given by

\[
F(s, K, n) = \sum_{m=0}^{n} (-1)^{m} \binom{n}{m} \left( \frac{s - mK - 1}{n - 1} \right). \tag{5}
\]

We can extend this formula to \( x_{j} \in \mathbb{R}^{d} \) for higher dimension \( d \) by considering each coordinate of \( x_{j} \) individually. Let \( s_{l} \) be the corresponding sum of the \( l \)-th coordinate of \( x_{j} \); then the total number of combinations whose weighted mean is \( x_{j} \) is

\[
N_{j} = \prod_{l=1}^{d} F(s_{l}, K, n). \tag{6}
\]

The maximum number of variables \( y_{ijk} \) which will be introduced for a particular \( z_{j} \) is \( nK^{d} \). So for any \( j \) where \( N_{j} \) exceeds \( nK^{d} + 1 \), it is better to use \( y, z \) variables, rather than introducing individual \( w_{hk}, h = 1, \ldots, N_{j} \), for the combinations. The \( x_{j} \) for which the \( w \) are preferable correspond to the corners of the possible support grid \( S \). Figure 4 illustrates these \( x_{j} \) for \( n = 3 \) and \( n = 4 \) for the grid corresponding to the digits data set.

In Figure 4, we do note that for large \( n \), the majority of the grid will prefer \( y, z \) variables. Additionally, the areas where \( w \) variables are preferable overlap heavily with the same areas of the grid where, as in Figure 3, we can reduce the number of \( y \) variables. We therefore anticipate that in this scenario, the hybrid approach will not show significant improvement over LP (reduced). We will therefore conclude our computational experiments with an example in which LP (hybrid) does show a significant advantage.
4.2 Digit Computations: No Usage of A Priori Knowledge

We first present the LP sizes and solving times for computations where we do not take advantage of the a priori knowledge of the barycenter support set structure. Each run is the computation of a barycenter for four handwritten variants of the same digits, with the chosen digits’ computations being representative of a typical sample of that digit. These computations follow the discussion in the beginning Section 4; they give a baseline comparison for the four LP formulations when there is a lot of repetition in the weighted means, but no a priori expert knowledge on the structure of the possible support $S$. Recall that $S$ is much smaller than $S^*$ for such data. It is expected that LP (general) will perform poorly in this scenario, as it is unable to benefit from the underlying structure.

Table 3 contains the number of variables and constraints produced by the formulations and Table 4 contains the solving times of the LPs after setup. The size of the support of each measure varies significantly based on the digit (with digit 1 being the smallest), which is the reason for the significant difference in running and solving times among the different digits.

| Digit | LP (original) | LP (reduced) | LP (general) | LP (hybrid) |
|-------|---------------|--------------|--------------|-------------|
|       | Rows          | Columns      | Rows         | Columns     |
| 0     | 10,598        | 1,317,437    | 10,598       | 740,116     |
| 1     | 3,083         | 144,200      | 3,083        | 82,834      |
| 3     | 10,999        | 1,212,100    | 10,569       | 679,509     |
| 7     | 7,944         | 640,974      | 7,944        | 353,378     |
| 9     | 8,787         | 918,280      | 8,787        | 500,554     |

Table 3. Number of constraints and variables for four representative measures of each digit with each formulation without using a priori knowledge.

Fig. 4. Support points in $S$ for a set of $n = 3$ and $n = 4$ measures supported in a $16 \times 16$ grid. The dark support points indicate those where $w$ variables are preferable over $y, z$ variables in an implementation of LP (hybrid).
### Table 4. Solving times, in seconds, for four measures of each digit with each formulation without using a priori knowledge. (*) For three of the digits, the LP solver was unable to complete the setup on the laptop due to memory constraints.

| Digit | LP (original) | LP (reduced) | LP (general) | LP (hybrid) |
|-------|---------------|--------------|--------------|-------------|
| 0     | 151.69        | 110.31       | *            | 93.83       |
| 1     | 2.69          | 1.82         | 323.15       | 1.46        |
| 3     | 163.57        | 76.06        | *            | 69.42       |
| 7     | 53.50         | 24.54        | 6488.2       | 22.44       |
| 9     | 86.06         | 44.92        | *            | 36.14       |

As seen in Table 3, the number of variables (columns) in the LP is reduced significantly when moving from LP (original) to LP (reduced) as is expected from the earlier analysis. This is roughly a 44% reduction in number of variables. The lower reduction than in the theoretical analysis is due to the sparsity of the measures; not all measures are of size $p_{\text{max}}$. In Table 3 we also include the size of LP (general), which has fewer constraints (rows), but significantly more variables, as expected.

The reduction in LP size, moving from LP (reduced) to LP (hybrid), is relatively small. This is not surprising, because LP (reduced) is already quite small. As the mass of the digits is densest in the middle of the grids, where there are also the highest numbers of combinations that produce the same support points, $w$ variables are only being introduced for a few of the extreme values of the support grid, as anticipated in view of Figure 4.

The LP solution times do improve significantly in each of the smaller formulations. The solving times for LP (reduced) are often about half of the solving time for the original formulation. Despite not being significantly smaller in size, the improvement in solution times continues, though less dramatically, moving to LP (hybrid).

### 4.3 Digit Computations: Using A Priori Knowledge of Grid

We now turn to the versions of the LPs that make use of the a priori knowledge of the support structure. We recall that the primary goal in this approach is to avoid exponential setup costs by not operating on $S^*$. LP (general) is not competitive with the other three LPs in this case as it always produces an exponential number of variables. While the a priori knowledge produced in Section 4.1 can be used to implement a hybrid approach, we only report on LP (original) and LP (reduced). Table 5 lists the problem sizes; Table 6 includes the solving times as before.

In Table 5, we first note that, as expected, the LP sizes are larger than those given in Table 3. This is because this implementation allows for slightly larger LPs for a significant tradeoff in shorter setup time. For a fair comparison of the two approaches, we have to include the setup time. To this end, we show the total running time, including setup and solution of the LP, for LP (original) without a priori knowledge, and LP (original) and LP (reduced) with a priori knowledge, in Table 7.

The reduced formulation, LP (reduced), typically yields about a 13% reduction in number of variables as seen in Table 5. Yet, we see a significant improvement in both solution times and total running times, as displayed in Tables 6 and 7.

17
| Digit | LP (original) | LP (reduced) |
|-------|---------------|---------------|
|       | Rows          | Columns       | Rows          | Columns       |
| 0     | 15,406        | 1.946,083     | 15,406        | 1.692,982     |
| 1     | 3,371         | 158,600       | 3,371         | 126,409       |
| 3     | 15,343        | 1.711,660     | 15,343        | 1,494,751     |
| 7     | 13,268        | 1,089,521     | 13,268        | 942,719       |
| 9     | 15,323        | 1,637,240     | 15,323        | 1,420,895     |

Table 5. Number of constraints and variables for four measures of each digit using a priori knowledge.

| Digit | LP (original) | LP (reduced) |
|-------|---------------|---------------|
|       |               |               |
| 0     | 414.82        | 216.46        |
| 1     | 3.34          | 3.25          |
| 3     | 272.33        | 178.08        |
| 7     | 222.82        | 114.61        |
| 9     | 355.24        | 159.90        |

Table 6. Solving times, in seconds, for four measures of each digit using a priori knowledge.

| Digit | LP (original) | LP (original) | LP (reduced) |
|-------|---------------|---------------|---------------|
|       | No a priori   | With a priori | With a priori |
| 0     | 763.41        | 425.34        | 225.57        |
| 1     | 7.28          | 3.51          | 3.36          |
| 3     | 608.19        | 278.93        | 183.74        |
| 7     | 123.48        | 225.42        | 116.79        |
| 9     | 313.26        | 360.81        | 164.66        |

Table 7. Total running times, in seconds, for four measures of each digit, including setup times.
4.4 Computation: Strength of Hybrid Formulation

For our final computational experiment, we exhibit an example where the hybrid approach performs significantly better than the other three formulations. In this example, we have 5 measures of equal support size 28; that is, \( n = 5 \) and \(|P_i| = 28\). Each measure is supported on a \( 5 \times 5 \) grid with the same coordinates in all five measures, and an additional three points which have different coordinates in each measure. Figure 5 exhibits an example. Informally, these measures combine features of the previous examples: a part of the support is highly structured (the \( 5 \times 5 \) grid), and the other support points lie in general position. In all measures, the three points are sufficiently far from the grid to avoid generating any duplicates in \( S \) inside the grid boundaries. Masses for each measure were randomly generated and differ in distribution among the possible support points in each measure. LP solution times for this example are shown in Table 8.

![Fig. 5. A sample support set of a measure where the hybrid approach is preferable.](image)

| LP (original) | LP (reduced) | LP (general) | LP (hybrid) |
|--------------|-------------|--------------|-------------|
| 43,143       | 1222.74     | 880.78       | 386.73      |

Table 8. LP solving times, in seconds, for \( n=5 \) where each measure has 28 support points including a \( 5 \times 5 \) regular grid.

In this example, all the alternative formulations presented in this paper are preferable to the original formulation by a significant margin. However, LP (hybrid) vastly outperforms all the others. This is because of the mixed structure of the support of the measures. Due to its ability to choose \( y, z \) variables for the grid part of the support set and \( w \) variables for the combinations involving the support points in general points, the hybrid approach finds a smallest set of variables to represent the problem. In contrast, LP (reduced) is restricted to \( y, z \) variables, which is not an optimal approach for data in general position, and LP (general) is restricted to \( w \) variables, which is not an optimal approach for the grid data.
Acknowledgments

We thank Natalia Villegas Franco for running and visualizing the computations on the Denver crime data in Section 3.2; see Figure 1. Borgwardt gratefully acknowledges support through Grant 524210 Polyhedral Theory in Data Analytics, Collaboration Grant for Mathematicians, of the Simons Foundation.

References

1. M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43 (2):904–924, 2011.
2. E. Anderes, S. Borgwardt, and J. Miller. Discrete Wasserstein Barycenters: Optimal Transport for Discrete Data. *Mathematical Methods of Operations Research*, 84 (2):389–409, 2016.
3. M. Beiglböck, P. Henry-Labordere, and F. Penkner. Model-independent bounds for option prices – a mass transport approach. *Finance and Stochastics*, 17 (3):477–501, 2013.
4. J. Bigot and T. Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *eprint arXiv:1212.2562*.
5. E. Boissard, T. Le Gouic, and J.-M. Loubes. Distribution’s template estimate with Wasserstein metrics. *Bernoulli*, 21 (2):740–759, 2015.
6. S. Borgwardt. Strongly Polynomial 2-Approximations for Discrete Wasserstein Barycenters. *eprint arXiv:1704.05491*, 2017.
7. G. Carlier and I. Ekeland. Matching for teams. *Economic Theory*, 42 (2):397–418, 2010.
8. G. Carlier, A. Oberman, and E. Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modeling and Numerical Analysis*, 49 (6):1621–1642, 2015.
9. P-A. Chiaporri, R. McCann, and L. Nesheim. Hedonic price equilibria, stable matching and optimal transport; equivalence, topology and uniqueness. *Economic Theory*, 42 (2):317–354, 2010.
10. C. Cotar, G. Friesecke, and C. Klüppelberg. Density functional theory and optimal transportation with coulomb cost. *Communications on Pure and Applied Mathematics*, 66 (4):548–599, 2013.
11. M. Cuturi and A. Doucet. Fast Computation of Wasserstein Barycenters. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 685–693. JMLR Workshop and Conference Proceedings, 2014.
12. E. del Barrio, J.A. Cuesta- Albertos, C. Matrán, and A. Mayo-Iscar. Robust clustering tools based on optimal transportation. *Statistics and Computing*, https://doi.org/10.1007/s11222-018-9800-z, 2018.
13. A. Galichon, P. Henry-Labordère, and N. Touzi. A stochastic control approach to non-arbitrage bounds given marginals, with an application to lookback options. *Annals of Applied Probability*, 24 (1):312–336, 2014.
14. P. Kleinenschmidt and H. Schumann. A strongly polynomial algorithm for the transportation problem. *Mathematical Programming*, 68(1-3):1–13, 1995.
15. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
16. Y. Mileyko, S. Mukherjee, and J. Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12), 2011.
17. J. Miller. Transportation Networks and Matroids: Algorithms through Circuits and Polyhedrality, 2016. Ph.D. thesis, University of California Davis.
18. E. Munch, K. Turner, P. Bendich, S. Mukherjee, J. Mattingly, and J. Harer. Probabilistic Fréchet means for time varying persistence diagrams. *Electronic Journal of Statistics*, 9:1173–1204, 2015.
19. B. Pass. On the local structure of optimal measures in the multi-marginal optimal transportation problem. *Calculus of Variations and Partial Differential Equations*, 43 (3-4):529–536, 2011.

20. B. Pass. Uniqueness and Monge Solutions in the Multimarginal Optimal Transportation Problem. *SIAM Journal on Mathematical Analysis*, 43 (6):2758–2775, 2011.

21. B. Pass. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis*, 264 (4):947–963, 2013.

22. B. Pass. Multi-marginal optimal transport and multi-agent matching problems: Uniqueness and structure of solutions. *Discrete and Continuous Dynamical Systems A*, 34 (4):1623–1639, 2014.

23. J. Rabin, G. Peyre, J. Delon, and M. Bernot. Wasserstein Barycenter and its Application to Texture Mixing. *Scale Space and Variational Methods in Computer Vision. Lecture Notes in Computer Science*, 6667:435–446, 2012.

24. E. Tardos. A strongly polynomial algorithm to solve combinatorial linear programs. *Operations Research*, 34(2):250–256, 1986.

25. A. Trouvé and L. Younes. Local Geometry of Deformable Templates. *SIAM Journal on Mathematical Analysis*, 37 (1):17–59, 2005.

26. K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. Fréchet Means for Distributions of Persistence Diagrams. *Discrete and Computational Geometry*, 52(1):44–70, 2014.

27. C. Villani. *Topics in Optimal Transportation*, volume 58. 2003.

28. C. Villani. *Optimal transport: old and new*, volume 338. 2009.

29. J. Ye, P. Wu, J. Z. Wang, and J. Li. Fast Discrete Distribution Clustering Using Wasserstein Barycenter With Sparse Support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, 2017.

30. Y. Zemel and V. Panaretos. Fréchet Means and Procrustes Analysis in Wasserstein Space. *eprint arXiv:1701.06876*, 2017.