# Asymmetric Factorization for Low-Rank PSD Learning: When Is the Relaxation Exact?

**Enliang Hu**[*1,5,2] **Juho Kannala**[2] **Guoying Zhao**[2]
**Quanming Yao**[3] **Kun Yue**[4]

[1]Yunnan Key Laboratory of Modern Analytical Mathematics and Applications
[2]Aalto University [3]Tsinghua University
[4]Yunnan Key Laboratory of Intelligent Systems and Computing
[5]Key Laboratory of Education Informatization for Nationalities, Ministry of Education, Yunnan Normal University

## Abstract

Low-rank factorization is a standard way to scale optimization in machine learning by replacing large matrix variables with compact factors. For positive semidefinite (PSD) variables, the symmetric Burer–Monteiro factorization (sBMF) writes $Z = XX^\top$ with a single low-rank factor $X$. A recent asymmetric alternative (aBMF) writes $Z = XY^\top$ and adds a quadratic penalty $(\gamma/2)\|X - Y\|_F^2$ to encourage symmetry. This split is attractive because it yields a biconvex objective with alternating convex subproblems, but its practical value depends strongly on how the penalty parameter $\gamma$ is chosen.

We study a unified regularized aBMF framework and derive an explicit spectral lower bound on $\gamma$ that guarantees exactness at critical points: under the stated assumptions, any aBMF critical point satisfying this bound has equal factors and yields an sBMF critical point. These results extend earlier guarantees to general transpose-symmetric convex losses with shared convex regularizers or constraints. Specializing our framework to symmetric nonnegative matrix factorization (sNMF), we disprove the earlier positive-penalty conjecture with a counterexample. Our main theorem also yields conditions under which every $\gamma > 0$ ensures factor agreement at critical points of the sNMF relaxation.

## 1 Introduction

Low-rank representations reduce storage and computation in machine learning [14, 45, 25, 30]. For example, low-rank adaptation (LoRA) [18] reduces the number of trainable parameters and memory requirements during fine-tuning. Low-rank structure also arises naturally in many ML problems that are formulated as, or relaxed to, semidefinite programs (SDPs), including matrix sensing [49], matrix completion [25, 52, 43], convex neural networks [1, 38], neural network verification [7, 26, 5], graph inference [8], and combinatorial optimization and clustering [36, 37, 24]. We consider the convex PSD-constrained problem

$$\min_{Z \in S_n^+} f(Z),$$

where $f(\cdot)$ is a proper, closed, convex function and $S_n^+ = \{Z \in \mathbb{R}^{n\times n} \mid Z = Z^\top,\ Z \succeq 0\}$ denotes the cone of symmetric PSD matrices. Generic SDP solvers, including interior-point methods, become costly for large matrix variables [33].

*Corresponding author (ynel.hu@gmail.com).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Selected penalty guarantees under each work's assumptions. Exact factor agreement and approximate KKT transfer are distinct guarantees.

| Work | Model | Penalty guarantee |
|---|---|---|
| Zhu et al. [53], Li et al. [29] | sNMF, $h = \delta_+$ | Critical-point condition and initialization-based fixed-penalty bounds |
| Hu and Kwok [20] | Smooth convex PSD learning, $h = 0$ | Block-smoothness threshold for exact factor agreement |
| Han et al. [15] | Smooth convex SDP with linear equalities, $h = 0$ | Approximate KKT transfer to the symmetric factorized problem |
| This paper | Transpose-symmetric convex $f$; shared convex $h$ | Pointwise spectral condition; sublevel-set and initialization-based bounds |

The symmetric Burer–Monteiro factorization (sBMF) imposes low rank directly [6, 21, 38, 49, 36]. Parameterizing $Z = XX^\top$ with $X \in \mathbb{R}^{n\times r}$ and $r \ll n$ yields

$$\min_{X\in\mathbb{R}^{n\times r}} f(XX^\top). \tag{1}$$

This reduces the number of variables from $\mathcal{O}(n^2)$ to $\mathcal{O}(nr)$, but the objective is generally nonconvex in $X$, even when $f$ is convex in $Z$. First-order methods are inexpensive per iteration, but their convergence can depend strongly on step sizes and initialization.

To include regularization and constraints, we consider

$$\text{sBMF:} \quad \min_{X\in\mathbb{R}^{n\times r}} \quad f(XX^\top) + h(X), \tag{2}$$

where $h(\cdot)$ is a regularizer or an indicator function encoding constraints.

**Asymmetric splitting and the penalty selection problem.** Recent work replaces $XX^\top$ by $XY^\top$ and penalizes factor disagreement to obtain convex subproblems in each factor [53, 20, 29, 15]. For the regularized sBMF model (2), this asymmetric Burer–Monteiro factorization (aBMF) gives

$$\text{aBMF:} \quad \min_{X,Y} \quad \mathcal{F}(X,Y;\gamma) = f(XY^\top) + \tfrac{1}{2}h(X) + \tfrac{1}{2}h(Y) + \frac{\gamma}{2}\left\|X-Y\right\|_F^2, \tag{3}$$

where $\gamma > 0$. The objective is biconvex when $h$ is convex: $\mathcal{F}$ is convex in $X$ for fixed $Y$ and convex in $Y$ for fixed $X$ (see Proposition 1). This structure enables efficient alternating or splitting schemes with convex subproblems [22, 12, 44, 40, 42].

The penalty parameter $\gamma$ controls factor agreement. A small penalty may leave $X$ and $Y$ unequal, in which case $XY^\top$ need not be symmetric or PSD. A large penalty can slow alternating updates by keeping each factor close to the other. We seek sufficient penalties for factor agreement and use these bounds to guide their selection.

**Exactness of the asymmetric relaxation.** Our main result gives a spectral sufficient condition under which a critical point of aBMF (3) has equal factors and therefore corresponds to a critical point of sBMF (2). The spectral threshold depends on the point being considered. We further derive sublevel-set and initialization-based sufficient bounds. This extends the sNMF (4) analysis to general transpose-symmetric convex losses and shared convex regularizers and constraints; see Remarks 1–2.

## 2 Related Work

Zhu et al. [53] replace the symmetric factorization $XX^\top$ by $XY^\top$ for the specific problem of symmetric nonnegative matrix factorization (sNMF):

$$\min_{X\in\mathbb{R}^{n\times r}} \frac{1}{2}\left\|XX^\top - A\right\|_F^2 + \delta_+(X), \tag{4}$$

where $\delta_+(X)$ enforces nonnegativity ($X \geq 0$). They introduce the penalty $\frac{\gamma}{2}\|X-Y\|_F^2$ and derive sufficient conditions for a critical point $(\bar{X},\bar{Y})$ of the fixed-penalty problem to satisfy $\bar{X} = \bar{Y}$. These conditions can be conservative in practice [53].

For the same sNMF problem (4), Li et al. [29] establish fixed-penalty exactness and convergence results and also propose the adaptive penalty update

$$\gamma_k = \tau_{k-1}\gamma_{k-1}, \qquad \tau_{k-1} = \frac{\|X_{k-1}\|_F^2 + \|Y_{k-1}\|_F^2}{2|\langle X_{k-1}, Y_{k-1}\rangle|}, \tag{5}$$

where $\tau_{k-1} \geq 1$ whenever the denominator is nonzero. When applied to general aBMF, however, this rule cannot distinguish equal factors from nonzero opposite factors: $\tau_{k-1} = 1$ holds both when $X_{k-1} = Y_{k-1} \neq 0$ and when $X_{k-1} = -Y_{k-1} \neq 0$. In the latter case, the penalty remains unchanged despite factor disagreement. Near opposite-factor pairs, the multiplier can remain close to one despite a large mismatch, as the scalar runs in Sections 4.1 and B.1 illustrate. Nonnegativity excludes such nonzero opposite-factor pairs in the original sNMF setting.

More generally, multiplicative updates $\gamma_k = c_{k-1}\gamma_{k-1}$ with $c_{k-1} \geq 1$ are common penalty heuristics [35]. Their effectiveness depends on the initialization and growth rate: insufficient penalties may retain unequal-factor stationary points, whereas excessive penalties can slow alternating updates.

Hu and Kwok [20] study (3) with $h(\cdot) = 0$ using alternating (accelerated) gradient descent. For the $X$-block update without extrapolation, a sufficient penalty condition when $X_{k-1} \neq Y_{k-1}$ is

$$\gamma_k > \left[ L_{k-1}^{f_X} + \frac{2\,\mathrm{tr}\left((Y_{k-1} - X_{k-1})^\top \partial_X f(X_{k-1}Y_{k-1}^\top)\right)}{\|Y_{k-1} - X_{k-1}\|_F^2} \right]_+ . \tag{6}$$

This gives $\|X_k - Y_{k-1}\|_F^2 < \|X_k - X_{k-1}\|_F^2$, which implies $X_k - Y_k \to 0$ if $X_k - X_{k-1} \to 0$. This estimate depends on the alternating (accelerated) gradient descent and does not automatically extend to other alternating schemes.

Han et al. [15] study low-rank SDP splitting with an ADMM-based method. Their accuracy-dependent sufficient bound, of order $\Omega(1/\sqrt{\epsilon})$, transfers approximate KKT points to the symmetric factorized problem.

Our analysis concerns exact factor agreement at critical points with shared convex regularizers or constraints. Table 1 distinguishes these guarantees.

**Contributions.**

• **Spectral condition for exactness.** Theorem 1 gives a pointwise spectral sufficient condition for factor agreement at critical points. It extends the sNMF exactness analysis to general transpose-symmetric convex losses with shared convex regularizers or constraints.

• **Sublevel-set and initialization-based bounds.** Corollary 1 gives sufficient penalty bounds based on sublevel-set geometry and initialization, ensuring factor agreement at critical points in the corresponding sublevel sets.

• **Positive-penalty conditions for sNMF.** A counterexample shows that a positive penalty alone need not ensure convergence to equal factors (Remark 2). Corollary 2 gives sufficient conditions under which $\gamma > 0$ ensures factor agreement at critical points.

**Notation.** For a matrix $A = [a_{ij}]$, we write $\|A\|_F = (\sum_{ij} a_{ij}^2)^{1/2}$ for its Frobenius norm and define $([A]_+)_{ij} = \max(a_{ij}, 0)$ elementwise. For square $A$, $\mathrm{tr}(A)$ denotes its trace. We use $|\cdot|$ for absolute value and $^T$ for transpose. For matrices $A$ and $B$, $\langle A, B\rangle = \mathrm{tr}(A^T B)$ denotes their inner product. We use $\partial$ for subdifferentiation, with subscripts specifying the variable. When $f$ is differentiable, $\partial_Z f(Z)$ denotes its unique gradient. In nonsmooth formulas, a matrix-valued occurrence denotes a specified subgradient; membership and set-equality statements retain the set interpretation. Finally, $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest eigenvalues of a symmetric matrix.

## 3 Biconvex Relaxation for sBMF

We study when a critical point of the aBMF problem (3) also corresponds to a critical point of the sBMF problem (2). Proofs are deferred to Appendix A.

### 3.1 Biconvexity

The convexity of each block subproblem follows from the biconvexity of $\mathcal{F}(X, Y; \gamma)$ in (3).

**Definition 1.** *(biconvexity [12]) A function $\phi(X, Y)$ is biconvex if $\phi(\cdot, Y)$ (resp. $\phi(X, \cdot)$) is convex for every fixed $Y$ (resp. $X$).*

**Proposition 1.** *If $f(\cdot)$ and $h(\cdot)$ are convex, then $\mathcal{F}(X, Y; \gamma)$ in* (3) *is biconvex in terms of $X$ and $Y$.*

Each block subproblem is therefore convex. For the exactness results, a pair $(\bar{X}, \bar{Y})$ is critical for (3) if there exist $G \in \partial f(\bar{X}\bar{Y}^\top)$, $H_X \in \partial h(\bar{X})$, and $H_Y \in \partial h(\bar{Y})$ such that

$$G\bar{Y} + \gamma(\bar{X} - \bar{Y}) + \tfrac{1}{2}H_X = 0, \quad G^\top\bar{X} - \gamma(\bar{X} - \bar{Y}) + \tfrac{1}{2}H_Y = 0. \tag{7}$$

Both equations must hold with the same loss subgradient $G$, which is unique when $f$ is differentiable.

### 3.2 Penalty Bounds for Factor Agreement

If a critical point $(\bar{X}, \bar{Y})$ of problem (3) satisfies $\bar{X} = \bar{Y}$, then $\bar{X}$ (equivalently, $\bar{Y}$) is also a critical point of problem (2) [53, 15]. This reduces the main question to identifying conditions on $\gamma$ under which the asymmetric formulation enforces $\bar{X} = \bar{Y}$ at criticality. Once such conditions are available, the asymmetric relaxation becomes exact relative to the symmetric formulation. We first introduce the diameter of a sublevel set.

**Definition 2.** *The $c$-*sublevel set *of a function $f$ is the set: $\mathcal{L}_f(c) \equiv \{Z \mid f(Z) \le c\}$. The diameter of $\mathcal{L}_f(c)$ is defined as $d_{\mathcal{L}_f(c)} = \sup\{\|Z_1 - Z_2\|_F : Z_1, Z_2 \in \mathcal{L}_f(c)\}$.*

We now state the assumptions on $h$ and $f$ of (3).

**Assumption 1.** *The regularizer $h$ is proper, closed, and $\sigma_h$-strongly convex, with $\sigma_h \ge 0$: for all $X_1, X_2 \in \operatorname{dom}\partial h$ and $H_i \in \partial h(X_i)$, $\operatorname{tr}[(X_1 - X_2)^\top(H_1 - H_2)] \ge \sigma_h \|X_1 - X_2\|_F^2$.*

**Assumption 2.** *The loss $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ is finite-valued and convex. Additional conditions are invoked as needed: (a) $f(Z) \to \infty$ as $\|Z\|_F \to \infty$ (coercivity); (b) $\partial f(Z^\top) = \{G^\top : G \in \partial f(Z)\}$; (c) for some $\sigma_f \ge 0$, $\operatorname{tr}[(Z_1 - Z_2)^\top\big(\partial_Z f(Z_1) - \partial_Z f(Z_2)\big)] \ge \sigma_f \|Z_1 - Z_2\|_F^2$, $\forall Z_1, Z_2 \in \mathbb{R}^{n\times n}$, for every choice of the displayed subgradients; (d) $f$ is differentiable and, for some $l_f \ge 0$, $\operatorname{tr}[(Z_1 - Z_2)^\top\big(\partial_Z f(Z_1) - \partial_Z f(Z_2)\big)] \le l_f \|Z_1 - Z_2\|_F^2$, $\forall Z_1, Z_2 \in \mathbb{R}^{n\times n}$.*

Condition (a) makes every nonempty sublevel set bounded and hence of finite diameter. Condition (b) expresses transpose symmetry; for $f(Z) = \|Z - A\|_F^2$, it is equivalent to $A = A^\top$. Conditions (c,d) express strong convexity and smoothness [4, 35]; for convex differentiable $f$, (d) is equivalent to $l_f$-Lipschitz continuity of $\partial_Z f$. **Ordinary convexity is included by taking $\sigma_f = 0$ or $\sigma_h = 0$.**

At a critical point $(\bar{X}, \bar{Y})$ of (3), the following bounds on $\gamma$ control the distance between $\bar{X}$ and $\bar{Y}$ and give a sufficient condition on $\gamma$ for equality.

**Theorem 1.** *Under Assumptions 1 and 2(b,c), let $(\bar{X}, \bar{Y})$ be a critical point of* (3) *and use its common loss subgradient from* (7)*. For any $\epsilon > 0$, $\left\|\bar{X} - \bar{Y}\right\|_F^2 \le \epsilon$ holds when*

$$\gamma \ge \frac{1}{2}\frac{\operatorname{tr}\left((\bar{X} - \bar{Y})^\top\partial_Z f(\bar{X}\bar{Y}^\top)(\bar{X} - \bar{Y})\right)}{\epsilon} - \frac{\sigma_h}{4}. \tag{8}$$

*In particular, $\bar{X} = \bar{Y}$ holds when*

$$\gamma > \lambda_{\max}\Big(\frac{\partial_Z f(\bar{X}\bar{Y}^\top) + (\partial_Z f(\bar{X}\bar{Y}^\top))^\top}{4}\Big) - \frac{\sigma_h}{4}. \tag{9}$$

Since Assumptions 1 and 2 include ordinary convexity as the special case $\sigma_h = 0$ or $\sigma_f = 0$, Theorem 1 also applies to settings where $f$ or $h$ is merely convex, such as a transpose-symmetric extension of the hinge-loss model in (35).

Conversely, every critical point $\bar{X}$ of sBMF gives a joint critical point $(\bar{X}, \bar{X})$ of aBMF. Indeed, at $Z = \bar{X}\bar{X}^\top$, condition (b) allows a loss subgradient $G$ to be replaced by $(G + G^\top)/2$. When (9) holds at every aBMF critical point, critical points of the two formulations correspond under the identification $\bar{X} \leftrightarrow (\bar{X}, \bar{X})$.

**Geometric interpretation.** Write $U = (X+Y)/2$, $D = X - Y$, and $\mathcal{D} = \{(V,V) : V \in \mathbb{R}^{n\times r}\}$. In the product Frobenius norm, $(U,U)$ is the orthogonal projection of $(X,Y)$ onto $\mathcal{D}$, with squared distance $\|D\|_F^2/2$. The penalty is thus $\gamma$ times this squared distance. Moreover, $(XY^\top + YX^\top)/2 = UU^\top - DD^\top/4$: disagreement subtracts a PSD term from $UU^\top$. At a critical point, let $D = \bar{X} - \bar{Y}$ and $S = (G+G^\top)/2$, using the common loss subgradient $G$ in (7). Lemma 1 gives

$$\operatorname{tr}\big[D^\top\big((2\gamma + \sigma_h/2)I - S\big)D\big] \le 0.$$

The penalty and strong convexity of $h$ supply the shift $2\gamma+\sigma_h/2$. Condition (9) makes the displayed matrix positive definite, excluding nonzero disagreement at that critical point.

**Remark 1.** *Theorem 1 extends the sNMF exactness analysis of Zhu et al. [53], Li et al. [29] to general transpose-symmetric convex losses and shared proper, closed, convex regularizers, including indicators of closed convex constraint sets. It provides an exactness guarantee for the regularized extension discussed in Zhu et al. [53, Section 3.2].*

Let $Y_0 \in \operatorname{dom} h$ be the initialization of $Y$ and write $Z_0 = Y_0 Y_0^\top$. The following corollary gives sublevel-set and initialization-based sufficient bounds on $\gamma$.

**Corollary 1.** *Under Assumptions 1 and 2, let $(\bar{X},\bar{Y})$ be a critical point of* (3)*. If $f(\bar{X}\bar{Y}^\top) \le f(Z_0)$, then $\bar{X} = \bar{Y}$ when*

$$\gamma > \frac{1}{2}(\|\partial_Z f(Z_0)\|_F + l_f d_{\mathcal{L}_f(f(Z_0))}) - \frac{\sigma_h}{4}. \tag{10}$$

*Alternatively, assume $h \ge \underline{h} > -\infty$, $Y_0 \in \operatorname{dom} h$, $\mathcal{F}(\bar{X},\bar{Y};\gamma) \le \mathcal{F}(Y_0,Y_0;\gamma)$, $\sigma_f > 0$, and $\partial_Z f(\dot{Z}) = 0$. Then $\bar{X} = \bar{Y}$ if*

$$\gamma > \frac{l_f}{\sqrt{2\sigma_f}}\sqrt{[f(Z_0) + h(Y_0) - \underline{h} - f(\dot{Z})]_+} - \frac{\sigma_h}{4}. \tag{11}$$

Bound (10) depends on the diameter of the loss sublevel set $\mathcal{L}_f(f(Z_0))$, whereas (11) depends on the initialization and problem constants. For sNMF (4), where $f(Z) = \frac{1}{2}\|A - Z\|_F^2$ and $h(\cdot) = \delta_+(\cdot)$, we have $l_f = \sigma_f = 1$, $\sigma_h = 0$, and $f(\dot{Z}) = 0$ with $\dot{Z} = A$. Taking $\underline{h} = 0$, (11) becomes

$$\gamma > \frac{1}{\sqrt{2}}\sqrt{\frac{1}{2}\|Z_0 - A\|_F^2 + \delta_+(Y_0)},$$

which further reduces to $\gamma > \frac{1}{2}\|Z_0 - A\|_F$ when $\delta_+(Y_0) = 0$. This gives an initialization-based sufficient bound on the corresponding objective sublevel set.

The following structured cases of sNMF (4) extend an observation in [53].

**Corollary 2.** *Let $A = A^\top$ and $\gamma > 0$, and suppose that $(\bar{X},\bar{Y})$ is a critical point of* (3) *with $f(Z) = \frac{1}{2}\|Z - A\|_F^2$ and $h = \delta_+$. Then $\|\bar{X}\|_F = \|\bar{Y}\|_F$. Moreover, $\bar{X} = \bar{Y}$ if either of the following holds: (i) $\gamma \ge -\lambda_{\min}(A)$ and both factors have pairwise orthogonal columns (including the single-column case); or (ii) $A \ge 0$ entrywise and both factors have a single column.*

For the nonnegative version of the scalar example in [53], namely $\min_{x,y\ge 0}\frac{1}{2}(xy-1)^2 + \frac{\gamma}{2}(x-y)^2$, equal factor norms imply $x = y$. Corollary 2 extends this observation to the stated structured cases.

**Remark 2.** *Zhu et al. [53, Section 4.1] conjectured that, for any positive penalty,* SymHALS *converges to equal factors that yield a critical point of sNMF* (4)*. Li et al. [29, Conclusion] also raised the positive-penalty question. The following $2\times 2$ example shows that a positive penalty alone need not ensure factor agreement for symmetric nonnegative input matrices:*

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \gamma = \frac{3}{8}, \quad \bar{X} = \frac{1}{4}\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}, \quad \bar{Y} = \frac{1}{4}\begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}.$$

*Both factors are strictly positive but unequal. With $G = \bar{X}\bar{Y}^\top - A$ and $D = \bar{X} - \bar{Y}$, direct calculation gives*

$$G\bar{Y} + \gamma D = 0, \qquad G^\top \bar{X} - \gamma D = 0.$$

*Thus, $(\bar{X},\bar{Y})$ is a critical point of the penalized problem. Since each block objective is $\gamma$-strongly convex, exact* AM *initialized at $X_0 = Y_0 = \bar{Y}$ reaches $(\bar{X},\bar{Y})$ after one sweep and remains there. Here $r = n = 2$ and $A$ is indefinite. The example answers the positive-penalty question negatively for symmetric nonnegative input matrices. Remark 3 in Appendix A provides the full verification, including the failure of factor agreement for* HAM *and* SymHALS*.*

### 3.3 Alternating-Convex Algorithms for aBMF

Algorithm 1 alternates factor updates for (3) with a penalty update. The critical-point exactness conditions in Section 3.2 are independent of these algorithmic choices.

The convex subproblems in Steps 3 and 4 can be solved by standard methods. We consider three update schemes [53, 20, 29, 13, 48].

**(i) Alternating Minimization (AM)**

$$X_k = \arg\min_{X\in\mathbb{R}^{n\times r}} \mathcal{F}(X, Y_{k-1}; \gamma_{k-1}),\ Y_k = \arg\min_{Y\in\mathbb{R}^{n\times r}} \mathcal{F}(X_k, Y; \gamma_{k-1}). \tag{12}$$

**(ii) Hierarchical Alternating Minimization (HAM)**

HAM first updates all columns of $X = [x^{(1)}, \dots, x^{(r)}]$, then all columns of $Y = [y^{(1)}, \dots, y^{(r)}]$. For sNMF, this uses the same columnwise minimization principle as SymHALS [53], which instead alternates the two factors within each column. Within each sweep, the first equation below is applied for $i = 1, \dots, r$, followed by the second equation for $i = 1, \dots, r$:

$$\begin{aligned} x_k^{(i)} &= \arg\min_{x^{(i)}\in\mathbb{R}^{n\times 1}} \mathcal{F}([x_k^{(<i)}, x^{(i)}, x_{k-1}^{(>i)}], Y_{k-1}; \gamma_{k-1}) \\ y_k^{(i)} &= \arg\min_{y^{(i)}\in\mathbb{R}^{n\times 1}} \mathcal{F}(X_k, [y_k^{(<i)}, y^{(i)}, y_{k-1}^{(>i)}]; \gamma_{k-1}). \end{aligned} \tag{13}$$

**(iii) Alternating Accelerated Proximal Gradient (AAPG) Descent**

$$X_k = \arg\min_{X\in\mathbb{R}^{n\times r}} Q_X(X; \hat{X}_{k-1}, Y_{k-1}, \gamma_{k-1}),\ Y_k = \arg\min_{Y\in\mathbb{R}^{n\times r}} Q_Y(Y; X_k, \hat{Y}_{k-1}, \gamma_{k-1}), \tag{14}$$

where $Q_X(X; \hat{X}_{k-1}, Y_{k-1}, \gamma_{k-1}) = \langle \partial_X f(\hat{X}_{k-1}Y_{k-1}^\top), X - \hat{X}_{k-1}\rangle + \frac{L_{k-1}^{f_X}}{2}\left\|X - \hat{X}_{k-1}\right\|_F^2 + \frac{\gamma_{k-1}}{2}\|X - Y_{k-1}\|_F^2 + \frac{1}{2}h(X)$ and $\hat{X}_{k-1} = X_{k-1} + \omega_{k-1}^X(X_{k-1} - X_{k-2})$, where $\omega_{k-1}^X$ denotes the extrapolation weight and $L_{k-1}^{f_X}$ is a local smoothness constant for $f_X(X) = f(XY_{k-1}^\top)$. For AAPG, assume that $f$ is differentiable and that the indicated blockwise local smoothness constants exist. The $Y$ model is defined analogously with $X_k$ fixed, and uses $\frac{1}{2}h(Y)$. For the first step, set $\hat{X}_0 = X_0$ and $\hat{Y}_0 = Y_0$. Step sizes and extrapolation require algorithm-specific safeguards.

**Algorithm 1** Alternating convex optimization for aBMF.

1: **Initialization:** $X_0$, $Y_0$, $\gamma_0$, and the maximum iteration number $K$.
2: **for** $k = 1, \dots, K$ **do**
3: Update $X_k \approx \arg\min_{X\in\mathbb{R}^{n\times r}} \mathcal{F}(X, Y_{k-1}; \gamma_{k-1})$;
4: Update $Y_k \approx \arg\min_{Y\in\mathbb{R}^{n\times r}} \mathcal{F}(X_k, Y; \gamma_{k-1})$;
5: Update $\gamma_k$;
6: **if** a stopping criterion is satisfied **then**
7: **break**
8: **end if**
9: **end for**
10: **return** $\frac{1}{2}(X_k + Y_k)$.

**Dynamic update of $\gamma$.** We use the critical-point bound in Theorem 1 to motivate a practical penalty update. Under the theorem's hypotheses, an unequal-factor critical point cannot satisfy (8) with $\epsilon = \nu\left\|\bar{X} - \bar{Y}\right\|_F^2$ and $0 < \nu < 1$: this would imply $\left\|\bar{X} - \bar{Y}\right\|_F^2 \le \nu\left\|\bar{X} - \bar{Y}\right\|_F^2$, a contradiction.

Let $D_k = X_k - Y_k$ and let $G_k = \partial_Z f(X_k Y_k^\top)$ denote the selected loss subgradient. For $D_k \neq 0$, define the directional and norm-based candidates

$$\check{\gamma}_k = \Big[\frac{\mathrm{tr}(D_k^\top G_k D_k)}{2\nu\|D_k\|_F^2} - \frac{\sigma_h}{4}\Big]_+ + \varepsilon_0, \qquad \hat{\gamma}_k = \Big[\frac{1}{2}\|G_k\| - \frac{\sigma_h}{4}\Big]_+,$$

where $[a]_+ = \max\{a, 0\}$ and $\varepsilon_0 > 0$. Here the matrix norm is chosen to dominate the spectral norm; the Frobenius norm is one admissible choice. The norm-based candidate uses $\lambda_{\max}((G_k + G_k^\top)/2) \le \|G_k\|$; neither candidate uniformly dominates the other.

Step 5 combines them through

$$\gamma_k = \max\big\{\min\{\gamma_{k-1}, \hat{\gamma}_k\}, \check{\gamma}_k\big\}, \tag{15}$$

which ensures

$$\varepsilon_0 \le \check{\gamma}_k \le \gamma_k \le \max\{\gamma_{k-1}, \check{\gamma}_k\}.$$

Thus, $\gamma_k \le \gamma_{k-1}$ when $\check{\gamma}_k \le \gamma_{k-1}$; otherwise, $\gamma_k = \check{\gamma}_k$. The update (15) has a continuation-like interpretation: it reuses the current factors while adjusting the penalty, allowing decreases when permitted by the directional safeguard. We initialize $\gamma_0 > 0$ and retain $\gamma_{k-1}$ when $D_k = 0$, avoiding the undefined directional quotient.

### 3.4 Convergence with a Fixed Penalty

We analyze exact AM (12) with a fixed penalty. The dynamic update (15) is not covered by the following theorem.

**Theorem 2.** *Suppose Assumptions 1 and 2(a,c) hold, $f$ is continuously differentiable, and $X_0 = Y_0 \in \operatorname{dom} h$. For any fixed $\gamma > 0$, exact* AM *generates a bounded sequence, its successive differences tend to zero, and every accumulation point is critical for* (3)*. If, in addition, Assumption 2(b,d) holds, $\sigma_f > 0$, $h \ge \underline{h} > -\infty$, and $\gamma$ satisfies* (11)*, every accumulation point has equal factors and gives a critical point of* (2)*.*

The blockwise descent inequalities also bound the smallest squared step over the first $K$ iterations. If $\mathcal{F}_{\inf}$ is a lower bound on the objective, then

$$\min_{1\le k\le K}\{\|X_k - X_{k-1}\|_F^2 + \|Y_k - Y_{k-1}\|_F^2\} \le \frac{2(\mathcal{F}(X_0, Y_0; \gamma) - \mathcal{F}_{\inf})}{\gamma K}.$$

Full-sequence convergence and local rates require further conditions, such as the KL property of the full objective $\mathcal{F}$, together with sufficient-decrease, relative-error, and continuity conditions [10, 48].

**Computational cost.** The update (15) uses the loss subgradient, its norm, and a directional quadratic form. For quadratic sNMF with the Frobenius norm candidate, these can be evaluated without forming a dense residual, using matrix–factor products and $r \times r$ Gram matrices. Their cost is $\mathcal{O}(\operatorname{nnz}(A)r + nr^2)$ for sparse $A$, or $\mathcal{O}(n^2 r + nr^2)$ for dense $A$, with $\|A\|_F^2$ precomputed. Reusing products from the factor updates can reduce the additional cost.

### 3.5 Relationship to ADMM

ADMM combines multiplier updates with quadratic penalties [47, 19]. With an explicit constraint $X = Y$ [32], feasibility gives factor agreement, and the multiplier cancels when the two stationarity equations are added. In Han et al. [15], as in our Max-Cut formulation (Appendix B.5), the multipliers act on the linear SDP constraints. For fixed multipliers, the corresponding augmented Lagrangian terms can be absorbed into the loss, so Theorem 1 applies to the resulting objective.

## 4 Experiments

This section reports a scalar example, Top-$k$ PSD approximation, sNMF clustering, and a nonconvex neural-network surrogate. The second scalar example and further applications are reported in Appendix B. The comparisons examine penalty selection within alternating schemes. Project repository: `https://github.com/ynelhu/aBMF`.

**Methods and setup.** We use the AM, HAM, and AAPG updates described in Section 3.3. The subscripts indicate the penalty strategies: $\mathrm{adp}$ (**ad**a**p**tive strategy) uses (5); $\mathrm{agd}$ (**a**l**g**orithm-**d**ependent) uses the rule inspired by (6); $\mathrm{pca}$ (**p**roportional-to-**c**omputational-**a**ccuracy) labels the accuracy-inspired schedules motivated by Han et al. [15]; and $\mathrm{our}$ denotes the proposed strategy in (15). We implement the $\mathrm{adp}$, $\mathrm{agd}$, and $\mathrm{pca}$ strategies in the same alternating solvers, with adaptations where needed. Comparisons within each solver focus on penalty selection.

The application experiments were implemented in MATLAB on a server with 64 CPU cores and 512 GB RAM; reported times are wall-clock times. Unless specified otherwise, the stopping criterion is a relative objective change below $10^{-4}$ or $3 \times 10^3$ iterations. The default initialization

for unconstrained application experiments is $X_0 = \mathrm{randn}(n, r)$, $Y_0 = X_0$, and $\gamma_0 = 10^{-5}$. The toy examples and neural-network surrogate use the settings stated below. Runtime comparisons are interpreted together with the attained objective values and errors.

### 4.1 A Scalar Example for Theoretical Verification

Consider the scalar problem

$$\min_{x\in\mathbb{R}} x^2. \tag{16}$$

Its unique critical point is $x^\star = 0$. Writing $f(xx^\top) = x^2$ (equivalently, $f(z) = z$, $\partial_z f(z) = 1$, and $\sigma_f = l_f = 0$), the corresponding biconvex relaxation becomes

$$\min_{x,y\in\mathbb{R}} \mathcal{F}(x, y; \gamma) = xy + \frac{\gamma}{2}(x - y)^2. \tag{17}$$

Applying AM (12) to (17) gives

$$x_k = \frac{\gamma_{k-1} - 1}{\gamma_{k-1}} y_{k-1}, \quad y_k = \frac{\gamma_{k-1} - 1}{\gamma_{k-1}} x_k. \tag{18}$$

For fixed $\gamma > 0$ and $y_0 \neq 0$, iteration (18) converges to zero if $\gamma > 1/2$. This is an AM stability threshold, not a necessary condition for critical-point exactness: for $\gamma \neq 1/2$ the origin is the only critical point, whereas at $\gamma = 1/2$ every pair with $x = -y$ is critical. For $0 < \gamma < 1/2$, the split objective is unbounded below. Since $f(z) = z$ is not coercive, the convergence statement here follows directly from the recurrence, not Theorem 2.

We compare four schedules. In this scalar experiment, a sweep first computes $(x_k, y_k)$ using $\gamma_{k-1}$ and then updates $\gamma_k$; the penalty plots record this updated value.

$\mathrm{AM}_{\mathrm{agd}}$ uses a $Y$-block adaptation of (6): $\gamma_k = [2(x_k - y_k)x_k/(x_k - y_k)^2]_+ + \varepsilon_0$ for unequal factors. Since the second-block optimum satisfies $x_k = \gamma_{k-1}(x_k - y_k)$, this becomes $\gamma_k = 2\gamma_{k-1} + \varepsilon_0$ in exact arithmetic. The resulting rapid growth concerns this adaptation to AM.

$\mathrm{AM}_{\mathrm{adp}}$ evaluates (5) after each sweep: $\gamma_k = \gamma_{k-1}(x_k^2 + y_k^2)/(2|x_k y_k|)$. Its multiplier equals one at $x_k = -y_k \neq 0$, despite factor disagreement.

$\mathrm{AM}_{\mathrm{pca}}$ sets $\gamma_k = \min\{1/\sqrt{|x_k - y_k|}, 10^3\}$, with the cap used at zero disagreement and no positive lower floor. A mismatch greater than four gives $\gamma_k < 1/2$, below the fixed-penalty stability threshold.

$\mathrm{AM}_{\mathrm{our}}$ illustrates both the spectral bound and the dynamic rule. Theorem 1 gives $\gamma > 1/2$, which coincides with the fixed-penalty AM stability threshold here. For $x_k \neq y_k$, $\check{\gamma}_k = 1/(2\nu) + \varepsilon_0$ and $\hat{\gamma}_k \equiv 1/2$, so

$$\gamma_k = \max\{\min\{\gamma_{k-1}, 1/2\}, 1/(2\nu) + \varepsilon_0\} = 1/(2\nu) + \varepsilon_0 > 1/2. \tag{19}$$

The scalar implementation evaluates this constant directly, avoiding the directional quotient.

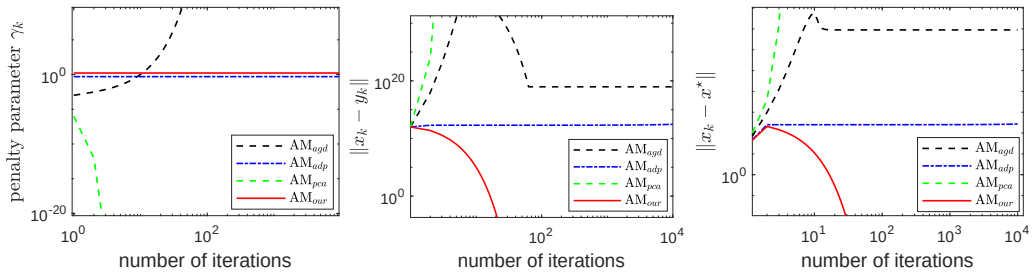
We use $y_0 = 100$, $\gamma_0 = 10^{-5}$, $\varepsilon_0 = 10^{-3}$, $\nu = 0.3$, and a budget of $10^4$ sweeps. In Figure 1, $\mathrm{AM}_{\mathrm{our}}$ sets $\gamma_k \approx 1.6677$ after the first sweep and approaches the origin. $\mathrm{AM}_{\mathrm{adp}}$ remains just below $1/2$ with large, oppositely signed factors, while $\mathrm{AM}_{\mathrm{agd}}$ stagnates far from the minimizer as its penalty grows. $\mathrm{AM}_{\mathrm{pca}}$ lowers the penalty as the mismatch grows and stops on overflow at sweep 5. Thus only $\mathrm{AM}_{\mathrm{our}}$ approaches the origin in this run.

### 4.2 Top-$k$ PSD Approximation

Symmetric low-rank product approximation has been studied in Liu et al. [31]. In the baseline names, GD and GN denote gradient descent and Gauss–Newton. We compare symmetric and asymmetric factorizations for

$$\min_{Z\succeq 0,\, \mathrm{rank}(Z)\leq r} \frac{1}{2}\|A - Z\|_F^2, \qquad A = A^\top \succeq 0.$$

For a fixed positive penalty, the exact $X$-update is $X = (A+\gamma I)Y(Y^\top Y+\gamma I)^{-1}$. Thus $\mathrm{span}(X) = \mathrm{span}((A + \gamma I)Y)$, so the column space follows a shifted subspace iteration. For PSD $A$ with a positive eigengap, increasing $\gamma$ moves the ratio $(\lambda_{r+1}(A) + \gamma)/(\lambda_r(A) + \gamma)$ toward one and can




(a) Updated penalty $\gamma_k$. (b) Factor gap $|x_k - y_k|$. (c) Solution error $|x_k - x^\star|$.

Figure 1: Penalty schedules for (16). Curves may leave the plotted range before the corresponding runs terminate. Only the iterates of $\mathrm{AM}_{\mathrm{our}}$ approach the origin.

Table 2: Top-$k$ PSD approximation: means over five trials. Lower is better for time and both error measures.

| $(n, r)$ | Method | Time (s) | Rel. objective | Subspace error | Final $\gamma$ |
|---|---|---|---|---|---|
| $(10^4, 50)$ | sBMF-GD | 3.979 | $1.98\times10^{-5}$ | $1.25\times10^{-2}$ | – |
| | sBMF-GN | 0.472 | $2.61\times10^{-7}$ | $1.45\times10^{-3}$ | – |
| | $\mathrm{AM}_{\mathrm{adp}}$ | 1.551 | $3.28\times10^{-6}$ | $5.16\times10^{-3}$ | 33.77 |
| | $\mathrm{AM}_{\mathrm{our}}$ | 0.678 | $4.62\times10^{-7}$ | $1.95\times10^{-3}$ | 6.90 |
| $(5\times10^4, 1000)$ | sBMF-GD | 2332.627 | $1.99\times10^{-5}$ | $5.57\times10^{-2}$ | – |
| | sBMF-GN | 225.145 | $2.85\times10^{-7}$ | $6.68\times10^{-3}$ | – |
| | $\mathrm{AM}_{\mathrm{adp}}$ | 2586.158 | $1.48\times10^{-5}$ | $4.82\times10^{-2}$ | 143.91 |
| | $\mathrm{AM}_{\mathrm{our}}$ | 738.091 | $2.79\times10^{-6}$ | $2.10\times10^{-2}$ | 30.51 |

slow subspace separation. Table 2 reports results averaged over five paired trials. $\mathrm{AM}_{\mathrm{our}}$ uses the penalty update (15) with $\nu = 0.9$ and $\varepsilon_0 = 10^{-6}$. The Frobenius-norm term is evaluated using low-rank matrix products. It reduces mean runtime relative to $\mathrm{AM}_{\mathrm{adp}}$ by factors of 2.29 and 3.50 in the two configurations, respectively, with smaller final errors; sBMF-GN remains best on all three measures.

### 4.3 Clustering via Symmetric NMF

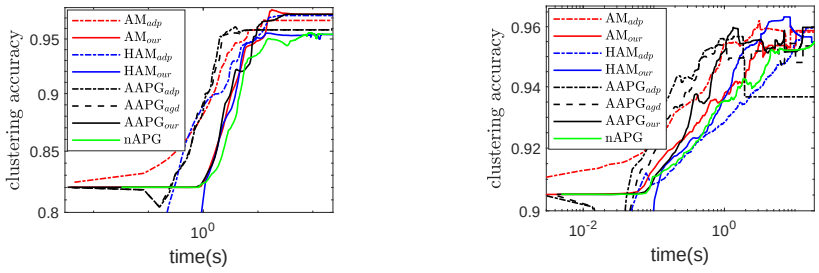
Symmetric nonnegative matrix factorization (sNMF) is used for graph-based clustering [29, 51]. Problem (4) fits (2) with $f(Z) = \frac{1}{2}\|A - Z\|_F^2$ and $h(X) = \delta_+(X)$, the nonnegative-orthant indicator ($\sigma_h = 0$). We use **OPTDIGITS** $(5,620$ samples, 10 classes$)$[2] and **COIL-20** $(1,440$ images, 20 objects$)$,[3] with rank equal to the number of classes. Each affinity matrix is a symmetrized 10-nearest-neighbor graph: $A = A^\top \in \{0,1\}^{n\times n}$, with $A_{ij} = 1$ exactly when samples $i$ and $j$ are connected. Following the rowwise maximum-entry rule in [53], we assign $\mathrm{label}(i) = \arg\max_j [X_\star + Y_\star]_{ij}$. For nAPG [28], an accelerated proximal gradient method applied to the symmetric nonconvex problem, we set $Y_\star = X_\star$ in this rule.

Figure 2 compares clustering accuracy against elapsed time. On OPTDIGITS, $\mathrm{AM}_{\mathrm{our}}$ and $\mathrm{AAPG}_{\mathrm{our}}$ reach the highest late-time accuracy, although their adaptive counterparts improve earlier; $\mathrm{HAM}_{\mathrm{adp}}$ attains a higher plateau than $\mathrm{HAM}_{\mathrm{our}}$. On COIL-20, the two AM variants attain similar final accuracy, while the intermediate-time advantage of $\mathrm{HAM}_{\mathrm{our}}$ over $\mathrm{HAM}_{\mathrm{adp}}$ does not persist.

Among the APG-based methods, $\mathrm{AAPG}_{\mathrm{our}}$ reaches high accuracy earlier than nAPG and attains higher late-time accuracy on both datasets. $\mathrm{AAPG}_{\mathrm{adp}}$ and $\mathrm{AAPG}_{\mathrm{agd}}$ also improve earlier, but their later behavior differs: on COIL-20, $\mathrm{AAPG}_{\mathrm{adp}}$ drops to a plateau below nAPG, whereas $\mathrm{AAPG}_{\mathrm{our}}$ maintains high accuracy. Thus asymmetric splitting can offer a better accuracy–time trade-off than direct symmetric optimization with nAPG, depending on the penalty rule and dataset. Similar empirical advantages have been reported in [29, 53].

[2] https://doi.org/10.24432/C50P49

[3] https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php




(a) OPTDIGITS ($n = 5620, r = 10$). (b) COIL-20 ($n = 1440, r = 20$).

Figure 2: Clustering accuracy versus elapsed time; the horizontal axes use a logarithmic scale.

### 4.4 Nonconvex Neural-Network Verification Surrogate

We study a PSD-lifted nonconvex ReLU verification surrogate [9, 5]. Appendix B.6 describes its loss $f_{\mathrm{nc}}$, residuals, and settings for 120 matched runs per method. Asymmetric methods alternate gradient steps with Armijo backtracking.

For this nonconvex experiment, $\mathrm{AM}_{\mathrm{our}}$ uses the following nondecreasing variant:

$$\gamma_k = \max\left\{\gamma_{k-1}, \frac{1}{2}\|\partial_Z f_{\mathrm{nc}}(Z_k)\|_F + \epsilon_\gamma\right\}, \tag{20}$$

where $Z_k = \mathrm{sym}(X_k Y_k^\top) = (X_k Y_k^\top + Y_k X_k^\top)/2$, $\epsilon_\gamma = 10^{-6}$, and the starting penalty is 10. The loss and lift-based residuals are evaluated at $Z_k$. The reported objective excludes the factor-agreement penalty. Success is assessed using the residual test in Appendix B.6. Table 3 shows similar asymmetric means and $100\%$ success, versus $66.7\%$ for sBMF, which has a smaller inequality residual and enforces PSD by construction. The surrogate lies outside Theorem 1; success is not a robustness certificate.

Table 3: Nonconvex verification surrogate: pooled means over 120 runs per method. Success is assessed using the residual diagnostic defined in the appendix.

| Method | Objective | Affine | Inequality | Comp. | Rel. disagr. | PSD violation | Success |
|---|---|---|---|---|---|---|---|
| sBMF | 4.100 | $2.95\times10^{-2}$ | $1.19\times10^{-2}$ | $2.21\times10^{-2}$ | – | $2.24\times10^{-15}$ | 66.7% |
| aBMF-fixed | $-1.225$ | $2.51\times10^{-2}$ | $4.05\times10^{-2}$ | $2.21\times10^{-2}$ | $1.67\times10^{-3}$ | $2.07\times10^{-5}$ | 100% |
| $\mathrm{AM}_{\mathrm{adp}}$ | $-1.225$ | $2.51\times10^{-2}$ | $4.05\times10^{-2}$ | $2.21\times10^{-2}$ | $1.61\times10^{-3}$ | $2.24\times10^{-5}$ | 100% |
| $\mathrm{AM}_{\mathrm{our}}$ | $-1.228$ | $2.51\times10^{-2}$ | $4.05\times10^{-2}$ | $2.21\times10^{-2}$ | $1.65\times10^{-3}$ | $2.21\times10^{-5}$ | 100% |

## 5 Conclusion

We derive a spectral sufficient condition for factor agreement at critical points of regularized aBMF, together with sublevel-set and initialization-based bounds. These results extend the sNMF (4) exactness analysis to general transpose-symmetric convex losses with shared convex regularizers. For sNMF, every positive penalty ensures factor agreement at critical points in structured cases, whereas a counterexample rules out this guarantee in general.

The spectral and initialization-based bounds are sufficient conditions and need not identify the smallest penalty ensuring factor agreement. The scalar experiments distinguish factor agreement from solution accuracy: an insufficient penalty can leave a large mismatch, while rapid penalty growth can stall alternating updates. The proposed rule avoids these effects in the two tests.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62266055), the Yunnan Key Laboratory of Intelligent Systems and Computing (No. 202549CE340006), the Yunnan Key Laboratory of Digital Intelligence Supervision and Governance (No. 202749CE340006), and the Innovation Team Project (No. 013010323004). We thank the three anonymous reviewers and the area chair for their constructive feedback. The authors declare no competing interests.

## References


[1] Burak Bartan and Mert Pilanci. Neural spectrahedra and semidefinite lifts: global convex optimization of degree-two polynomial activation neural networks in polynomial-time. *Mathematical Programming*, 213(1-2):737–769, 2025. doi: 10.1007/s10107-024-02153-5.

[2] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, 72(3):239–257, 2007.

[3] Srinadh Bhojanapalli, Nicolas Boumal, Prateek Jain, and Praneeth Netrapalli. Smoothed analysis for low-rank solutions to semidefinite programs in quadratic penalty form. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 3243–3270. PMLR, 2018.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] Robin A. Brown, Edward Schmerling, Navid Azizan, and Marco Pavone. A unified view of SDP-based neural network verification through completely positive programming. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9334–9355. PMLR, 28–30 Mar 2022.

[6] Samuel Burer and Renato D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003. doi: 10.1007/s10107-002-0352-8.

[7] Hong-Ming Chiu, Hao Chen, Huan Zhang, and Richard Y. Zhang. SDP-CROWN: Efficient bound propagation for neural network verification with tightness of semidefinite programming. In *Forty-second International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10449–10468. PMLR, 2025.

[8] Murat A. Erdogdu, Yash Deshpande, and Andrea Montanari. Inference in graphical models via semidefinite programming hierarchies. In *Advances in Neural Information Processing Systems 30*, pages 416–424, 2017.

[9] Mahyar Fazlyab, Manfred Morari, and George J. Pappas. An introduction to neural network analysis via semidefinite programming. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 6341–6350, 2021.

[10] Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015. doi: 10.1007/s10957-014-0642-3.

[11] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42 (6):1115–1145, 1995.

[12] J. Gorski, F. Pfeuffer, and K. Klamroth. Biconvex sets and optimization with biconvex functions: A survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007.

[13] L. Grippo and M. Sciandrone. On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Operations Research Letters*, 26(3):127–136, 2000.

[14] Andi Han, Jiaxiang Li, Wei Huang, Mingyi Hong, Akiko Takeda, Pratik Jawanpuria, and Bamdev Mishra. SLTrain: a sparse plus low rank approach for parameter and memory efficient pretraining. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[15] Qiushi Han, Chenxi Li, Zhenwei Lin, Caihua Chen, Qi Deng, Dongdong Ge, Huikang Liu, and Yinyu Ye. A low-rank ADMM splitting approach for semidefinite programming. arXiv preprint arXiv:2403.09133, 2024. URL `https://arxiv.org/abs/2403.09133`.

[16] F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015. doi: 10.1145/2827872.

[17] S.C.H. Hoi, R. Jin, and M.R. Lyu. Learning nonparametric kernel matrices from pairwise constraints. In *Proceedings of the 24th International Conference on Machine Learning*, pages 361–368, Corvallis, Oregon, USA, June 2007.

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[19] En-Liang Hu and James T. Kwok. Efficient kernel learning from side information using ADMM. In Francesca Rossi, editor, *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 1408–1414, 2013.

[20] En-Liang Hu and James T Kwok. Low-rank matrix learning using biconvex surrogate minimization. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3517–3527, 2019.

[21] Michel Journée, Francis Bach, Pierre-Antoine Absil, and Rodolphe Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5): 2327–2351, 2010. doi: 10.1137/080731359.

[22] Ragodaya Deepthi Kadiyala, Deepti Kakarla, and Neetu Chikyal. Hybrid distributed online and alternating convex optimization-based initial value accelerated hybrid precoding for millimeter wave multiple-input-multiple-output. *International Journal of Communication Systems*, 37(3): e5660, 2024. doi: 10.1002/dac.5660.

[23] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[24] B. Kulis, A. C. Surendran, and J. C. Platt. Fast low-rank semidefinite programming for embedding and clustering. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pages 235–242, San Juan, Puerto Rico, 2007.

[25] Minsu Kwon and Ho-Jin Choi. Learning low-rank representation for matrix completion. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 161–164, Los Alamitos, CA, USA, February 2020. IEEE Computer Society.

[26] Jianglin Lan, Benedikt Brückner, and Alessio Lomuscio. A semidefinite relaxation based branch-and-bound method for tight neural network verification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14946–14954, Jun. 2023.

[27] Soeren Laue. A hybrid algorithm for convex semidefinite optimization. In *International Conference on Machine Learning*, 2012.

[28] Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. In *Advances in Neural Information Processing Systems 28*, 2015.

[29] Xiao Li, Zhihui Zhu, Qiuwei Li, and Kai Liu. A provable splitting approach for symmetric nonnegative matrix factorization. *IEEE Transactions on Knowledge and Data Engineering*, 35 (3):2206–2219, 2023. doi: 10.1109/TKDE.2021.3125947.

[30] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013. doi: 10.1109/TPAMI.2012.88.

[31] Xin Liu, Zaiwen Wen, and Yin Zhang. An efficient Gauss–Newton algorithm for symmetric low-rank product matrix approximations. *SIAM Journal on Optimization*, 25(3):1571–1608, 2015.

[32] Songtao Lu, Mingyi Hong, and Zhengdao Wang. A nonconvex splitting method for symmetric nonnegative matrix factorization: Convergence analysis and optimality. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2572–2576, 2017.

[33] Anirudha Majumdar, Georgina Hall, and Amir Ali Ahmadi. Recent scalability improvements for semidefinite programming with applications in machine learning, control, and robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:331–360, 2020.

[34] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80):2287–2322, 2010. URL https://jmlr.org/papers/v11/mazumder10a.html.

[35] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer, 2006.

[36] Faniriana Rakoto Endor and Irène Waldspurger. Benign landscape for Burer–Monteiro factorizations of MaxCut-type semidefinite programs. arXiv preprint arXiv:2411.03103, 2024.

[37] Martin Royer. Adaptive clustering through semidefinite programming. In *Advances in Neural Information Processing Systems 30*, pages 1793–1801, 2017.

[38] Arda Sahiner, Tolga Ergen, Batu Ozturkler, John M. Pauly, Morteza Mardani, and Mert Pilanci. Scaling convex neural networks with Burer-Monteiro factorization. In *International Conference on Learning Representations*, 2024.

[39] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

[40] Sohil Shah, Abhay Kumar Yadav, Carlos D. Castillo, David W. Jacobs, Christoph Studer, and Tom Goldstein. Biconvex relaxation for semidefinite programming in computer vision. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 717–735, 2016.

[41] S. Shalev-Shwartz, A. Gonen, and O. Shamir. Large-scale convex minimization with a low-rank constraint. In *International Conference on Machine Learning*, 2011.

[42] Xinyue Shen, Steven Diamond, Madeleine Udell, Yuantao Gu, and Stephen Boyd. Disciplined multi-convex programming. In *2017 29th Chinese Control And Decision Conference (CCDC)*, pages 895–900, 2017. doi: 10.1109/CCDC.2017.7978647.

[43] N. Srebro, J.D.M. Rennie, and T.S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, volume 17, 2005.

[44] Quoc Tran-Dinh. Proximal alternating penalty algorithms for nonsmooth constrained convex optimization. *Computational Optimization and Applications*, 72(1):1–43, 2019.

[45] Minghua Wang, Qiang Wang, Danfeng Hong, Swalpa Kumar Roy, and Jocelyn Chanussot. Learning tensor low-rank representation for hyperspectral anomaly detection. *IEEE Transactions on Cybernetics*, 53(1):679–691, 2023.

[46] Po-Wei Wang, Wei-Cheng Chang, and J. Zico Kolter. The Mixing method: low-rank coordinate descent for semidefinite programming with diagonal constraints. *arXiv preprint arXiv:1706.00476*, 2017. URL https://arxiv.org/abs/1706.00476.

[47] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78(1):29–63, 2019.

[48] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

[49] Baturalp Yalcin, Ziye Ma, Javad Lavaei, and Somayeh Sojoudi. Semidefinite programming versus Burer-Monteiro factorization for matrix sensing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10702–10710, 2023.

[50] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. Network representation learning with rich text information. In *IJCAI*, volume 2015, pages 2111–2117, 2015.

[51] Zhirong Yang, Jukka Corander, and Erkki Oja. Low-rank doubly stochastic matrix decomposition for cluster analysis. *Journal of Machine Learning Research*, 17(187):1–25, 2016.

[52] Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. arXiv preprint arXiv:1605.07051, 2016.

[53] Zhihui Zhu, Xiao Li, Kai Liu, and Qiuwei Li. Dropping symmetry for fast symmetric nonnegative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[54] J. Zhuang, I.W. Tsang, and S.C.H. Hoi. A family of simple non-parametric kernel learning algorithms. *Journal of Machine Learning Research*, 12:1313–1347, 2011.

# Asymmetric Factorization for Low-Rank PSD Learning: When Is the Relaxation Exact?

## Supplementary Material

### Notation and Assumptions for Appendix

| Symbol | Description |
|---|---|
| $X, Y \in \mathbb{R}^{n\times r}$ | Factor matrices in low-rank decomposition |
| $Z = XY^\top$ | Product of factors representing original matrix |
| $\bar{X}, \bar{Y}$ | Factors at a critical point of $\mathcal{F}(X, Y; \gamma)$ |
| $\dot{Z}$ | Critical point of $f(Z)$ |
| $f(Z)$ | Finite-valued convex loss |
| $h(X), h(Y)$ | Regularization/constraint terms, with monotone subdifferential $\partial h$ |
| $\mathcal{F}(X, Y; \gamma)$ | Penalized objective: $f(XY^\top) + \frac{\gamma}{2}\|X - Y\|_F^2 + \frac{1}{2}(h(X) + h(Y))$ |
| $\partial_X f, \partial_Y f, \partial_Z f$ | Subdifferentials with respect to $X$, $Y$, and $Z$; unique gradients when differentiable, or specified selections otherwise. |
| $\sigma_f$ | Strong convexity parameter of $f$, an ordinary convex function $f$ is simply the special case $\sigma_f = 0$. |
| $l_f$ | Smoothness (Lipschitz gradient) constant of $f$ |
| $\sigma_h$ | Strong convexity parameter of $h$ |
| $\mathcal{L}_f(c)$ | The *c-sublevel set* of a function $f$: $\mathcal{L}_f(c) \equiv \{Z \mid f(Z) \le c\}$ |
| $d_{\mathcal{L}_f(c)}$ | The diameter of $\mathcal{L}_f(c)$: $d_{\mathcal{L}_f(c)} = \sup\{\|Z_1 - Z_2\|_F : Z_1, Z_2 \in \mathcal{L}_f(c)\}$. |
| $\delta_+(X)$ | Indicator function of nonnegative orthant |
| $\|\cdot\|_F$ | Frobenius norm |
| $\lambda_{\max}(A)$ | Largest eigenvalue of a symmetric matrix $A$ |
| $\rho(A)$ | Spectral radius of matrix $A$ |

Table 4: Notation used throughout the appendix.

**Assumption 1.** *$h : \mathbb{R}^{n\times r} \to \mathbb{R} \cup \{+\infty\}$ is proper, closed, and $\sigma_h$-strongly convex, with $\sigma_h \ge 0$. Thus, for $H_i \in \partial h(X_i)$,*

$$\mathrm{tr}[(X_1 - X_2)^\top (H_1 - H_2)] \ge \sigma_h \|X_1 - X_2\|_F^2 .$$

**Assumption 2.** *$f : \mathbb{R}^{n\times n} \to \mathbb{R}$ is convex and finite everywhere. The following conditions are invoked where specified:*

*(a)* $\lim_{\|Z\|_F\to\infty} f(Z) = +\infty$*; that is, $f$ is coercive.*

*(b)* $\partial f(Z^\top) = \{G^\top : G \in \partial f(Z)\}$.

*(c) For some $\sigma_f \ge 0$,* $\mathrm{tr}[(Z_1 - Z_2)^\top \big(\partial_Z f(Z_1) - \partial_Z f(Z_2)\big)] \ge \sigma_f \|Z_1 - Z_2\|_F^2$, $\forall Z_1, Z_2 \in \mathbb{R}^{n\times n}$, *for every choice of the displayed subgradients.*

*(d) $f$ is differentiable and, for some $l_f \ge 0$,* $\mathrm{tr}[(Z_1 - Z_2)^\top \big(\partial_Z f(Z_1) - \partial_Z f(Z_2)\big)] \le l_f \|Z_1 - Z_2\|_F^2$, $\forall Z_1, Z_2 \in \mathbb{R}^{n\times n}$.

For convex differentiable $f$ on $\mathbb{R}^{n\times n}$, condition (d) is equivalent to $\|\partial_Z f(Z_1) - \partial_Z f(Z_2)\|_F \le l_f \|Z_1 - Z_2\|_F$; we use this equivalent form in the proofs.

## A Details for Proofs

Throughout this section, $\gamma > 0$. For the nonconvex composite objectives, criticality means that the limiting subdifferential contains zero. As in the main text, $\partial_Z f$ denotes the unique gradient when $f$ is differentiable; otherwise a matrix-valued occurrence denotes a specified selection, while membership statements use the subdifferential set. At a critical point $(\bar{X}, \bar{Y})$ of (3), one can choose a *common* $G \in \partial f(\bar{X}\bar{Y}^\top)$ and $H_X \in \partial h(\bar{X})$, $H_Y \in \partial h(\bar{Y})$ such that

$$G\bar{Y} + \gamma(\bar{X} - \bar{Y}) + \tfrac{1}{2}H_X = 0, \qquad G^\top \bar{X} - \gamma(\bar{X} - \bar{Y}) + \tfrac{1}{2}H_Y = 0. \tag{21}$$

The subgradient $G$ used in the bounds below is this common selection. Separate blockwise subgradient selections need not give (21) when $f$ is nonsmooth.

Use the smooth maps $(X, Y) \mapsto (XY^\top, X, Y)$ for aBMF and $X \mapsto (XX^\top, X)$ for sBMF. Since $f$ is finite and convex, every outer horizon subgradient has zero first component. The adjoint derivatives preserve the remaining components, so the chain-rule qualification holds. Convexity of the outer functions gives equality in the chain rule. Including the smooth penalty for aBMF, we obtain (21) and the sBMF stationarity condition $(G + G^\top)X + H = 0$, where $G \in \partial f(XX^\top)$ and $H \in \partial h(X)$.

At an sBMF critical point $X$, transpose symmetry and convexity give $S = (G + G^\top)/2 \in \partial f(XX^\top)$. Thus $(X, X)$ satisfies (21) with common loss subgradient $S$ and $H_X = H_Y = H$. Conversely, adding the two equations at a diagonal aBMF critical point gives sBMF stationarity with $H = (H_X + H_Y)/2 \in \partial h(X)$.

**Proposition 1.** *If $f$ and $h$ are convex and $\gamma \geq 0$, then $\mathcal{F}(X, Y; \gamma)$ in* (3) *is biconvex.*

### A.1 Proof of Proposition 1

*Proof.* Let $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 = 1$, and fix $Y \in \mathbb{R}^{n\times r}$. By convexity of $f$,

$$f\big((\alpha_1 X_1 + \alpha_2 X_2)Y^\top\big) = f\big(\alpha_1 X_1 Y^\top + \alpha_2 X_2 Y^\top\big) \leq \alpha_1 f(X_1 Y^\top) + \alpha_2 f(X_2 Y^\top).$$

Thus $f(XY^\top)$ is convex in $X$. The terms $h(X)/2$ and $\gamma \|X - Y\|_F^2 /2$ are also convex in $X$, while $h(Y)/2$ is constant. Hence $\mathcal{F}(\cdot, Y; \gamma)$ is convex. The same argument with $X$ fixed proves convexity in $Y$, completing the proof. □

**Lemma 1.** *Under Assumptions 1 and 2(b,c), let $(\bar{X}, \bar{Y})$ be a critical point of* (3)*, with the common subgradient $G$ in* (21)*. Then*

$$(2\gamma + \tfrac{1}{2}\sigma_h)\left\|\bar{X} - \bar{Y}\right\|_F^2 \leq \operatorname{tr}\big((\bar{X} - \bar{Y})^\top G(\bar{X} - \bar{Y})\big).$$

### A.2 Proof of Lemma 1

*Proof.* Set $D = \bar{X} - \bar{Y}$ and $Z = \bar{X}\bar{Y}^\top$, and use the common subgradient $G \in \partial f(Z)$ and the selections $H_X \in \partial h(\bar{X})$, $H_Y \in \partial h(\bar{Y})$ in (21). Subtracting its second equation from the first gives

$$2\gamma D = G^\top \bar{X} - G\bar{Y} - \tfrac{1}{2}(H_X - H_Y).$$

Taking the inner product with $D$ and applying Assumption 1, we obtain

$$\begin{aligned} 2\gamma \|D\|_F^2 &= \langle D, G^\top \bar{X} - G\bar{Y}\rangle - \tfrac{1}{2}\langle D, H_X - H_Y\rangle \\ &\leq \langle D, G^\top \bar{X} - G\bar{Y}\rangle - \tfrac{1}{2}\sigma_h \|D\|_F^2 . \end{aligned}$$

Thus

$$(2\gamma + \tfrac{1}{2}\sigma_h)\|D\|_F^2 \leq \langle D, G^\top \bar{X} - G\bar{Y}\rangle. \tag{22}$$

To bound the right-hand side, decompose $G = S + T$, where $S = (G + G^\top)/2$ is symmetric and $T = (G - G^\top)/2$ is skew-symmetric. Since $S^\top = S$ and $T^\top = -T$,

$$G^\top \bar{X} - G\bar{Y} = (S - T)\bar{X} - (S + T)\bar{Y} = SD - T(\bar{X} + \bar{Y}).$$

Taking the inner product with $D$ and using cyclicity of the trace gives

$$\begin{aligned} \langle D, G^\top \bar{X} - G\bar{Y}\rangle &= \operatorname{tr}(D^\top SD) - \operatorname{tr}\big(T(\bar{X} + \bar{Y})D^\top\big) \\ &= \operatorname{tr}(D^\top SD) - \operatorname{tr}\big(T(\bar{X}\bar{X}^\top - \bar{Y}\bar{Y}^\top + Z^\top - Z)\big) \\ &= \operatorname{tr}(D^\top SD) - \tfrac{1}{2}\langle G - G^\top, Z - Z^\top\rangle. \end{aligned}$$

In the last equality, the terms involving $\bar{X}\bar{X}^\top$ and $\bar{Y}\bar{Y}^\top$ vanish because the trace of a product of a symmetric matrix and a skew-symmetric matrix is zero. The remaining term uses $T = (G - G^\top)/2$ and the definition of the matrix inner product.

By Assumption 2(b), $G^\top \in \partial f(Z^\top)$. Applying Assumption 2(c) to the subgradients $G$ and $G^\top$ therefore gives

$$\langle G - G^\top, Z - Z^\top \rangle \geq \sigma_f \left\| Z - Z^\top \right\|_F^2 .$$

It follows that

$$\begin{aligned} \langle D, G^\top \bar{X} - G\bar{Y} \rangle &\leq \operatorname{tr}(D^\top S D) - \tfrac{1}{2}\sigma_f \left\| Z - Z^\top \right\|_F^2 \\ &\leq \operatorname{tr}(D^\top S D) = \operatorname{tr}(D^\top G D), \end{aligned} \tag{23}$$

where the second inequality uses $\sigma_f \geq 0$, and the last equality uses $\operatorname{tr}(D^\top T D) = 0$. Combining (22) and (23) proves the claim. □

**Theorem 1.** *Under Assumptions 1 and 2(b,c), let $(\bar{X}, \bar{Y})$ be a critical point of* (3)*, and let $G$ be its common subgradient in* (21)*. For $\epsilon > 0$, $\left\| \bar{X} - \bar{Y} \right\|_F^2 \leq \epsilon$ holds if*

$$\gamma \geq \frac{\operatorname{tr}((\bar{X} - \bar{Y})^\top G (\bar{X} - \bar{Y}))}{2\epsilon} - \frac{\sigma_h}{4}.$$

*In particular, $\bar{X} = \bar{Y}$ if*

$$\gamma > \lambda_{\max}\left( \frac{G + G^\top}{4} \right) - \frac{\sigma_h}{4}.$$

### A.3 Proof of Theorem 1

*Proof.* Write $D = \bar{X} - \bar{Y}$, and let $G$ be the common loss subgradient. By Lemma 1,

$$(2\gamma + \tfrac{1}{2}\sigma_h) \left\| D \right\|_F^2 \leq \operatorname{tr}(D^\top G D).$$

Since $\gamma > 0$ and $\sigma_h \geq 0$, division gives

$$\left\| D \right\|_F^2 \leq \frac{\operatorname{tr}(D^\top G D)}{2\gamma + \sigma_h/2}.$$

To guarantee $\left\| D \right\|_F^2 \leq \epsilon$, it therefore suffices that

$$\frac{\operatorname{tr}(D^\top G D)}{2\gamma + \sigma_h/2} \leq \epsilon, \qquad \text{or equivalently} \qquad \gamma \geq \frac{\operatorname{tr}(D^\top G D)}{2\epsilon} - \frac{\sigma_h}{4}.$$

This proves the first assertion.

For exact factor agreement, set $S = (G + G^\top)/2$. Since the skew-symmetric part of $G$ contributes zero to the quadratic form,

$$\begin{aligned} (2\gamma + \tfrac{1}{2}\sigma_h) \left\| D \right\|_F^2 &\leq \operatorname{tr}(D^\top G D) = \operatorname{tr}(D^\top S D) \\ &\leq \lambda_{\max}(S) \left\| D \right\|_F^2 . \end{aligned} \tag{24}$$

The last inequality applies the largest-eigenvalue bound to each column of $D$. Rearranging gives

$$\big(2\gamma + \tfrac{1}{2}\sigma_h - \lambda_{\max}(S)\big) \left\| D \right\|_F^2 \leq 0.$$

Thus $D = 0$ whenever the coefficient is positive, namely when

$$\gamma > \tfrac{1}{2}\lambda_{\max}(S) - \tfrac{1}{4}\sigma_h = \lambda_{\max}\left( \frac{G + G^\top}{4} \right) - \frac{\sigma_h}{4}.$$

This is the claimed exactness condition. □

**A quadratic check.** For $f(Z) = \operatorname{tr}(CZ)$ with $C = C^\top$ and $h(X) = \frac{1}{2}\sigma \left\| X \right\|_F^2$, the stationarity equations (21) imply

$$\big((2\gamma + \tfrac{1}{2}\sigma) I - C\big)(\bar{X} - \bar{Y}) = 0.$$

Thus $\gamma > \frac{1}{2}\lambda_{\max}(C) - \frac{1}{4}\sigma$ excludes unequal factors, in agreement with Theorem 1.

**Corollary 1.** *Under Assumptions 1 and 2, let $Y_0 \in \operatorname{dom} h$ and $Z_0 = Y_0 Y_0^\top$. A critical point $(\bar{X}, \bar{Y})$ satisfying $f(\bar{X}\bar{Y}^\top) \leq f(Z_0)$ has $\bar{X} = \bar{Y}$ if*

$$\gamma > \tfrac{1}{2}\big( \|\partial_Z f(Z_0)\|_F + l_f d_{\mathcal{L}_f(f(Z_0))} \big) - \tfrac{1}{4}\sigma_h.$$

*Alternatively, suppose $\sigma_f > 0$, $h \geq \underline{h} > -\infty$, and $\partial_Z f(\dot{Z}) = 0$. A critical point satisfying $\mathcal{F}(\bar{X}, \bar{Y}; \gamma) \leq f(Z_0) + h(Y_0)$ has $\bar{X} = \bar{Y}$ if*

$$\gamma > \frac{l_f}{\sqrt{2\sigma_f}} \sqrt{[f(Z_0) + h(Y_0) - \underline{h} - f(\dot{Z})]_+} - \tfrac{1}{4}\sigma_h.$$

### A.4 Proof of Corollary 1

*Proof.* Set $\bar{Z} = \bar{X}\bar{Y}^\top$ and $G = \partial_Z f(\bar{Z})$, and denote the spectral radius by $\rho$. Since $G + G^\top$ is symmetric,

$$\begin{aligned}\lambda_{\max}\left(\frac{G+G^\top}{4}\right) &\le \tfrac{1}{4}\rho(G+G^\top) \le \tfrac{1}{4}\|G+G^\top\|_F \\ &\le \tfrac{1}{2}\|G\|_F. \end{aligned} \tag{25}$$

The triangle inequality and Assumption 2(d) give

$$\begin{aligned}\tfrac{1}{2}\|G\|_F &\le \tfrac{1}{2}\big(\|\partial_Z f(Z_0)\|_F + \|\partial_Z f(\bar{Z}) - \partial_Z f(Z_0)\|_F\big) \\ &\le \tfrac{1}{2}\big(\|\partial_Z f(Z_0)\|_F + l_f\|\bar{Z} - Z_0\|_F\big) \\ &\le \tfrac{1}{2}\big(\|\partial_Z f(Z_0)\|_F + l_f d_{\mathcal{L}_f(f(Z_0))}\big). \end{aligned} \tag{26}$$

For the last inequality, the assumed loss bound $f(\bar{Z}) \le f(Z_0)$ places both points in $\mathcal{L}_f(f(Z_0))$, which is bounded by coercivity. Combining (25)–(26) with Theorem 1 proves the first assertion.

For the second assertion, smoothness and $\partial_Z f(\dot{Z}) = 0$ yield

$$\tfrac{1}{2}\|\partial_Z f(\bar{Z})\|_F = \tfrac{1}{2}\|\partial_Z f(\bar{Z}) - \partial_Z f(\dot{Z})\|_F \le \tfrac{1}{2} l_f \|\bar{Z} - \dot{Z}\|_F.$$

Strong convexity gives

$$\|\bar{Z} - \dot{Z}\|_F^2 \le \frac{2}{\sigma_f}\big(f(\bar{Z}) - f(\dot{Z})\big).$$

Using these two inequalities in (25), we obtain

$$\begin{aligned}\lambda_{\max}\left(\frac{G+G^\top}{4}\right) &\le \frac{l_f}{\sqrt{2\sigma_f}}\sqrt{[f(\bar{Z}) - f(\dot{Z})]_+} \\ &\le \frac{l_f}{\sqrt{2\sigma_f}}\sqrt{[\mathcal{F}(\bar{X},\bar{Y};\gamma) - \underline{h} - f(\dot{Z})]_+} \\ &\le \frac{l_f}{\sqrt{2\sigma_f}}\sqrt{[f(Z_0) + h(Y_0) - \underline{h} - f(\dot{Z})]_+}. \end{aligned}$$

The second inequality uses $h \ge \underline{h}$ and $\gamma > 0$; the third uses the stated objective-sublevel condition and $\mathcal{F}(Y_0, Y_0; \gamma) = f(Z_0) + h(Y_0)$. Theorem 1 now proves the second assertion. Equivalently, replacing $h$ by $h - \underline{h}$ reduces this part to a nonnegative regularizer without changing any stationary point. The two assertions use different sublevel assumptions. □

**Corollary 2.** *Let $A = A^\top$ and $\gamma > 0$, and suppose that $(\bar{X}, \bar{Y})$ is a critical point of (3) with $f(Z) = \frac{1}{2}\|Z - A\|_F^2$ and $h = \delta_+$. Then $\|\bar{X}\|_F = \|\bar{Y}\|_F$. Moreover, $\bar{X} = \bar{Y}$ if either of the following holds: (i) $\gamma \ge -\lambda_{\min}(A)$ and both factors have pairwise orthogonal columns (including the single-column case); or (ii) $A \ge 0$ entrywise and both factors have a single column.*

### A.5 Proof of Corollary 2

*Proof.* Write $\bar{X} = [\bar{x}_1, \dots, \bar{x}_r]$, $\bar{Y} = [\bar{y}_1, \dots, \bar{y}_r]$, $D = \bar{X} - \bar{Y}$, and $G = \bar{X}\bar{Y}^\top - A$. For each $i = 1, \dots, r$, the columnwise stationarity equations are

$$G\bar{y}_i + \gamma(\bar{x}_i - \bar{y}_i) + n_{X,i} = 0, \qquad G^\top \bar{x}_i - \gamma(\bar{x}_i - \bar{y}_i) + n_{Y,i} = 0,$$

where $n_{X,i} \in N_{\mathbb{R}^n_+}(\bar{x}_i)$ and $n_{Y,i} \in N_{\mathbb{R}^n_+}(\bar{y}_i)$. The factor $1/2$ on the indicator is absorbed into the normal vectors. Taking inner products with $\bar{x}_i$ and $\bar{y}_i$, respectively, and using $\bar{x}_i^\top n_{X,i} = \bar{y}_i^\top n_{Y,i} = 0$, gives

$$\gamma(\|\bar{x}_i\|_2^2 - \bar{x}_i^\top \bar{y}_i) = -\bar{x}_i^\top G\bar{y}_i = \gamma(\|\bar{y}_i\|_2^2 - \bar{x}_i^\top \bar{y}_i).$$

Since $\gamma > 0$, corresponding columns have equal norms. Summing their squared norms gives $\|\bar{X}\|_F = \|\bar{Y}\|_F$.

**Case 1: a single column and $\gamma \geq -\lambda_{\min}(A)$.** Write $\bar{x} = \bar{X}$, $\bar{y} = \bar{Y}$, and $d = \bar{x} - \bar{y}$. The equality of norms gives

$$d^\top \bar{x} = \tfrac{1}{2}\|d\|_2^2, \qquad \bar{y}^\top d = -\tfrac{1}{2}\|d\|_2^2.$$

Hence

$$\begin{aligned} d^\top(\bar{x}\bar{y}^\top - A)d &= (d^\top \bar{x})(\bar{y}^\top d) - d^\top A d \\ &= -\tfrac{1}{4}\|d\|_2^4 - d^\top A d \\ &\leq -\tfrac{1}{4}\|d\|_2^4 - \lambda_{\min}(A)\|d\|_2^2 \\ &\leq -\lambda_{\min}(A)\|d\|_2^2. \end{aligned}$$

Lemma 1 therefore gives

$$(2\gamma + \lambda_{\min}(A))\|d\|_2^2 \leq 0.$$

Since $2\gamma + \lambda_{\min}(A) \geq \gamma > 0$, we obtain $\bar{x} = \bar{y}$.

**Case 2: column-orthogonal factors and $\gamma \geq -\lambda_{\min}(A)$.** Using the rank-one decomposition $\bar{X}\bar{Y}^\top = \sum_{j=1}^r \bar{x}_j \bar{y}_j^\top$, let $d_i = \bar{x}_i - \bar{y}_i$ and $A_i = A - \sum_{j \neq i} \bar{x}_j \bar{y}_j^\top$. As in Case 1, equal column norms give

$$(d_i^\top \bar{x}_i)(\bar{y}_i^\top d_i) = -\tfrac{1}{4}\|d_i\|_2^4.$$

For $j \neq i$, orthogonality within both factors and nonnegativity give

$$(d_i^\top \bar{x}_j)(\bar{y}_j^\top d_i) = -(\bar{y}_i^\top \bar{x}_j)(\bar{y}_j^\top \bar{x}_i) \leq 0.$$

Consequently,

$$\begin{aligned} d_i^\top G d_i &= d_i^\top(\bar{x}_i\bar{y}_i^\top - A_i)d_i \\ &= -\tfrac{1}{4}\|d_i\|_2^4 - d_i^\top A d_i + \sum_{j \neq i}(d_i^\top \bar{x}_j)(\bar{y}_j^\top d_i) \\ &\leq -\tfrac{1}{4}\|d_i\|_2^4 - \lambda_{\min}(A)\|d_i\|_2^2. \end{aligned}$$

Summing over $i$ gives $\mathrm{tr}(D^\top G D) \leq -\lambda_{\min}(A)\,\|D\|_F^2$. Applying Lemma 1 to the full factors yields

$$(2\gamma + \lambda_{\min}(A))\,\|D\|_F^2 \leq 0.$$

As in Case 1, $2\gamma + \lambda_{\min}(A) \geq \gamma > 0$, so $\bar{X} = \bar{Y}$.

**Case 3: a single column and $A \geq 0$ entrywise.** Write $\bar{x} = \bar{X}$, $\bar{y} = \bar{Y}$, and let $a = \|\bar{x}\|_2^2 = \|\bar{y}\|_2^2$. The block optimality conditions give

$$(a+\gamma)\bar{x} = [(A+\gamma I)\bar{y}]_+ = (A+\gamma I)\bar{y}, \qquad (a+\gamma)\bar{y} = [(A+\gamma I)\bar{x}]_+ = (A+\gamma I)\bar{x},$$

where the last equality in each relation follows from the entrywise nonnegativity of $A$, $\bar{x}$, and $\bar{y}$. Adding and subtracting these equations, with $v = \bar{x} + \bar{y}$ and $w = \bar{x} - \bar{y}$, gives

$$Av = av, \qquad Aw = -(a+2\gamma)w.$$

Since $A \geq 0$ entrywise, $A|w| \geq |Aw| = (a+2\gamma)|w|$ entrywise. Multiplying by $v^\top \geq 0$ and using $A = A^\top$ yields

$$(a+2\gamma)v^\top|w| \leq v^\top A|w| = (Av)^\top|w| = a v^\top|w|.$$

Since $\gamma > 0$, we have $v^\top|w| = 0$. Moreover, $v \geq |w|$ entrywise, so $\|w\|_2^2 \leq v^\top|w| = 0$. Thus $\bar{x} = \bar{y}$. □

**Remark 3** (Positive penalties need not ensure factor agreement)**.** *We verify the $2 \times 2$ example in Remark 2. For the sNMF relaxation, take $\gamma = 3/8$ and*

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \bar{X} = \frac{1}{4}\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}, \qquad \bar{Y} = \frac{1}{4}\begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}.$$

*Here $A$ is symmetric, binary, and has zero diagonal, and $r = n = 2$. Both factors are strictly positive but unequal. Setting $K = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$ gives $G = \bar{X}\bar{Y}^\top - A = \frac{3}{8}K$ and $D = \bar{X} - \bar{Y} = \frac{1}{2}K$, so*

$$G\bar{Y} + \gamma D = 0, \qquad G^\top \bar{X} - \gamma D = 0.$$

*Thus $(\bar{X}, \bar{Y})$ is a critical point. Moreover, $\mathcal{F}(\bar{X}, \bar{Y}; \gamma) = 15/32$, whereas every equal-factor pair has objective value at least $1/2$. Indeed, for the unit vector $q = (1, -1)^\top/\sqrt{2}$, we have $q^\top A q = -1$, so*

$$\mathcal{F}(W, W; \gamma) = \tfrac{1}{2}\left\|WW^\top - A\right\|_F^2 \geq \tfrac{1}{2}\left(q^\top WW^\top q + 1\right)^2 \geq \tfrac{1}{2} > \tfrac{15}{32} \qquad (W \geq 0).$$

*Hence no equal-factor pair is globally optimal in this instance.*

*The failure also occurs from equal initial factors. Each block objective is $\gamma$-strongly convex, so the stationarity conditions above identify its unique minimizer. Exact* AM *initialized at $X_0 = Y_0 = \bar{Y}$ therefore gives $(X_k, Y_k) = (\bar{X}, \bar{Y})$ for every $k \geq 1$.*

*For* HAM *in* (13)*, the same initialization and column-update order $x_1, x_2, y_1, y_2$ give, after the first two updates,*

$$\tilde{X} = \begin{pmatrix} 9/16 & 41/128 \\ 7/16 & 87/128 \end{pmatrix}, \qquad \tilde{Y} = \bar{Y}, \qquad \mathcal{F}(\tilde{X}, \tilde{Y}; \gamma) = \frac{8175}{16384} < \frac{1}{2}.$$

*Subsequent exact column updates cannot increase $\mathcal{F}$, so continuity excludes any equal-factor accumulation point.*

*For* SymHALS *[53], use the same initialization and the column-update order $x_1, y_1, x_2, y_2$. After the first three column updates, direct calculation gives*

$$\tilde{x}_1 = \frac{(9, 7)^\top}{16}, \qquad \tilde{y}_1 = \frac{(32, 76)^\top}{113},$$

$$\tilde{x}_2 = \frac{(1147, 2449)^\top}{3616}, \qquad \tilde{y}_2 = \frac{(3, 1)^\top}{4}.$$

*For $\tilde{X} = [\tilde{x}_1, \tilde{x}_2]$ and $\tilde{Y} = [\tilde{y}_1, \tilde{y}_2]$, we have*

$$\mathcal{F}(\tilde{X}, \tilde{Y}; \gamma) = \frac{12958425}{26150912} < \frac{1}{2}.$$

*Subsequent exact column updates cannot increase $\mathcal{F}$. Since every equal-factor pair has objective value at least $1/2$, continuity excludes any equal-factor accumulation point. Thus a positive penalty alone does not ensure convergence to equal factors in general sNMF.*

**Theorem 2.** *Suppose Assumptions 1 and 2(a,c) hold, $f$ is continuously differentiable, and $X_0 = Y_0 \in \operatorname{dom} h$. For any fixed $\gamma > 0$, exact* AM *generates a bounded sequence, its successive differences tend to zero, and every accumulation point is critical for* (3)*. If, in addition, Assumption 2(b,d) holds, $\sigma_f > 0$, $h \geq \underline{h} > -\infty$, and $\gamma$ satisfies* (11)*, every accumulation point has equal factors and gives a critical point of* (2)*.*

### A.6 Proof of Theorem 2

*Proof.* Write $\mathcal{F}_k = \mathcal{F}(X_k, Y_k; \gamma)$ and $Z_0 = Y_0 Y_0^\top$, so $\mathcal{F}_0 = f(Z_0) + h(Y_0)$. We proceed in three steps.

(i) **Existence of an accumulation point.** We first show that the full objective is coercive without a lower bound on $h$. Convexity and coercivity of the finite-valued loss imply that $f(Z) \geq a\|Z\|_F - b$ for some $a > 0$ and $b \in \mathbb{R}$. This linear bound follows by applying convexity along rays outside a bounded sublevel set. Since $h$ is proper, closed, and convex, it admits an affine minorant $h(X) \geq \langle P, X\rangle + c$. Write $R^2 = \|X\|_F^2 + \|Y\|_F^2$. The identity

$$R^2 = \|X - Y\|_F^2 + 2\mathrm{tr}(XY^\top) \leq \|X - Y\|_F^2 + 2\sqrt{n}\|XY^\top\|_F$$

and Cauchy–Schwarz give

$$\begin{aligned}\mathcal{F}(X, Y; \gamma) &\geq a\|XY^\top\|_F + \tfrac{\gamma}{2}\|X - Y\|_F^2 - \tfrac{1}{2}\|P\|_F(\|X\|_F + \|Y\|_F) + c - b \\ &\geq \kappa R^2 - \frac{\|P\|_F}{\sqrt{2}}R + c - b, \qquad \kappa = \min\left\{\frac{\gamma}{2}, \frac{a}{2\sqrt{n}}\right\} > 0.\end{aligned}$$

Thus $\mathcal{F}$ is proper, closed, and coercive, so its sublevel sets are compact and $\mathcal{F}_{\inf} = \min_{X,Y} \mathcal{F}(X, Y; \gamma)$ is finite. Each block objective is also $\gamma$-strongly convex; hence its exact minimizer exists and is unique. Blockwise strong convexity gives

$$\mathcal{F}(X_{k-1}, Y_{k-1}; \gamma) - \mathcal{F}(X_k, Y_{k-1}; \gamma) \geq \frac{\gamma}{2} \|X_k - X_{k-1}\|_F^2 ,$$
$$\mathcal{F}(X_k, Y_{k-1}; \gamma) - \mathcal{F}(X_k, Y_k; \gamma) \geq \frac{\gamma}{2} \|Y_k - Y_{k-1}\|_F^2 .$$

Adding these inequalities yields

$$\mathcal{F}_{k-1} - \mathcal{F}_k \geq \frac{\gamma}{2} \big( \|X_k - X_{k-1}\|_F^2 + \|Y_k - Y_{k-1}\|_F^2 \big).$$

Since $\mathcal{F}_k \geq \mathcal{F}_{\inf}$, summing this inequality yields

$$\sum_{k=1}^{\infty} \big( \|X_k - X_{k-1}\|_F^2 + \|Y_k - Y_{k-1}\|_F^2 \big) \leq \frac{2(\mathcal{F}_0 - \mathcal{F}_{\inf})}{\gamma} < \infty. \tag{27}$$

In particular, both successive differences tend to zero.

All iterates lie in the compact sublevel set $\{(X, Y) : \mathcal{F}(X, Y; \gamma) \leq \mathcal{F}_0\}$, so they form a bounded sequence and an accumulation point exists. Take any convergent subsequence $(X_{k_j}, Y_{k_j}) \to (X_\star, Y_\star)$.

(ii) **The accumulation point is critical for** (3)**.** To prove stationarity, choose $H_{X,k} \in \partial h(X_k)$ and $H_{Y,k} \in \partial h(Y_k)$ from the block optimality conditions

$$0 = \partial_Z f(X_k Y_{k-1}^\top) Y_{k-1} + \gamma(X_k - Y_{k-1}) + \tfrac{1}{2} H_{X,k}, \tag{28}$$
$$0 = \partial_Z f(X_k Y_k^\top)^\top X_k - \gamma(X_k - Y_k) + \tfrac{1}{2} H_{Y,k}. \tag{29}$$

Define

$$R_k = \partial_Z f(X_k Y_k^\top) Y_k - \partial_Z f(X_k Y_{k-1}^\top) Y_{k-1} + \gamma(Y_{k-1} - Y_k).$$

The same selections give $(R_k, 0) \in \partial \mathcal{F}(X_k, Y_k; \gamma)$. The products $X_k Y_k^\top$ and $X_k Y_{k-1}^\top$ lie in a common compact set, and their difference tends to zero by boundedness of $X_k$ and (27). The loss gradient is bounded and uniformly continuous on this set because $f$ is continuously differentiable. Together with $Y_k - Y_{k-1} \to 0$, this gives $R_k \to 0$; no Lipschitz estimate is needed for this step.

Along the chosen subsequence, (27) also gives $Y_{k_j - 1} \to Y_\star$. The selected subgradients in (28) and (29) therefore converge to

$$H_{X,\star} = -2\big(\partial_Z f(X_\star Y_\star^\top) Y_\star + \gamma(X_\star - Y_\star)\big),$$
$$H_{Y,\star} = -2\big(\partial_Z f(X_\star Y_\star^\top)^\top X_\star - \gamma(X_\star - Y_\star)\big).$$

The graph of the convex subdifferential $\partial h$ is closed, so $H_{X,\star} \in \partial h(X_\star)$ and $H_{Y,\star} \in \partial h(Y_\star)$. These limits satisfy (21), proving that $(X_\star, Y_\star)$ is a critical point of (3).

(iii) **Transfer to a critical point of** (2)**.** Under the additional hypotheses for exactness, let $\dot{Z}$ minimize $f$, so $\partial_Z f(\dot{Z}) = 0$. Lower semicontinuity gives $\mathcal{F}(X_\star, Y_\star; \gamma) \leq \mathcal{F}_0 = f(Z_0) + h(Y_0)$. The second part of Corollary 1 now applies and yields $X_\star = Y_\star$. The diagonal stationarity argument above shows that $X_\star$ is a critical point of (2).

□

## B More Experiments and Implementation Details

Algorithm names and penalty subscripts follow the definitions in Section 4.

### B.1 The Second Toy Example for Dynamic Penalty Updates

Consider

$$\min_{x\in\mathbb{R}} \frac{1}{2}(x^2+d)^2. \tag{30}$$

Its unique solution is $x^\star = 0$ for $d \geq 0$, corresponding to $f(xx^\top) = \frac{1}{2}(xx^\top + d)^2$ (clearly, $f(z) = \frac{1}{2}(z+d)^2$, $\partial_z f(z) = z + d$, $\sigma_f = l_f = 1$). Its biconvex relaxation is

$$\min_{x,y\in\mathbb{R}} \mathcal{F}(x,y,\gamma) = \frac{1}{2}(xy+d)^2 + \frac{\gamma}{2}(x-y)^2. \tag{31}$$

Applying AM (12) to (31) yields

$$x_k = \frac{\gamma_{k-1}-d}{\gamma_{k-1}+y_{k-1}^2} y_{k-1}, \qquad y_k = \frac{\gamma_{k-1}-d}{\gamma_{k-1}+x_k^2} x_k. \tag{32}$$

When $\gamma_{k-1}$ is small relative to $d$, $x_k^2$, and $y_{k-1}^2$, the updates approximately satisfy $x_k \simeq -d/y_{k-1}$ and $y_k \simeq -d/x_k \simeq y_{k-1}$. Thus, progress can be slow even in exact arithmetic. Conversely, a penalty much larger than these quantities gives update factors close to one and can also slow progress.

We use the same after-sweep convention as in Section 4.1 and compare the following rules:

i) $\mathrm{AM}_{\mathrm{agd}}$: the $Y$-block adaptation of (6), with $D_k = x_k - y_k \neq 0$, is

$$\gamma_k = \left[x_k^2 + \frac{2D_k(x_k y_k + d)x_k}{D_k^2}\right]_+ + \varepsilon_0.$$

The second-block optimality condition gives $(x_k y_k + d)x_k = \gamma_{k-1} D_k$; hence $\gamma_k = x_k^2 + 2\gamma_{k-1} + \varepsilon_0$ in exact arithmetic. The implementation evaluates the quotient above, whose numerical accuracy deteriorates when subtracting nearly equal factors.

ii) $\mathrm{AM}_{\mathrm{adp}}$ updates $\gamma_k = \gamma_{k-1}(x_k^2 + y_k^2)/(2|x_k y_k|)$. Near nonzero opposite factors, the multiplier is close to one despite a large mismatch.

iii) $\mathrm{AM}_{\mathrm{pca}}$: **an accuracy-inspired schedule**. Motivated by the accuracy-dependent penalty scale in [15], we use the following capped heuristic:

$$\gamma_k = \min\left\{\frac{1}{\sqrt{|x_k - y_k|}}, 10^3\right\}.$$

At zero disagreement the cap is used. This schedule approaches $10^3$ as disagreement vanishes. It is an AM comparison, not an implementation of the ADMM method or its KKT accuracy criterion in [15].

iv) $\mathrm{AM}_{\mathrm{our}}$: **the proposed directional candidate**. For $x_k \neq y_k$, let $q_k = x_k y_k + d$. The exact second-block update gives

$$q_k = \frac{\gamma_{k-1}(d + x_k^2)}{\gamma_{k-1} + x_k^2} \geq 0, \qquad \check{\gamma}_k = \frac{q_k}{2\nu} + \varepsilon_0.$$

For $\nu = 0.3 < 1/2$, we have $\check{\gamma}_k > q_k/2 = \hat{\gamma}_k$, so (15) gives $\gamma_k = \check{\gamma}_k$ whenever the factors differ. The scalar implementation evaluates this simplified expression directly, including when the factors become zero in floating-point arithmetic.

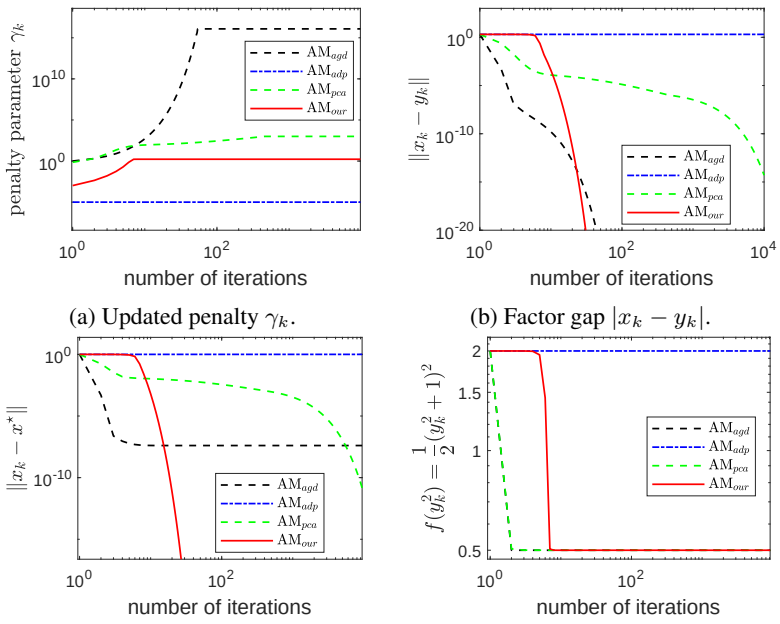
Experimental settings:

$$d = 1, \quad x_0 = 1, \quad y_0 = -1, \quad \gamma_0 = 10^{-5}, \quad \varepsilon_0 = 10^{-3}, \quad \nu = 0.3, \quad \mathrm{MaxIter} = 10^4.$$

All four methods complete $10^4$ sweeps. Figure 3 reports the updated penalty, factor gap $|x_k - y_k|$, solution error $|x_k - x^\star|$, and symmetric objective $f(y_k^2) = \frac{1}{2}(y_k^2+1)^2$, whose minimum is $1/2$. The objective is evaluated at the updated second factor for every method.

$\mathrm{AM}_{\mathrm{agd}}$ and $\mathrm{AM}_{\mathrm{pca}}$ reduce the objective earlier than $\mathrm{AM}_{\mathrm{our}}$, but their later behavior differs. For $\mathrm{AM}_{\mathrm{agd}}$, the penalty reaches about $1.21 \times 10^{16}$: the factors nearly agree, while the solution error stalls at about $4.02 \times 10^{-8}$. $\mathrm{AM}_{\mathrm{pca}}$ reaches its cap $10^3$ and continues toward the origin more slowly. Near zero, each block update then has a contraction factor close to $(10^3 - 1)/10^3 = 0.999$. Its solution error reaches about $5.09 \times 10^{-12}$ by sweep $10^4$.

$\mathrm{AM}_{\mathrm{our}}$ starts more slowly, then rapidly reduces both the factor gap and solution error while its penalty approaches $1.6677$. Its solution error is below $10^{-12}$ by sweep 22. In contrast, $\mathrm{AM}_{\mathrm{adp}}$ keeps a penalty near $10^{-5}$ and remains close to $(1, -1)$, with a factor gap near two and an objective near two. The nearly coincident objective curves for the other three methods at later iterations conceal substantial differences in solution error.




(a) Updated penalty $\gamma_k$.

(b) Factor gap $|x_k - y_k|$.

(c) Solution error $|x_k - x^\star|$.

(d) Symmetric objective $f(y_k^2)$.

Figure 3: Penalty schedules for (30) with $d = 1$. Small factor gaps and near-minimal objective values need not imply comparable solution errors. Zeros are omitted on the logarithmic axes.

Table 5: Recorded values after $10^4$ sweeps for the second scalar example (rounded). The zeros for $\mathrm{AM}_{\mathrm{our}}$ arise in floating-point arithmetic.

| Method | $\gamma_k$ | $\lvert x_k - x^\star \rvert$ | $\lvert x_k - y_k \rvert$ |
|---|---|---|---|
| $\mathrm{AM}_{\mathrm{agd}}$ | $1.2148 \times 10^{16}$ | $4.0195 \times 10^{-8}$ | $6.6174 \times 10^{-24}$ |
| $\mathrm{AM}_{\mathrm{adp}}$ | $1.0000014 \times 10^{-5}$ | $0.9999833$ | $1.9999800$ |
| $\mathrm{AM}_{\mathrm{pca}}$ | $10^3$ | $5.0931 \times 10^{-12}$ | $5.0931 \times 10^{-15}$ |
| $\mathrm{AM}_{\mathrm{our}}$ | $1.6677$ | $0$ | $0$ |

Together, the scalar examples in Sections 4.1 and B.1 illustrate the distinction between factor agreement and optimization accuracy. Insufficient penalties can leave a large mismatch, whereas rapid penalty growth can stall progress even when the factors nearly agree. The proposed update avoids both behaviors in these tests and reaches high accuracy with a moderate penalty.

### B.2 Low-rank Matrix Recovery

Low-rank matrix recovery arises in recommender systems [39], collaborative filtering [41], and computer vision [2]. We use matrix completion to examine penalty updates with partially observed data.

Given an incomplete observed matrix $\hat{O}$ and an observation indicator $\hat{\Omega} \in \{0,1\}^{m\times n}$, matrix completion seeks a low-rank matrix $M \in \mathbb{R}^{m\times n}$ that fits the observed entries. A nuclear-norm-regularized formulation [34] is

$$\min_{M\in\mathbb{R}^{m\times n}} \frac{1}{2}\|\hat{\Omega}\odot(M-\hat{O})\|_F^2 + \frac{\eta}{2}\|M\|_*, \tag{33}$$

where $\|\cdot\|_*$ denotes the nuclear norm, $\odot$ denotes element-wise multiplication, and $\eta > 0$ controls regularization. After multiplying its objective by two, problem (33) has the equivalent PSD formulation [27, 43]

$$\min_{Z\in\mathbb{S}_+^{m+n}} \frac{1}{2}\|\Omega\odot(Z-O)\|_F^2 + \frac{\eta}{2}\operatorname{tr}(Z), \tag{34}$$

where

$$Z = \begin{bmatrix} V & M \\ M^\top & W \end{bmatrix}, \qquad O = \begin{bmatrix} \mathbf{0} & \hat{O} \\ \hat{O}^\top & \mathbf{0} \end{bmatrix}, \qquad \Omega = \begin{bmatrix} \mathbf{0} & \hat{\Omega} \\ \hat{\Omega}^\top & \mathbf{0} \end{bmatrix}.$$

The substitution $Z = XX^\top$ gives the regularized sBMF formulation (2) with

$$f(XX^\top) = \frac{1}{2}\|\Omega\odot(XX^\top - O)\|_F^2, \qquad h(X) = \frac{\eta}{2}\|X\|_F^2.$$

A fixed factor width $r$ additionally restricts $\operatorname{rank}(Z) \le r$.

**Experimental setup.** We use the MovieLens datasets [16][4], whose statistics are listed in Table 6. The original ratings use a five-star scale, with half-star increments in MovieLens-10M. We normalize each rating matrix by its largest absolute entry, and set $\eta = 10^{-3}$ and $r = 2$ for all three datasets. Testing RMSE is evaluated on the normalized rating scale:

$$\mathrm{RMSE} = \left(\frac{1}{|\mathcal{I}_{\mathrm{test}}|}\sum_{(i,j)\in\mathcal{I}_{\mathrm{test}}} (\widehat{M}_{ij} - \hat{O}_{ij})^2\right)^{1/2},$$

where $\mathcal{I}_{\mathrm{test}}$ is the set of test entries, $\widehat{M}$ contains the predicted normalized ratings, and $\hat{O}$ denotes the normalized observations. For each user, we use $70\%$ of the ratings for training and the remaining $30\%$ for testing. All methods use a maximum of $3{,}000$ iterations. We stop earlier if the objective value change over three consecutive iterations falls below $10^{-4}$. For MovieLens-10M, we also impose a four-hour wall-clock limit per run. Results are averaged over $50$ runs for MovieLens-100K and MovieLens-1M, and $10$ runs for MovieLens-10M.

**Results.** Table 7 reports testing RMSE and runtime. $\mathrm{AM}_{\mathrm{our}}$ and $\mathrm{HAM}_{\mathrm{our}}$ attain the lowest displayed mean RMSE on all three datasets: $0.19$, $0.18$, and $0.17$, respectively. $\mathrm{AM}_{\mathrm{adp}}$ ties these values on MovieLens-1M and MovieLens-10M at the reported precision. Within the AM group, $\mathrm{AM}_{\mathrm{our}}$ also has the shortest mean runtime on every dataset, giving speedups of approximately $1.59$, $2.31$, and $1.63$ over $\mathrm{AM}_{\mathrm{adp}}$ on 100K, 1M, and 10M, respectively.

$\mathrm{HAM}_{\mathrm{our}}$ achieves lower mean RMSE than $\mathrm{HAM}_{\mathrm{adp}}$ on all three datasets. Its runtime advantage is confined to MovieLens-1M, where it is approximately $2.09$ times faster; on 100K and 10M, its mean runtime is slightly higher. $\mathrm{HAM}_{\mathrm{pca}}$ is the fastest HAM variant on 10M, but its RMSE is substantially larger ($0.42$ versus $0.17$ for $\mathrm{HAM}_{\mathrm{our}}$).

Within the AAPG group, the proposed update offers no clear advantage over the adaptive rule. $\mathrm{AAPG}_{\mathrm{our}}$, $\mathrm{AAPG}_{\mathrm{adp}}$, and $\mathrm{AAPG}_{\mathrm{agd}}$ have the same displayed mean RMSE on 100K and 1M, with $\mathrm{AAPG}_{\mathrm{adp}}$ being slightly faster. On 10M, $\mathrm{AAPG}_{\mathrm{our}}$ and $\mathrm{AAPG}_{\mathrm{adp}}$ both achieve an RMSE of $0.19$ with nearly identical runtimes, whereas $\mathrm{AAPG}_{\mathrm{agd}}$ takes less time but has a higher RMSE of $0.36$. $\mathrm{AAPG}_{\mathrm{pca}}$ has the shortest runtime in its block on 10M, at the cost of a much higher RMSE of $0.79$.

[4] https://grouplens.org/datasets/movielens/

Overall, $\mathrm{AM}_{\mathrm{our}}$ provides the clearest accuracy–time advantage: on each dataset, it is the fastest among the methods attaining the lowest displayed mean RMSE. The improvements are less consistent for HAM and AAPG, indicating that the benefit of the penalty update depends on the underlying solver. These comparisons describe the reported means and do not establish statistical significance.

Table 6: Statistics of the MovieLens datasets.

| Dataset | #Users | #Movies | #Training | #Testing |
|---|---|---|---|---|
| 100K | 943 | 1,682 | 70,012 | 29,988 |
| 1M | 6,040 | 3,449 | 700,029 | 299,685 |
| 10M | 69,878 | 10,677 | 6,997,029 | 3,003,025 |

Table 7: Testing RMSE on normalized ratings and running time in seconds on MovieLens. Entries report the mean $\pm$ standard deviation. Lower means are better. Bold entries mark the lowest displayed mean within each solver block, including ties.

| Method | RMSE | | | Time (s) | | |
|---|---|---|---|---|---|---|
| | 100K | 1M | 10M | 100K | 1M | 10M |
| $\mathrm{AM}_{\mathrm{pca}}$ | $0.30 \pm 0.31$ | $0.25 \pm 0.24$ | $0.47 \pm 0.30$ | $18.94 \pm 16.56$ | $1010.71 \pm 957.29$ | $14450.56 \pm 24.25$ |
| $\mathrm{AM}_{\mathrm{adp}}$ | $0.20 \pm 0.01$ | $\mathbf{0.18 \pm 0.01}$ | $\mathbf{0.17 \pm 0.01}$ | $12.56 \pm 8.86$ | $174.81 \pm 136.86$ | $10962.04 \pm 4695.71$ |
| $\mathrm{AM}_{\mathrm{our}}$ | $\mathbf{0.19 \pm 0.00}$ | $\mathbf{0.18 \pm 0.00}$ | $\mathbf{0.17 \pm 0.00}$ | $\mathbf{7.88 \pm 1.44}$ | $\mathbf{75.57 \pm 7.53}$ | $\mathbf{6723.75 \pm 292.08}$ |
| $\mathrm{HAM}_{\mathrm{pca}}$ | $0.44 \pm 0.47$ | $0.50 \pm 0.55$ | $0.42 \pm 0.39$ | $42.52 \pm 43.78$ | $2380.97 \pm 2451.69$ | $\mathbf{13117.92 \pm 3609.45}$ |
| $\mathrm{HAM}_{\mathrm{adp}}$ | $0.21 \pm 0.07$ | $0.39 \pm 0.48$ | $0.18 \pm 0.01$ | $\mathbf{19.47 \pm 17.29}$ | $499.90 \pm 478.35$ | $13914.98 \pm 1490.56$ |
| $\mathrm{HAM}_{\mathrm{our}}$ | $\mathbf{0.19 \pm 0.00}$ | $\mathbf{0.18 \pm 0.00}$ | $\mathbf{0.17 \pm 0.00}$ | $21.10 \pm 6.45$ | $\mathbf{239.05 \pm 70.40}$ | $14287.33 \pm 623.20$ |
| $\mathrm{AAPG}_{\mathrm{pca}}$ | $12.69 \pm 7.94$ | $7.61 \pm 5.51$ | $0.79 \pm 0.01$ | $75.72 \pm 44.57$ | $821.53 \pm 573.78$ | $\mathbf{1544.25 \pm 171.94}$ |
| $\mathrm{AAPG}_{\mathrm{adp}}$ | $\mathbf{0.20 \pm 0.00}$ | $\mathbf{0.19 \pm 0.00}$ | $\mathbf{0.19 \pm 0.00}$ | $\mathbf{31.32 \pm 4.91}$ | $\mathbf{324.29 \pm 11.50}$ | $14400.54 \pm 1.16$ |
| $\mathrm{AAPG}_{\mathrm{agd}}$ | $\mathbf{0.20 \pm 0.00}$ | $\mathbf{0.19 \pm 0.00}$ | $0.36 \pm 0.26$ | $31.78 \pm 4.60$ | $327.45 \pm 13.77$ | $10518.11 \pm 6253.22$ |
| $\mathrm{AAPG}_{\mathrm{our}}$ | $\mathbf{0.20 \pm 0.00}$ | $\mathbf{0.19 \pm 0.00}$ | $\mathbf{0.19 \pm 0.00}$ | $31.87 \pm 2.97$ | $338.09 \pm 15.12$ | $14401.85 \pm 1.87$ |

### B.3 Link Prediction on Graph

Link prediction infers unobserved edges from partial graph observations [23, 50]. We consider a PSD affinity matrix learned from pairwise supervision [17, 54].

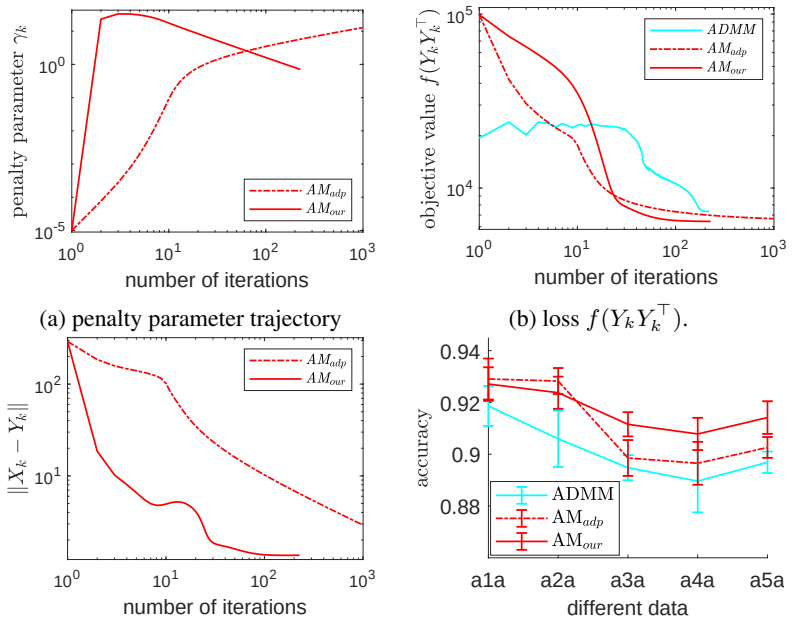
Consider the set of observed pair labels $\mathcal{T} = \{(i, j, t_{ij})\}$, where $t_{ij} = 1$ for positive pairs and $t_{ij} = -1$ for negative pairs. The goal is to learn a PSD matrix $Z$ consistent with the observed labels and regularized by the graph Laplacian $L$:

$$\min_{Z \in \mathbb{S}^n_+} \quad \mathrm{tr}(LZ) + \eta \sum_{(i,j,t_{ij}) \in \mathcal{T}} \ell_{hinge}\left(t_{ij} Z_{ij}\right) \tag{35}$$

where $\ell_{hinge}(u) = \max(1 - u, 0)$ is the hinge loss. To apply Theorem 1 to asymmetric factors, the loss must be extended off the symmetric cone in a transpose-symmetric way. One such extension replaces each $Z_{ij}$ in the loss by $(Z_{ij} + Z_{ji})/2$; this leaves the PSD problem unchanged. The exactness theorem applies to this extension, while the nonsmooth hinge model is outside the smooth-loss convergence theorem.

We conduct experiments on *a1a–a5a* from the *Adult* datasets[5] and report representative results in Figure 4. Subplots (4a)–(4c) correspond to *a5a*, while Figure 4d reports the average prediction accuracy over 10 repeated runs on each dataset. $\mathrm{AM}_{\mathrm{our}}$ has higher mean accuracy than $\mathrm{AM}_{\mathrm{adp}}$ on *a3a–a5a*, but slightly lower accuracy on *a1a* and *a2a*; both exceed the ADMM baseline in the displayed means. On *a5a*, $\mathrm{AM}_{\mathrm{our}}$ has smaller factor disagreement than $\mathrm{AM}_{\mathrm{adp}}$ over most iterations. The trajectory panels use iteration counts rather than elapsed time. The ADMM curves denote an alternating direction method of multipliers baseline.

[5] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/




(a) penalty parameter trajectory

(b) loss $f(Y_kY_k^\top)$.

(c) difference of $X_k$ and $Y_k$.

(d) prediction accuracy.

Figure 4: Link prediction on the Adult datasets. In panel (d), each point is the mean prediction accuracy over 10 runs, and the error bars extend one standard deviation on either side of the mean.

### B.4 Matrix Sensing on Synthetic Data

Consider the linear SDP

$$\min_{Z\in\mathbb{S}^n_+} \operatorname{tr}(C_0Z), \quad \text{s.t.} \quad \operatorname{tr}(C_iZ) = b_i, \quad i = 1,\dots,m, \tag{36}$$

where each $C_i \in \mathbb{R}^{n\times n}$ is symmetric. Restricting $\operatorname{rank}(Z) \le r$ yields the factorized formulation

$$\min_{X\in\mathbb{R}^{n\times r}} \operatorname{tr}(C_0XX^\top), \quad \text{s.t.} \quad \operatorname{tr}(C_iXX^\top) = b_i, \quad i = 1,\dots,m. \tag{37}$$

Bhojanapalli et al. [3] study the squared-penalty model

$$\min_{X\in\mathbb{R}^{n\times r}} \operatorname{tr}(C_0XX^\top) + \frac{\eta}{2}\sum_{i=1}^m \big(\operatorname{tr}(C_iXX^\top) - b_i\big)^2, \tag{38}$$

where $\eta > 0$ controls the penalty on constraint violations. Problem (38) has the symmetric form (2) with

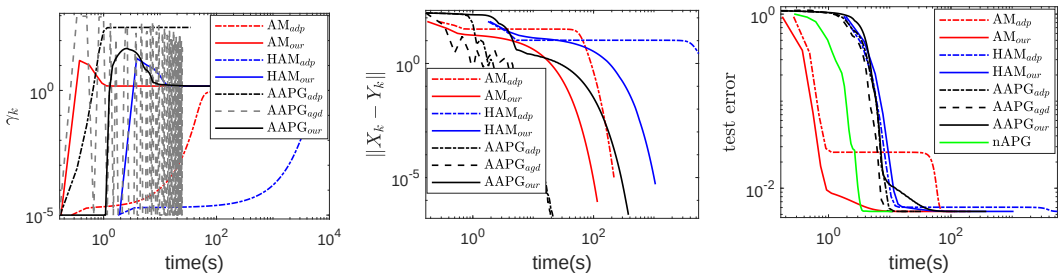
$$f(Z) = \operatorname{tr}(C_0Z) + \frac{\eta}{2}\sum_{i=1}^m \big(\operatorname{tr}(C_iZ) - b_i\big)^2$$

and $h = 0$. Replacing $XX^\top$ by $XY^\top$ and adding the factor-agreement penalty gives (3).

**Experimental setup.** We consider a PSD recovery problem with partially observed entries. In (38), we set $C_0 = I_n$, and each $C_i$ represents an observed entry of the symmetric decision matrix $Z$, with observed value $b_i$. We generate noisy data

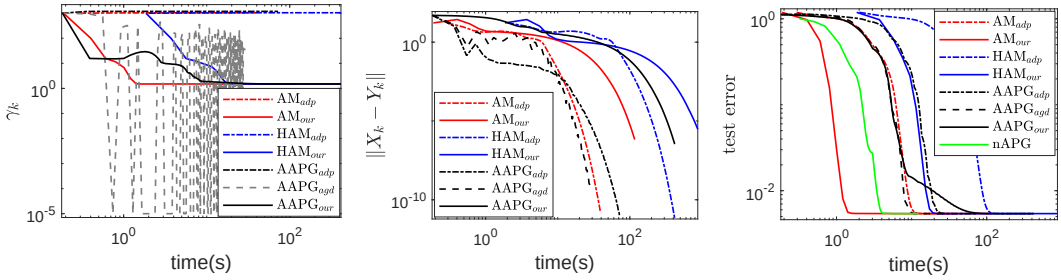
$$Z_\star = X_\star X_\star^\top + S,$$

where the entries of $X_\star$ and $S$ are drawn independently from $N(0,1)$ and $N(0,0.05)$, respectively. The noisy matrix $Z_\star$ need not be symmetric or PSD. We sample $15n\log(n)$ entries for training and use the remaining entries for testing; the symmetric formulation constrains the fitted matrix to be PSD. We also include nAPG [28], an accelerated proximal gradient method applied directly to the symmetric problem (38).




(a) Penalty parameter trajectory. (b) Factor difference $\|X_k - Y_k\|_F$. (c) Recovery error on test data.

Figure 5: Results for matrix sensing on synthetic data with initial $\gamma_0 = 10^{-5}$.




(a) Penalty parameter trajectory. (b) Factor difference $\|X_k - Y_k\|_F$. (c) Recovery error on test data.

Figure 6: Results for matrix sensing on synthetic data with initial $\gamma_0 = 10^{3}$.

**Results.** Figures 5 and 6 compare the penalty parameter, factor difference $\|X_k - Y_k\|_F$, and test recovery error against elapsed time, with initial penalties $\gamma_0 = 10^{-5}$ and $\gamma_0 = 10^{3}$, respectively. Despite this large difference in initialization, $\text{AM}_{\text{our}}$, $\text{HAM}_{\text{our}}$, and $\text{AAPG}_{\text{our}}$ settle at similar penalty values of order one. By contrast, $\text{AM}_{\text{adp}}$ and $\text{HAM}_{\text{adp}}$ increase their penalties slowly from the small initialization and retain large values from the large initialization. $\text{AAPG}_{\text{agd}}$ exhibits pronounced penalty oscillations in both runs. These observations suggest that the proposed update reduces the dependence of the terminal penalty on its initial value.

The main differences in recovery concern the time needed to reach a low test error, since the final errors are comparable. With $\gamma_0 = 10^{-5}$, $\text{AM}_{\text{our}}$ achieves the fastest initial error reduction and avoids the intermediate plateau of $\text{AM}_{\text{adp}}$. With $\gamma_0 = 10^{3}$, it reaches the low-error plateau first. $\text{HAM}_{\text{our}}$ also reaches low test error earlier than $\text{HAM}_{\text{adp}}$ in both runs. The advantage does not extend uniformly to the AAPG variants: $\text{AAPG}_{\text{adp}}$ and $\text{AAPG}_{\text{agd}}$ reach the final error plateau earlier than $\text{AAPG}_{\text{our}}$. Direct symmetric optimization with nAPG is also competitive, particularly in Figure 5, where it reaches the low-error plateau before $\text{AM}_{\text{our}}$.

The factor-difference curves show a different ordering. For $\gamma_0 = 10^{-5}$, $\text{HAM}_{\text{our}}$ substantially reduces the factor difference, whereas $\text{HAM}_{\text{adp}}$ retains a large mismatch over most of the run. For $\gamma_0 = 10^{3}$, however, the adaptive variants generally drive the factors closer together more quickly than their proposed counterparts. Thus, faster factor agreement does not necessarily translate into faster recovery on the test data. The benefits of the proposed update are clearest for AM and HAM in recovery speed, while its effect on factor agreement depends on the underlying algorithm and initialization.

### B.5 Standard Max-Cut SDP

We consider the standard Max-Cut relaxation [11] $\max_{Z \succeq 0,\, \mathrm{diag}(Z)=\mathbf{1}} \frac{1}{4}\langle L, Z\rangle$ on Gset graphs[6], where $L$ is the graph Laplacian. The asymmetric formulation handles $\mathrm{diag}(XY^\top) = \mathbf{1}$ with an augmented Lagrangian. The symmetric baselines are projected gradient descent (sBMF-PGD) and the Mixing method (sBMF-Mixing) [46]. We run G55 at rank 50 and G67 at ranks 50 and 100, with

[6] https://web.stanford.edu/~yyye/yyye/Gset/

ten paired trials per configuration. All methods share the same row-normalized Gaussian initialization and 100 random rounding hyperplanes within each trial. The experiments use MATLAB on the CU cluster, with 64 CPU cores allocated per trial.

$\mathrm{AM}_{\mathrm{our}}$ uses (15) with $\nu = 0.9$, $\varepsilon_0 = 10^{-6}$, and $\sigma_h = 0$. After each pair of factor updates, we fix the updated unscaled multiplier $\lambda_k$ and augmented-Lagrangian penalty $\rho_k$. The minimization-form augmented loss is

$$f_k(Z) = -\tfrac{1}{4}\langle L_{\mathrm{work}}, Z\rangle + \lambda_k^\top\big(\mathrm{diag}(Z) - \mathbf{1}\big) + \tfrac{\rho_k}{2}\big\|\mathrm{diag}(Z) - \mathbf{1}\big\|_2^2.$$

Its gradient with respect to $Z$, evaluated at $X_k Y_k^\top$, is

$$G_k = \partial_Z f_k(X_k Y_k^\top) = -\tfrac{1}{4}L_{\mathrm{work}} + \mathrm{Diag}\big(\lambda_k + \rho_k\big(\mathrm{diag}(X_k Y_k^\top) - \mathbf{1}\big)\big)\,,$$

where $L_{\mathrm{work}}$ is the internally scaled Laplacian. The Frobenius norm and directional quotient reuse the sparse Laplacian and cached matrix–factor products; no dense $n \times n$ gradient is formed. The resulting $\gamma_k$ is used in the next factor sweep. All methods allow at most 3000 iterations or sweeps. AM stops when the relative feasible-objective change is below $10^{-6}$, relative factor disagreement is below $10^{-3}$, and the raw equality residual is below $10^{-6}$; the symmetric methods retain their own stopping tests.

Table 8 reports means over ten trials. We evaluate the SDP objective and rounded cut using the row-normalized factor $U$ obtained from $(X+Y)/2$, and measure the equality residual on the raw product $XY^\top$. In the three configurations, $\mathrm{AM}_{\mathrm{our}}$ reduces mean runtime relative to $\mathrm{AM}_{\mathrm{adp}}$ by factors of 2.56, 1.37, and 1.24, respectively, while attaining higher mean SDP objectives and cuts. The symmetric methods attain higher mean cuts, and sBMF-Mixing is faster in all three configurations. $\mathrm{AM}_{\mathrm{our}}$ is faster than sBMF-PGD on G55, but slower on G67.

All 30 $\mathrm{AM}_{\mathrm{our}}$ runs meet the joint stopping test; all $\mathrm{AM}_{\mathrm{adp}}$ and sBMF-PGD runs reach the iteration cap. The largest final raw equality residual for $\mathrm{AM}_{\mathrm{our}}$ is $2.96 \times 10^{-9}$, and the largest relative factor disagreement is $2.42 \times 10^{-4}$. Its 50,358 recorded penalty updates satisfy (15), with 60 decreases and no increases. The directional safeguard never determines the update in this batch, so these results do not isolate its benefit.

Table 8: Standard Max-Cut SDP: means over ten matched trials. Higher SDP objectives and rounded cuts are better; lower time and equality residual are better.

| Graph, rank | Method | Time (s) | SDP objective | Rounded cut | Equality residual |
|---|---|---|---|---|---|
| G55, 50 | sBMF-PGD | 13.29 | 11038.58 | 9883.5 | 0 |
| | sBMF-Mixing | 3.47 | 11039.00 | 9892.4 | 0 |
| | $\mathrm{AM}_{\mathrm{adp}}$ | 29.21 | 11020.44 | 9852.8 | $1.2\times10^{-9}$ |
| | $\mathrm{AM}_{\mathrm{our}}$ | 11.41 | 11034.70 | 9876.3 | $2.5\times10^{-9}$ |
| G67, 50 | sBMF-PGD | 20.26 | 7741.50 | 6282.6 | 0 |
| | sBMF-Mixing | 18.10 | 7741.90 | 6283.0 | 0 |
| | $\mathrm{AM}_{\mathrm{adp}}$ | 40.22 | 7717.07 | 6221.4 | $6.5\times10^{-10}$ |
| | $\mathrm{AM}_{\mathrm{our}}$ | 29.40 | 7729.50 | 6255.2 | $3.9\times10^{-10}$ |
| G67, 100 | sBMF-PGD | 30.83 | 7741.58 | 6289.6 | 0 |
| | sBMF-Mixing | 22.49 | 7741.95 | 6289.0 | 0 |
| | $\mathrm{AM}_{\mathrm{adp}}$ | 66.08 | 7717.65 | 6228.6 | $7.0\times10^{-10}$ |
| | $\mathrm{AM}_{\mathrm{our}}$ | 53.38 | 7729.68 | 6263.2 | $3.6\times10^{-10}$ |

### B.6 Neural-Network Surrogate and Evaluation

**Classifier and lifted variables.** We use a fixed ReLU classifier for digits 3 and 8, with 249 training samples, 108 test samples, $d = 12$ PCA-standardized input features, and $m = 16$ hidden units. The classifier assigns labels $+1$ to digit 3 and $-1$ to digit 8 and correctly classifies 107 of the 108 test samples. The perturbation box for an input $x_0$ is $[\ell, u]$, where $\ell = x_0 - \epsilon\mathbf{1}$ and $u = x_0 + \epsilon\mathbf{1}$ in feature space. Let $W_1 \in \mathbb{R}^{m\times d}$ and $b_1 \in \mathbb{R}^m$ be the hidden layer parameters, and let $w_2 \in \mathbb{R}^m$ and $b_2 \in \mathbb{R}$ be the output layer parameters. With $W_+ = \max(W_1, 0)$ and $W_- = \min(W_1, 0)$ taken entrywise, the preactivation bounds are

$$\ell_s = W_+\ell + W_-u + b_1, \qquad u_s = W_+u + W_-\ell + b_1.$$

For the symmetric lift $Z \in \mathbb{S}^{1+d+m}$, let 0 index the homogenizing coordinate and let $\mathcal{I}_x$ and $\mathcal{I}_a$ index the input and activation coordinates. Define

$$t = Z_{00}, \quad x = Z_{0,\mathcal{I}_x}^\top, \quad a = Z_{0,\mathcal{I}_a}^\top, \quad s = W_1 x + b_1 t, \quad Q = Z_{\mathcal{I}_x,\mathcal{I}_x}, \quad R = Z_{\mathcal{I}_a,\mathcal{I}_a}.$$

The equality residual vector $e(Z)$ collects

$$t - 1, \qquad a_i \quad (u_{s,i} \le 0), \qquad a_i - s_i \quad (\ell_{s,i} \ge 0,\ u_{s,i} > 0).$$

The inequality residual vector $g(Z)$ collects the following expressions, each constrained to be nonpositive in the underlying model:

$$\begin{aligned}
&\ell_j t - x_j, \quad x_j - u_j t, \quad Q_{jj} - (\ell_j + u_j) x_j + \ell_j u_j t && (j = 1, \dots, d),\\
&-a_i, \quad s_i - a_i && (i = 1, \dots, m),\\
&a_i - \alpha_i s_i + \alpha_i \ell_{s,i} t, \quad \alpha_i = \frac{u_{s,i}}{u_{s,i} - \ell_{s,i}} && (\ell_{s,i} < 0 < u_{s,i}),\\
&R_{ii} - u_{s,i} a_i && (u_{s,i} > 0).
\end{aligned}$$

**Loss and diagnostics.** Let $y_0 \in \{-1, 1\}$ be the label of $x_0$ and $c(Z) = a \odot (a - s)$, where $\odot$ denotes entrywise multiplication. The surrogate loss is

$$\begin{aligned}
f_{\mathrm{nc}}(Z) = y_0(w_2^\top a + b_2 t) + \lambda \|Z\|_F^2 + \rho_{\mathrm{eq}} \|e(Z)\|_2^2 \\
+ \rho_{\mathrm{ineq}} \|[g(Z)]_+\|_2^2 + \rho_{\mathrm{comp}} \|c(Z)\|_2^2,
\end{aligned}$$

where $[\cdot]_+$ is applied entrywise, $\lambda = 10^{-3}$, $\rho_{\mathrm{eq}} = \rho_{\mathrm{ineq}} = 50$, and $\rho_{\mathrm{comp}} = 100$. The complementarity penalty is nonconvex. We use $Z = XX^\top$ for sBMF and $Z = (XY^\top + YX^\top)/2$ for aBMF. The asymmetric objective adds $(\gamma/2)\|X - Y\|_F^2$ to $f_{\mathrm{nc}}(Z)$, whereas Table 3 reports $f_{\mathrm{nc}}(Z)$ alone.

The reported residuals are

$$\begin{aligned}
&r_{\mathrm{aff}} = \|e(Z)\|_\infty, && r_{\mathrm{ineq}} = \|[g(Z)]_+\|_\infty, && r_{\mathrm{comp}} = \|c(Z)\|_2,\\
&r_{\mathrm{dis}} = \frac{\|X - Y\|_F}{\max\{1, \|X\|_F, \|Y\|_F\}}, && r_{\mathrm{PSD}} = [-\lambda_{\min}(Z)]_+.
\end{aligned}$$

A run passes if $f_{\mathrm{nc}}(Z)$ is finite and $\max\{r_{\mathrm{aff}}, r_{\mathrm{ineq}}, r_{\mathrm{comp}}, r_{\mathrm{dis}}\} \le 0.05$, with $r_{\mathrm{dis}}$ omitted for sBMF. The PSD residual is reported separately and does not enter this test.

**Experimental protocol.** We select three correctly classified test inputs at the 25th, 50th, and 75th percentiles of the nominal signed margin (test indices 25, 74, and 29). We use radii $\epsilon \in \{0.6, 0.8\}$ and factor widths $r \in \{8, 16\}$, with ten matched trials per combination, for 120 runs per method. The first factor column is $[1; x_0; \max(W_1 x_0 + b_1, 0)]$, and the remaining columns are initialized with independent $10^{-3} N(0, 1)$ entries. Asymmetric methods start with $Y_0 = X_0$.

All runs in Table 3 reach the 3,000-iteration cap. The fixed baseline uses $\gamma = 10$, the multiplicative baseline caps $\gamma$ at $10^4$, and $\mathrm{AM}_{\mathrm{our}}$ applies its penalty update before the first factor step.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the scope of the spectral exactness condition, the initialization-based bounds, and the sNMF counterexample and structured cases; see Theorem 1, Corollaries 1–2, and Remark 2.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper identifies the penalty bounds as sufficient, restricts the convergence theorem to fixed-penalty exact AM, and explains that the nonconvex NN diagnostic is not a robustness certificate; see Sections 3.4, 4.4, and 5.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions are stated with the theoretical results in Section 3.2, and complete proofs are provided in Appendix A.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The formulations, update rules, benchmark sources, and main experimental settings are given in Sections 3.3 and 4 and Appendix B. The theoretical claims have complete proofs, and the public code provides a reproduction route for the numerical benchmarks covered by the release described below.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The public repository at `https://github.com/ynelhu/aBMF` provides runnable code for the scalar tests, Top-$k$ approximation, NN surrogate, and Max-Cut, plus an AM example for MovieLens. The NN package includes fixed inputs and classifier weights, and sources of the public benchmark datasets are cited and linked in the paper.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen)?

   Answer: [Yes]

   Justification: Section 4 and Appendix B report data splits, preprocessing, parameter settings, initialization, stopping criteria, and evaluation procedures. Implementation details for the released benchmarks are provided in the accompanying code.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 7 reports means and standard deviations over 50 runs for MovieLens-100K and MovieLens-1M and 10 runs for MovieLens-10M. Other aggregate tables state their trial counts and report means; these comparisons are not presented as formal significance tests.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4 specifies the MATLAB implementation and CPU server with 64 cores and 512 GB RAM. The experiments report wall-clock runtimes, trial counts, and iteration budgets; Section B.5 also gives the CPU allocation per trial. Costs of preliminary experiments are not separately itemized.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, the work complies with the NeurIPS Code of Ethics. It studies optimization theory and existing public benchmarks, without new human-subject data collection or release of high-risk models.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational work on low-rank optimization. The experiments assess optimization behavior on public benchmarks, and no new deployment with direct societal impacts is proposed.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work does not release high-risk models, scraped datasets, or other assets requiring controlled access or misuse safeguards.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The datasets and comparison methods are cited and linked, but asset-specific licenses and terms of use are not enumerated for every external resource.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released implementations include runnable examples and experiment-specific usage instructions. The NN benchmark also provides the fixed inputs and classifier weights used in its experiments.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work uses existing public benchmark data and does not involve crowdsourcing or human-participant studies.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human-participant study or crowdsourcing was conducted, so IRB approval is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not components of the proposed formulations, optimization algorithms, or experiments. They assisted language editing and checking mathematical arguments; the authors remain responsible for the scientific content.