Online Homotopy Algorithm for a Generalization of the LASSO

A. Hofleitner, T. Rabbani, L. El Ghaoui, and A. Bayen

***Abstract*—The LASSO is a widely used shrinkage method for linear regression. We propose an online homotopy algorithm to solve a generalization of the LASSO in which the l_1 regularization is applied on a linear transformation of the solution, allowing to input prior information on the structure of the problem and to improve interpretability of the results. The algorithm takes advantage of the sparsity of the solution for computational efficiency and is promising for mining large datasets.**

***Index Terms*— LASSO.**

I. Introduction and Related Work

Least-Squares regression with l_1-norm regularization is known as the LASSO algorithm [1]. It has generated significant interest in the statistics [1], [2], signal processing [3]–[5] and machine learning [6], [7] communities, in particular for estimation problems. Adding a l_1-penalty usually leads to sparse solutions, which is a desirable property used to achieve model selection, data compression, or to obtain interpretable results. The LASSO can be solved using interior-point methods [8], iterative thresholding algorithms [9], [10], feature-sign search [11], bound optimization methods [12], incremental methods [13] or gradient projection algorithms [14]. Homotopy algorithms compute the regularization path [15], [16] or perform online updates [17]–[19]. They are particularly efficient when the solution is very sparse [20], [21].

The article extends the results of [18], [19] with the following contributions: (i) the algorithm updates the solution as a new batch of p observations is received (previous work only considered updates with one measurement at a time), (ii) the online algorithm solves the LASSO when an affine transformation of the estimate is sparse.

Problem Statement: At estimation step n, we are given a set I_n of training examples or observations $(y_i, a_i) \in \mathbb{R} \times \mathbb{R}^m, i \in I_n$. We wish to fit a linear model to estimate the response y_i as a function of $x \in \mathbb{R}^m$. A linear function of the solution, $K_1 x$, with $K_1 \in \mathbb{R}^{k \times m}$ is expected to be sparse, representing inherent structure of the problem or trend filtering [22], [23]. To achieve this property, we add an l_1 penalty on $K_1 x$ and solve the following optimization problem:

$$\min_{x \in \mathbb{R}^m} \frac{1}{2} \sum_{i=1}^{n} \left(a_i^T x - y_i \right)^2 + \mu_n \| K_1 x \|_1. \tag{1}$$

Manuscript received April 12, 2012; revised January 29, 2013 and April 16, 2013; accepted April 19, 2013. Date of publication April 22, 2013; date of current version November 18, 2013. The authors are supported in part by the National Science Foundation, via Grants SES-0835550 and CMMI-0966842. Recommended by Associate Editor A. Chiuso.

A. Hofleitner is with the Electrical Engineering and Computer Science, UC Berkeley, CA (e-mail: aude@eecs.berkeley.edu).

T. Rabbani is with the Mechanical Engineering, UC Berkeley, CA (e-mail: trabbani@berkeley.edu).

L. El Ghaoui is with the Electrical Engineering and Computer Science, Industrial Engineering and Operations Research, UC Berkeley, CA (e-mail: elghaoui@berkeley.edu).

A. Bayen is with the Electrical Engineering and Computer Science, Civil and Environmental Engineering, UC Berkeley, CA (e-mail: bayen@berkeley.edu).

Color versions of one or more of the figures in this technical note are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2013.2259373

The regularization parameter μ_n may depend on the number of measurements $|I_n|$. For example, we can choose $\mu_n = |I_n|\mu_0$ as in [18] or $\mu_n = \sqrt{|I_n|}\mu_0$ as in [24]. Both the dependency of μ_n on $|I_n|$ and the parameter μ_0 are chosen using a validation metric, computed on a dataset which is not used to train the model. The validation procedure is described in more details in the result section (Section IV). It chooses a trade-off between the structure imposed by the regularization, and the fit to the data. After receiving p new observations, the algorithm updates the solution of (1) without having to *completely* re-solve the problem. As done in the Elastic Net [25], we investigate the addition of an l_2 regularization term to (1) to improve estimation capabilities by leveraging prior information $\hat{x}$ on the value of the solution.

Organization of the Article: Section II reviews an existing homotopy algorithm which solves the LASSO recursively [18], [19]. Section III presents the extension of the algorithm to update the solution with p observations and a l_1 penalization on a linear function of the solution. We apply the algorithm on a traffic estimation problem from streaming probe data in Section IV and discuss possible extensions of this work in Section V.

II. The LASSO Problem

The LASSO problem [1] is defined as follows:

$$x = \arg\min_{x \in \mathbb{R}^m} \frac{1}{2} \sum_{i=1}^{n} \left(a_i^T x - y_i \right)^2 + \mu_n \| x \|_1. \tag{2}$$

There is a global minimum at x if and only if the sub differential of the objective function at x contains the 0-vector. The sub differential of the l_1-norm at x is the following set:

$$\partial \|x\|_1 = \left\{ v \in \mathbb{R}^m : \begin{cases} v_i = \operatorname{sgn}(x_i) & \text{if} \quad |x_i| > 0 \\ v_i \in [-1, 1] & \text{if} \quad x_i = 0 \end{cases} \right\}$$

where $\operatorname{sgn}(\cdot)$ is the sign function. Let $A \in \mathbb{R}^{|I_n| \times m}$ be the matrix whose i^{th} row is equal to a_i^T, and let $y = (y_i)_{i \in I_n}^T$ be the vector of response variables. The optimality conditions for (2) are given by $A^T(Ax - y) + \mu_n v = 0, v \in \partial \|x\|_1$.

We define the *active set* a (resp. *non active set* na) as the set of indices representing non-zero (resp. zero) elements of x. The matrix A_a (resp. A_{na}) is a selection of the columns of A in a (resp. in na). The non-zero coordinates of x are in x_a and x_{na} is the 0-vector. The index a_i (resp. na_i) references the i^{th} coordinate of the active (resp. non active) set. Since $v \in \partial \|x\|_1$, $v_{a_i} = \operatorname{sgn}(x_{a_i})$ and $v_{na_i} \in [-1, 1]$. If the solution is unique, $A_a^T A_a$ is non-singular;[1] we rewrite the optimality conditions as $x_a = (A_a^T A_a)^{-1}(A_a^T y - \mu_n v_a)$ and $-\mu_n v_{na} = A_{na}^T (A_a x_a - y)$. If we know the active set and the signs of the coefficients of the solution, thus the vector v_a, we can compute the solution x in closed form. When observations come sequentially, a homotopy algorithm [18], [19] solves the LASSO problem recursively by considering the following problem:

$$x(t, \mu) = \arg\min_{x \in \mathbb{R}^m} \frac{1}{2} \left\| \begin{pmatrix} A \\ t a_{n+1}^T \end{pmatrix} x - \begin{pmatrix} y \\ t y_{n+1} \end{pmatrix} \right\|_2^2 + \mu \| x \|_1.$$

Adding (resp. removing) a point is equivalent to computing the path from $t = 0$ to $t = 1$ (resp. from $t = 1$ to $t = 0$). Varying the regularization parameter is equivalent to computing the path from $\mu = \mu_n$ to $\mu = \mu_{n+1}$.

[1]The Elastic Net [25] ensures the uniqueness of the solution without requiring $A^T A$ to be non-singular.

III. Recursive Lasso With p New Observations, l_2 and Linear l_1 Regularizations

We consider a least square estimation problem, for which we want a linear transform of the solution, $K_1 x$ for $K_1 \in \mathbb{R}^{k\times m}$, to be sparse. We update the estimation as we receive p new observations $(y^{\text{new}}, A^{\text{new}}) \in \mathbb{R}^p \times \mathbb{R}^{p\times m}$. We assume a priori information (e.g. historical estimate) $\hat{x} \in \mathbb{R}^m$ on the solution, which is used as additional regularization when the matrix A is not full column rank or is ill conditioned (see [25] for details). We assume that the matrix K_1 is full row rank, which is the case for numerous applications including total variation regularization. Each row of K_1 corresponds to an information on the sparsity structure of the solution. We define $K_2 \in \mathbb{R}^{m-k\times m}$ such that $K = (K_1^T K_2^T)^T$ is non singular. For example, K_2 is such that the columns of K_2^T form a basis for the null-space of K_1. We define a change of variable $z = Kx$, new data matrices $B = AK^{-1}$, $B_{\text{new}} = A_{\text{new}}K^{-1}$ and $\hat{z} = K\hat{x}$. We propose an algorithm that updates the solution z of

$$\min_{z\in\mathbb{R}^m} \frac{1}{2}\left\|\begin{pmatrix} B \\ tB^{\text{new}}\end{pmatrix} z - \begin{pmatrix} y \\ ty^{\text{new}}\end{pmatrix}\right\|_2^2 + \mu\left\|(I_k 0_{k\times m-k})z\right\|_1 + \frac{\lambda}{2}\|z-\hat{z}\|_2^2 \quad (3)$$

as we (i) vary t to add or remove observations and (ii) vary μ to change the weight of the l_1 regularization. The l_1 penalization is on the first k coordinates of z, denoted *regularized indices*. The last $m-k$ indices are in the active set and are referred to as the *non-regularized indices*.

A. Add p Observations

At $t = 0$, we know the solution $z(0, \mu_n)$ and thus the active set and signs of the regularized indices of z. Let v_{a_i} be the sign of $z_{a_i}(0)$ for the regularized indices and define $v_{a_i} = 0$ for the non-regularized indices. The data matrices with the new observations are indicated with a tilde: $\tilde{B} = (B^T B^{\text{new}^T})^T$ and $\tilde{y} = (y^T y^{\text{new}^T})^T$. The optimality conditions of (3) read

$$\tilde{B}_a^T\left(\tilde{B}_a z_a(t) - \tilde{y}\right) + (t^2-1)B_a^{\text{new}^T}\left(B_a^{\text{new}} z_a(t) - y^{\text{new}}\right) + \mu_n v_a + \lambda\left(z_a(t) - \hat{z}_a\right) = 0 \quad (4)$$

$$\tilde{B}_{na}^T\left(\tilde{B}_a z_a(t) - \tilde{y}\right) + (t^2-1)B_{na}^{\text{new}^T}\left(B_a^{\text{new}} z_a(t) - y^{\text{new}}\right) + \mu_n w_{na}(t) - \lambda\hat{z}_{na} = 0 \quad (5)$$

where $w_{na}(t)$ is a vector with coordinates in $[-1,1]$. We notice that, at $t = 0$, $z_a(\cdot)$ and $w_{na}(\cdot)$ are continuous in t. Let t^* to be the largest $t \in [0,1]$ such that: (i) for all $t \in [0, t^*)$, for all i in the regularized indices, $\text{sgn}(z_a(t)) = \text{sgn}(z_a(0))$ and (ii) for all $t \in [0, t^*)$, for all i in the non-active set, $|w_{na_i}(t)| < 1$. On this interval, v_{a_i} is the sign of $z_{a_i}(t)$ and Equations (4)–(5) are valid.

We compute $Q = (\tilde{B}_a^T\tilde{B}_a + \lambda I_{|a|})^{-1}$ from its previous value without the p new observations using the Woodbury matrix identity (p rank update). We define $\tilde{z}_a = Q(\tilde{B}_a^T\tilde{y} + \lambda\hat{z}_a - \mu v_a)$ and $\alpha = t^2 - 1$. We consider the singular value decomposition of $B_a^{\text{new}}QB_a^{\text{new}^T} = \Gamma^T\Sigma\Gamma$ and define the rotated data $\bar{B}^{\text{new}} = \Gamma B^{\text{new}}$ and $\bar{y}^{\text{new}} = \Gamma y^{\text{new}}$,, the rotated error $\overline{E} = \overline{B}_a^{\text{new}}\tilde{z}_a - \bar{y}^{\text{new}}$ and $U = Q\overline{B}_a^{\text{new}^T}$.

Proposition 1 (Solution Path as We Add p Observations): For $t \in [0, t^*)$, $z_a(\cdot)$ is continuous in t and given by

$$z_a(t) = \tilde{z}_a - (t^2-1)U\left(I + (t^2-1)\Sigma\right)^{-1}\overline{E}. \quad (6)$$

Let t^0 be the smallest[2] $t \in [0,1]$ such that a coordinate of $z_a(t)$ equals zero, t^+ (resp. t^-) the smallest[2] $t \in [0,1]$ which sets a coordinate of $w_{na}(t)$ to 1 (resp. to -1). The transition point t^* is defined as $t^* = \min(t^0, t^+, t^-)$ and can be computed by solving p-degree polynomial equations on a bounded interval.

Proof: For $t \in [0, t^*)$, we use (4) and the Woodbury matrix identity to write $(Q^{-1} + \alpha B_a^{\text{new}^T}B_a^{\text{new}})^{-1}$ as $Q - \alpha U(I + \alpha\Sigma)^{-1}\Gamma B_a^{\text{new}}Q$. It follows that:

$$\begin{aligned} z_a(t) &= \tilde{z}_a - \alpha U(I+\alpha\Sigma)^{-1}\bar{B}_a^{\text{new}}\tilde{z}_a \\ &\quad + \alpha\left(Q - \alpha U(I+\alpha\Sigma)^{-1}\Gamma B_a^{\text{new}}Q\right)B_a^{\text{new}^T}y^{\text{new}} \\ z_a(t) &= \tilde{z}_a - \alpha U(I+\alpha\Sigma)^{-1}\bar{B}_a^{\text{new}}\tilde{z}_a \\ &\quad + \alpha\left(U\bar{y}^{\text{new}} - \alpha U(I+\alpha\Sigma)^{-1}\Sigma\bar{y}^{\text{new}}\right) \\ z_a(t) &= \tilde{z}_a - \alpha U(I+\alpha\Sigma)^{-1}\bar{B}_a^{\text{new}}\tilde{z}_a + \alpha U(I+\alpha\Sigma)^{-1}\bar{y}^{\text{new}} \end{aligned}$$

which proves (6). The computation of t^0, t^+ and t^- is given by Lemma 1 and 2. ∎

We denote by $U_{i,j}$ the element of U on line i and column j and by U_i the i^{th} line of U, σ_i is the i^{th} singular value in Σ and $\overline{E}_i$ is the i^{th} coordinate of $\overline{E}$.

Lemma 1 (Computation of t^0): Let $t^0_{a_i}$ be the smallest value of $t \in [0,1]$ which sets the i^{th} coordinate of z_a (in the regularized indices) to zero. It is given by $t^0_{a_i} = \sqrt{\alpha^0_{a_i} + 1}$ where $\alpha^0_{a_i}$ is the smallest real valued solution in the interval $[-1, 0]$ of the p degree polynomial equation in α, which can be solved numerically

$$0 = \tilde{z}_{a_i}\prod_{l=1}^{p}(1+\alpha\sigma_l) - \alpha\sum_{j=1}^{p}U_{i,j}\bar{E}_j\prod_{l\neq j}(1+\alpha\sigma_l)$$

If the polynomial equation does not have real valued solutions in $[-1, 0]$, we set $t^0_{a_i} = 1$. It follows that t^0 is the smallest value of $t^0_{a_i}$ in the interval [0, 1].

Proof: Setting the i^{th} coordinate of z_a to zero in (6), we have

$$\begin{aligned} 0 &= \tilde{z}_{a_i} - \alpha U_i(I+\alpha\Sigma)^{-1}\overline{E} \\ 0 &= \tilde{z}_{a_i} - \alpha\sum_{j=1}^{p}\frac{U_{i,j}\bar{E}_j}{1+\alpha\sigma_j} \\ 0 &= \tilde{z}_{a_i}\prod_{l=1}^{p}(1+\alpha\sigma_l) - \alpha\sum_{j=1}^{p}U_{i,j}\bar{E}_j\prod_{l\neq j}(1+\alpha\sigma_l) \end{aligned}$$

∎

We denote by c_i the i^{th} column of $\tilde{B}_{na}$, d_i is the i^{th} row of $\bar{B}_{na}^{\text{new}}$ and $d_{i,j}$ the element of $\bar{B}_{na}^{\text{new}}$ on the i^{th} row and j^{th} column. We also denote by f_i the i^{th} element of $\tilde{B}_{na}^T\tilde{e} - \lambda\hat{z}_{na}$ and $\tilde{e} = \tilde{B}_a\tilde{z}_a - \tilde{y}$.

Lemma 2 (Computation of t^+ and t^-): The smallest value of t that sets the i^{th} coordinate of w_{na} to 1 (resp. to -1) is denoted $t^+_{na_i}$ (resp. $t^-_{na_i}$). It is given by $t^+_{na_i} = \sqrt{\alpha^+_{na_i} + 1}$ (resp. $t^-_{na_i} = \sqrt{\alpha^-_{na_i} + 1}$) where $\alpha^+_{na_i}$ (resp. $\alpha^+_{na_i}$) is the smallest real valued solution in the interval $[-1, 0]$ of the p degree polynomial equation in α^+ (resp. in α^-)

$$\begin{aligned} &(-\mu - f_i)\prod_{l=1}^{p}\left(1+\alpha^+\sigma_l\right) \\ &\quad = \alpha^+\sum_{j=1}^{p}\overline{E}_j\left(d_{i,j} - c_i^T\tilde{B}_aU_j\right)\prod_{l\neq j}\left(1+\alpha^+\sigma_l\right) \\ &(\mu - f_i)\prod_{l=1}^{p}\left(1+\alpha^-\sigma_l\right) \\ &\quad = \alpha^-\sum_{j=1}^{p}\overline{E}_j\left(d_{i,j} - c_i^T\tilde{B}_aU_j\right)\prod_{l\neq j}\left(1+\alpha^-\sigma_l\right). \end{aligned}$$

If the polynomial equation does not have real valued solutions in $[-1, 0]$, we set $t^+_{na_i} = 1$ (resp. $t^-_{na_i} = 1$). It follows that t^+ (resp. t^-) is the smallest value of $t^+_{na_i}$ (resp. $t^-_{na_i}$) in the interval [0, 1].

[2] If no such t exists, we set t^0 (resp. t^+ and t^-) to 1.

Proof: From (6), we notice that

$$\begin{aligned} B_a^{\text{new}} z_a(t) - y^{\text{new}} &= B_a^{\text{new}} \tilde{z}_a - \alpha \Gamma^T (I + \alpha\Sigma)^{-1} \overline{E} - y^{\text{new}} \\ &= \Gamma^T \overline{E} - \alpha \Gamma^T \Sigma (I + \alpha\Sigma)^{-1} \overline{E} \\ &= \Gamma^T (I + \alpha\Sigma)^{-1} \overline{E}. \end{aligned}$$

We also have $\tilde{B}_a z_a(t) - \tilde{y} = \tilde{e} - \alpha \tilde{B}_a U (I + \alpha\Sigma)^{-1} \overline{E}$ We rewrite (5) as

$$\begin{aligned} 0 =& \tilde{B}_{na}^T \left(\tilde{e} - \alpha \tilde{B}_a U (I + \alpha\Sigma)^{-1} \overline{E} \right) + \mu w_{na}^T - \lambda \hat{z}_{na} \\ &+ \alpha B_{na}^{\text{new}^T} \Gamma^T (I + \alpha\Sigma)^{-1} \overline{E} \\ -\mu w_{na}(t) =& \tilde{B}_{na}^T \tilde{e} - \lambda \hat{z}_{na} \\ &+ \alpha \left(\bar{B}_{na}^{\text{new}T} - \tilde{B}_{na}^T \tilde{B}_a U \right) (I + \alpha\Sigma)^{-1} \overline{E}. \end{aligned}$$

We obtain the values of $t_{na_i}^+$ (resp. $t_{na_i}^-$) by solving the p degree polynomial equation in α^+ (resp. α^-) on the interval $[-1, 0]$

$$\begin{aligned} &(-\mu - f_i) \prod_{l=1}^{p} (1 + \alpha^+ \sigma_l) \\ &\quad = \alpha^+ \sum_{j=1}^{p} \overline{E}_j \left(d_{i,j} - c_i^T \tilde{B}_a U_j \right) \prod_{l \neq j} \left(1 + \alpha^+ \sigma_l \right) \\ &(\mu - f_i) \prod_{l=1}^{p} \left(1 + \alpha^- \sigma_l \right) \\ &\quad = \alpha^- \sum_{j=1}^{p} \overline{E}_j \left(d_{i,j} - c_i^T \tilde{B}_a U_j \right) \prod_{l \neq j} \left(1 + \alpha^- \sigma_l \right) \end{aligned}$$

■

Lemma 3 (Update of the Active Set): When we reach a transition point, the active set and signs of the regularized indices are updated as follows: (i) if $t^* = t^0$, we remove the corresponding coordinate from the active set, (ii) if $t^* = t^+$ (resp. $t^* = t^-$), we add the coordinate to the active set and set its sign to positive (resp. to negative).

Proof: If $t^* = t^0$, let a_i be such that $z_{a_i}(t^*) = 0$. The subgradient of $\|(I_k 0_{k \times m-k}) z\|_1$ with respect to the coordinate a_i is in the interval $[-1, 1]$, hence we remove the coordinate from the active set.

If $t^* = t^+$, let na_i be such that $w_{na_i}(t^*) = 1$. For $t > t^*$, the optimality condition for the coordinate na_i cannot be satisfied with the current active set because w_{na_i} is bounded by 1. If we let the coordinate na_i of the solution take non-zero values, we can rewrite the optimality condition as $f(z_a(t)) + \beta z_{na_i}(t) = 0$, where $f(z_a(t)) < 0$ and β is a positive term which depends on the norm of the column na_i of B and B^{new} and on λ. This proves that z_{na_i} takes positive value and adding the index to the active set provides a solution, thus *the* solution (strict concavity). ■

Algorithm 1 updates the solution when t varies from $t = 0$ to $t = 1$. The same algorithm is relevant to remove p observations by finding the transition points as t decreases from 1 to 0.

Algorithm 1 Update of the solution as we add p observations

Initialize the active set a, non active set na and signs of the regularized indices v_a.

$t = 0$

while $t < 1$ **do**

Compute t^0, t^+ and t^- as the smallest value of $t_{a,i}^0$, $t_{na,i}^+$ and $t_{na,i}^-$ in $(t, 1]$ (Lemma 1–2).

$t = \min(t^0, t^+, t^-)$

if $t > 1$ **then**

break;

else

Update the active set and sign of the regularized indices according to the transition point.

end if

Update the matrix Q to account for the updated active set (rank 1 update).

end while

B. Update the Regularization Parameter

The computation of the regularization path is detailed in [16] and in [25] for the Elastic Net. To solve the problem of interest 3, it is necessary to define the *non-regularized indices*, as done in Section III-A and set $v_{a_i} = 0$ for these indices. With this convention, we refer the reader to [16] and [25] to compute the solution of (3).

Remark 1 (Leveraging the Sparsity Structure): We efficiently update Q when the active or non active set change or when we add/remove observations with low rank updates. We actually update the Cholesky factorization of Q which provides better numerical stability to the algorithm than updating Q directly [26].

Remark 2 (Complexity): The complexity of the algorithm depends on the number of transitions and the size of the active set. The theoretical bound on the number of transitions is 3^k, where k is the number of rows of K_1. In practice, it is much smaller because successive estimates are expected to have a similar support.

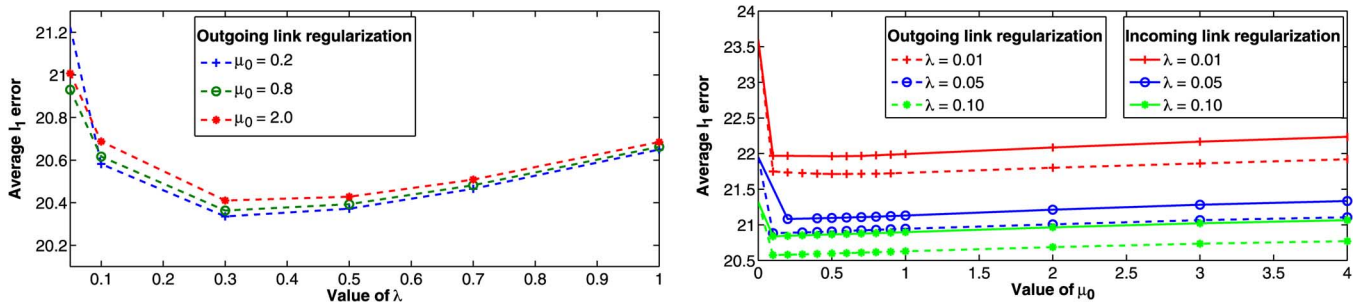
IV. Large Scale Traffic Estimation on an Arterial Network

A. Experimental Setting

We apply the algorithm to arterial traffic estimation on a sub network of San Francisco, CA totalling more than 800 links (12.6 kilometers of roadway). Dedicated sensing infrastructure is rarely available on arterial networks and we use data collected by the *Mobile Millennium* system [27] from a fleet of 500 vehicles which report their location every minute. The duration between two successive location reports ξ_1 and ξ_2 is an observation of the travel time y_i on the path from ξ_1 to ξ_2. We use a conditional random field algorithm [28] to reconstruct the trajectory between ξ_1 and ξ_2. Each trajectory (path) is converted in a vector $a_i \in [0, 1]^m$, where m is the number of links in the network. The j^{th} coordinate of a_i, denoted $a_{i,j}$, is the fraction of the link traveled by the probe vehicle. It is computed as the distance traveled on the link divided by the length of the link.[3] In particular, $a_{i,j} = 0$ if the vehicle did not travel on link j and $a_{i,j} = 1$ if the vehicle fully traversed link j.

The solution x^n represents the average travel time on each link of the network at time t^n. We add an l_1 regularization on the spatial variations of the travel times for several reasons. First, it improves estimation capabilities by exploiting a-priori information on the structure of the solution. Traffic signals cause important variation on the travel time experienced on a link and regularization is important to prevent overfitting. Second, it exhibits the inherent spatial structure of traffic by noticing the area where traffic conditions actually change. Finally, the sparse structure of the solution enables an efficient update of the estimate as new measurements are received.

[3]The coefficients $a_{i,j}$ can account for the fact that travel time on a fraction of the link does not vary proportionally with the distance traveled as vehicles are more likely to experience delays close to signalized intersections [29].



(a) (b)

Fig. 1. Variation of the l_1 error in function of the parameters (a) λ and (b) μ_0. The figures indicate the importance of the additional l_2 regularization to improve the accuracy of the estimation.

Fig. 2. Geographical representation of the traffic estimation results. The color of each link varies with the pace (green for small paces, red for large paces). The pins indicate the intersections for which not all outgoing links have the same estimated pace (inverse of the speed).

We estimate the average travel times $x = K^{-1}z$ by solving (3). We use historical mean travel times for the l_2 regularization $\hat{x}$. For each new travel time observation, we increase the regularization parameter from $|I_n|\mu_0$ to $|I_{n+1}|\mu_0$ and add the new observation $(y_{n+1}, a_{n+1}) \in \mathbb{R} \times \mathbb{R}^m$. The parameters of the l_1 and l_2 regularizations (respectively μ_0 and λ) are chosen via cross-validation as described in the following Section. Observations remain relevant only for a *limited period of time* T.[4] When observations become obsolete, the algorithm updates the regularization parameter and removes the old observations (decrease t from 1 to 0).

We investigate two potential choices for the full row rank matrix K_1, which represents the prior information on the spatial structure of the estimate. The first one encourages all the incoming links of an intersection to have the same pace (inverse of velocity), the second one encourages the outgoing links to have the same pace. To each junctions j with n_j incoming links (resp. outgoing links), corresponds $n_j - 1$ rows in K_1. The k^{th} such row encourages the k and $k+1$ incoming (resp. outgoing) links of the junction to have the same pace (travel time on the link divided by the length of link).

[4]Typically, T is in the order of five to fifteen minutes

B. Validation Framework

At time t_n we compute the estimate x_n corresponding to the observations in I_n. We compute the prediction error $e_n = |a_{n+1}x_n - y_{n+1}|$ and analyze the effect of the choice of the parameters λ and μ_0 as well as the choice of matrix K_1 in Fig. 1. The numerical results indicate that both the l_1 and l_2 regularization improve the results for a wide range of λ and μ_0. As the error is not very sensitive to the choice of these parameters, they can be calibrated off-line using cross-validation. Fig. 1 (right) also indicate that the choice of the regularization matrix K_1 influences the accuracy. The regularization on the outgoing links always provide better results than the choice of regularization on the incoming links. The algorithm does not provide an automated way to choose the optimal matrix K_1 even though this is part of current research interest.

The results can also be represented as a traffic map (see Fig. 2) with colors representing the pace of the vehicles: green for smallest pace, i.e., fastest speed, red for largest pace. We indicate the intersections for which we detect spatial variation of the pace by a pin. The pins tend to cluster in a few regions of the network, indicating regions with important spatial variations in the traffic conditions.

Imposing and exploiting a sparsity structure on the solution limits the computational cost of traffic estimation on large networks as the algorithm leverages the sparsity of the solution in the algorithm. The number of transition points and active indices remain small throughout the algorithm with an average of 0.5 transition points per estimate update (addition of new data points and variation of the regularization parameter) and 20 active regularized indices for a network with 815 links.

V. Conclusion

The article presents an online-algorithm to update the solution of linear regression problems with a large class of l_1 and l_2 regularizations. The l_1-norm improves the accuracy and computational efficiency of the estimation as well as the interpretability of the results by exhibiting and exploiting the underlying sparsity structure of the problem. The l_2-norm increases the robustness of the estimator and limits numerical issues. The algorithm provides the ability to i) impose sparsity on a linear function of the state, ii) update the solution online by computing a homotopy as new measurements are available (or old ones become obsolete). The potential of the algorithm is demonstrated on real-time traffic estimation problem from streaming probe vehicle data in large urban networks. It provides accurate estimation capabilities and a better understanding of the spatial variations of traffic across the network.

ACKNOWLEDGMENT

The authors would like to thank T. Hunter for providing the filtered probe trajectories from the raw measurements of the probe vehicles used to produce the numerical experiment.

REFERENCES

[1] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
[2] Y. Dodge, *Statistical Data Analysis Based on the l_1-Norm and Related Methods*. : Birkhauser, 2002.
[3] R. Baraniuk, "Compressive sensing," *IEEE Signal. Proc. Mag.*, vol. 24, no. 4, p. 118, 2007.
[4] E. Candès, "Compressive sampling," *Congress of Mathematicians*, vol. 3, pp. 1433–1452, 2006.
[5] J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Information Theory*, vol. 50, no. 6, pp. 1341–1344, 2004.
[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
[7] A. Ng, "Feature selection, l_1 vs. l_2 regularization, rotational invariance," in *21st International Conference on Machine Learning*, 2004, p. 78, ACM.
[8] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l_1-regularized least squares," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2008.
[9] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.
[10] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.
[11] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," *Advances in Neural Information Processing Systems*, vol. 19, p. 801, 2007.
[12] M. Figueiredo and R. Nowak, "A bound optimization approach to wavelet-based image deconvolution," in *International Conference on Image Processing*, 2005, vol. 2, IEEE.
[13] D. Bertsekas, "Incremental gradient, subgradient, proximal methods for convex optimization: A survey," *Optimization for Machine Learning*, p. 85, 2011.
[14] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2008.
[15] M. Osborne, B. Presnell, and B. Turlach, "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, vol. 20, no. 3, p. 389, 2000.
[16] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–451, 2004.
[17] M. Salman Asif and J. Romberg, "Dynamic updating for l_1 regularization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 421–434, 2010.
[18] P. Garrigues and L. El Ghaoui, "An homotopy algorithm for the Lasso with online observations," *Neural Information Processing Systems*, vol. 21, 2008.
[19] A. Hofleitner, L. E. Ghaoui, and A. Bayen, "Online least-squares estimation of time varying systems with sparse temporal evolution and application to traffic estimation," in *50th Conference on Decision and Control*, Dec. 2011, IEEE.
[20] I. Drori and D. Donoho, "Solution of l_1 minimization problems by LARS/homotopy methods," in *International Conference on Acoustics, Speech and Signal Processing*, 2006, vol. 3, IEEE.
[21] I. Loris, "On the performance of algorithms for the minimization of l_1-penalized functionals," *Inverse Problems*, vol. 25, no. 3, pp. 35 008–35 023, 2009.
[22] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Convex optimization with sparsity-inducing norms," *Optimization for Machine Learning*, pp. 19–53, 2011.
[23] S. Kim, K. Koh, S. Boyd, and D. Gorinevsky, "l_1 trend filtering," *SIAM Rev.*, vol. 51, no. 2, pp. 339–360, 2009.
[24] K. Knight and W. Fu, "Asymptotics for lasso-type estimators," *Annals of Statistics*, pp. 1356–1378, 2000.
[25] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2003.
[26] G. Golub and C. Van Loan, *Matrix Computations*. : Johns Hopkins Univ Press, 1996, vol. 3.
[27] A. Bayen, J. Butler, and A. Patire *et al.*, Mobile Millennium Final Report University of California, Berkeley, UCB-ITS-CWP-2011-6, Tech. Rep., 2011.
[28] T. Hunter, T. Moldovan, M. Zaharia, S. Merzgui, J. Ma, M. J. Franklin, P. Abbeel, and A. M. Bayen, "Scaling the mobile millennium system in the cloud," in *2nd ACM Symposium on Cloud Computing*, 2011, vol. 28, ser. SOCC'11, pp. 1–8, ACM.
[29] A. Hofleitner, R. Herring, P. Abbeel, and A. Bayen, "Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network," *IEEE Trans. Intelligent Transportation Systems*, 2012.

On the Controllability Properties of Circulant Networks

Marzieh Nabi-Abdolyousefi and Mehran Mesbahi

***Abstract*—This paper examines the controllability of a group of first order agents, adopting a weighted consensus-type coordination protocol over a circulant network. Specifically, it is shown that a circulant network with Laplacian eigenvalues of maximum algebraic multiplicity q is controllable from q nodes. Our approach leverages on the Cauchy–Binet formula, which in conjunction with the Popov–Belevitch–Hautus test, leads to new insights on structural aspects of network controllability.**

***Index Terms*—Circulant graphs, coordination algorithms, network controllability.**

I. INTRODUCTION

Recently, controllability and observability of networked dynamic systems adopting consensus-type coordination algorithm has attracted the attention of researchers in distinct disciplines [1]–[5]. Network controllability arises in situations where a networked system is influenced or observed by an external entity, a scenario that is of importance in networked robotic systems, human-swarm interaction, and network security [6]–[8], as well as in areas such as quantum networks [9], [10]. In this direction, an intriguing conjecture by Godsil [1], [9] states that the ratio of graphs that are uncontrollable from any set of nodes to the

Manuscript received August 02, 2010; revised February 25, 2012; accepted April 21, 2013. Date of publication April 24, 2013; date of current version November 18, 2013. This work was supported by AFOSR grant FA9550-09-1-0091 and NSF grant CMMI-0856737. Recommended by Associate Editor M. Egerstedt.

M. Nabi-Abdolyousefi is with the Palo Alto Research Center (PARC), a Xerox Company, Palo Alto, CA 94304 USA. (e-mail: marzieh.nabi@gmail.com).

M. Mesbahi is with the Department of Aeronautics and Astronautics, University of Washington, Seattle, WA 98195-2400 USA (e-mail: mesbahi@uw.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2013.2259992