EECS289A Introduction to Machine Learning, Project T Final Note

Han Liu, Peng Tan, Dilu Xu, Jinyan Zhao
Department of Civil Engineering
Universitu of California, Berkeley
Berkeley, CA 94704
{han_liu, tanpeng, diluxu, jinyan_zhao}@berkeley.edu

December 1, 2020

1 Introduction

While neural network sheds new insights in image processing in multiple applications including, image denoting, image blurring, image restoration, etc, conventional methods revoking linear algebra still remain active for 2 reasons, 1) linear algebra approach allows closer connection between theory and applications. 2) some machine learning ideas perfectly overlap and therefore empowers conventional image processing approaches. In this note, we are going to study one particular area of image reconstruction, namely image reconstruction to build a bridge between image processing and various types of least squares (least square regression problem, constrained least square problem), and demonstrate how to train a learning-based model through solving some linear programming problems. There are large amount of background knowledge in this note and we don't expect the hypothetical students to understand them thoroughly. However, we aim to demonstrate a particular application of least squares method, specifically LASSO regression and enhance the students' understanding of least square problem by an unfamiliar topic. We believe that applying what they learned in the class to unfamiliar problems is an important skill for the students to achieve a success career in industry. The background knowledge is provided to the students and the students will find out that the tasks left for them to do are what they have learned in previous weeks and 16B classes.

2 Background

Image reconstruction is currently a very active area in digital image processing because it makes possible to overcome some of the inherent resolution limitations of low-cost imaging sensors, and generate high-resolution counterparts,

this turns out to be essentially critical in areas such as medical imaging and satellite imaging, where diagnosis or analysis from low-quality images can be extremely difficult. Conventional definition of super-resolution task is considered as an inverse problem of recovering the original high-resolution images based on reasonable assumption or prior knowledge which builds a connection between high-resolution image with the low-resolution image. Super-resolution image reconstruction is hard because it is literally an ill-posed problem, meaning that there are insufficient number of low-resolution images, unknown blurring operations and limited prior knowledge or constraints. Therefore, it is logical to propose regularization methods to stabilize the inversion of this ill-posed problem.

This section focuses on the problem of recovering the super-resolution version from a given low-resolution image, it does not intend to propose a novel method but to reproduce the algorithm from a well-elaborated paper written by Yang et al. (2010). The algorithm introduced in the paper is a learning-based method that relies on patches from the input image. However, instead of working directly with the image patch pairs sampled from high-resolution images, a compact representation was learned for both high and low-resolution image in order to capture the cooccurrence prior and to improve the speed of the algorithm. This approach is motivated by recent results in sparse signal representation, which suggests that the linear relationships among high-resolution signals can be accurately recovered from their low-dimensional projections.

The core portion of this algorithm is based on iteratively solving two convex optimization problem: an L_1 regularized least squares problem and an L_2 constrained least squares problem. Yang et al. (2010) was published on an image processing journal and therefore does not draw significant amount of attention on convex optimization, and this is where I am going to probe in more details. The remainder of this section is organized as follows: the new super-resolution algorithm would be introduced with more emphasis on the how this method is associated with particular types of convex optimization problems; an efficient sparse coding algorithm proposed by Lee et al. (2007) will be covered and specialized for super-resolution problems given before; eventually, Python and Matlab code were implemented to realize this method and a reconstruction example on X-ray tomography of granular material were discussed to verify the performance of this method.

3 Image Super-Resolution via Sparse Representation

The basic idea of image super-resolution via sparse representation is that: under certain conditions, any sufficiently sparse linear representation of a high-resolution image patch can be recovered almost perfectly form the low-resolution image patch. To be more precise, let $\boldsymbol{D} \in \mathbb{R}^{n \times K}$ be an overcomplete dictionary of K

represented features. Therefore, an image patch $\boldsymbol{x}$ can be written as $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \in \mathbb{R}^K$ is a vector with very few nonzero entries. In practice, we might only observe a small set of measurement $\boldsymbol{y}$ of $\boldsymbol{x}$.

$$\boldsymbol{y} = L\boldsymbol{D}\boldsymbol{\alpha}_0$$

where $L \in \mathbb{R}^{k \times n}$ with $k < n$ is a projection matrix. In the context of image super-resolution, $\boldsymbol{x}$ is a high-resolution image patch, while $\boldsymbol{y}$ is its low-resolution counterpart. An implementation detail here is that, as we hope to retrieve as much as useful information from corrupted images, very often high-frequency components, we could instead extract features from low-resolution images as $\boldsymbol{y}$. The remaining issue is that if the dictionary $\boldsymbol{D}$ is overcomplete, the equatioin $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}$ is underdetermined for the unknown coefficients $\boldsymbol{\alpha}$. The equation $\boldsymbol{y} = L\boldsymbol{D}\boldsymbol{\alpha}$ is even more dramatically underdetermined. It is natural to regularize this problem such that, under mild conditions, the sparsest solution $\boldsymbol{\alpha}$ to this equation will be unique. In our setting, instead of directly computing the sparse representation of the high-resolution patch, we work with two coupled dictionaries: $\boldsymbol{D}_h$ for hiigh-resolution patches, and $\boldsymbol{D}_l$ for low-resolution ones. The sparse representation of a low-resolution patch in terms of $\boldsymbol{D}_l$ will be directly used to recover the corresponding high-resolution patch from $\boldsymbol{D}_h$. It is possible to allow patch pairs to have the same sparse representation with respect to $\boldsymbol{D}_h$ and $\boldsymbol{D}_l$ by concatenating them with proper normalization and learning the resulted simultaneously.

4 Image Super-Resolution From Sparsity

Two constraints are modeled for the problem if given a low-resolution image $\boldsymbol{Y}$, recover a higher resolution image $\boldsymbol{X}$. (1) reconstruction constraints requires that the recovered $\boldsymbol{X}$ should be consistent with the input $\boldsymbol{Y}$. (2) sparsity prior assumes that the high-resolution patches can be sparsely represented in an appropriated chosen overcomplete dictionary, and their sparse representations can be recovered from the low resolution observation. Combined these two constraints, for each input low-resolution patch $\boldsymbol{y}$, the problem is to find a sparse representation with respect to $\boldsymbol{D}_l$ so that the corresponding high-resolution patch could be recovered from $\boldsymbol{D}_h$ with same coefficients. Therefore, the problem of finding the sparest representation of $\boldsymbol{y}$ can be formulated as:

$$\min ||\boldsymbol{\alpha}||_0 \quad s.t \quad ||F\boldsymbol{D}_l\boldsymbol{\alpha} - F\boldsymbol{y}||_2^2 \leq \epsilon$$

where F is a feature extraction operator. Donoho (2008) believes that as long as the desired coefficients $\boldsymbol{\alpha}$ are sufficiently sparse, they can be efficiently recovered by instead minimizing the l_1 norms:

$$\min ||\boldsymbol{\alpha}||_1 \quad s.t \quad ||F\boldsymbol{D}_l\boldsymbol{\alpha} - F\boldsymbol{y}||_2^2 \leq \epsilon$$

Or an equivalent Lagrange formulation:

$$\min_{\boldsymbol{\alpha}} ||F\boldsymbol{D}_l\boldsymbol{\alpha} - F\boldsymbol{y}||_2^2 + \lambda||\boldsymbol{\alpha}||_1$$

This is essentially a linear regression regularized with l_1 norm knowned as Lasso regression, where λ balances sparsity of the solution and fidelity of the approximation to $\boldsymbol{y}$. This is the first convex optimization problem that would be addressed here, and an efficient algorithm will be given in the next section. Solving Lasso regression separately for high-resolution patch and low-resolution patch does not guarantee the compatibility between adjacent patches, to handle this limitation, the super-resolution reconstruction $\boldsymbol{D}_h\boldsymbol{\alpha}$ of patch $\boldsymbol{y}$ is constrained to closely agree with the previously computed adjacent high-resolution patches:

$$\begin{aligned} \min ||\boldsymbol{\alpha}||_1 \quad s.t. \quad & ||F\boldsymbol{D}_l\boldsymbol{\alpha} - F\boldsymbol{y}||_2^2 \leq \epsilon_1 \\ & ||P\boldsymbol{D}_h\boldsymbol{\alpha} - \boldsymbol{w}||_2^2 \leq \epsilon_2 \end{aligned}$$

Where the matrix P extracts the region of overlap between the current target patch and previously reconstructed high-resolution image, and $\boldsymbol{w}$ contains the values of the previously reconstructed high-resolution image on the overlap. The constrained optimization can be similarly reformulated as:

$$\min_{\boldsymbol{\alpha}} ||\tilde{\boldsymbol{D}}\boldsymbol{\alpha} - \tilde{\boldsymbol{y}}||_2^2 + \lambda||\boldsymbol{\alpha}||_1$$

where $\tilde{\boldsymbol{D}} = \begin{bmatrix} F\boldsymbol{D}_l \\ P\boldsymbol{D}_h \end{bmatrix}$ and $\tilde{\boldsymbol{y}} = \begin{bmatrix} F\boldsymbol{y} \\ \boldsymbol{w} \end{bmatrix}$.

5 Dictionary Training

This learning-based algorithm is a sparse coding problem to explore sparse representations of the signals with respect to an overcomplete dictionary $\boldsymbol{D}$. A particular formulation to learn a compact dictionary while guarantees sparse representation is given:

$$\begin{aligned} \boldsymbol{D} = \arg\min_{\boldsymbol{D},Z} . & ||X - \boldsymbol{D}\boldsymbol{Z}||_2^2 + \lambda||\boldsymbol{Z}||_1 \\ & s.t.\ ||D_i||_2^2 \leq 1, i = 1, 2, \cdots, K \end{aligned}$$

where the l_1 norm $||\boldsymbol{Z}||_1$ is to enforce sparsity, and the l_2 norm constraints on the columns of $\boldsymbol{D}$ remove the scaling ambiguity (otherwise, the cost can always be reduce by dividing $\boldsymbol{Z}$ by $c > 1$ and multiplying $\boldsymbol{D}$ by $c > 1$). This constrained least square problem is the second convex optimization problem we are going to solve. Note that this is not convex if both $\boldsymbol{D}$ and $\boldsymbol{Z}$ are free to vary, but is convex in one of them with the other fixed. The optimization performs in an iteratively manner over $\boldsymbol{Z}$ and $\boldsymbol{D}$:

(1) Initialize $\boldsymbol{D}$ with a Gaussian random matrix, with each column unit normalized.
(2) Fix $\boldsymbol{D}$, update Z by

$$Z = \arg\min_{Z} ||X - \boldsymbol{D}z||_2^2 + \lambda||Z||_1$$

which can be solved efficiently through linear programming.
(3) Fix Z, update $\boldsymbol{D}$ by

$$\boldsymbol{D} = \arg\min_{\boldsymbol{D},Z} ||X - \boldsymbol{DZ}||_2^2 \quad s.t.\ ||D_i||_2^2 \leq 1, i = 1, 2, \cdots, K$$

which is quadratically constrained quadratic programming that could be solved in many optimization packages, providing objective function, gradient and Hessian.
(4) Iterate between (2) and (3) until converge.

Solving above constrained least square problem separately does not arrive the same coefficient for high-resolution patch and low-resolution patch. Instead, we can combine the learning objectives and force the high-resolution and low-resolution representations to share the same codes:

$$\min_{\boldsymbol{D}_h,\boldsymbol{d}_l,Z} \frac{1}{N}||X^h - \boldsymbol{D}_h Z||_2^2 + \frac{1}{M}||Y^l - \boldsymbol{D}_l Z||_2^2 + \lambda(\frac{1}{N} + \frac{1}{M})||Z||_1$$

where N and M are the dimensions of the high-resolution and low-resolution image patches in vector form or equivalently:

$$\min_{\boldsymbol{D}_h,\boldsymbol{D}_l,Z} ||X_c - \boldsymbol{D}_c Z||_2^2 + \hat{\lambda}||Z||_1$$

where $X_c = \begin{bmatrix} \frac{1}{\sqrt{N}} X^h \\ \frac{1}{\sqrt{M}} Y^l \end{bmatrix}$ and $\boldsymbol{D}_c = \begin{bmatrix} \frac{1}{\sqrt{N}} \boldsymbol{D}_h \\ \frac{1}{\sqrt{M}} \boldsymbol{D}_l \end{bmatrix}$ Thus the same learning strategy outlined before can be used to train the two dictionaries simultaneously.

6 An Efficient Sparse Coding Algorithm

6.1 L_1-Regularized Least Squares

The feature-sign search algorithm is presented here, which is aimed to solve the following equivalent optimization problem:

$$\min_x ||y - Ax||^2 + \gamma||x||_1$$

If we know the signs (positive, zero or negative) of the x_i at the optimal value, we can replace each of the term $||x_i||_1$ with either x_i if $(x_i > 0)$, $-x_i$ (if $x_i < 0$), or 0 (if $x_i = 0$). Consider only nonzero coefficients, this reduces to a standard, unconstrained quadratic optimization problem, which can be solved analytically and efficiently. The presented algorithm therefore attempts to search or guess iteratively the signs of the coefficients x_i.

Conclusion 1: if the current coefficients x_c are consistent with the active set and sign vector, but are not optimal for the problem at the start of step 3, the feature-sign step is guaranteed to strictly reduce the objective. *Proof sketch*:

Algorithm 1 Feature-Sign Search Algorithm

1. Initialize $x = \mathbf{0}$, $\theta = \mathbf{0}$ and *active set* $\{\}$, where $\theta_i \in \{-1, 0, 1\}$ denotes $\text{sign}(x_i)$.

2. From zero coefficients of x, select $i = \arg\max_i |\frac{\partial ||y-Ax||^2}{\partial x_i}|$. Activate x_i (add i to the activate set) only if it locally improves the objective, namely:

if $\frac{\partial ||y-Ax||^2}{\partial x_i} > \gamma$, then set $\theta_i = -1$, *active set* $= \{i\} \cup$ *active set*.

if $\frac{\partial ||y-Ax||^2}{\partial x_i} < -\gamma$, then set $\theta_i = 1$, *active set* $= \{i\} \cup$ *active set*.

3. Feature-sign step:

Let $\hat{A}$ be a submatrix of A that contains only the columns corresponding to the *active set*.

Let $\hat{x}$ and $\hat{\theta}$ be subvectors of x and θ correpondinig to the *active set*.

Compute the analytical solution to the resulting unconstrained quadratic programming. $\hat{x}_{new} = (\hat{A}^T\hat{A})^{-1}(\hat{A}^T y - \gamma\frac{\hat{\theta}}{2})$.

Perform a discrete line search on the closed line segment from $\hat{x}$ to $\hat{x}_{new}$:

Check the objective value at $\hat{x}_{new}$ and all points where any coefficient changes sign.

Update $\hat{x}$ to the point with the lowest objective value.

Remove zero coefficients of $\hat{x}$ from the activate set and update $\theta = \text{sign}(x)$.

4. Check the optimality conditions:

(a) Optimality condition for nonzero coefficients: $\frac{\partial ||y-Ax||^2}{\partial x_j} + \gamma\text{sign}(x_j) = 0$, $\forall x_j \neq 0$. If condition (a) is not satisfied, go to Step 3; else check condition (b).

(b) Optimality condition for zero coefficients: $|\frac{\partial ||y-Ax||^2}{\partial x_j}| \leq \gamma$, $\forall x_j = 0$. If condition (b) is not satisfied, go to Step 2; otherwise return x as the solution.

consider a smooth quadratic function $f(x) = ||y - Ax||^2 + \gamma\theta^T x$, since x_c is not an optimal point of f, then $f(x_{new}) < f(x_c)$. (i) if x_{new} is consistent with the given active set and sign vector, x_{new} strictly decreases the objective. (ii) if x_{new} is not consistent with the given active set and sign vector, let x_d beiing the zero-crossing point between x_{new} and x_c through linear interpolation, then $x_d < x_c$ by convexity of f. Therefore, the discrete line search described in Step 3 ensures a decrease in the objective value.

Conclusion 2: if the current coefficients x_c are optimal at start of Step 2, but are not optimal for the problem (optimality condition (a) is satisfied, while optimality condition (b) is not), the feature-sign step is guaranteed to strictly reduce the objective. *Proof sketch*: from the optimality condition (b), there is some i, such that $|\frac{\partial||y-Ax||^2}{\partial x_i}| > \gamma$, this i-th coefficient is activated and i is added to the *activate set*. Since all other coefficients satisfy optimality condition (a), the gradient of objective function has direction consistent with the sign of activated x_i. Therefore, the line search direction from x_c to x_{new} at feature-sign step is also coherent with the sign of the activated x_i. From conclusion 1, either x_new is consistent, or the first zero-crossing from x_c to x_{new} has lower objective value.

6.2 Solve Constrained Least Squares with Lagrange Dual

After Z is solved (each column of Z is the solution of a L_1 regularized least squares), a method is presented here for solving optimization problem over D, this will be useful to alternatively find out compact dictionary with sparse representation Z. Following problem will be discussed in this section:

$$\min_D ||X - DZ||_F^2$$
$$s.t. \sum_{i=1}^{k} D_{i,j}^2 \leq c, \forall j = 1, \cdots, n.$$

Or equivalently consider the Lagrangian:

$$\mathcal{L}(D, \boldsymbol{\lambda}) = \text{trace}((X - DZ)^T(X - DZ)) + \sum_{j=1}^{n} \boldsymbol{\lambda}_j (\sum_{i=1}^{k} D_{i,j}^2 - c)$$

where each $\boldsymbol{\lambda}_j \geq 0$ is a dual variable. Minimizing over D analytically, we obtain the Lagrange dual:

$$\mathcal{D}(\boldsymbol{\lambda}) = \min_D \mathcal{L}(D, \boldsymbol{\lambda}) = \text{trace}(X^T - XZ^T(ZZ^T + \Lambda)^{-1}(XZ^T)^T - c\Lambda)$$

where $\Lambda = \text{diag}(\boldsymbol{\lambda})$. The gradient and Hessian of $\mathcal{D}(\boldsymbol{\lambda})$ are computed as follows:

$$\frac{\mathcal{D}(\boldsymbol{\lambda})}{\partial \lambda_i} = ||XZ^T(ZZ^T + \Lambda)^{-1}e_i||^2 - c$$

$$\frac{\partial^2 \mathcal{D}(\boldsymbol{\lambda})}{\partial \lambda_i \partial \lambda_j} = -2(ZZ^T + \Lambda)^{-1}(XZ^T)^T XZ^T (ZZ^T + \Lambda)^{-1})_{i,j}((ZZ^T + \Lambda)^{-1})_{i,j}$$

where $e_i \in \mathbb{R}^n$ is the i-th unit vector. Now, the Lagrange dual problem could be optimized using Newton's method or conjugate gradient. After maximizing $\mathcal{D}(\boldsymbol{\lambda})$, the optimum dictionary D is obtained:

$$D^T = (ZZ^T + \Lambda)^{-1}(XZ^T)^T$$

.

7 Experimental Results

In this section, we demonstrated the super-resolution results obtained by applying the previously mentioned methods on X-ray tomography images on granular material. This is an excellent research candidate because it has highly repetitive, regular patterns, hence requires only a small batch of training images. In image super-resolution, the patch size as 5×5 pixels for both low- and high-resolution images. The two dictionaries for high-resolution and low-resolution image patches are trained from 10000 patch pairs randomly sampled from an X-ray tomography dataset, the low-resolution images were generated via bicubic interpolation using high-resolution counterparts. For high-resolution images, the image patches were randomly cropped as a $(25, 1)$ vector; for low-resolution images, feature extractor filters, i.e. gradient and Laplacian were first applied in both horizontal and vertical direction to discard smooth parts and leave high-frequency textured regions (this results an image patch of size $(100, 1)$). Therefore, the input training data $X \in \mathbb{R}^{125 \times 10000}$, where The dictionary size is initialized as 512 while might be smaller in the end because columns with zero norm would be abandoned. The choice of λ depends upon the level of noise in the input image, for this experiment, setting $\lambda = 0.2$ would generally yield satisfactory results. The effects of λ on the recovered image given the input is that: the noisier the data, the larger the value of λ should be. The results were evaluated both via visual inspection and qualitative Root Mean Square Error (RMSE). Both Python and Matlab codes are attached as Appendix, it is better to be written in a more consistent manner in one single programming language, and I did so in Python. The problem I encountered is Python (**scipy.optimize.minimize**) does not provide robust quadratic programming solvers for the constrained least square problem mentioned in previous sections. However, it can be very effectively solved using **fmincon** routine with Matlab. So, the Matlab scripts were used to learn and obtain compact dictionaries, and Python scripts were used for super-resolution in practice with the trained dictionaries. For the sake of completeness, I also implemented Python counterparts for the learning stage and might come back and embark on programming solvers if I am free at some time. Majority of code is re-written based on Yang et al. (2010) to fit the purpose of this project, the feature-sign algorithm and constrained least square optimization algorithm were implemented based on Lee et al. (2007).

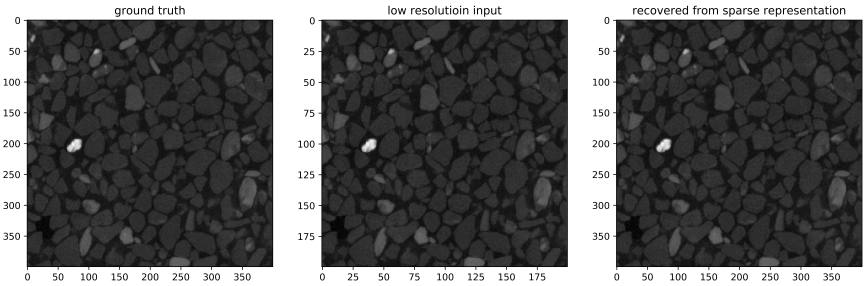

Figure 1: performance of super-resolution with sparse representation

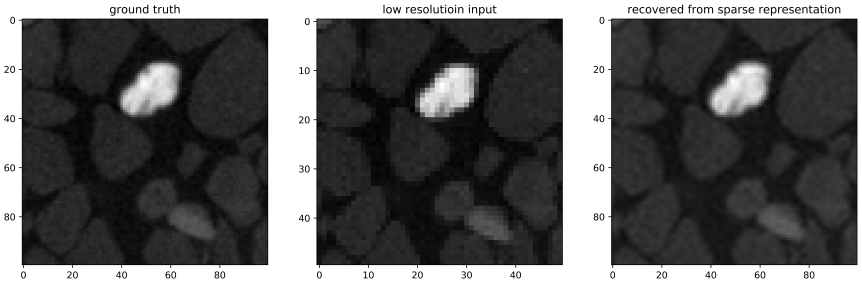

Figure 2: zoomed in: performance of super-resolution with sparse representation

The sparse representation method has satisfactory performance as there are visually no differences between the ground true high-resolution image and the recovered one. The RMSE is obtained after the image is whitened, RMSE for recovered image via sparse representation method is **2.98**, and RMSE for linear bicubic interpolation is **3.16**.

8 Conclusion

This note presented a novel approach toward single image super-resolution based upon sparse representations in terms of coupled dictionaries jointly trained from high-and low-resolution image patch pairs. Two convex optimization problems are associated, the sparse coding representations are obtained by solving Lasso regularization and the dictionaries were optimized by solving a constrained least square problem. These two convex optimization problems are solved alternatively until converge. Experimental results demonstrate the effectiveness of the sparsity as a prior for patch-based super-resolution task. The results were presented to check whether the algorithm is successfully implemented or not and it is not aimed to tune hyper-parameters like the optimal dictionary size and magnitude of λ.

The students are not required to remember the method of pre-processing of images but they should appreciate the importance of feature engineering in machine learning problems. Similarly, the students are not required to remember the exact method of the convex optimization algorithms we used in this note, however, understanding the general concept of optimization will be very useful for their later study.

References

[1] Yang, J., Wright, J., Huang, T. S., & Ma, Y. (2010). Image super-resolution via sparse representation. IEEE transactions on image processing, 19(11), 2861-2873.

[2] Lee, H., Battle, A., Raina, R., & Ng, A. Y. (2007). Efficient sparse coding algorithms. In Advances in neural information processing systems (pp. 801-808).