Shrinkage Without Thresholding: ℓ_1 Norm Minimization using the Landweber Iteration

Andrew E. Yagle
Department of EECS, The University of Michigan, Ann Arbor, MI 48109-2122

Abstract— **We use the Landweber iteration to compute sparse solutions to the underdetermined linear system of equations y=Ax using iterative reweighted least squares (IRLS). We show that shrinkage, not thresholding, is the key to minimizing the LASSO functional. We also show how the Landweber iteration without thresholding can be used instead of linear programming for basis pursuit, and how to accelerate convergence of the Landweber iteration in this case.**

Keywords— **Sparse reconstruction**
Phone: 734-763-9810. Fax: 734-763-1503.
Email: aey@eecs.umich.edu. **EDICS: 2-REST.**

I. INTRODUCTION

A. Problem Statement

The goal is to compute a sparse solution to the underdetermined (M<<N) linear system of equations

$$y = Ax; \quad \mathbf{A}: M \times N; \quad \mathbf{x}: \text{N vector}; \quad \mathbf{y}: \text{M vector}.$$

If A has random entries, then ℓ_1 norm minimization of x tends to sparsify the solution x to $y = Ax$ as much as possible. This is currently of great interest in compressed sensing, since many real-world signals and images have sparse (mostly zero) representations in an appropriate basis, such as a set of wavelet or curvelet basis functions. Another case we consider in this paper is A consists of randomly-selected rows k of the DFT matrix F_{kn}=$(1/\sqrt{N})$exp(−j2πnk/N).

Sparse solutions can be computed using *matching pursuit*, in which the columns of A most highly correlated with $y - A\hat{x}_i$ are successively chosen to minimize the residual $y - A\hat{x}_{i+1}$. Here $\hat{x}_i$ is the approximation to x after i iterations. This produces a sparse solution that approximately satisfies $y = Ax$, but it does not always find the optimally sparse solution.

The ℓ_1-norm solution can be computed using *basis pursuit (BP)*, in which linear programming is used to minimize $||x||_1$. This has been done in the geophysics literature since the 1960s, for deconvolution of sparse spike trains in reflection seismology.

The LASSO functional can be used when zero-mean Gaussian white noise is present in the data. The minimum LASSO solution can be computed using gradient methods, coordinate descent, thresholded Landweber iteration, or iterative reweighted-norm least-squares. Here we combine the latter two.

B. Review of Basic Landweber Iteration

The Landweber iteration solves $y = Ax$ iteratively:

$$x^{N+1} = x^N + A^H(y - Ax^N); \quad x^0 = 0. \quad (1)$$

Also called a Van Cittert iteration, this formula is derived as a Neumann series applied to $(I - A^H A)$.

To analyze it, rewrite the Landweber iteration as

$$x^{N+1} = (I - A^H A)x^N + A^H y. \quad (2)$$

This iteration can be rewritten as the summation

$$x^N = \sum_{i=0}^{N-1} (I - A^H A)^i A^H y. \quad (3)$$

This summation can be written in closed form as

$$\begin{aligned} x^N &= [I - (I - A^H A)^N][I - (I - A^H A)]^{-1} A^H y \\ &= [I - (I - A^H A)^N][(A^H A)^{-1} A^H y]. \end{aligned} \quad (4)$$

If A has the SVD $A = U \,\text{diag}[\sigma_n] V$, then $A^H A$ has the eigendecomposition $A^H A = V^H \,\text{diag}[\sigma_n^2] V$ and

$$Vx^N = \text{diag}[1 - (1 - \sigma_n^2)^N] V[(A^H A)^{-1} A^H y]. \quad (5)$$

From this formula the following facts are evident:

- Vx^N components converge as the rising-exponential function $1 - (1 - \sigma_n^2)^N \to 1$ as $N \to \infty$ if $0 < \sigma_n^2 < 2$;
- Convergence is slow for components $\sigma_n^2 \approx 0$ or 2;
- If the squared singular values $\sigma_n^2 < 2$, then x^N converges to least-squares solution $(A^H A)^{-1} A^H y$;
- If some $\sigma_n = 0$, then x^N converges to $A^* y$, where A^* is an appropriate pseudoinverse of A;
- If some $\sigma_n = 1$, then the associated singular vector component of x^N converges in one iteration;
- If each column of A is normalized to one, this is equivalent to coordinate descent without updating.

The convergence condition $\sigma_n^2 < 2$ can be guaranteed by scaling $y = Ax$ to y/a=$(A/a)x$ with constant a

$$a = \sum\sum |A_{ij}|^2 = Tr[A^H A] = \sum \sigma_n^2 > \sigma_n^2. \quad (6)$$

II. Regularized Landweber Iteration

A. Tikhonov Regularization

We wish to have weighted quadratic regularization:

$$\min_x \{||y - Ax||_2^2 + \lambda^2 ||Dx||_2^2\} \tag{7}$$

where the penalty term stabilizes the solution.

This is equivalent to the least-squares solution to

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} A \\ \lambda D \end{bmatrix} x \tag{8}$$

which is easily seen to be the pseudoinverse

$$\hat{x} = (A^H A + \lambda^2 D^H D)^{-1} A^H y. \tag{9}$$

Applying the Landweber iteration to this gives

$$\begin{aligned} x^{N+1} &= x^N + [\,A^H \quad \lambda D^H\,] \left(\begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} A \\ \lambda D \end{bmatrix} x^N \right) \\ &= x^N + A^H(y - Ax^N) - \lambda^2 (D^H D) x^N. \end{aligned} \tag{10}$$

The sole effect is subtraction of an additional term.

B. LASSO Cost Functional Regularization

The LASSO functional is a sum of mixed norms

$$\min_x \{||y - Ax||_2^2 + \lambda ||x||_1\} \tag{11}$$

where the penalty term sparsifies the solution.

We may apply the regularized Landweber to the iterative reweighted least squares (IRLS) functional

$$\min ||y - Ax||_2^2 + \lambda \sum x_n^2 / |x_n^N| \tag{12}$$

making D a diagonal matrix with $D_n = \sqrt{\lambda}/\sqrt{|x_n^N|}$. The regularized Landweber iteration then becomes

$$x^{N+1} = x^N + A^H(y - Ax^N) - \lambda \,\mathrm{sign}[x^N]. \tag{13}$$

This is the Landweber iteration with shrinkage, but without thresholding. Normally *soft thresholding* is used with the Landweber iteration, which subtracts $\lambda \,\mathrm{sign}[x^{N+1}]$ (shrinkage), but also sets $x^{N+1}{=}0$ if $|x^{N+1}| < \lambda$ (thresholding). Here, we only subtract $\lambda \,\mathrm{sign}[x^N]$ (shrinkage of x^N) and do not threshold.

To avoid problems when $x_n^N{=}0$, we should replace $|x_n^N|$ with $|x_n^N|{+}\epsilon$ above for very small $\epsilon > 0$. This rounds off the LASSO term $||x||_1$ slightly at $x{=}0$. The Landweber-with-shrinkage iteration still converges since $\lambda \,\mathrm{sign}[x^N]$ is a bounded driving term.

C. Basis Pursuit Regularization

In basis pursuit we minimize $||x||_1$ subject to the constraint $y = Ax$. This can be viewed as a LASSO functional as $\lambda \to 0$. This can be solved using linear programming, or by using IRLS as described next.

The minimum-norm least-squares solution to the underdetermined system $y = Ax$ has the properties:

- The solution x is the one minimizing $||x||_2^2$;
- $x = A^H (AA^H)^{-1} y$ (another pseudoinverse);
- x can also be computed by applying a Landweber iteration to $y = Ax$, as noted above.

We can minimize $||x||_1$ by rewriting $y = Ax$ as

$$y = Ax = (AD)(D^{-1}x), \; D = \mathrm{diag}[\sqrt{|x_n|}]. \tag{14}$$

A minimum-norm solution to $y{=}(AD)z$ minimizes

$$||D^{-1}x||_2^2 = \sum x_n^2 / |x_n| = ||x||_1 \tag{15}$$

so applying the Landweber iteration to $y = (AD)z$ and setting $x = D^{-1}z$ computes x minimizing $||x||_1$.

D. Basis Pursuit Acceleration

The Landweber iteration applied to y=$(AD)z$ has its convergence behavior determined by eigenvalues of I–$(AD)^H(AD)$. These are in turn the eigenvalues of I–$A \,\mathrm{diag}[|x_n^N|] A^H$ augmented with ones, since the matrices AB and BA have same nonzero eigenvalues.

This creates a problem, since as many $x_n^N \to 0$ matrix $A \,\mathrm{diag}[|x_n^N|] A^H$ becomes close to singular, and convergence of the Landweber iteration will be slow. So we again replace $|x_n^N|$ with $|x_n^N|{+}\epsilon$ above for very small $\epsilon > 0$. This bounds the matrix $A \,\mathrm{diag}[|x_n^N|] A^H$ away from singularity, accelerating convergence.

E. Basis Pursuit Acceleration: Random Frequencies

Let A be randomly-chosen rows of the DFT matrix F_{kn}=$(1/\sqrt{N})$exp(–j2πnk/N), Then $AA^H = I$ and

$$A \,\mathrm{diag}[|x_n^N| + \epsilon] A^H = A \,\mathrm{diag}[|x_n^N|] A^H + \epsilon I. \tag{16}$$

By the Cauchy interlace theorem, the eigenvalues of $A \,\mathrm{diag}[|x_n^N|] A^H$ will (in this case) interlace those of $F \,\mathrm{diag}[|x_n^N|] F^H$, which are $\{|x_n^N|\}$. Hence the components Vx^N will converge as $1 - (1 - \sigma_n^2)^N$, where

$$0 < \epsilon < \sigma_n^2 < \epsilon + \max |x_n^N|. \tag{17}$$

The closer σ_n^2 are to unity, the faster the Landweber iteration will converge. So including even a small ϵ speeds up convergence. And the iteration can be initialized by setting unobserved frequencies to zero, and then computing the inverse DFT of the result.