Multiplicative Updates for Nonnegative Least Squares

Donghui Chen

School of Securities and Futures
Southwestern University of Finance and Economics

November 18, 2013

Joint work with Matt Brand, Mitsbushi Electronic Research Lab *what really matters is the wisdom he teaches you, ...*

– Sofia Pauca

Outline

Objective function

Nonnegative Least Squares

$$\underset{x}{\operatorname{argmin}}\ F(x) = \underset{x}{\operatorname{argmin}}\ ||Ax - b||_2^2 \quad s.t. \quad x \geq 0, \tag{1}$$

Objective function

Nonnegative Least Squares

$$\underset{x}{\operatorname{argmin}}\ F(x) = \underset{x}{\operatorname{argmin}}\ ||Ax - b||_2^2 \quad s.t. \quad x \geq 0, \tag{1}$$

Because

$$\begin{aligned} ||Ax - b||_2^2 &= (Ax - b)^T(Ax - b) \\ &= x^T(A^TA)x - \underbrace{b^T(Ax)}_{\text{scalar}} - \underbrace{(Ax)^Tb}_{\text{scalar}} + \underbrace{b^Tb}_{\text{constant}} \\ &= x^T(A^TA)x - x^T(A^Tb) - x^T(A^Tb) + b^Tb \\ &= x^T(A^TA)x - 2x^T(A^Tb) + b^Tb \end{aligned}$$

Objective function

Nonnegative Least Squares

$$\operatorname*{argmin}_x F(x) = \operatorname*{argmin}_x ||Ax - b||_2^2 \quad s.t. \quad x \geq 0, \tag{1}$$

Because

$$\begin{aligned}
||Ax - b||_2^2 &= (Ax - b)^T(Ax - b) \\
&= x^T(A^TA)x - \underbrace{b^T(Ax)}_{\text{scalar}} - \underbrace{(Ax)^Tb}_{\text{scalar}} + \underbrace{b^Tb}_{\text{constant}} \\
&= x^T(A^TA)x - x^T(A^Tb) - x^T(A^Tb) + b^Tb \\
&= x^T(A^TA)x - 2x^T(A^Tb) + b^Tb
\end{aligned}$$

Hence, solving Equation (1) is equivalent to solving

$$\operatorname*{argmin}_x F(x) = \operatorname*{argmin}_x \frac{1}{2}x^TQx - x^Th \quad s.t. \quad x \geq 0, \tag{2}$$

with $Q = A^TA$ and $h = A^Tb$.

Multiplicative NNLS Iteration

Theorem (Multiplicative NNLS Iteration)

Nonnegative least squares objective function $F(x)$ *in Equation* (2) *is monotonically decreasing under the multiplicative update*

$$x_i^{k+1} = x_i^k \left[\frac{2(Q^- x^k)_i + h_i^+ + \delta}{(|Q| x^k)_i + h_i^- + \delta} \right], \tag{3}$$

with $\delta > 0$, $Q^- = -\min(Q, 0)$, $|Q| = \mathit{abs}(Q)$, $h^+ = \max(h, 0)$, $h^- = -\min(h, 0)$.

M. E. Daube-Witherspoon, G. Muehllehner, in *IEEE Trans. on Medical Imaging*, 1986.

D. Lee, S. Seung, in *Nature*, 1999

Multiplicative NNLS Iteration

Theorem (Multiplicative NNLS Iteration)

Nonnegative least squares objective function $F(x)$ *in Equation* (2) *is monotonically decreasing under the multiplicative update*

$$x_i^{k+1} = x_i^k \left[\frac{2(Q^- x^k)_i + h_i^+ + \delta}{(|Q|x^k)_i + h_i^- + \delta} \right], \quad (3)$$

with $\delta > 0$, $Q^- = -\min(Q, 0)$, $|Q| = \textit{abs}(Q)$, $h^+ = \max(h, 0)$, $h^- = -\min(h, 0)$.

Remark: If Q and h have only nonnegative components and $\delta = 0$, above iteration reduces to

$$x_i^{k+1} = x_i^k \left[\frac{h_i}{(Qx^k)_i} \right],$$

which is called image space reconstruction algorithm (ISRA). Lee ad Seung generalize the ISRA idea to NMF.

M. E. Daube-Witherspoon, G. Muehllehner, in *IEEE Trans. on Medical Imaging*, 1986.

D. Lee, S. Seung, in *Nature*, 1999

Gradient Descent Property

The multiplicative update (3) is an element-wise iterative gradient descent method.

$$
\begin{aligned}
x_i^{k+1} - x_i^k &= \left[\frac{2(Q^- x^k)_i + h_i^+ + \delta}{(|Q|x^k)_i + h_i^- + \delta}\right] x_i^k - x_i^k \\
&= \left[\frac{2(Q^- x^k)_i + h_i^+ - (|Q|x^k)_i - h_i^-}{(|Q|x^k)_i + h_i^- + \delta}\right] x_i^k \\
&= -\left[\frac{(Qx^k)_i - h_i}{(|Q|x^k)_i - h_i^- + \delta}\right] x_i^k \\
&= -\left[\frac{x_i^k}{(|Q|x^k)_i - h_i^- + \delta}\right] ((Qx^k)_i - h_i) \\
&= -\gamma_k \nabla(F(x^k)),
\end{aligned}
$$

where the step-size $\gamma_k = \left[\frac{x_i^k}{(|Q|x^k)_i - h_i^- + \delta}\right]$, and $\nabla(F(x)) = Qx^k - h$.

What if $\delta = 0$?

Suppose

$$Q = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad h = 0,$$

with initial guess,

$$\begin{aligned} x^0 &= (\frac{2}{3}, \frac{4}{3}), \\ x^1 &= (\frac{4}{3}, \frac{2}{3}), \\ x^2 &= (\frac{2}{3}, \frac{4}{3}), \quad \cdots \end{aligned}$$

However, the optimal solution is

$$x^* = (r, r), r \in \mathcal{R}.$$

iterations by (3) with $\delta = 0$

Positive δ

Suppose

$$Q = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad h = 0,$$

with initial guess,

$$\begin{aligned} x^0 &= (\frac{2}{3}, \frac{4}{3}), \\ &\vdots \\ x^\infty &= (1, 1), \end{aligned}$$

iterations by (3) with $\delta = 1$

Convergence Analysis

Definition (Auxiliary Function)

For positive vectors, x, y, an auxiliary function, $G(x, y)$, of $F(x)$, has the following two properties

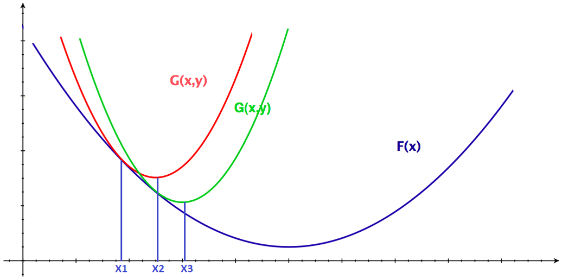
- $F(x) < G(x, y)$ *if* $x \neq y$;
- $F(x) = G(x, x)$

Convergence Analysis

Definition (Auxiliary Function)

For positive vectors, x, y, an auxiliary function, $G(x,y)$, of $F(x)$, has the following two properties

- $F(x) < G(x,y) \quad \textit{if} \quad x \neq y$;
- $F(x) = G(x,x)$



Convergence Analysis contd.

Lemma

Assume $G(x, y)$ *is an auxiliary function of* $F(x)$*, then* $F(x)$ *is strictly decreasing under the update*

$$x^{k+1} = \operatorname*{argmin}_x G(x, x^k),$$

if and only if $x^{k+1} \neq x^k$.

Convergence Analysis contd.

Lemma

Assume $G(x, y)$ *is an auxiliary function of* $F(x)$*, then* $F(x)$ *is strictly decreasing under the update*

$$x^{k+1} = \operatorname*{argmin}_x G(x, x^k),$$

if and only if $x^{k+1} \neq x^k$*.*

Proof:
By the definition of an auxiliary function $G(x, y)$, if $x^{k+1} \neq x^k$, we have

$$F(x^{k+1}) < G(x^{k+1}, x^k) \leq G(x^k, x^k) = F(x^k).$$

The equality attains if and only if $x^{k+1} = x^k$.

□

Convergence Analysis contd.

Lemma

For any positive vectors, x, y, *define the diagonal matrix,* $D(y)$, *with diagonal element*

$$D_{ii} = \frac{(|Q|y)_i + h_i^- + \delta}{y_i}, \quad i = 1, 2, \cdots, n$$

where $\delta > 0$. *The function*

$$G(x, y) = F(y) + (x - y)^T \nabla F(y) + \frac{1}{2}(x - y)^T D(y)(x - y)$$

is an auxiliary function for

$$F(x) = \frac{1}{2}x^T Qx - x^T h.$$

Review

Theorem (Multiplicative NNLS Iteration)

Nonnegative least squares objective function $F(x)$

$$\underset{x}{\operatorname{argmin}}\ F(x) = \underset{x}{\operatorname{argmin}}\ \frac{1}{2} x^T Q x - x^T h \quad s.t. \quad x \geq 0,$$

is monotonically decreasing under the multiplicative update

$$x_i^{k+1} = x_i^k \left[\frac{2(Q^- x^k)_i + h_i^+ + \delta}{(|Q| x^k)_i + h_i^- + \delta} \right],$$

with $\delta > 0$, $Q^- = -\min(Q, 0)$, $|Q| = \textit{abs}(Q)$, $h^+ = \max(h, 0)$, $h^- = -\min(h, 0)$.

Review contd.

Suppose

$$Q = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad h = 0,$$

with initial guess,

$$\begin{aligned} x^0 &= (\frac{2}{3}, \frac{4}{3}), \\ &\vdots \\ x^\infty &= (1, 1), \end{aligned}$$

iterations by (3) with $\delta = 1$

Sparse Solution?

If a sparse solution is expected, it is recommended to add a regularization term to the original least squares problem,

$$\underset{x}{\operatorname{argmin}}\ \hat{F}(x) = \underset{x}{\operatorname{argmin}}\ ||Ax - b||_2^2 + \lambda||x||_1, \quad x \geq 0, \lambda > 0 \qquad (4)$$

with nonnegative λ as the regularization parameter.

Sparse Solution?

If a sparse solution is expected, it is recommended to add a regularization term to the original least squares problem,

$$\underset{x}{\operatorname{argmin}}\ \hat{F}(x) = \underset{x}{\operatorname{argmin}}\ ||Ax - b||_2^2 + \lambda||x||_1, \quad x \geq 0, \lambda > 0 \qquad (4)$$

with nonnegative λ as the regularization parameter.

Theorem

The objective function $\hat{F}(x)$ in (4) is monotonically decreasing under the multiplicative update

$$x_i^{k+1} = x_i^k \left[\frac{2(Q^- x^k)_i + h_i^+}{(|Q|x^k)_i + h_i^- + \lambda} \right], \qquad (5)$$

with $\lambda > 0$.

Sparse Solution cont.

Suppose

$$Q = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad h = 0,$$

with initial guess,

$$\begin{aligned} x^0 &= (\frac{2}{3}, \frac{4}{3}), \\ &\vdots \\ x^\infty &= (0, 0), \end{aligned}$$

iterations by (5) with $\lambda = 2$

Image Labelling

$$f(x) := \sum_{a=1}^{K} \sum_{i} \left(\frac{\eta}{2} \sum_{j \in \mathcal{N}(i)} \omega_{ij} (x_{ia} - x_{ja})^2 + d_{ia} x_{ia} \right)$$

with constraints

$$\forall i, \quad \sum_{a=1}^{K} x_{ia} = 1, \quad x_{ia} \geq 0,$$

- x_{ia} is the probability of pixel i belongs to labelling set a
- K is the number of labelling sets
- ω_{ij} is the weight between adjacent pixel i and j,

$$\omega_{ij} := \frac{I_i^T I_j}{|I_i| \cdot |I_j|} = \cos(\theta),$$

  where $I.$ is the image value
- $\mathcal{N}(i)$ represents the neighbours of pixel i
- η is a parameter controlling the spatial smoothness
- d_{ia} is the cost of label a at each pixel

M. Rivera, O. Dalmau, and J. Tago, in *ICPR*, pp.1-5, 2008.

Image Labelling: Matrix d

- Mixture Gaussian
  - Assume the data points were drawn from N independent Gaussian distributions with mean μ_l and covariance Σ_l.
  - Compute the Mahalanobis distance between each pixel i and these Gaussian distributions.

$$d_{ia} = \sum_l (x_i - \mu_{la})^T \Sigma_{la}^{-1} (x - \mu_{la}) + \log(\Sigma_{la})$$

C. Chang, C. Lin, LIBSVM, 2001.

Image Labelling: Matrix d

- Mixture Gaussian
  - Assume the data points were drawn from N independent Gaussian distributions with mean μ_l and covariance Σ_l.
  - Compute the Mahalanobis distance between each pixel i and these Gaussian distributions.

$$d_{ia} = \sum_l (x_i - \mu_{la})^T \Sigma_{la}^{-1} (x - \mu_{la}) + \log(\Sigma_{la})$$

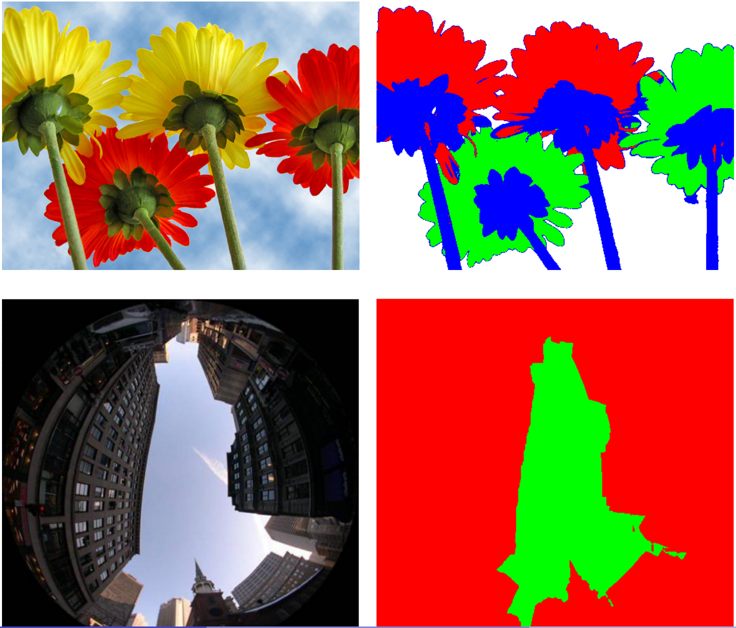
- Support Vector Machine (SVM)
  - Using SVM to find the support vectors for each labelling set.
  - Compute the decision function.

$$d_{ia} = \sum_l \alpha_{la} K(x_i, SV_{ia}) + b_a,$$

where $K(*,*)$ is the kernel function in SVM, α_{la} is the coefficients, and b_a is the bias for labelling set a.

C. Chang, C. Lin, LIBSVM, 2001.

Image Labelling contd.



Conclusion

Introduced a new algorithm along with its convergence analysis for the NNLS problem

$$\underset{x}{\operatorname{argmin}}\ F(x) = \underset{x}{\operatorname{argmin}}\ ||Ax - b||_2^2 \quad s.t. \quad x \geq 0,$$

$$x_i^{k+1} = x_i^k \left[\frac{2(Q^- x^k)_i + h_i^+ + \delta}{(|Q|x^k)_i + h_i^- + \delta} \right],$$

where $Q = A^T A$ and $h = A^T b$.

Happy Birthday, Bob!

Happy Birthday, Bob!