Non-negative Matrix Factorization: Projected Gradient Descent

Problem Setup

We are given a data matrix $A \in \mathbb{R}^{N \times M}$ and seek a rank-R nonnegative factorization

$$A \approx WH, \quad W \in \mathbb{R}_+^{N \times R}, \quad H \in \mathbb{R}_+^{R \times M}.$$

We minimize the squared Frobenius loss

$$L(W, H) = \|A - WH\|_F^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} \big(A_{ij} - (WH)_{ij}\big)^2.$$

Gradient Derivation

Rewrite the loss using traces:

$$L(W, H) = \mathrm{Tr}\Big((A - WH)^\top (A - WH)\Big) = \mathrm{Tr}(A^\top A) - 2\,\mathrm{Tr}(A^\top WH) + \mathrm{Tr}\big(H^\top W^\top WH\big).$$

Using standard matrix calculus identities, we obtain the gradients

$$\nabla_W L(W, H) = -2AH^\top + 2WHH^\top,$$

$$\nabla_H L(W, H) = -2W^\top A + 2W^\top WH.$$

Projected Gradient Updates

Given a step size $\alpha > 0$ and the elementwise projection onto the nonnegative orthant

$$[X]_+ \;\coloneqq\; \max(X, 0) \quad \text{(applied entrywise)},$$

the projected gradient descent iterations are

$$W^{t+1} = \Big[W^t - \alpha\, \nabla_W L(W^t, H^t)\Big]_+ = \Big[W^t + 2\alpha\big(A(H^t)^\top - W^t H^t (H^t)^\top\big)\Big]_+,$$

$$H^{t+1} = \Big[H^t - \alpha\, \nabla_H L(W^t, H^t)\Big]_+ = \Big[H^t + 2\alpha\big((W^t)^\top A - (W^t)^\top W^t H^t\big)\Big]_+.$$

Algorithm (Pseudo-code)

Algorithm 1 Projected Gradient Descent for NMF

Input: $A \in \mathbb{R}^{N\times M}$, rank R, step size $\alpha > 0$, max iters T
Initialize $W^0 \in \mathbb{R}_+^{N\times R}$, $H^0 \in \mathbb{R}_+^{R\times M}$ (e.g., random nonnegative)
for $t = 0, 1, 2, \dots, T-1$ **do**
  Compute gradients:
$$G_W^t = -2A(H^t)^\top + 2W^tH^t(H^t)^\top, \quad G_H^t = -2(W^t)^\top A + 2(W^t)^\top W^tH^t$$
  Gradient steps:
$$\widetilde{W}^{t+1} = W^t - \alpha\, G_W^t, \quad \widetilde{H}^{t+1} = H^t - \alpha\, G_H^t$$
  Project to nonnegativity (entrywise):
$$W^{t+1} = [\,\widetilde{W}^{t+1}\,]_+, \quad H^{t+1} = [\,\widetilde{H}^{t+1}\,]_+$$
end for
Output: W^T, H^T

Notes

- The projection $[X]_+$ is applied elementwise: negative entries are set to zero.
- Step size α may be tuned (e.g., via backtracking or fixed small values).
- Convergence can be monitored using the loss $\|A - WH\|_F^2$ over iterations.