More Applications of Optimality Condition Theory

Yinyu Ye
Department of Management Science and Engineering
Stanford University
Stanford, CA 94305, U.S.A.

http://www.stanford.edu/~yyye

Sparse-Least-Squares: Quasi-Norm Regularization

Consider the problem:

$$\text{Minimize}_x \quad f_p(\mathbf{x}) := \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_p^p \tag{1}$$

where data $A \in R^{m \times n}, \mathbf{b} \in R^m$, and parameter $0 \leq p < 1$.

$\|\mathbf{x}\|_p$ with $0 < p < 1$ is called quasi-norm of vector $\mathbf{x}$. When $p = 0$:

$$\|\mathbf{x}\|_0^0 := \|\mathbf{x}\|_0 := |\{j : \ x_j \neq 0\}|$$

that is, the number of nonzero entries in $\mathbf{x}$.

More general model: for $q \geq 1$

$$\text{Minimize}_x \quad f_{qp}(\mathbf{x}) := \|A\mathbf{x} - \mathbf{b}\|_q^q + \lambda \|\mathbf{x}\|_p^p$$

Constrained Quasi-Norm Minimization

One may consider another related problem:

$$
\begin{array}{ll}
\text{Minimize} & \|\mathbf{x}\|_p^p = \sum_{1\le j\le n} |x_j|^p \\
\text{Subject to} & A\mathbf{x} = \mathbf{b}.
\end{array} \qquad (2)
$$

Or

$$
\begin{array}{ll}
\text{Minimize} & p(\mathbf{x}) = \sum_{1\le j\le n} x_j^p \\
\text{Subject to} & A\mathbf{x} = \mathbf{b}, \\
& \mathbf{x} \ge \mathbf{0},
\end{array} \qquad (3)
$$

Application and Motivation

The original goal is to minimize $\|\mathbf{x}\|_0^0 = |\{j : x_j \neq 0\}|$, the size of the support set of $\mathbf{x}$, for

- Sparse image reconstruction
- Sparse signal recovering
- Compressed sensing

which is known to be an NP-Hard problem.

Theory of Constrained L_p Minimization I

Theorem 1 *(The first order bound) Let $\mathbf{x}^*$ be any local minimizer of (1) and*

$$\ell_j = \left(\frac{\lambda p}{2\|\mathbf{a}_j\| \sqrt{f_p(\mathbf{x}^*)}} \right)^{\frac{1}{1-p}},$$

where $\mathbf{a}_j$ is the jth column of A. Then, the following property holds:

$$\text{for each } j, \qquad x_j^* \in (-\ell_j, \ell_j) \quad \Rightarrow \quad x_j^* = 0.$$

Moreover, the number of nonzero entries in $\mathbf{x}^*$ is bounded by

$$\|\mathbf{x}^*\|_0 \le \min\left(m, \frac{f_p(\mathbf{x}^*)}{\lambda \ell^p} \right);$$

where $\ell = \min\{\ell_j\}$.

Sketch of Proof

Let $\mathbf{x}^*$ be a local minimizer. Then it remains a minimizer after eliminating those variables whose values are zeros. For the nonzero-value variables, they must still satisfy the first-order KKT conditions:

$$2\mathbf{a}_j^T(A\mathbf{x}^* - \mathbf{b}) + \lambda p(|x_j^*|^{p-1} \cdot \mathrm{sign}(x_j^*)) = 0.$$

Thus,

$$|x_j^*|^{1-p} \geq \frac{\lambda p}{2\|\mathbf{a}_j\|\|A\mathbf{x}^* - \mathbf{b}\|} \geq \frac{\lambda p}{2\|\mathbf{a}_j\|\sqrt{f_p(\mathbf{x}^*)}}.$$

Now we show the second part of the theorem. Again,

$$\lambda\|\mathbf{x}^*\|_p^p \leq \|A\mathbf{x}^* - \mathbf{b}\| + \lambda\|\mathbf{x}^*\|_p^p = f_p(\mathbf{x}^*).$$

From the first part of this theorem, any nonzero entry of $\mathbf{x}^*$ is bounded from below by ℓ so that we have the desired result.

Theory of Constrained L_p Minimization II

Theorem 2 *(The second order bound) Let $\mathbf{x}^*$ be any local minimizer of (1), and* $\kappa_j = \left(\frac{\lambda p(1-p)}{2\|\mathbf{a}_j\|^2} \right)^{\frac{1}{2-p}}, j \in \mathcal{N}$. *Then the following property holds:*

$$\text{for each } j, \qquad x_j^* \in (-\kappa_j, \kappa_j) \quad \Rightarrow \quad x_j^* = 0.$$

Again, we remove zero-value variables from $\mathbf{x}^*$ and the remain variables must still satisfy the second-order KKT condition for a local minimizer of (1):

$$\nabla^2 f_p(\mathbf{x}) = 2A^T A - \lambda p(1-p)\text{Diag}(|x_j^*|^{p-2}) \succeq \mathbf{0}.$$

Then all diagonal entries of the Hessian must be nonnegative, which gives the proof.

Theory of Constrained L_p Minimization III

- The first-order theorem indicates that the lower the objective value, the sparser the solution cardinality bound. Also, for λ sufficiently large but finite, the number of nonzero entries in any local minimizer reduces to 0.
- The result of the second-order theorem depends only on λ and p. In practice, one would typically choose $p = 1/2$.
- The two theorems establish relations between model parameters $p,\ \lambda$ and the desired degree of sparsity of the solution. In particular, it gives a guidance on how to choose the combination of λ and p.
- Later, we would show that a second-order KKT solution of (1) would be relatively easy to compute, either in theory or practice.

Sparse Portfolio Selection: Quasi-Norm Regularization

Recall the modern portfolio selection problem:

$$\begin{array}{ll} \text{minimize} & \mathbf{x}^T V \mathbf{x} \\ \text{subject to} & \mathbf{r}^T \mathbf{x} \geq \mu, \\ & \mathbf{e}^T \mathbf{x} = 1, \ \mathbf{x} \ \geq \ \mathbf{0}, \end{array}$$

where expect-value vector $\mathbf{r}$ and co-variance matrix V are given, and $\mathbf{e}$ is the vector of all ones.

In shorting-allowed models, constraint $\mathbf{x} \ \geq \ \mathbf{0}$ is dropped; and it is replaced by $\|\mathbf{x}\|_1 \leq 1 + \delta$ for some $\delta > 0$, where δ controls the leverage of the portfolio.

But the final solution of the model are typically dense ...

A Quasi-Norm Regularized Model

We now consider

$$\begin{array}{ll} \text{minimize} & \mathbf{x}^T V \mathbf{x} + \mathbf{c}^T \mathbf{x} + \lambda \|\mathbf{x}\|_p \\ \text{subject to} & \mathbf{e}^T \mathbf{x} = 1,\ \mathbf{x} \ \geq \ \mathbf{0}, \end{array}$$

where we removed the linear expectation constraint for simplicity. Also for simplicity, we fix $p = 1/2$ in the analysis.

One may consider more complicated regularization model:

$$\begin{array}{ll} \text{minimize} & \mathbf{x}^T V \mathbf{x} + \mathbf{c}^T \mathbf{x} + \lambda \|\mathbf{x}\|_p \\ \text{subject to} & \mathbf{e}^T \mathbf{x} = 1,\ \|\mathbf{x}\| \leq 1 + \delta. \end{array}$$

Theory of the Quasi-Norm Regularized Model

Theorem 3 *(The second order theorem) Let* $\mathbf{x}^*$ *be any second-order KKT solution (after removing zero-value entries),* P^* *be the support of* $\mathbf{x}^*$ *and* $K = |P^*|$*, and* V^* *be the corresponding covariance sub-matrix. Furthermore, let*

$$\kappa_j = V^*_{jj} - \frac{2}{K}(V^*\mathbf{e})_j + \frac{1}{K^2}(\mathbf{e}^T V^*\mathbf{e}),\ j \in P^*,$$

which are the diagonal entries of matrix $\left(1 - \frac{1}{K}\mathbf{e}\mathbf{e}^T\right) V^* \left(1 - \frac{1}{K}\mathbf{e}\mathbf{e}^T\right)$*. Then the following properties hold:*

- $(K-1)K^{3/2} \leq \frac{4}{\lambda}\sum_{j\in P^*}\kappa_j$.
- *If there is* $\kappa_j = 0$*, then* $K = 1$ *and* $x^*_j = 1$*; otherwise,*

$$x^*_j \geq \left(\frac{\lambda(1-\frac{1}{K})^2}{4\kappa_j}\right)^{2/3}.$$

Proof

In the proof, we only consider variables $j \in P^*$. The second-order condition requires that the Hessian of the Lagrangian function

$$V^* - \frac{\lambda}{4}\text{Diag}\left[(x_j^*)^{-3/2}\right]$$

must be positive semidefinite in the null space of $\mathbf{e} \in R^K$. Or, the projected Hessian matrix

$$\left(I - \frac{1}{K}\mathbf{e}\mathbf{e}^T\right)\left(V^* - \frac{\lambda}{4}\text{Diag}\left[(x_j^*)^{-3/2}\right]\right)\left(I - \frac{1}{K}ee^T\right) \succeq \mathbf{0},$$

must be positive semidefinite.

Thus, the jth diagonal entry of the projected Hessian matrix

$$\kappa_j - \frac{\lambda}{4}\left((x_j^*)^{-3/2}\left(1 - \frac{2}{K}\right) + \frac{\sum_k (x_k^*)^{-3/2}}{K^2}\right) \geq 0, \qquad (4)$$

and the trace of projected Hessian matrix

$$\sum_k \kappa_k - \frac{\lambda}{4}\frac{K-1}{K}\sum_k (x_k^*)^{-3/2} \geq 0.$$

The quantity $\sum_k (x_k^*)^{-3/2}$, with $\sum_k x_k^* = 1,\ x_k^* \geq 0$ achieves its minimum at $x_k^* = 1/K$ for all k with the minimum value $K \cdot K^{3/2}$. Thus,

$$\frac{\lambda}{4}(K-1)K^{3/2} \leq \sum_k \kappa_k,$$

or

$$(K-1)K^{3/2} \leq \frac{4\sum_k \kappa_k}{\lambda},$$

which complete the proof of the first item.

Again, from (4) we have

$$\frac{\lambda}{4}\left((x_j^*)^{-3/2}\left(1-\frac{2}{K}\right)+\frac{\sum_k (x_k^*)^{-3/2}}{K^2}\right) \leq \kappa_j.$$

Or

$$\frac{\lambda}{4}\left((x_j^*)^{-3/2}\left(1-\frac{1}{K}\right)^2+\frac{\sum_{k,k\neq j}(x_k^*)^{-3/2}}{K^2}\right)\leq \kappa_j,$$

which implies

$$\frac{\lambda}{4}(x_j^*)^{-3/2}\left(1-\frac{1}{K}\right)^2\leq \kappa_j.$$

Hence, if any $\kappa_j = 0$, we must have $K = 1$ and x_j^* is the only non-zero entry in $\mathbf{x}^*$ so that $x_j^* = 1$. Otherwise, we have the desired second statement in the Theorem.

The Hardness I

Question: As for deciding the global minimal value, is $L_2 - L_p$ easier to solve than $L_2 - L_0$?

Theorem 4 *The global minimal value of either the $L_q - L_p$ or constrained L_p minimization problem is strongly NP-hard to decide for any given $0 \leq p < 1$, $q \geq 1$ and $\lambda > 0$.*

The NP-Hard Class: a class of problems don't have any "provably efficient" algorithm up to now. To prove a problem, say A, is NP-Hard, it is efficient to reduce a known NP-Hard problem B to solving problem A. In other words, if you can solve A efficiently, then you can solve problem B efficiently.

More Precise Statement

Theorem 5 *It is (strongly) NP-hard to decide the global minimal objective value of problem:*

$$\begin{array}{ll} \textit{Minimize} & p(\mathbf{x}) = \sum_{1 \le j \le n} x_j^p \\ \textit{Subject to} & A\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \ge \mathbf{0}, \end{array}$$

or

$$\begin{array}{ll} \textit{Minimize} & p(\mathbf{x}) = \sum_{1 \le j \le n} |x_j|^p \\ \textit{Subject to} & A\mathbf{x} = \mathbf{b}. \end{array}$$

Proof of the NP-hardness

An instance of the partition problem can be described as follows: given a set S of integers or rational numbers $\{a_1, a_2, \ldots, a_n\}$, is there a way to partition S into two disjoint subsets S_1 and S_2 such that the sum of the numbers in S_1 equals the sum of the numbers in S_2?

Let vector $\mathbf{a} = (a_1, a_2, \ldots, a_n) \in R^n$. Then, we consider the following reduced minimization problem in form (3):

$$\begin{array}{ll} \text{Minimize} & P(\mathbf{x}, \mathbf{y}) = \sum_{1 \le j \le n} (x_j^p + y_j^p) \\ \text{Subject to} & \mathbf{a}^T(\mathbf{x} - \mathbf{y}) = 0, \\ & x_j + y_j = 1, \ \forall j, \\ & \mathbf{x}, \mathbf{y} \ge \mathbf{0}. \end{array}$$

Proof continued

From the strict concavity of the objective function,

$$x_j^p + y_j^p \geq x_j + y_j = 1, \ \forall j,$$

and they are equal if and only if $(x_j = 1, y_j = 0)$ or $(x_j = 0, y_j = 1)$. Thus, $P(\mathbf{x}, \mathbf{y}) = n$ for any (continuous) feasible solution; and if there is a feasible solution pair $(\mathbf{x}, \mathbf{y})$ such that $P(\mathbf{x}, \mathbf{y}) \leq n$, it must be true $x_j^p + y_j^p = 1 = x_j + y_j$ for all j so that $(\mathbf{x}, \mathbf{y})$ is a binary solution, $(x_j = 1, y_j = 0)$ or $(x_j = 0, y_j = 1)$, which generates an equitable partition of the entries of $\mathbf{a}$.

On the other hand, if the entries of $\mathbf{a}$ has an equitable partition, then the reduced problem must have a binary solution pair $(\mathbf{x}, \mathbf{y})$ such that $P(\mathbf{x}, \mathbf{y}) = n$. Therefore, it is NP-hard to decide if there is a feasible solution $(\mathbf{x}, \mathbf{y})$ such that its objective value $P(\mathbf{x}, \mathbf{y}) = n$.

Proof continued

For the same partition problem, consider the following reduced minimization problem in form (2):

$$\begin{array}{ll} \text{Minimize} & \sum_{1\le j\le n} (|x_j|^p + |y_j|^p) \\ \text{Subject to} & \mathbf{a}^T(\mathbf{x}-\mathbf{y}) = 0, \\ & x_j + y_j = 1,\ \forall j. \end{array}$$

Note that this problem has no non-negativity constraints on variables $(\mathbf{x},\mathbf{y})$. However, for any feasible solution $(\mathbf{x},\mathbf{y})$ of the problem, we still have

$$|x_j|^p + |y_j|^p \ge x_j + y_j = 1,\ \forall j.$$

This is because when $x_j + y_j = 1$, the minimal value of $|x_j|^p + |y_j|^p$ is 1, and it equals 1 if and only if $(x_j = 1, y_j = 0)$ or $(x_j = 0, y_j = 1)$.