On Extensions of LARS by Information Geometry : Convex Objectives and ℓ_p-Norm

Masahiro Yukawa*† and Shun-ichi Amari*
* Brain Science Institute, RIKEN, Japan
† Dept. Electrical and Electronic Engineering, Niigata University, Japan

***Abstract*—This paper addresses extensions of the Least Angle Regression (LARS) algorithm from two different aspects: (i) from quadratic to more general objectives, and (ii) from ℓ_1-norm to ℓ_p-norm for $p < 1$. The equiangular vector, which is the key of LARS, is reproduced in connection with the Riemannian metric induced by the objective function, thereby making the extensions feasible. It is shown, in the case of $p < 1$, that two types of trajectory — the c-trajectory and the λ-trajectory — need to be distinguished by revealing the discontinuity of the λ-trajectory.**

I. Introduction

The sparsity-seeking property of ℓ_1-norm has widely been studied and used in the signal processing and statistical communities (there also exist literatures on this topic in the information theory community). The value of this property has been explored for sparse signal recovery as well as model selection. In particular, one of the main researches on this topic in signal processing is *compressed sensing* (see [1]–[3] among many others), while that in statistics is *Lasso* [4], which stands for *least absolute shrinkage and selection operator*. The models of the two researches have similarity but are different from at least two perspectives. A typical formulation of compressed sensing is given as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \|\boldsymbol{\beta}\|_1 \ \text{s.t.}\ \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta} = \boldsymbol{y}, \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} \in \mathbb{R}^d$ are given under the condition $n \gg d$. Namely, a highly underdetermined linear system is assumed, and it is required to find the minimum ℓ_1-norm solution among the infinitely many solutions. It has been widely known that the problem in (1) is a reasonable convex relaxation of the following sparse optimization:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \|\boldsymbol{\beta}\|_0 \ \text{s.t.}\ \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta} = \boldsymbol{y}, \tag{2}$$

where $\|\cdot\|_0$ counts the number of nonzero components; the solutions of (1) and (2) coincide under certain conditions ensuring sufficient sparsity of the ℓ_0 solution. On the other hand, the original formulation of Lasso is given as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \left\|\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{y}\right\|_2^2 \ \text{s.t.}\ \|\boldsymbol{\beta}\|_1 \le c,\ c > 0, \tag{3}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} \in \mathbb{R}^d$ are given under the condition $n < d$. In contrast to the case of compressed sensing, an overdetermined linear system is assumed, and it is required to minimize the discrepancy under a restricted model complexity.

Efron *et al.* have proposed the *Least Angle Regression (LARS)* algorithm [5], which constructs the solution path (or the trajectory) of Lasso efficiently. The key of LARS is the use of *equiangular vectors* and the trajectory consists of piecewise straight lines given by the equiangular vectors. In [6], LARS has been applied to the problem in (1) in the underdetermined case. The ultimate goal in this work is to extend LARS to the case of non-quadratic (convex) cost function and/or the case of ℓ_p-norm for $p < 1$. Unfortunately, however, the notion of equiangular vector cannot be extended straightforwardly to those cases, since the trajectory is no longer piecewise straight lines in the $\boldsymbol{\beta}$-coordinates. In the case of $p = 1$, an extension of LARS based on information geometry [7] has been presented in [8] using *geodesics in dually flat spaces* in place of straight lines (note: the trajectory is piecewise straight lines in the dual coordinates). The case of $p < 1$ is of particular interest because of its more attractive properties compared to the case of $p = 1$ [9], [10]. To the best of authors' knowledge, no result has been reported regarding an extension of LARS to the case of $p < 1$.

In this paper, we distinguish two problems which we call *the c-trajectory problem* and *the λ-trajectory problem*. *The c-trajectory problem* is the problem of finding a solution path of convex cost minimization under an ℓ_p-norm constraint; a simple example is the problem in (3), where c is a control parameter. *The λ-trajectory problem*, on the other hand, is the problem of finding a solution path of (unconstrained) regularized cost minimization; its corresponding counterpart of (3) is given as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \left\|\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{y}\right\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,\ \lambda > 0. \tag{4}$$

The problems in (3) and (4) are equivalent under an appropriate correspondence of c and λ. It should be remarked however that the two problems should be distinguished when we address an extension to the case of the ℓ_p-norm constraint for $p < 1$.

In the first part of this paper, we consider the case of $p = 1$. We present an extended LARS algorithm with a minimum notion of information geometry. A link between the trajectory and *the dual geodesic projection* is presented with the classical sufficient condition for unconstrained optimality. The proposed algorithm is derived through simple differential equations. A simple linear regression problem is considered as an example, and the equiangular vector of LARS is derived in connection with the Riemannian metric induced by the cost function (the metric is in fact Euclidean as the cost is quadratic in this

case). Another interesting problem with a non-quadratic cost function is also introduced.

In the second part of this paper, we consider the case of $p < 1$. The λ-trajectory is shown to be a subset of the c-trajectory, and therefore the two trajectories should be distinguished. Some properties of those trajectories are separately investigated, including *discontinuity of the λ-trajectory*. Some discussion on the construction of the trajectories is presented with an illustrative example of the c-trajectory.

II. c-Trajectory and λ-Trajectory Problems

Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a strictly-convex twice-differentiable function defined on the n dimensional Euclidean space $\mathbb{R}^n$, and $F_p : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{\beta} \mapsto \frac{1}{p}\|\boldsymbol{\beta}\|_p^p = \frac{1}{p}\sum_{i=1}^n |\beta_i|^p$, for $p > 0$. In the case of LARS (or Lasso), φ is quadratic and $p = 1$; see (3). In this case, the strict convexity corresponds to overdetermined systems ($n < d$).

Consider the following two problems.

Problem 1 (c-Trajectory Problem): Find the trajectory of solutions (a solution path) of the following problem (for all $c \geq 0$):

$$\mathcal{P}_c : \min_{\boldsymbol{\beta}\in\mathbb{R}^n} \varphi(\boldsymbol{\beta}) \text{ subject to } F_p(\boldsymbol{\beta}) \leq c. \tag{5}$$

Problem 2 (λ-Trajectory Problem): Find the trajectory of solutions of the following problem (for all $\lambda \geq 0$):

$$\mathcal{P}_\lambda : \min_{\boldsymbol{\beta}\in\mathbb{R}^n} \varphi(\boldsymbol{\beta}) + \lambda F_p(\boldsymbol{\beta}). \tag{6}$$

Let $\boldsymbol{\beta}_c^*$ denote a solution of (5) and $\boldsymbol{\beta}_\lambda^*$ a solution of (6). It is clear that $\boldsymbol{\beta}_c^*|_{c=0} = \boldsymbol{\beta}_\lambda^*|_{\lambda=\infty} = \mathbf{0}_n$ and $\boldsymbol{\beta}_c^*|_{c=\infty} = \boldsymbol{\beta}_\lambda^*|_{\lambda=0} = \boldsymbol{\beta}^* := \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^n} \varphi(\boldsymbol{\beta})$, where $\mathbf{0}_n \in \mathbb{R}^n$ stands for the length-n zero vector. If F_p is convex (i.e., $p \geq 1$), due to Lagrangian duality, the solution of (5) can be obtained by solving (6) (with the λ value chosen properly) which is unconstrained and thus relatively easy to solve in general. The c-trajectory and the λ-trajectory thus coincide with each other. This however does no longer hold when F_p is nonconvex (i.e., $p < 1$), as will be seen in Section IV.

III. The Case of $p = 1$

In this section, we assume that $p = 1$ so that F_p is convex and it also promotes the sparsity of $\boldsymbol{\beta}$ (i.e., it has a feature to attract each component of $\boldsymbol{\beta}$ to zero). As already mentioned, the c-trajectory and the λ-trajectory coincide in this case, and hence we simply call it the trajectory. A simple approach to construct the trajectory will be shown below.

A. Derivative of the Trajectory

Let us give some definitions first.

Definition 1: Given a continuous convex function $f : \mathbb{R}^n \to \mathbb{R}$, the *subdifferential* of f at any $\boldsymbol{y} \in \mathbb{R}^n$, defined as $\partial f(\boldsymbol{y}) := \{\boldsymbol{a} \in \mathbb{R}^n : \langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{a}\rangle + f(\boldsymbol{y}) \leq f(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathbb{R}^n\}$, is nonempty, where $\langle \cdot, \cdot \rangle$ stands for the standard inner product. An element of the subdifferential $\partial f(\boldsymbol{y})$ is called a subgradient of f at $\boldsymbol{y}$.

The function F_1 is nondifferentiable but continuous and convex, and hence it has nonempty *subdifferential*. A subgradient of F_1 at $\boldsymbol{\beta} := [\beta_1, \beta_2, \cdots, \beta_n]^\mathsf{T} \in \mathbb{R}^n$, which we denote by $\boldsymbol{\nabla} F_1(\boldsymbol{\beta})$, is given as follows:

$$[\boldsymbol{\nabla} F_1(\boldsymbol{\beta})]_i \begin{cases} = \operatorname{sgn}(\beta_i) & \text{if } \beta_i \neq 0 \\ \in [-1, 1] & \text{if } \beta_i = 0 \end{cases} \quad (i = 1, 2, \cdots, n), \tag{7}$$

where $[\cdot]_i$ stands for the ith component of a vector and $\operatorname{sgn}(\cdot)$ the signum function. Due to the convexity of φ and F_1, $\boldsymbol{\beta}_c^* \in \{\boldsymbol{\beta} \in \mathbb{R}^n : F_1(\boldsymbol{\beta}) \leq c\}$ is a solution of $\mathcal{P}_c$ if and only if there exist a subgradient $\boldsymbol{\nabla} F_1(\boldsymbol{\beta}_c^*)$ and an $\alpha_c \geq 0$ such that

$$\boldsymbol{\nabla}\varphi(\boldsymbol{\beta}_c^*) + \alpha_c \boldsymbol{\nabla} F_1(\boldsymbol{\beta}_c^*) = \mathbf{0}_n. \tag{8}$$

Clearly, $\boldsymbol{\beta}_c^*$ is also a solution of $\mathcal{P}_\lambda$ for $\lambda = \alpha_c$. For any function $f : \mathbb{R}^n \to \mathbb{R}$ differentiable in terms of $\beta_i \neq 0$, we define

$$\boldsymbol{\nabla}_{\mathcal{A}}\, f(\boldsymbol{\beta}) := \begin{bmatrix} \dfrac{\partial}{\partial \beta_{i_1}} f(\boldsymbol{\beta}) \\ \dfrac{\partial}{\partial \beta_{i_2}} f(\boldsymbol{\beta}) \\ \vdots \\ \dfrac{\partial}{\partial \beta_{i_m}} f(\boldsymbol{\beta}) \end{bmatrix} \in \mathbb{R}^m$$

where $\mathcal{A} := \{i_1, i_2, \cdots, i_m\} := \{i \in \{1, 2, \cdots, n\} : \beta_i \neq 0\}$ denotes the *set of active indices* for $\boldsymbol{\beta} := [\beta_1, \beta_2, \cdots, \beta_n]^\mathsf{T}$ (m is the number of nonzero components of $\boldsymbol{\beta}$). Then, from (8), we immediately have the following:

$$\boldsymbol{\nabla}_{\mathcal{A}}\varphi(\boldsymbol{\beta}_\lambda^*) + \lambda \boldsymbol{\nabla}_{\mathcal{A}} F_1(\boldsymbol{\beta}_\lambda^*) = \mathbf{0}_m. \tag{9}$$

Differentiating (9) with respect to λ, we thus obtain

$$\begin{aligned} &[\boldsymbol{\nabla}_{\mathcal{A}}\boldsymbol{\nabla}_{\mathcal{A}}\varphi(\boldsymbol{\beta}_\lambda^*) + \lambda \boldsymbol{\nabla}_{\mathcal{A}}\boldsymbol{\nabla}_{\mathcal{A}} F_1(\boldsymbol{\beta}_\lambda^*)]\dot{\boldsymbol{\beta}}_{\lambda,\mathcal{A}}^* \\ &= \boldsymbol{\nabla}_{\mathcal{A}}\boldsymbol{\nabla}_{\mathcal{A}}\varphi(\boldsymbol{\beta}_\lambda^*)\dot{\boldsymbol{\beta}}_{\lambda,\mathcal{A}}^* = -\boldsymbol{\nabla}_{\mathcal{A}} F_1(\boldsymbol{\beta}_\lambda^*), \end{aligned} \tag{10}$$

where $\dot{\boldsymbol{\beta}}_{\lambda,\mathcal{A}}^* := d\boldsymbol{\beta}_{\lambda,\mathcal{A}}^*/d\lambda$ with $\boldsymbol{\beta}_{\lambda,\mathcal{A}}^*$ obtained by eliminating all the zero components from $\boldsymbol{\beta}_\lambda^*$. Note here that $\boldsymbol{\nabla}_{\mathcal{A}}\boldsymbol{\nabla}_{\mathcal{A}} F_1(\boldsymbol{\beta}) = \boldsymbol{O}_m$, where $\boldsymbol{O}_m$ denotes the $m \times m$ zero matrix. Since the strict convexity of φ guarantees the positive definiteness (thereby nonsingularity) of $\boldsymbol{\nabla}_{\mathcal{A}}\boldsymbol{\nabla}_{\mathcal{A}}\varphi(\boldsymbol{\beta}_\lambda^*)$, we have

$$\dot{\boldsymbol{\beta}}_{\lambda,\mathcal{A}}^* = -\left[\boldsymbol{\nabla}_{\mathcal{A}}\boldsymbol{\nabla}_{\mathcal{A}}\varphi(\boldsymbol{\beta}_\lambda^*)\right]^{-1} \boldsymbol{\nabla}_{\mathcal{A}} F_1(\boldsymbol{\beta}_\lambda^*). \tag{11}$$

B. Trajectory and Dual Geodesic Projection

From the discussion in the previous subsection, we obtain the following observations.

Observation 1:

1) The trajectory is a piecewise smooth continuous curve.
2) Suppose that all the components of $\boldsymbol{\beta}$ are nonzero. Along the piecewise smooth curve, the direction of the gradient of φ keeps constant (see (8)). Consider the coordinate transform $\boldsymbol{\eta} := \boldsymbol{\nabla}\varphi(\boldsymbol{\beta})$, which is one to one due to the strict convexity of φ, and it is called the Legendre transform. The $\boldsymbol{\eta}$-coordinates are called the *dual affine coordinates*. Equation (8) suggests that $\boldsymbol{\eta}_\lambda^* := \boldsymbol{\nabla}\varphi(\boldsymbol{\beta}_\lambda^*)$

draws a straight line in the dual coordinates and it is called a *dual geodesic*. The same applies to the case where there exist zero components in $\boldsymbol{\beta}$; in this case we may consider the lower dimensional space.

3) Suppose that all the components of $\boldsymbol{\beta}$ are nonzero. Let $B_\lambda(\ni \boldsymbol{\beta}^*_\lambda)$ stands for the face of the polyhedron $\{\boldsymbol{\beta} \in \mathbb{R}^n : F_1(\boldsymbol{\beta}) \leq F_1(\boldsymbol{\beta}^*_\lambda)\}$. It is then clear that $\langle \boldsymbol{\nabla} F_1(\boldsymbol{\beta}^*_\lambda), \boldsymbol{\beta}^*_\lambda - \boldsymbol{y}\rangle = 0$ for any $\boldsymbol{y} \in B_\lambda$ (with the standard inner product $\langle \cdot, \cdot \rangle$); in words $\boldsymbol{\nabla} F_1(\boldsymbol{\beta}^*_\lambda)$ is orthogonal to B_λ. It is therefore seen that the curve of $\boldsymbol{\beta}^*_\lambda$ is also *orthogonal to* B_λ in the sense of the Riemannian metric; i.e.,

$$\left\langle \dot{\boldsymbol{\beta}}^*_\lambda, \boldsymbol{\beta}^*_\lambda - \boldsymbol{y} \right\rangle_{\boldsymbol{\nabla}\boldsymbol{\nabla}\varphi(\boldsymbol{\beta}^*_\lambda)} = 0$$

for any $\boldsymbol{y} \in B_\lambda$. This implies that the trajectory gives the *dual geodesic projection* of $\boldsymbol{\beta}^*$ onto the hypersurface B_λ. The same applies to the case where there exist zero components in $\boldsymbol{\beta}$.

C. Extended LARS Algorithm

We now extend the LARS algorithm. In analogy with LARS, we set the initial point to $\boldsymbol{\beta}_0 := \mathbf{0}_n$. We now need to find in which direction to step. Viewing the trajectory with $\mathcal{P}_c$, the direction is given by the negative Minkowskian-gradient of φ. The Minkowskian-gradient of φ at $\boldsymbol{\beta}$ is defined as such a vector $\boldsymbol{a} := [a_1, a_2, \cdots, a_n]^\mathsf{T} \in \mathbb{R}^n$ that maximizes $\langle \boldsymbol{\nabla}\varphi(\boldsymbol{\beta}), \boldsymbol{a}\rangle$ subject to $F_1(\boldsymbol{a}) = 1$. Assume for simplicity that there exists a unique i^* that maximizes $|\eta_i|$ among $i = 1, 2, \cdots, n$, i.e. $\mathcal{I} := \mathrm{argmax}_{i=1,2,\cdots,n} |\eta_i| = \{i^*\}$, where $\boldsymbol{\eta} := [\eta_1, \eta_2, \cdots, \eta_n]^\mathsf{T} := \boldsymbol{\nabla}\varphi(\boldsymbol{\beta})$. The Minkowskian-gradient is then given as

$$a_i = \begin{cases} \mathrm{sgn}\ \eta_i & \text{if } i = i^* \\ 0 & \text{otherwise.} \end{cases}$$

One can thus depart the initial point $\mathbf{0}_n$ in the negative Minkowskian-gradient direction, i.e. $-\boldsymbol{a}$. Taking a careful look at (8) together with (7), one can readily see that the trajectory should move along the i^*th coordinate until $|\eta_{i^*}|$ becomes no longer the unique maximum among $|\eta_i|$s (i.e., only the i^*th component of $\boldsymbol{\beta}$ should be decreased, or increased, depending on the sign of η_{i^*}). At some point, we will have $|\eta_{i^*}| = |\eta_{j^*}| = \max\{|\eta_1|, |\eta_2|, \cdots, |\eta_n|\}$ for some $j^* \neq i^*$. This is a turning point of the trajectory, and we redefine $\mathcal{I} := \{i^*, j^*\}$. The direction to step on the i^*-j^* plane is given by

$$-\dot{\boldsymbol{\beta}}_{n,\mathcal{I}} := \left[\boldsymbol{\nabla}_\mathcal{I}\boldsymbol{\nabla}_\mathcal{I}\varphi(\boldsymbol{\beta}_{n,\mathcal{I}})\right]^{-1} \boldsymbol{\nabla}_\mathcal{I} F_1(\boldsymbol{\beta}_{n,\mathcal{I}}), \tag{12}$$

where $\boldsymbol{\beta}_{n,\mathcal{I}} := [\beta_{n,i^*}, \beta_{n,j^*}]^\mathsf{T}$ and $\boldsymbol{\nabla}_\mathcal{I}$ denotes the partial derivatives only in terms of the components corresponding to the indices in $\mathcal{I}$ (in this case, $\boldsymbol{\nabla}_\mathcal{I} := [\frac{\partial}{\partial\beta_{i^*}}, \frac{\partial}{\partial\beta_{j^*}}]^\mathsf{T}$). Although $\frac{\partial}{\partial\beta_{j^*}} F_1(\boldsymbol{\beta}_{n,\mathcal{I}})$ is not uniquely determined (see (7)), it is determined as

$$\frac{\partial}{\partial\beta_{j^*}} F_1(\boldsymbol{\beta}_{n,\mathcal{I}}) = \mathrm{sgn}(-\eta_{j^*}) = \mathrm{sgn}\left(-[\boldsymbol{\nabla}\varphi(\boldsymbol{\beta}_{n,\mathcal{I}})]_{j^*}\right).$$

The negative sign in (12) is due to the fact that λ decreases as we step along the trajectory in our way. The trajectory curve from the turning point (to the next turning point if $n \geq 3$) is constructed by updating $\boldsymbol{\beta}_{n,\mathcal{I}}$ as follows:

$$\boldsymbol{\beta}_{n+1,\mathcal{I}} := \boldsymbol{\beta}_{n,\mathcal{I}} + \mu[\boldsymbol{\nabla}_\mathcal{I}\boldsymbol{\nabla}_\mathcal{I}\varphi(\boldsymbol{\beta}_{n,\mathcal{I}})]^{-1}\boldsymbol{\nabla} F_1(\boldsymbol{\beta}_{n,\mathcal{I}}), \tag{13}$$

where $\mu > 0$ is a small constant. The next turning point comes when we will have $|\eta_{i^*}| = |\eta_{j^*}| = |\eta_{k^*}| = \max\{|\eta_1|, |\eta_2|, \cdots, |\eta_n|\}$ for some $k^* \notin \{i^*, j^*\}$, and in this case we redefine $\mathcal{I} := \{i^*, j^*, k^*\}$. One can follow the above idea until the trajectory finally reaches $\boldsymbol{\beta}^*(:= \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^n} \varphi(\boldsymbol{\beta}))$. In addition to the trajectory, we can also obtain the λ value at each point on the trajectory based on (8) as follows:

$$\lambda = |[\boldsymbol{\nabla}\varphi(\boldsymbol{\beta}_n)]_{i^*}|.$$

D. Example 1: Quadratic Function

A typical example is the linear regression. Consider the following linear model:

$$y_t = \langle \boldsymbol{\beta}_\mathrm{o}, \boldsymbol{x}_t\rangle + n_t,\ t = 1, 2, \cdots, d,$$

where $\boldsymbol{x}_t \in \mathbb{R}^n$ are called the design vectors (assumed available), $\boldsymbol{\beta}_\mathrm{o}$ the explanatory vector to be estimated, y_t the response variables which are observed, and n_t the ambient noise subject to the Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Consider the statistical model where the joint probability of y_1, y_2, $\cdots$, y_d given $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\cdots$, $\boldsymbol{x}_d$, and $\boldsymbol{\beta}$ is given by

$$p(y_t|\boldsymbol{x}_t, \boldsymbol{\beta}) = \exp\left(-\frac{1}{2}\sum_{t=1}^d (y_t - \langle \boldsymbol{\beta}, \boldsymbol{x}_t\rangle)^2\right).$$

In order to maximize this probability, one can alternatively minimize its negative log likelihood, which is given as follows:

$$\begin{aligned}\varphi(\boldsymbol{\beta}) :=& -\log(p(y_t|\boldsymbol{x}_t, \boldsymbol{\beta})) = \frac{1}{2}\sum_{t=1}^d (y_t - \langle \boldsymbol{\beta}, \boldsymbol{x}_t\rangle)^2 \\ =& \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}\right\|_2^2,\ \boldsymbol{\beta} \in \mathbb{R}^n,\end{aligned} \tag{14}$$

where $\boldsymbol{y} := [y_1, y_2, \cdots, y_d]^\mathsf{T} \in \mathbb{R}^d$ and $\boldsymbol{X} := [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_d] \in \mathbb{R}^{n\times d}$. This is a typical cost function in least squares problems. Lasso formulates the linear regression problem as the constrained minimization problem (5) for the quadratic function in (15) and $p = 1$ (the ℓ_1-norm constraint) under the assumption of an overdetermined system ($n < d$). If $\mathrm{rank}(\boldsymbol{X}) = n$, $\varphi(\boldsymbol{\beta})$ is strictly convex with its unique global minimizer given by $\boldsymbol{\beta}^* = (\boldsymbol{X}\boldsymbol{X}^\mathsf{T})^{-1}\boldsymbol{X}\boldsymbol{y}$. In this case, we have

$$\varphi(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\mathsf{T}\boldsymbol{G}(\boldsymbol{\beta} - \boldsymbol{\beta}^*),\ \boldsymbol{\beta} \in \mathbb{R}^n, \tag{15}$$

where

$$\boldsymbol{G} := \boldsymbol{X}\boldsymbol{X}^\mathsf{T} = \boldsymbol{\nabla}\boldsymbol{\nabla}\varphi(\boldsymbol{\beta}),\ \forall\boldsymbol{\beta} \in \mathbb{R}^n.$$

Namely the metric is constant over $\mathbb{R}^n$. Also we define

$$\boldsymbol{G}_\mathcal{I} := \boldsymbol{\nabla}_\mathcal{I}\boldsymbol{\nabla}_\mathcal{I}\varphi(\boldsymbol{\beta}) \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|},\ \forall\boldsymbol{\beta} \in \mathbb{R}^n.$$

Let $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_{|\mathcal{I}|}\}$ be the standard basis of $\mathbb{R}^{|\mathcal{I}|}$. Then we have

$$\left|\left\langle \boldsymbol{e}_i, \dot{\boldsymbol{\beta}}_{n,\mathcal{I}} \right\rangle_{\boldsymbol{G}_\mathcal{I}}\right| = \left|\boldsymbol{e}_i^\mathsf{T} \boldsymbol{\nabla}_\mathcal{I} F_1(\boldsymbol{\beta}_{n,\mathcal{I}})\right| = 1. \tag{16}$$

LARS exploits the property in (16). As the negative differential $-\dot{\boldsymbol{\beta}}_{n,\mathcal{I}}$ is constant from a turning point to the next one, the trajectory is a straight line in the $\boldsymbol{\beta}$-coordinates (as well as in the $\boldsymbol{\eta}$-coordinates).

E. Example 2: Non-Quadratic Function

We consider the 0-1 response case (i.e., $y_t = \pm 1$) with the joint probability of y_1, y_2, $\cdots$, y_d ($n < d$) given $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\cdots$, $\boldsymbol{x}_d$, and $\boldsymbol{\beta}$ is given by

$$\begin{aligned} p(y_t|\boldsymbol{x}_t, \boldsymbol{\beta}) &= \prod_{t=1}^{d} \frac{\exp(y_t \langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle)}{1 + \exp(\langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle)} \\ &= \prod_{t=1}^{d} \exp\left(y_t \langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle - \psi(\boldsymbol{\beta}, \boldsymbol{x}_t)\right), \end{aligned} \tag{17}$$

where

$$\psi(\boldsymbol{\beta}, \boldsymbol{x}_t) := \log\left(1 + \exp(\langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle)\right).$$

In this case, the negative log likelihood is given as follows:

$$\varphi(\boldsymbol{\beta}) = \sum_{t=1}^{d} \left[\log\left(1 + \exp\langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle\right) - y_t \langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle\right].$$

Its gradient and Hessian are given respectively as follows:

$$\boldsymbol{\nabla}\varphi(\boldsymbol{\beta}) = \sum_{t=1}^{d} \left[\frac{\exp\langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle}{1 + \exp\langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle} - y_t\right] \boldsymbol{x}_t \tag{18}$$

$$\boldsymbol{\nabla}\boldsymbol{\nabla}\varphi(\boldsymbol{\beta}) = \sum_{t=1}^{d} \zeta(\boldsymbol{x}_t, \boldsymbol{\beta}) \boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T}, \tag{19}$$

where

$$\begin{aligned} &\zeta(\boldsymbol{x}_t, \boldsymbol{\beta}) := \exp\langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle \\ &\times \frac{(1 + \langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle^2)(1 + \exp\langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle) + \langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle^2 \exp\langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle}{[1 + \exp\langle \boldsymbol{\beta}, \boldsymbol{x}_t \rangle]^2}. \end{aligned}$$

It is clear that $\zeta(\boldsymbol{x}_t, \boldsymbol{\beta}) > 0$ for any $\boldsymbol{x}_t, \boldsymbol{\beta} \in \mathbb{R}^n$, and hence the Hessian is positive definite (and thus $\varphi(\boldsymbol{\beta})$ is strictly convex[1]) provided that $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_d$ spans $\mathbb{R}^n$.

IV. The Case of $p < 1$

We now consider the case of $p < 1$, where the function F_p is *nonconvex*. In this case, unlike the case of $p = 1$, the nice relation between the trajectory and the dual geodesic projection does no longer hold, although each point on the c-trajectory can be characterized as a dual geodesic projection.

[1]Let $f : \mathbb{R}^N \to \mathbb{R}$ be twice differentiable. Then, if $\boldsymbol{\nabla}\boldsymbol{\nabla} f(\boldsymbol{\beta})$ is positive definite, then f is strictly convex. The converse is not true in general; the converse is true in the particular case of quadratic functions [11].

A. Difference Between c-Trajectory and λ-Trajectory

We first show the following observation.

Observation 2: For $0 < p < 1$, the λ-trajectory is a proper subset of the c-trajectory.

Sketch of Proof : Given an arbitrary $\lambda \geq 0$, let $\boldsymbol{\beta}_\lambda^*$ is a solution of $\mathcal{P}_\lambda$. Define $c_\lambda := F_p(\boldsymbol{\beta}_\lambda^*)$. Assume that $\boldsymbol{\beta}_\lambda^*$ is not a solution of $\mathcal{P}_{c_\lambda}$; i.e., there exists some $\widehat{\boldsymbol{\beta}}$ such that $F_p(\widehat{\boldsymbol{\beta}}) \leq c_\lambda = F_p(\boldsymbol{\beta}_\lambda^*)$ and $\varphi(\widehat{\boldsymbol{\beta}}) < \varphi(\boldsymbol{\beta}_\lambda^*)$. It can then immediately be shown that

$$\varphi(\widehat{\boldsymbol{\beta}}) + \lambda F_p(\widehat{\boldsymbol{\beta}}) < \varphi(\boldsymbol{\beta}_\lambda^*) + \lambda F_p(\boldsymbol{\beta}_\lambda^*),$$

which contradicts the fact that $\boldsymbol{\beta}_\lambda^*$ is a solution of $\mathcal{P}_\lambda$. This verifies that $\boldsymbol{\beta}_\lambda^*$ is a solution of $\mathcal{P}_{c_\lambda}$. It will be shown in the following subsections that the c-trajectory is continuous at $\boldsymbol{\beta} = \mathbf{0}_n$ whereas *the λ-trajectory is discontinuous at $\boldsymbol{\beta} = \mathbf{0}_n$* for $0 < p < 1$. □

The observation above implies that we need to distinguish the two trajectories for $0 < p < 1$. In analogy with the case of $p = 1$, it holds that $\boldsymbol{\beta}_c^*|_{c=0} = \boldsymbol{\beta}_\lambda^*|_{\lambda=\infty} = \mathbf{0}_n$ and $\boldsymbol{\beta}_c^*|_{c=\infty} = \boldsymbol{\beta}_\lambda^*|_{\lambda=0} = \boldsymbol{\beta}^* := \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^n} \varphi(\boldsymbol{\beta})$.

B. c-Trajectory

Observation 3: For any $c \leq F_p(\boldsymbol{\beta}^*)$, it holds that $F_p(\boldsymbol{\beta}_c^*) = c$; in words $\boldsymbol{\beta}_c^*$ lies on the boundary of the feasible set of $\mathcal{P}_c$.

Proof : Assume that $F_p(\boldsymbol{\beta}_c^*) < c$. Then there exists $\boldsymbol{\beta}_\alpha := \alpha\boldsymbol{\beta}_c^* + (1-\alpha)\boldsymbol{\beta}^*$, $\alpha \in [0, 1)$, such that $F_p(\boldsymbol{\beta}_\alpha) = c$. The strict convexity of φ ensures that $\boldsymbol{\beta}^*$ is the unique global minimizer of φ over $\mathbb{R}^n$. Hence it follows that $\varphi(\boldsymbol{\beta}_\alpha) \leq \alpha\varphi(\boldsymbol{\beta}_c^*) + (1-\alpha)\varphi(\boldsymbol{\beta}^*) < \varphi(\boldsymbol{\beta}_c^*)$, indicating that $\boldsymbol{\beta}_c^*$ is not a solution of $\mathcal{P}_c$. This verifies that $F_p(\boldsymbol{\beta}_c^*) = c$. □

Observation 3 implicitly ensures the *continuity* of the c-trajectory at $\boldsymbol{\beta} = \mathbf{0}_n$. We define the F_p-gradient of φ at $\boldsymbol{\beta}$ is defined as such a vector $\boldsymbol{a} := [a_1, a_2, \cdots, a_N]^\mathsf{T} \in \mathbb{R}^n$ that maximizes $\langle \boldsymbol{\nabla}\varphi(\boldsymbol{\beta}), \boldsymbol{a} \rangle$ subject to $F_p(\boldsymbol{a}) = 1/p$. Then we obtain the following observation.

Observation 4: Assume that there exists a unique i^* that maximizes $|\eta_i|$ among $i = 1, 2, \cdots, n$, where $\boldsymbol{\eta} := [\eta_1, \eta_2, \cdots, \eta_n]^\mathsf{T} := \boldsymbol{\nabla}\varphi(\boldsymbol{\beta})$. Then, at the point $\boldsymbol{\beta} = \mathbf{0}_n$, the F_p-gradient of φ for any $0 < p < 1$ coincides with its Minkowskian gradient.

Proof : We mention that the F_p-gradient of φ at $\boldsymbol{\beta}$ can be defined also as such a vector $\boldsymbol{a} := [a_1, a_2, \cdots, a_N]^\mathsf{T} \in \mathbb{R}^n$ that maximizes $\langle \boldsymbol{\nabla}\varphi(\boldsymbol{\beta}), \boldsymbol{a} \rangle$ subject to $F_p(\boldsymbol{a}) \leq 1/p$, instead of $F_p(\boldsymbol{a}) = 1/p$. The claim is readily verified by noting that $F_p(\boldsymbol{a}) \leq 1/p \Rightarrow F_1(\boldsymbol{a}) \leq 1$. □

Observation 4 suggests that the c-trajectory for $0 < p < 1$ leaves the point $\mathbf{0}_n$ in the same direction as for $p = 1$.

C. λ-Trajectory

We consider a simple one-dimensional case of the λ-trajectory problem: let $\varphi(\beta) := \frac{1}{2}(\beta - 2)^2$ and $F_{1/2}(\beta) := 2|\beta|^{1/2}$. The graphs of the cost function for several values of λ are illustrated in Fig. 1. It is seen that each graph has two local minima, one at $\beta = 0$ and the other at some point $\beta_\lambda \in [1.333, 1.8]$. Also it can be seen that, although β_λ is the unique global minimizer for $\lambda = 0.3$ and $\lambda = 0.5$, both β_λ

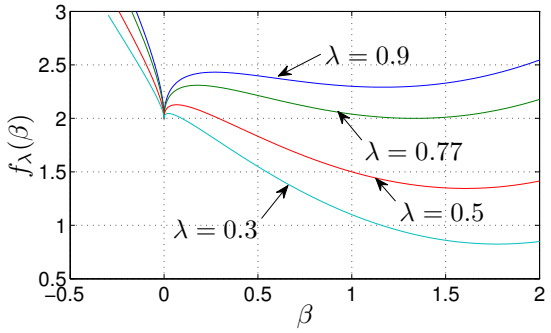

Fig. 1. Graphs of $f_\lambda(\beta) := \frac{1}{2}(\beta - 2)^2 + 2\lambda|\beta|^{1/2}$.

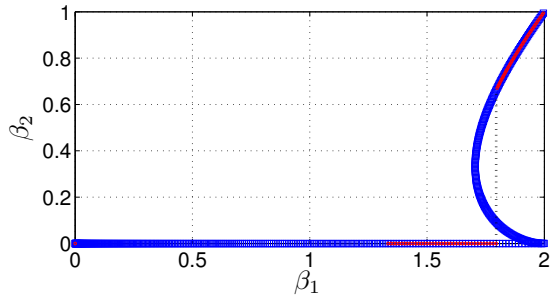

Fig. 2. The c-trajectory (blue) and the λ-trajectory (red) for $\varphi(\boldsymbol{\beta}) := \frac{1}{2}\left\|\boldsymbol{\beta} - [2,1]^\mathsf{T}\right\|_2^2$ and $p = 1/2$.

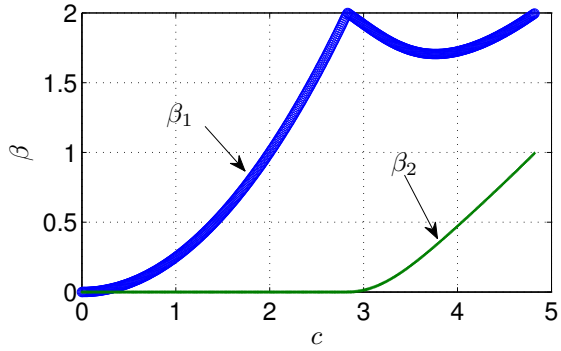

Fig. 3. Correspondence between c and β.

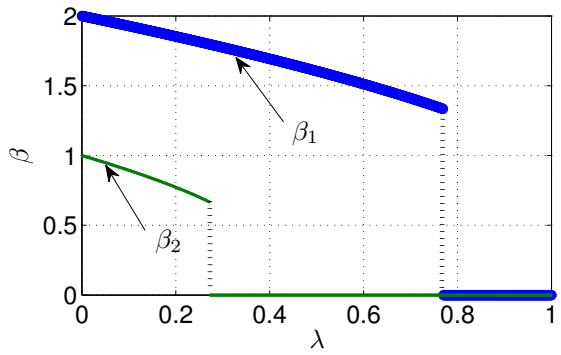

Fig. 4. Correspondence between λ and β.

and the origin are the global minimizers for $\lambda = 0.77$, and β_λ is no longer the global minimizer for $\lambda = 0.9$. This suggests that the λ-trajectory in this simple example is *discontinuous* and it 'jumps' from the continuous curve of β_λ to the origin; the 'jump' occurs at $\lambda \approx 0.77$.

A question arises: *could the λ-trajectory have any other discontinuous points?* To answer it simply, let us consider the following special two-dimensional case: $\varphi(\boldsymbol{\beta}) := \frac{1}{2}\left\|\boldsymbol{\beta} - [2,1]^\mathsf{T}\right\|_2^2$ and $F_{1/2}(\beta) := 2(|\beta_1|^{1/2} + |\beta_2|^{1/2})$. In this case, $\varphi(\boldsymbol{\beta}) + \lambda F_{1/2}(\boldsymbol{\beta}) = f_1(\beta_1) + f_2(\beta_2)$, where $f_1(\beta) := \frac{1}{2}(\beta - 2)^2 + 2\lambda\,|\beta|^{1/2}$, $\beta \in \mathbb{R}$, and $f_2(\beta) := \frac{1}{2}(\beta - 1)^2 + 2\lambda\,|\beta|^{1/2}$, $\beta \in \mathbb{R}$. One can therefore minimize $f_1(\beta_1)$ and $f_2(\beta_2)$ separately. From the observation for the one-dimensional case, it can be expected that both β_1 and β_2 jump from zero to some values at different λs. The c-trajectory and the λ-trajectory for this example are depicted in Fig. 2. The c-trajectory is continuous in this case. On the other hand, the λ-trajectory is as follows: it jumps from $\mathbf{0}_n$ to $[1.333, 0]$, moves along the β_1-coordinate up to $[1.797, 0]$, jumps again to $[1.797, 0.666]$, and then follows the c-trajectory up to $[2, 1]$. This exemplifies that the λ-trajectory may have multiple discontinuous points in general. Figures 3 and 4 depict the correspondence between c and β, and that between λ and β, respectively. Figure 5 depicts the graph of α_c.

For a more general case where $\varphi(\boldsymbol{\beta}) := \frac{1}{2}\left\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\right\|_2^2$, the λ-trajectory behaves as follows: (i) $\boldsymbol{\beta}_\lambda^*$ jumps from $\mathbf{0}_n$ to some point on the i^*th coordinate ($i^* := \mathrm{argmax}_{i=1,2}\,|\eta_i|$ assuming it is unique), (ii) moves along the coordinate in the direction of increasing $|\beta_{i^*}|$, and (iii) at some point the j^*th element β_{j^*} jumps, where $j^* \in \{1, 2\} \setminus \{i^*\}$. We can indeed verify the following observation.

Observation 5: For $0 < p < 1$, the λ-trajectory is always discontinuous at $\boldsymbol{\beta} = \mathbf{0}_n$.

Sketch of Proof : Recall first that φ is supposed to be differentiable over $\mathbb{R}^n$. The function $F_p(\boldsymbol{\beta})$ can be expressed as $F_p(\boldsymbol{\beta}) = \sum_{i=1}^n \psi(\beta_i)$, where $\psi(\beta) := \frac{1}{p}\,|\beta|^p$ for $\beta \in \mathbb{R}$. It can be verified that $\lim_{\beta \uparrow 0} \frac{d}{d\beta}\psi(\beta) = \infty$ and $\lim_{\beta \downarrow 0} \frac{d}{d\beta}\psi(\beta) = -\infty$. This implies that the function $\varphi(\boldsymbol{\beta}) + \lambda F_p(\boldsymbol{\beta})$ has a local minimum at $\boldsymbol{\beta} = \mathbf{0}_n$ for any $\lambda > 0$ and thus the λ-trajectory is discontinuous at $\boldsymbol{\beta} = \mathbf{0}_n$. □

D. Discussion

At any solution $\boldsymbol{\beta}_c^*$ of $\mathcal{P}_c$, a contour of φ should touch a contour of F_p at $\boldsymbol{\beta}_c^*$; in other words, the two contours should share the same tangent plane. Therefore it holds that

$$\boldsymbol{\nabla}_{\mathcal{A}}\varphi(\boldsymbol{\beta}_c^*) = -\alpha_c \boldsymbol{\nabla}_{\mathcal{A}} F_p(\boldsymbol{\beta}_c^*),\ \ \alpha_c \geq 0. \tag{20}$$

Comparing (9) and (20), it is seen that α_c plays a similar role to λ. It should however be mentioned that α_c is *not* monotone in terms of c. Differentiating (20) with respect to c, it follows that

$$[\boldsymbol{\nabla}_{\mathcal{A}}\boldsymbol{\nabla}_{\mathcal{A}}\varphi(\boldsymbol{\beta}_c^*) + \alpha_c \boldsymbol{\nabla}_{\mathcal{A}}\boldsymbol{\nabla}_{\mathcal{A}} F_p(\boldsymbol{\beta}_c^*)]\dot{\boldsymbol{\beta}}_{c,\mathcal{A}}^* = -\dot{\alpha}_c \boldsymbol{\nabla}_{\mathcal{A}} F_p(\boldsymbol{\beta}_c^*), \tag{21}$$

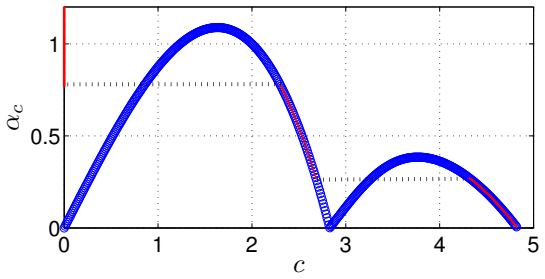

Fig. 5. The graphs of α_c (blue) and λ (red).

where $\dot{\boldsymbol{\beta}}^*_{c,\mathcal{A}} := d\boldsymbol{\beta}^*_{c,\mathcal{A}}/dc$ with $\boldsymbol{\beta}^*_{c,\mathcal{A}}$ obtained by eliminating all the zero components from $\boldsymbol{\beta}^*_c$ and $\dot{\alpha}_c := \frac{d}{dc}\alpha_c$. Equation (21) may offer a clue to construct the c-trajectory as like (10) in the case of $p = 1$. However delicate care would be required at turning points, as will be explained below with the example shown in Fig. 2.

For $0 < p < 1$, the derivative of $\psi(\beta) := \frac{1}{p}|\beta|^p$ at a point $\beta \neq 0$ is given by

$$\psi'(\beta) = \mathrm{sgn}(\beta)\,|\beta|^{-(1-p)}\,. \tag{22}$$

Viewing the curving part (between $[2,1]^\mathsf{T}$ and $[2,0]^\mathsf{T}$) of the trajectory in Fig 2 with (20), it is seen that $\boldsymbol{\nabla}_{\mathcal{A}} F_p(\boldsymbol{\beta}^*_c) \to [1/\sqrt{2}, \infty]^\mathsf{T}$ as $\boldsymbol{\beta}^*_c \to [2,0]^\mathsf{T}$, and hence at the turning point $\boldsymbol{\nabla}_{\mathcal{A}}\varphi(\boldsymbol{\beta}^*_c) \sim [0,1]^\mathsf{T}$ with $\alpha_c = 0$. If we follow the trajectory from $[0,0]^\mathsf{T}$ to $[2,1]^\mathsf{T}$, the α_c value increases from zero to some positive value, then starts to decrease until it becomes zero at the turning point $[2,0]^\mathsf{T}$, and then again it increases up to some value and then keep decreasing until it becomes zero again at $[2,1]^\mathsf{T}$. Accordingly, the sign of $\dot{\alpha}_c$ changes from positive to negative between $[0,0]^\mathsf{T}$ and $[2,0]^\mathsf{T}$, it changes from negative to positive at the turning point $[2,0]^\mathsf{T}$, and then again it changes from positive to negative between $[2,0]^\mathsf{T}$ and $[2,1]^\mathsf{T}$. This causes the *cusp* at the turning point observed in Fig. 2. The large value of the second component of $\boldsymbol{\nabla}_{\mathcal{A}} F_p(\boldsymbol{\beta}^*_c)$ is scaled down by the multiplication of the inverse of $\boldsymbol{\nabla}_{\mathcal{A}}\boldsymbol{\nabla}_{\mathcal{A}}\varphi(\boldsymbol{\beta}^*_c) + \alpha_c \boldsymbol{\nabla}_{\mathcal{A}}\boldsymbol{\nabla}_{\mathcal{A}} F_1(\boldsymbol{\beta}^*_c)$. Also we need to care the changes of the sign of $\dot{\alpha}_c$ in the middle of adjacent turning points.

The use of the 'jump' phenomena of the λ-trajectory would bypass the aforementioned delicate problem of the c-trajectory. However there still remain some open issues in this approach. The first issue is: at what value of λ does a 'jump' phenomenon happen? Is this the point where the sign of $\dot{\alpha}_c$ changes? This would be denied by Figs. 2, 3, and 5. The second issue is: where to 'jump'? Is there any geometric connection between the points before and after the 'jump'? The λ value should be preserved by the 'jump', and therefore a possible approach would be to find a path along which the λ value (i.e., $|[\boldsymbol{\nabla}\varphi(\boldsymbol{\beta})]_1| \,/\, |[\boldsymbol{\nabla} F_p(\boldsymbol{\beta})]_1|$) is unchanged.

V. Conclusion

This paper has presented a study of extending LARS to a strictly-convex differentiable function and the ℓ_p-norm constraint for $p < 1$ with the distinction between the c-trajectory and the λ-trajectory. In the case of $p = 1$, the two trajectories coincide and an iterative algorithm to construct the trajectory has been presented based on simple differential equations. In the case of $p < 1$, it has been shown that the λ-trajectory is discontinuous and the two trajectories should be distinguished. There are therefore two possibilities for the extension: construct the c-trajectory or the λ-trajectory. The former involves the delicate problem at turning points because of the presence of cusps. The latter involves the problem of finding from where to where the λ-trajectory jumps.

Acknowledgment: The authors would like to thank Prof. I. Yamada of Tokyo Institute of Technology for the fruitful discussions. This work was partially supported by JSPS Grants-in-Aid (20760252).

References

[1] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[3] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Statistical Society. Siries B*, vol. 58, no. 1, pp. 267–288, 1996.

[5] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[6] D. L. Donoho and Y. Tsaig, "Fast solution of ℓ_1-norm minimization problems when the solution may be sparse," *IEEE Trans. Inform. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

[7] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Oxford University Press, 2000, vol. 191, translations of Mathematical Monograph.

[8] Y. Hirose and F. Komaki, "An extension of least angle regression based on the information geometry of dually flat spaces," *J. Computational and Graphical Statistics*, vol. 19, no. 4, pp. 1007–1023, 2010.

[9] Z. Xu, H. Zhang, Y. Wang, and X. Chang, "$L_{1/2}$ regularizer," *Science in China Series F: Inform. Sci.*, vol. 52, no. 1, pp. 1–9, Jan. 2009.

[10] Z. Xu, X. Chang, F. Xu, and H. Zhang, "$L_{1/2}$ regularization: A thresholding representation theory and a fast solver," submitted for publication.

[11] S. Boyd and L.Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004.