Non-Negative Matrix Factorization: Derivation of Update Rules & Convergence Proof

Following Lee & Seung (1999, 2001)

Reference Notes — Matt Jacob

Contents

1 Problem Setup

We are given a non-negative data matrix $V \in \mathbb{R}_{\geq 0}^{n \times m}$. We seek non-negative matrix factors

$$W \in \mathbb{R}_{\geq 0}^{n \times r}, \qquad H \in \mathbb{R}_{\geq 0}^{r \times m}$$

such that $V \approx WH$. The rank r is chosen so that $(n+m)r \ll nm$, making WH a compressed representation of V.

Key Insight

The non-negativity constraints $W \geq 0$, $H \geq 0$ are what distinguish NMF from PCA or SVD. They force the factorization to use only *additive* combinations, which leads to parts-based representations.

2 Two Objective Functions

Lee & Seung consider two divergence measures between V and WH.

2.1 Squared Euclidean Distance (Frobenius Norm)

Definition 1 (Euclidean Cost)**.**

$$\mathcal{L}_{\mathrm{EU}}(W, H) = \|V - WH\|_F^2 = \sum_{i=1}^{n} \sum_{\mu=1}^{m} (V_{i\mu} - (WH)_{i\mu})^2 \tag{1}$$

This is minimized when WH is a least-squares approximation to V.

2.2 Generalized KL Divergence (Poisson Likelihood)

Definition 2 (Divergence Cost)**.**

$$D(V \| WH) = \sum_{i=1}^{n} \sum_{\mu=1}^{m} \left[V_{i\mu} \log \frac{V_{i\mu}}{(WH)_{i\mu}} - V_{i\mu} + (WH)_{i\mu} \right] \tag{2}$$

This is non-negative and equals zero if and only if $V = WH$. It arises from a Poisson generative model: if $V_{i\mu} \sim \mathrm{Poisson}((WH)_{i\mu})$, then minimizing $D(V\|WH)$ is equivalent to maximizing the log-likelihood

$$\mathcal{F} = \sum_{i,\mu} [V_{i\mu} \log(WH)_{i\mu} - (WH)_{i\mu}] + \mathrm{const.}$$

Remark 1. *We will derive update rules for both objectives. The Euclidean case is simpler and builds intuition; the divergence case is what Lee & Seung use in the 1999 Nature paper.*

3 Matrix Calculus Preliminaries

Before deriving the updates, we need several matrix calculus results. Throughout, we treat W and H as collections of scalar variables W_{ia} and $H_{a\mu}$.

3.1 Expanding the Frobenius Norm

$$\begin{aligned}\|V - WH\|_F^2 &= \mathrm{Tr}\big[(V - WH)^T(V - WH)\big] \\ &= \mathrm{Tr}(V^T V) - 2\,\mathrm{Tr}(V^T WH) + \mathrm{Tr}(H^T W^T WH) \qquad (3)\end{aligned}$$

Key Insight

The trace identity $\|A\|_F^2 = \mathrm{Tr}(A^T A)$ converts a matrix norm into scalar traces, which are easy to differentiate. The key identity used here is $\mathrm{Tr}(A^T B) = \sum_{ij} A_{ij}B_{ij}$, which says the trace of a product is just the element-wise inner product.

3.2 Gradients of Trace Expressions

We need the following standard matrix derivative identities. For matrices A, B, X of compatible dimensions:

$$\frac{\partial}{\partial X}\mathrm{Tr}(AXB) = A^T B^T \qquad (4)$$

$$\frac{\partial}{\partial X}\mathrm{Tr}(X^T AXB) = AXB + A^T XB^T \qquad (5)$$

Why identity (4) holds: Write out $\mathrm{Tr}(AXB) = \sum_{i,j,k} A_{ij}X_{jk}B_{ki}$. Taking $\partial/\partial X_{jk}$ gives $\sum_i A_{ij}B_{ki} = (A^T)_{ji}(B^T)_{ik}$, which is the (j, k) entry of $A^T B^T$.

Why identity (5) holds: Write $\mathrm{Tr}(X^T AXB) = \sum_{i,j,k,l} X_{ji}A_{jk}X_{kl}B_{li}$. Taking $\partial/\partial X_{pq}$ and collecting terms (the variable X appears twice) gives two contributions, yielding the symmetric result.

3.3 Scalar-Level Derivatives

For element-wise derivations, recall that $(WH)_{i\mu} = \sum_{a=1}^{r} W_{ia}H_{a\mu}$. Therefore:

$$\frac{\partial (WH)_{i\mu}}{\partial H_{a\mu}} = W_{ia} \qquad (6)$$

$$\frac{\partial (WH)_{i\mu}}{\partial W_{ia}} = H_{a\mu} \qquad (7)$$

These are the "chain rule building blocks" used throughout.

4 Euclidean Update Rules

4.1 Gradient with Respect to H

Applying (4) and (5) to the expanded form (3):

$$\begin{aligned}\frac{\partial \mathcal{L}_{\mathrm{EU}}}{\partial H} &= \frac{\partial}{\partial H}\big[\mathrm{Tr}(V^T V) - 2\,\mathrm{Tr}(V^T WH) + \mathrm{Tr}(H^T W^T WH)\big] \\ &= 0 - 2W^T V + 2W^T WH \qquad (8)\end{aligned}$$

Term by term:

- $\mathrm{Tr}(V^TV)$ does not depend on H, so its derivative is 0.
- $-2\,\mathrm{Tr}(V^TWH)$: Use (4) with $A = V^TW$, $X = H$, $B = I$. Get $-2(V^TW)^TI^T = -2W^TV$.
- $\mathrm{Tr}(H^TW^TWH)$: Use (5) with $A = W^TW$, $X = H$, $B = I$. Since W^TW is symmetric, both terms equal W^TWH. Total: $2W^TWH$.

4.2 Gradient with Respect to W

By analogous computation (or by the symmetry $\|V - WH\|_F^2 = \|V^T - H^TW^T\|_F^2$):

$$\frac{\partial \mathcal{L}_{\mathrm{EU}}}{\partial W} = -2VH^T + 2WHH^T \tag{9}$$

4.3 Constructing Multiplicative Updates

The problem with additive gradient descent: The naive update $H \leftarrow H - \eta \nabla_H \mathcal{L}$ can make entries of H negative, violating the non-negativity constraint.

The multiplicative trick: Decompose the gradient into its positive and negative parts:

$$\nabla_H \mathcal{L} = \underbrace{2W^TWH}_{\text{positive part } [\nabla^+]} - \underbrace{2W^TV}_{\text{negative part } [\nabla^-]}$$

("positive" and "negative" refer to the sign of their contribution to the gradient, not the sign of the matrix entries—both W^TWH and W^TV have non-negative entries when $W, H, V \geq 0$.)

Now set the learning rate to be element-wise adaptive:

$$\eta_{a\mu} = \frac{H_{a\mu}}{(W^TWH)_{a\mu}}$$

The additive update becomes:

$$\begin{aligned} H_{a\mu} &\leftarrow H_{a\mu} - \eta_{a\mu} \cdot \left[(W^TWH)_{a\mu} - (W^TV)_{a\mu}\right] \\ &= H_{a\mu} - \frac{H_{a\mu}}{(W^TWH)_{a\mu}} \cdot \left[(W^TWH)_{a\mu} - (W^TV)_{a\mu}\right] \\ &= H_{a\mu} \cdot \frac{(W^TV)_{a\mu}}{(W^TWH)_{a\mu}} \qquad (10) \end{aligned}$$

Euclidean NMF Update Rules

$$H_{a\mu} \leftarrow H_{a\mu} \cdot \frac{(W^TV)_{a\mu}}{(W^TWH)_{a\mu}} \tag{H-update}$$

$$W_{ia} \leftarrow W_{ia} \cdot \frac{(VH^T)_{ia}}{(WHH^T)_{ia}} \tag{W-update}$$

Key Insight

Why non-negativity is preserved: If $H_{a\mu} \geq 0$, $W \geq 0$, and $V \geq 0$, then both W^TV and W^TWH have non-negative entries. A non-negative number times a non-negative ratio stays non-negative. The constraints are enforced *structurally*, with no projection or clipping needed.

Key Insight

Fixed points are stationary points: At convergence, $H_{a\mu}$ doesn't change, so the ratio equals 1, meaning $(W^T V)_{a\mu} = (W^T WH)_{a\mu}$, i.e., $\nabla_H \mathcal{L} = 0$. If $H_{a\mu} = 0$, the update keeps it at 0 regardless of the gradient—this satisfies the KKT complementary slackness condition for the non-negativity constraint.

5 Divergence Update Rules

Now we derive the updates for the KL divergence objective (2), which is the version in the 1999 *Nature* paper.

5.1 Gradient with Respect to $H_{a\mu}$

We work element-wise. Recall $(WH)_{i\mu} = \sum_b W_{ib} H_{b\mu}$.

$$\frac{\partial D}{\partial H_{a\mu}} = \sum_{i=1}^{n} \frac{\partial}{\partial H_{a\mu}} \left[V_{i\mu} \log \frac{V_{i\mu}}{(WH)_{i\mu}} - V_{i\mu} + (WH)_{i\mu} \right]$$
$$= \sum_{i=1}^{n} \left[-\frac{V_{i\mu}}{(WH)_{i\mu}} \cdot W_{ia} + W_{ia} \right] \tag{11}$$

Step by step:

- The $V_{i\mu} \log V_{i\mu}$ term is constant w.r.t. H, derivative is 0.
- $-V_{i\mu} \log(WH)_{i\mu}$: derivative of $\log(x)$ is $1/x$, chain rule via (6) gives $-V_{i\mu} \cdot W_{ia}/(WH)_{i\mu}$.
- $-V_{i\mu}$: constant, derivative is 0.
- $(WH)_{i\mu}$: derivative via (6) is W_{ia}.

Setting the gradient to zero at a fixed point:

$$\sum_i W_{ia} = \sum_i W_{ia} \cdot \frac{V_{i\mu}}{(WH)_{i\mu}}$$

5.2 Constructing the Multiplicative Update

Split the gradient (11) into positive and negative parts:

$$\frac{\partial D}{\partial H_{a\mu}} = \underbrace{\sum_i W_{ia}}_{\text{positive part}} - \underbrace{\sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}}_{\text{negative part}}$$

Set the adaptive learning rate $\eta_{a\mu} = H_{a\mu} / \sum_i W_{ia}$:

$$H_{a\mu} \leftarrow H_{a\mu} - \frac{H_{a\mu}}{\sum_i W_{ia}} \left[\sum_i W_{ia} - \sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}} \right]$$
$$= H_{a\mu} \cdot \frac{\sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}}{\sum_i W_{ia}} \tag{12}$$

5.3 The W Update

By the same procedure applied to $\partial D / \partial W_{ia}$:

$$W_{ia} \leftarrow W_{ia} \cdot \frac{\sum_\mu H_{a\mu} \frac{V_{i\mu}}{(WH)_{i\mu}}}{\sum_\mu H_{a\mu}} \tag{13}$$

Divergence NMF Update Rules (Lee & Seung 1999)

$$H_{a\mu} \leftarrow H_{a\mu} \cdot \frac{\sum_i W_{ia} V_{i\mu} / (WH)_{i\mu}}{\sum_i W_{ia}} \tag{H-update}$$

$$W_{ia} \leftarrow W_{ia} \cdot \frac{\sum_\mu H_{a\mu} V_{i\mu} / (WH)_{i\mu}}{\sum_\mu H_{a\mu}} \tag{W-update}$$

After the W-update, normalize each column of W to sum to 1.

Key Insight

The ratio $V_{i\mu}/(WH)_{i\mu}$ is the reconstruction error signal. If $(WH)_{i\mu}$ underestimates $V_{i\mu}$, the ratio > 1 and the responsible weights get boosted. If it overestimates, the ratio < 1 and weights shrink. This is the same correction mechanism as Richardson-Lucy deconvolution and EM for mixture models.

6 Convergence Proof (Euclidean Case)

We prove that the Euclidean update rule (10) monotonically decreases $\mathcal{L}_{\text{EU}}$. The proof uses an **auxiliary function** (the same technique used to prove EM convergence).

6.1 Auxiliary Function Method

Definition 3 (Auxiliary Function). *$G(h, h^t)$ is an auxiliary function for $F(h)$ if:*

(i) $G(h, h^t) \geq F(h)$ for all h (upper bound)

(ii) $G(h^t, h^t) = F(h^t)$ (touches F at the current point)

Lemma 4. *If G is an auxiliary function for F, then F is non-increasing under the update*

$$h^{t+1} = \arg\min_h G(h, h^t)$$

Proof.

$$F(h^{t+1}) \le G(h^{t+1}, h^t) \le G(h^t, h^t) = F(h^t)$$

The first inequality is property (i). The second holds because h^{t+1} minimizes $G(\cdot, h^t)$, so it is no worse than h^t. The equality is property (ii). □

Key Insight

This is exactly how EM convergence is proven. The E-step constructs an auxiliary function (the expected complete-data log-likelihood), and the M-step minimizes it. Each iteration is guaranteed to improve the objective. The NMF proof follows the same template.

6.2 Constructing the Auxiliary Function for NMF

We focus on the H-update, treating W as fixed. For clarity, consider a single column of H and the corresponding column of V. Let $h \in \mathbb{R}^r_{\ge 0}$ be the encoding vector (a column of H), and write the cost for a single data point as:

$$F(h) = \frac{1}{2}\|v - Wh\|^2 = \frac{1}{2}\sum_i \left(v_i - \sum_a W_{ia} h_a \right)^2$$

Expanding:

$$F(h) = \frac{1}{2}\left[v^T v - 2v^T Wh + h^T W^T Wh\right] \tag{14}$$

The quadratic term $h^T W^T Wh = \sum_{a,b}(W^T W)_{ab}\, h_a\, h_b$ couples different components of h through the off-diagonal entries of $W^T W$. The key idea is to find an auxiliary function that **decouples** these components.

Lemma 5 (Auxiliary Function for Euclidean NMF)**.** *Define*

$$G(h, h^t) = F(h^t) + (h - h^t)^T \nabla F(h^t) + \frac{1}{2}(h - h^t)^T K(h^t)(h - h^t) \tag{15}$$

where $K(h^t)$ is the diagonal matrix

$$K_{ab}(h^t) = \delta_{ab}\frac{(W^T W h^t)_a}{h^t_a} \tag{16}$$

Then $G(h, h^t)$ is an auxiliary function for $F(h)$.

Remark 2. *This is a second-order Taylor-like expansion, but with the true Hessian $W^T W$ replaced by the diagonal matrix $K(h^t)$. The diagonal structure means the minimization over h decouples into independent scalar problems for each h_a.*

Proof that G satisfies properties (i) and (ii).

Property (ii) is immediate: setting $h = h^t$ makes both correction terms vanish, giving $G(h^t, h^t) = F(h^t)$.

Property (i) requires showing $G(h, h^t) \ge F(h)$ for all h. Since F is quadratic in h, its exact second-order expansion (not an approximation) is:

$$F(h) = F(h^t) + (h - h^t)^T \nabla F(h^t) + \frac{1}{2}(h - h^t)^T W^T W(h - h^t)$$

Comparing with (15), we need:

$$(h - h^t)^T K(h^t)(h - h^t) \geq (h - h^t)^T W^T W(h - h^t) \quad \text{for all } h$$

This is equivalent to showing $K(h^t) - W^T W \succeq 0$ (positive semidefinite). Writing $\delta = h - h^t$:

$$\begin{aligned}
\delta^T K \delta - \delta^T W^T W \delta &= \sum_a \frac{(W^T W h^t)_a}{h_a^t} \delta_a^2 - \sum_{a,b} (W^T W)_{ab} \, \delta_a \, \delta_b \\
&= \sum_{a,b} (W^T W)_{ab} \left[\frac{h_b^t}{h_a^t} \delta_a^2 - \delta_a \delta_b \right] \qquad (17)
\end{aligned}$$

where we used $(W^T W h^t)_a = \sum_b (W^T W)_{ab} h_b^t$ to rewrite the diagonal term.

Now apply the inequality $x^2 y/z + z - 2x \geq 0$ for $y, z > 0$ (which follows from AM-GM: $x^2 y/z \geq 2|x|\sqrt{y} - y \cdot z/z$... more directly, for any a, b):

$$\frac{h_b^t}{h_a^t} \delta_a^2 + \frac{h_a^t}{h_b^t} \delta_b^2 \geq 2 \delta_a \delta_b$$

This is just the AM-GM inequality applied to $\delta_a \sqrt{h_b^t / h_a^t}$ and $\delta_b \sqrt{h_a^t / h_b^t}$. Since $(W^T W)_{ab} \geq 0$ (all entries are non-negative because $W \geq 0$), we can sum over all a, b with non-negative weights to obtain:

$$\sum_{a,b} (W^T W)_{ab} \left[\frac{h_b^t}{h_a^t} \delta_a^2 - \delta_a \delta_b \right] \geq 0$$

by the symmetrization argument (pair (a, b) with (b, a) and apply AM-GM to each pair). This establishes property (i). □ □

6.3 Minimizing the Auxiliary Function

Since G is quadratic in h with diagonal Hessian $K(h^t)$, we minimize by setting $\nabla_h G = 0$:

$$\begin{aligned}
\nabla_h G &= \nabla F(h^t) + K(h^t)(h - h^t) = 0 \\
h^{t+1} &= h^t - [K(h^t)]^{-1} \nabla F(h^t) \qquad (18)
\end{aligned}$$

Now substitute $\nabla F(h^t) = -W^T v + W^T W h^t$ and the diagonal form of K^{-1}:

$$\begin{aligned}
h_a^{t+1} &= h_a^t - \frac{h_a^t}{(W^T W h^t)_a} \left[-W^T v + W^T W h^t \right]_a \\
&= h_a^t - \frac{h_a^t}{(W^T W h^t)_a} \left[(W^T W h^t)_a - (W^T v)_a \right] \\
&= h_a^t \cdot \frac{(W^T v)_a}{(W^T W h^t)_a} \qquad (19)
\end{aligned}$$

Result

The multiplicative update rule

$$H_{a\mu} \leftarrow H_{a\mu} \cdot \frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}}$$

is exactly the minimizer of the auxiliary function $G(h, h^t)$, and therefore **monotonically decreases** the Euclidean cost $\|V - WH\|_F^2$ at every iteration.

6.4 Convergence Summary

Collecting the results:

Theorem 6 (Monotonic Convergence of Euclidean NMF)**.** *The Euclidean distance* $\|V - WH\|_F^2$ *is non-increasing under the update rules*

$$H_{a\mu} \leftarrow H_{a\mu} \cdot \frac{(W^T V)_{a\mu}}{(W^T WH)_{a\mu}}, \qquad W_{ia} \leftarrow W_{ia} \cdot \frac{(VH^T)_{ia}}{(WHH^T)_{ia}}$$

The cost is invariant under these updates if and only if W *and* H *are at a stationary point of the Lagrangian for the constrained optimization problem*

$$\min_{W \geq 0,\, H \geq 0} \|V - WH\|_F^2$$

satisfying the KKT conditions:

$$(W^T WH - W^T V)_{a\mu}\, H_{a\mu} = 0 \tag{20}$$

$$(WHH^T - VH^T)_{ia}\, W_{ia} = 0 \tag{21}$$

Proof. Monotonic decrease follows from Lemmas 4 and 5. The W-update proof is identical by symmetry (transpose the problem). The KKT conditions (20)–(21) follow from the fixed-point analysis: if $H_{a\mu} > 0$, the ratio must equal 1, so $\nabla_{H_{a\mu}} \mathcal{L} = 0$; if $H_{a\mu} = 0$, the update preserves this zero regardless of the gradient sign, which is precisely complementary slackness. □

Remark 3 (Local vs. Global Optima)**.** *The objective* $\|V - WH\|_F^2$ *is non-convex in* (W, H) *jointly (it is bilinear, hence the product creates non-convexity). The auxiliary function method guarantees convergence to a* ***local*** *minimum (or saddle point), not a global minimum. In practice, NMF is run with multiple random initializations.*

7 Convergence for the Divergence Case

The divergence case follows the same auxiliary function strategy. The construction is more involved because $F(h)$ is no longer quadratic, so we outline the key differences.

7.1 The Auxiliary Function

For the divergence objective, the auxiliary function uses **Jensen's inequality** on the log term. The concavity of log gives:

$$\log(WH)_{i\mu} = \log\left(\sum_a W_{ia} H_{a\mu}\right) \geq \sum_a \lambda_{ia}^{(a)} \log \frac{W_{ia} H_{a\mu}}{\lambda_{ia}^{(a)}}$$

where $\lambda_{ia}^{(a)} = W_{ia} H_{a\mu}^t / (WH^t)_{i\mu}$ are non-negative weights summing to 1 (they form a valid distribution over the index a).

Key Insight

This is the same Jensen's inequality trick used in the E-step of the EM algorithm. The λ values act like the "responsibilities" in a mixture model—they distribute each observed pixel $V_{i\mu}$ among the r hidden components proportionally to each component's contribution to the reconstruction.

7.2 Result

Minimizing this auxiliary function yields exactly the divergence update rules:

$$H_{a\mu} \leftarrow H_{a\mu} \cdot \frac{\sum_i W_{ia} V_{i\mu}/(WH)_{i\mu}}{\sum_i W_{ia}}$$

The proof of monotonic convergence is structurally identical to the Euclidean case: the auxiliary function touches the objective at the current point, upper-bounds it everywhere, and its minimizer gives the multiplicative update.

The full proof appeared in Lee & Seung (2001), "Algorithms for Non-negative Matrix Factorization," *NIPS*.

8 Summary: The Architecture of the Proof

1. **Write down the objective** (Euclidean or divergence).
2. **Compute the gradient** using matrix calculus or element-wise derivatives.
3. **Split the gradient** into positive and negative parts (both non-negative matrices when $V, W, H \geq 0$).
4. **Choose an adaptive learning rate** = current value / positive part, which converts additive gradient descent into a multiplicative ratio update.
5. **Prove monotonic convergence** by constructing an auxiliary function:
   - Euclidean case: replace the Hessian $W^T W$ with a diagonal upper bound → decouples the variables → minimizer gives the multiplicative update.
   - Divergence case: apply Jensen's inequality to the log term → decomposes the bound into separable terms → same structure.
6. **Verify KKT conditions** at fixed points to confirm convergence to constrained stationary points.

The core insight across both cases: **a carefully chosen diagonal majorization of the Hessian turns a coupled optimization problem into decoupled scalar updates that automatically preserve non-negativity.**

References: Lee, D.D. & Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788–791.
Lee, D.D. & Seung, H.S. (2001). Algorithms for Non-negative Matrix Factorization. *NIPS*.
Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *JRSS-B*, 39, 1–38.