Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Regularized coordinate minimization for quantized matrix factorization with application to image compression

Daniela Lupu
Universitatea Națională de Știință și Tehnologie Politehnica București
George T. Samoila
Universitatea Națională de Știință și Tehnologie Politehnica București
Adina M. Florea
Universitatea Națională de Știință și Tehnologie Politehnica București
Ion Necoara
ion.necoara@upb.ro

Universitatea Națională de Știință și Tehnologie Politehnica București

Research Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11599962/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Noname manuscript No.
(will be inserted by the editor)

Regularized coordinate minimization for quantized matrix factorization with application to image compression

Daniela Lupu · George T. Samoila · Adina M. Florea · Ion Necoara

Received: date / Accepted: date

Abstract This paper presents a regularized cyclic coordinate minimization method for solving nonconvex composite optimization problems having the objective function formed as the sum of two terms, one is twice continuously differentiable and the second term is simple and separable. We analyze the convergence behaviour of our coordinate minimization method, providing rates to first-order stationarity and objective residual depending on the assumptions on the problem. However, for optimization problems with bounded integer variables, the first-order stationarity does not measure the quality of a solution. Hence, on integer constrained problems we derive a discrete optimality criterion based on the decrease of the objective function and then prove convergence rates in this criterion or even finite termination for our algorithm with fixed and adaptive regularization parameters. Then, we show that our algorithmic framework can be efficiently applied for solving quantized matrix factorization

The research leading to these results has received funding from: project Romanian Hub for Artificial Intelligence - HRIA, Smart Growth, Digitization and Financial Instruments Program, 2021-2027, SMIS no. 2021-351416

Daniela Lupu
Automatic Control and Systems Engineering Department and PRECIS Research Institute, Politehnica Bucharest
E-mail: daniela.lupu@upb.ro

George T. Samoila
Automatic Control and Systems Engineering Department, Politehnica Bucharest
E-mail: samoilatibi.st@gmail.com

Adina M. Florea
Computer Science Department, Politehnica Bucharest
E-mail: adina.florea@upb.ro

Ion Necoara
Automatic Control and Systems Engineering Department, Politehnica Bucharest and Gheorghe Mihoc-Caius Iacob Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy
E-mail: ion.necoara@upb.ro

problems where the elements of the factor matrices are constrained to finite integer values, arising e.g., in lossy image compression. For this particular problem we derive an explicit expression of the discrete optimality measure and convergence estimates depending on the regularization parameters. We show numerically, on KODAK and CLIC 2024 datasets, that our method notably outperforms JPEG at low bit rates, achieving savings in bits per pixel without excessive degradation and remains comparable at higher bit rates. Moreover, we evaluate both the original and quantized compressed images in a classification task using 3 well-known convolutional networks AlexNet, ResNet50 and MobileNetV2 based on floating point and integer arithmetic representations. When comparing the two numerical representations of the convolutional networks, there is no substantial accuracy degradation on ImageNet dataset. The numerical results on image compression and classification using real data show the flexibility and efficiency of our algorithm compared to well-established methods from the literature. The code can be found on Github: rqmf-image-compression.

Keywords Regularized coordinate minimization · nonconvex optimization · discrete optimality · quantized matrix factorization · image compression

1 Introduction

In many applications ranging from signal processing (e.g., dimensionality reduction, data compression) [17] to distributed control [25], one needs to solve large-scale nonconvex composite optimization problems. However, for large-scale nonconvex optimization, the traditional methods based on full gradient/prox and on Hessian computations are prohibitive. In this case, a reasonable approach to solve such problems is to use (block) coordinate descent methods as each iteration is cheap since it involves optimizing a small-dimensional subproblem, see [4]. Comprehensive surveys on coordinate descent algorithms can be found in e.g., [23, 33].

The variants of (block) coordinate descent algorithms differ from each other in the way we define the local approximation function over which we optimize and the criterion of choosing at each iteration the subspace over which we minimize this local approximation function. For updating one (block) variable, while keeping the other variables fixed, two basic choices for the local approximation function are usually considered: (i) exact approximation function, leading to *coordinate minimization methods* [4,24], and (ii) quadratic approximation function, leading to *coordinate (proximal) gradient descent methods* [3, 24, 26, 30]. Furthermore, three classical criteria for choosing the coordinate search used often in these algorithms are the greedy, the cyclic and the random coordinate search, which differ significantly in the computational effort required to select the next coordinate. For cyclic coordinate search rates of convergence have been given only recently [3, 6]. Convergence rates for coordinate descent methods based on the Gauss-Southwell rule were given e.g.

in [30]. Finally, a third approach is based on stochastic coordinate descent, where the coordinate search is random. Recent complexity results on stochastic coordinate descent methods were obtained e.g., in [11, 26] in the convex settings and then extended to nonconvex problems as well, see e.g., [24].

Contributions. This paper deals with nonconvex composite optimization problems having the objective function formed as the sum of two terms, one is twice continuously differentiable and the second is simple and (block) separable, both of which may be nonconvex. Under these settings we consider a regularized cyclic multiblock coordinate minimization (RCCM) method that takes into account the composite form of the objective. RCCM minimizes at each iteration the original objective function along a subspace with user-determined dimension, together with a quadratic regularization term. We provide convergence analysis for RCCM, deriving worst-case bounds for achieving approximate first-order stationarity and objective residual. In the case of bounded integer variables, such a first-order stationarity criterion is satisfied at every feasible point and therefore does not yield information about solution quality. Hence, our convergence analysis but also the convergence results from [1, 2, 6] provide no meaningful guarantees for such problems. In order to overcome this limitation, we introduce a discrete optimality measure based on the decrease of the objective function (called *block-coordinate improvement gap*) that is a more appropriate optimality measure for integer constrained problems. We establish convergence rates in the block-coordinate improvement gap criterion and even finite termination for our algorithm with fixed and adaptive regularization parameters. Note that RCCM differs from the proximal alternating *linearized* algorithm from [6], which minimizes a quadratic approximation of the composite objective based on the linearization of the differentiable term and an estimate of the coordinatewise Lipschitz constant. On the other hand, RCCM uses the whole block objective and does not need to know any Lipschitz constant estimates. Our algorithm is related to the proximal alternating minimization method from [2], which considers alternating minimization with quadratic proximal regularization and assumes Lipschitz continuous gradient as [6]. In our work however, Lipschitz continuity assumption is removed, we give a cyclic multiblock formulation, we consider new adaptive regularization variants and investigate its application to integer-constrained problems, deriving new convergence results for this case.

Then, we show that quantized matrix factorization (QMF) problems fit into our algorithmic framework, leading to a method (called RQMF) with simple (closed-form) iterations and finite time convergence. QMF is an extension of the traditional matrix factorization [14] that explicitly incorporates quantization constraints into the optimization problem (i.e., the elements of the factor matrices are constrained to a set of bounded integer values), see [1,10,12]. The quantization step allows for a more effective low-rank approximation for data compression tasks, such as image compression, being useful when dealing with storing and transmission tasks, particularly when there are hardware limitations. For example, in lossy image compression, quantization reduces storage

size by eliminating redundant information and details that are less visually detected by humans. Our extensive numerical results using Kodak [20] and CLIC 2024 [28] datasets, confirm the efficiency of our coordinate minimization algorithm (RQMF) on solving QMF problems for lossy image compression, producing bits per pixel (bpp) savings without excessive degradation. Remarkably, our method (RQMF) outperforms numerically quantized compression algorithm in [1] and JPEG in several metrics (PSNR and SSIM) at low bit rates (< 0.3 bpp) and remains comparable at higher bit rates. Additionally, RQMF has also several important theoretical advantages compared to the compression algorithm in [1]: (i) due to regularization the objective function along RQMF iterates decreases strictly whenever the iterates change while [1] proves only nonincreasing objective values along their iterates; (ii) the result on convergence of the iterates to stationary points in [1][Proposition 2] is vacuous as any feasible integer point is a stationary point for QMF problem (see Remark 1 below), while we provide for RQMF method finite termination guarantees and obtain convergence bounds for a discrete optimality measure (called *column improvement gap*) which is significantly better than subdifferential stationarity.

We also investigate the impact of the quantized image compression in learning tasks such as classification. Driven by the need to improve deep neural networks (DNNs) speed and cost-effectiveness, especially when deployed on hardware with limited capabilities, there is growing demand for frameworks that support low-precision arithmetic. Hence, we consider both, floating point and integer arithmetic representations for some widely used convolutional neural network architectures such as AlexNet [22], ResNet50 [16] and MobileNetV2 [29]. For inference we use the original ImageNet validation dataset (50.000 images) [8], but also the quantized compressed ones and evaluate the models accuracy on the original DNNs and the post training quantized version of these networks [13, 32]. Comparing the two numerical representations of the convolutional networks AlexNet, ResNet50 and MobileNetV2, one can notice that there is no substantially accuracy degradations on (quantized and compressed) ImageNet dataset.

Hence, our key contributions can be summarized as follows:

(i) We consider a regularized cyclic coordinate minimization (RCCM) algorithm for nonconvex composite optimization problems, where the objective is the sum of a twice continuously differentiable term and a simple block-separable term, both potentially nonconvex. We also design new adaptive variants for the regularization parameters.

(ii) For RCCM we derive sufficient descent, a sublinear stationarity bound for bounded iterates, and local objective residual rates under the Kurdyka-Lojasiewic (KL) property. We further design an adaptive regularization yielding a sublinear estimate on the sum of block-coordinate improvement gaps evaluated at successive intermediate iterates. For problems with finite integer domains, we prove finite termination with explicit bound on the

block-coordinate improvement gap under fixed regularization and exact blockwise optimality under the adaptive regularization rule.

(iii) Then, we show that our algorithm can efficiently solve the quantized matrix factorization problem that can be used, e.g., in lossy image compression tasks. We prove finite termination with explicit bounds on the column improved gap and establish exact columnwise optimality for rational-grid data under sufficiently small fixed regularization parameters. Furthermore, we develop an inexpensive adaptive regularization rule based on explicit thresholds, guaranteeing finite termination and exact columnwise optimality for arbitrary real data. The numerical results on image compression and classification using real data show the flexibility and efficiency of our optimization algorithm.

Notations. We denote with $\mathbb{Z}$ and $\mathbb{R}$ the set of integer and real numbers and $\mathbb{Z}_{[\alpha,\beta]} = [\alpha, \beta] \cap \mathbb{Z}$. For a given vector $x \in \mathbb{R}^n$, its projection onto set $\mathcal{X}$ is denoted $[x]_{\mathcal{X}}$. For a scalar $y \in \mathbb{R}$, we denote $[y]_{\mathbb{Z}_{[\alpha,\beta]}} = \min(\beta, \max(\text{round}(y), \alpha)) = \text{clamp}(\alpha, \beta, (\text{round}(y))$, where $\text{round}(y)$ returns the integer value closest to y. Let $E \in \mathbb{R}^{n\times n}$ be a column permutation of the $n \times n$ identity matrix and let $E = [E_1, ..., E_N]$ be a decomposition of E into N submatrices, with $E_i \in \mathbb{R}^{n\times n_i}$, where $\sum_{i=1}^N n_i = n$. Then, any vector $x \in \mathbb{R}^n$ can be written uniquely as $x = \sum_{i=1}^N E_i x^{(i)}$, where $x^{(i)} = E_i^T x \in \mathbb{R}^{n_i}$. For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, $\nabla f(x) \in \mathbb{R}^n$ denotes the gradient of f at x, $\nabla_j f(x) \in \mathbb{R}$ denotes the jth component of the gradient, while $\nabla_{(i)} f(x) = E_i^T \nabla f(x) \in \mathbb{R}^{n_i}$ denotes the ith block component of the gradient. We denote $\nabla^2 f(x)$ the Hessian of f. We denote $\partial f(x)$ the limiting subdifferential (or simply the subdifferential) of f at x. For a squared matrix $A \in \mathbb{R}^{n\times n}$, $\text{Trace}(A) = \sum_{i=1}^n A_{ii}$ and $\|A\|_F^2 = \sum_{i,j=1}^n A_{ij}^2$. For a bounded integer set $\mathcal{D}$, $|\mathcal{D}|$ denotes its cardinality.

Content. In Section 2 we introduce our optimization problem and the main assumptions followed by the presentation of the regularized coordinate minimization scheme and its convergence analysis. In Section 3 we show that the quantized matrix factorization problem can be recast into our algorithmic framework, while Section 4 provides detailed numerical results for image compression and its efficiency in a classification task.

2 Regularized cyclic coordinate minimization for nonconvex composite optimization

In this section we present our problem of interest and the main assumptions, then we formulate the cyclic multiblock proximal coordinate minimization scheme and analyze its convergence properties.

2.1 Problem formulation

We consider the following nonconvex composite optimization problem (see also [6, 24]):

$$F^* = \min_{x \in \mathbb{R}^n} F(x) := f(x) + g(x), \tag{1}$$

where f is a differentiable (possibly nonconvex) function and g is a proper lower semi-continuous function (e.g., the indicator function of a nonconvex set). Such problems arise in many applications ranging e.g., from signal processing [17] to distributed control [25]. Throughout the paper the following assumptions are valid for the composite problem (1):

Assumption 1 *(i) Function f is twice continuously differentiable, function g is proper, lower semi-continuous and separable, i.e.:*

$$g(x) = \sum_{i=1}^{N} g_i(x^{(i)}),$$

and for any given $\hat{y}_i := (y^{(1)}, \cdots, y^{(i-1)}, y^{(i+1)}, \cdots y^{(N)})$, with $y^{(j)} \in dom\, g_j$ for all j, we have that the following minimum is attained and finite:

$$\min_{x^{(i)} \in \mathbb{R}^{n_i}} f(x^{(i)}, \hat{y}_i) + g_i(x^{(i)}) > -\infty.$$

(ii) A solution exists for (1) *(hence, $F^* > -\infty$).*

Optimization problems having the composite structure (1) permit to handle general coupling functions (e.g., $f(x^{(1)}, x^{(2)}) = \|(x^{(1)})^T A x^{(2)} - y\|^p$, with A linear operator and $p \geq 2$). In particular, this model satisfying Assumption 1 is suitable for applications where for the first term, although possibly smooth, the computations of its gradient or of the coordinatewise Lipschitz constants over a bounded set along subspaces generated by E_i's are expensive or even impossible; on the other hand the computation of the prox along a subspace is easy. Let us fix some notations. For a given counter k, we denote $x_{k,0} = x_k$ and then denote by

$$x_{k,i_k} = \left(x_{k+1}^{(1)}, \cdots, x_{k+1}^{(i_k)}, x_k^{(i_k+1)}, \cdots, x_k^{(N)}\right),$$
$$\hat{x}_{k,i_k} = \left(x_{k+1}^{(1)}, \cdots, x_{k+1}^{(i_k-1)}, x_k^{(i_k+1)}, \cdots, x_k^{(N)}\right).$$

2.2 Regularized coordinate minimization algorithm

In this section, we present the *Regularized Cyclic multiblock Coordinate Minimization* (RCCM) algorithm for solving problem (1), with f possibly nonseparable and nonconvex, while g is separable and nonconvex. The basic idea of our algorithm consists of updating the ith component of $x \in \mathbb{R}^n$ in a cyclic manner using a local model of the whole nonseparable objective function along

Algorithm 1 RCCM

Given the regularization parameters $M_i \geq 0 \ \ \forall i = 1 : N$ and a starting point $x_0 \in \operatorname{dom} g$
for $k \geq 0$ **do**
  for $i_k = 1 : N$ **do**
    Compute:

$$x_{k+1}^{(i_k)} \in \underset{x^{(i_k)} \in \mathbb{R}^{n_{i_k}}}{\arg\min} f(x^{(i_k)}, \hat{x}_{k,i_k}) + g_{i_k}(x^{(i_k)}) + \frac{M_{i_k}}{2}\|x^{(i_k)} - x_k^{(i_k)}\|^2 \qquad (2)$$

  end for
  Update: $x_{k+1} = x_{k,N} = \left(x_{k+1}^{(1)}, \cdots, x_{k+1}^{(N)}\right)$.
end for

a subspace in contrast to a quadratic approximation along that subspace as considered in [6].

It should be observed that Assumption 1 (i) ensures that the set of minima of subproblem (2) is nonempty and compact whenever $M_i > 0$, see [6] and, consequently, the RCCM algorithm is well defined in this case. We also assume in the sequel that the subproblem (2) is easy to solve, e.g., it has a closed-form solution (see the quantization-based matrix factorization application from the next section). Note that in RCCM algorithm we can also choose variable regularization parameters (i.e., M_{i_k} may vary with k based on some adaptive strategy). One can also notice that RCCM is different from e.g., proximal alternating linearized minimization (i.e., proximal coordinate gradient) algorithm from [6] in both, assumptions and iterations. Indeed, paper [6] assumes that the smooth term f has coordinatewise Lipschitz gradient, while we just need to assume twice differentiability of f. Moreover, [6] linearizes the smooth term f along a subspace and uses an estimate of the coordinatewise Lipschitz constant on that subspace to produce the next iterate. Our method however uses a local model of the whole objective along a subspace instead of a quadratic approximation and does not need to know any Lipschitz constant estimates. Note that, RCCM is related to the proximal alternating minimization scheme from [2], where the objective is minimized successively with respect to two blocks by adding quadratic regularization terms and the smooth term is still assumed to have coordinatewise Lipschitz gradient. Here we consider a cyclic update over N blocks, Lipschitz continuity is not required and we extend the convergence analysis to other optimality criteria that are adequate for example to problems with bounded integer variables.

2.3 Convergence analysis of RCCM algorithm

In this section we analyze the convergence behavior of the RCCM algorithm. Let us denote by $M_{\min} = \min(M_1, \cdots, M_N)$ and $M_{\max} = \max(M_1, \cdots, M_N)$. Next lemma gives sufficient decrease of the objective function of (1); the decrease is strict whenever the iterate changes.

Lemma 1 *If Assumption 1 holds and the regularization parameters $M_i > 0 \;\; \forall i = 1 : N$, then the iterates of the RCCM algorithm satisfy the following strict descent:*

$$F(x_{k+1}) \leq F(x_k) - \frac{M_{\min}}{2}\|x_{k+1} - x_k\|^2.$$

We also have $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\|^2 < \infty$ and thus $\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0$.

Proof First, one can notice that:

$$\sum_{i=1}^{N} f(x_{k,i}) - f(x_{k,i-1}) = f(x_{k+1}) - f(x_k).$$

Second, using the optimality of $x_{k+1}^{(i_k)}$ we obtain:

$$f(x_{k,i_k}) + g_{i_k}(x_{k+1}^{(i_k)}) + \frac{M_{i_k}}{2}\|x_{k+1}^{(i_k)} - x_k^{(i_k)}\|^2 \leq f(x_{k,i_k-1}) + g_{i_k}(x_k^{(i_k)}) \;\; \forall i_k = 1 : N.$$

Adding the previous relation from $i_k = 1 : N$ and combining with the first relation, we get

$$F(x_{k+1}) \leq F(x_k) - \sum_{i=1}^{N} \frac{M_i}{2}\|x_{k+1}^{(i)} - x_k^{(i)}\|^2,$$

which yields the first statement. Further, for any positive integer K, summing the previous descent relation from $k = 0$ to $k = N - 1$ we also get

$$\frac{M_{\min}}{2} \sum_{k=0}^{N} \|x_{k+1} - x_k\|^2 \leq F(x_0) - F(x_N) \leq F(x_0) - F^*.$$

Since $M_{\min} > 0$, the second statement also follows. □

Below we provide the convergence rates for the RCCM algorithm in first-order stationarity and function values, which match the usual rates for coordinate descent methods from the literature, see [2, 6, 24]. Our proofs follow similar arguments as in [2]. For completeness, we provide the proofs in appendix.

Lemma 2 *Let Assumption 1 hold and assume that the iterates of the RCCM algorithm, with the regularization parameters $M_i > 0 \; \forall i = 1 : N$, are bounded. Then, there exists $C > 0$ such that the following convergence rate in the first-order stationarity holds:*

$$\min_{k=0:K} \min_{s_{k+1} \in \partial F(x_{k+1})} \|s_{k+1}\|^2 \leq \frac{2C^2\,(F(x_0) - F^*)}{M_{\min}(K+1)}. \tag{3}$$

Note that RCCM algorithm does not require the knowledge of the coordinatewise Lipschitz constants of the differentiable function f, as for example proximal alternating linearized minimization algorithm in [6]. In some applications the first term, f, although differentiable might have expensive gradient evaluation or the corresponding coordinatewise Lipschitz constants might be difficult to estimate; on the other hand, if the computation of the proximal step from subproblem (2) is easy, then the RCCM algorithm can be used efficiently (see the next section for such proximal friendly functions).

Furthermore, improved (local) convergence rates can be derived for our algorithm when F satisfies additionally the Kurdyka-Lojasiewicz (KL) property [5], i.e., when there exist $\gamma, \epsilon > 0$ such that F satisfies for all x with $\text{dist}(x, \mathcal{X}) \leq \gamma$ and $F_* < F(x) < F_* + \epsilon$ the following inequality

$$F(x) - F_* \leq \sigma_q \text{dist}(0, \partial F(x))^q, \tag{4}$$

where $\mathcal{X}$ is a given compact set on which F takes a constant value F_*, $\sigma_q > 0$ and $q > 1$. Here, F_* needs not equal the global optimal value F^*. The KL property (4) holds for a large class of functions including semi-algebraic functions (e.g., real polynomial functions or functions whose graph is given as a finite union of sets defined by a finite number of polynomial equalities and inequalities); piecewise linear/quadratic functions such as $\|x\|_1, \|x\|_0, \gamma \sum_{i=1}^k |x_{[i]}|$, where $|x_{[i]}|$ is the i-th largest entry in x, $k \leq n$ and $\gamma \in (0, 1]$; the indicator function $\delta_\Delta(x)$, where $\Delta = \{x \in \mathbb{R}^n : e^T x = 1, x \geq 0\}$; least-squares problems with the smoothly clipped absolute deviation, see [6, 24]. Next, we state the improved (local) convergence rates under the KL property.

Lemma 3 *Let Assumption 1 hold, and assume that the sequence of iterates $(x_k)_{k\geq 0}$ generated by algorithm RCCM is bounded. Additionally, we assume that F satisfies the KL property* (4) *and that there exists $k_0 \geq 0$ such that, for every $k \geq k_0$, the iterates x_k belong to this KL neighborhood. Then, for every $k \geq k_0$, we have:*

$$F(x_{k+1}) - F_* \leq (F(x_k) - F_*) - \frac{M_{\min}}{2C^2} \left(\frac{F(x_{k+1}) - F_*}{\sigma_q} \right)^{2/q}. \tag{5}$$

It is worth noting that the convergence rate depends on the KL parameter $q > 1$. More precisely, let us denote $\Delta_k := F(x_k) - F_*$. Then, for all sufficiently large k, equation (5) becomes:

$$\Delta_{k+1} + \mu \Delta_{k+1}^{2/q} \leq \Delta_k, \qquad \mu := \frac{M_{\min}}{2C^2 \sigma_q^{2/q}} > 0. \tag{6}$$

When $q \in (1, 2)$, we have $\frac{2}{q} > 1$ and the recursion (6) yields a sublinear convergence rate in the objective residual of the form $\Delta_k = \mathcal{O}\left(k^{-\frac{q}{2-q}}\right)$ $\forall k \geq k_0$. Note that the sublinear rate in objective residual improves as the KL parameter q approaches 2 from below. When $q = 2$, the recursion (6) becomes $(1 + \mu)\Delta_{k+1} \leq \Delta_k$, and therefore the objective values converge linearly:

$\Delta_k \leq \left(\frac{1}{1+\mu}\right)^{k-k_0} \Delta_{k_0}\ \forall k \geq k_0$. When $q > 2$, we have $\frac{2}{q} < 1$. In this case, the recursion (6) implies a locally superlinear rate in the objective residual. Indeed, for sufficiently small Δ_k, we obtain $\Delta_{k+1} \leq \left(\frac{\Delta_k}{\mu}\right)^{q/2}$, where $q/2 > 1$. Consequently, once the iterates enter the KL neighborhood, the objective residual decreases faster than linearly, i.e., superlinearly.

However, in some nonconvex problems, the first-order stationarity (Lemma 2) or objective residual (Lemma 3) criteria do not adequately characterize the quality of the obtained solution. This limitation is particularly relevant for coordinate minimization methods applied to problems with discrete feasible sets such as QMF, see the next remark.

Remark 1 (Limitation of first-order stationarity on discrete domains) Suppose $g = 1_{\mathcal{D}}$ in (1), where $\mathcal{D} \subset \mathbb{Z}^n$ is finite and $1_{\mathcal{D}}$ is its indicator function. At every $x \in \mathcal{D}$, the limiting normal cone satisfies $\mathcal{N}_{\mathcal{D}}(x) = \mathbb{R}^n$. Hence $\partial F(x) = \partial(f + 1_{\mathcal{D}})(x) = \mathbb{R}^n$ and $\operatorname{dist}(0, \partial F(x)) = 0$ at *every* feasible integer point, regardless of its objective value. Thus, for this particular optimization problem, the preceding $\mathcal{O}(1/K)$ stationarity estimate although formally valid, is vacuous as a quality guarantee for a (local) optimum. Similarly, the KL property at isolated feasible points may not certify a useful integer solution as long as F_* in the local KL analysis is the objective value on a general stationary set, which may not be related to local/global optimal points; i.e., F_* needs not equal the global optimal value F^*. Hence, our previous convergence analysis but also the convergence results from [1,2,6] provide no meaningful guarantees for problems with bounded integer feasible sets. □

To overcome this limitation, we introduce a block-coordinate improvement gap that measures the maximum achievable decrease in the original objective function by optimizing over an individual block, following the general perspective of discrete block-coordinate optimization in [19].

Definition 1 [Block-coordinate improvement gap] The block-coordinate improvement gap is defined as:

$$\mathcal{G}(x) = \max_{i=1:N} \mathcal{G}_i(x),$$

where

$$\mathcal{G}_i(x) = F(x) - \min_{z^{(i)} \in \mathbb{R}^{n_i}} F(z^{(i)}, \hat{x}_i) \ \text{ and } \ \hat{x}_i = (x^{(1)}, \ldots, x^{(i-1)}, x^{(i+1)}, \ldots, x^{(N)}).$$

Note that from Assumption 1 it follows that $\mathcal{G}_i(x)$'s are well-defined. In the next sections we revisit the convergence analysis of the RCCM algorithm based on this optimality measure.

2.4 Convergence rate of RCCM in block-coordinate improvement gap under adaptive regularizers

In this section we propose an adaptive strategy for choosing the regularization parameters M_i's in RCCM algorithm. At iteration k, for the block i_k, let the unregularized update be

$$y_{k+1}^{(i_k)} \in \arg\min_{x^{(i_k)} \in \mathbb{R}^{n_{i_k}}} \left\{ f(x^{(i_k)}, \hat{x}_{k,i_k}) + g_{i_k}(x^{(i_k)}) \right\}.$$

Then, the corresponding i_k block-coordinate improvement gap is

$$\mathcal{G}_{i_k}(x_{k,i_k-1}) = F(x_{k,i_k-1}) - F(y_{k+1}^{(i_k)}, \hat{x}_{k,i_k}).$$

Let $\eta \in (0,1)$ be a given parameter. If $\mathcal{G}_{i_k}(x_{k,i_k-1}) > 0$, we choose the regularization parameter M_{k,i_k} satisfying

$$0 < M_{k,i_k} \leq \frac{2(1-\eta)\mathcal{G}_{i_k}(x_{k,i_k-1})}{\|y_{k+1}^{(i_k)} - x_k^{(i_k)}\|^2}. \tag{7}$$

Otherwise, if $\mathcal{G}_{i_k}(x_{k,i_k-1}) = 0$, we choose any $M_{k,i_k} > 0$. The corresponding block update in the adaptive RCCM algorithm just replaces M_{i_k} with M_{k,i_k} in the subproblem (2). Note that the adaptive choice (7) does not require evaluating the entire objective twice since only one block changes, all terms independent of that block cancel. Therefore, it can be implemented efficiently in applications. For example, for the QMF problem from the next section, its gap can be computed cheaply from quantities already used by the column update. The next theorem establishes a sufficient decrease property and a convergence rate in terms of the block-coordinate improvement gap.

Theorem 1 *Let Assumption 1 hold and consider the iterates generated by the RCCM algorithm with adaptive regularization parameters $M_{k,i_k} > 0$ satisfying (7) whenever the corresponding block-coordinate improvement gap is positive. Then, for any $k \geq 0$ and $i_k \in [1:N]$, we have the descent*

$$F(x_{k,i_k}) \leq F(x_{k,i_k-1}) - \eta\, \mathcal{G}_{i_k}(x_{k,i_k-1}) - \frac{M_{k,i_k}}{2}\|x_{k+1}^{(i_k)} - x_k^{(i_k)}\|^2. \tag{8}$$

Moreover, for any positive integer K, the following convergence rate holds:

$$\min_{k=0:K} \sum_{i=1}^{N} \mathcal{G}_i(x_{k,i-1}) \leq \frac{F(x_0) - F^*}{\eta(K+1)}. \tag{9}$$

Proof First, let us consider the case $\mathcal{G}_{i_k}(x_{k,i_k-1}) > 0$. By the optimality of $x_{k+1}^{(i_k)}$ in the regularized subproblem, we have

$$\begin{aligned} &f(x_{k+1}^{(i_k)}, \hat{x}_{k,i_k}) + g_{i_k}(x_{k+1}^{(i_k)}) + \frac{M_{k,i_k}}{2}\|x_{k+1}^{(i_k)} - x_k^{(i_k)}\|^2 \\ &\leq f(y_{k+1}^{(i_k)}, \hat{x}_{k,i_k}) + g_{i_k}(y_{k+1}^{(i_k)}) + \frac{M_{k,i_k}}{2}\|y_{k+1}^{(i_k)} - x_k^{(i_k)}\|^2. \end{aligned}$$

Using the separability of g, we can equivalently write

$$F(x_{k,i_k}) + \frac{M_{k,i_k}}{2}\|x_{k+1}^{(i_k)} - x_k^{(i_k)}\|^2 \leq F(y_{k+1}^{(i_k)}, \hat{x}_{k,i_k}) + \frac{M_{k,i_k}}{2}\|y_{k+1}^{(i_k)} - x_k^{(i_k)}\|^2.$$

Since $y_{k+1}^{(i_k)}$ is a minimizer of the unregularized block subproblem, from the definition of the block-coordinate improvement gap we obtain

$$\mathcal{G}_{i_k}(x_{k,i_k-1}) = F(x_{k,i_k-1}) - F(y_{k+1}^{(i_k)}, \hat{x}_{k,i_k}).$$

Combining the previous relations gives

$$\begin{aligned} &F(x_{k,i_k}) + \frac{M_{k,i_k}}{2}\|x_{k+1}^{(i_k)} - x_k^{(i_k)}\|^2 \\ &\leq F(x_{k,i_k-1}) - \mathcal{G}_{i_k}(x_{k,i_k-1}) + \frac{M_{k,i_k}}{2}\|y_{k+1}^{(i_k)} - x_k^{(i_k)}\|^2. \end{aligned}$$

Further, since $\mathcal{G}_{i_k}(x_{k,i_k-1}) > 0$, we have $y_{k+1}^{(i_k)} \neq x_k^{(i_k)}$. Using the adaptive choice of the regularization parameter (7) in the previous inequality, we obtain

$$\frac{M_{k,i_k}}{2}\|y_{k+1}^{(i_k)} - x_k^{(i_k)}\|^2 \leq (1-\eta)\mathcal{G}_{i_k}(x_{k,i_k-1}).$$

Substituting this inequality into the previous relation yields

$$F(x_{k,i_k}) + \frac{M_{k,i_k}}{2}\|x_{k+1}^{(i_k)} - x_k^{(i_k)}\|^2 \quad \leq F(x_{k,i_k-1}) - \eta\mathcal{G}_{i_k}(x_{k,i_k-1}),$$

which is (8). Next, if $\mathcal{G}_{i_k}(x_{k,i_k-1}) = 0$, then the current block $x_k^{(i_k)}$ is already a minimizer of the corresponding unregularized block subproblem, i.e., $y_{k+1}^{(i_k)} = x_k^{(i_k)}$. Since $M_{k,i_k} > 0$ and the quadratic regularization term is zero only at the current block it also follows $x_{k+1}^{(i_k)} = x_k^{(i_k)}$. Hence, the previous sufficient decrease inequality also holds in this case, with equality. This proves the first statement (8). To prove the second statement, we sum the sufficient decrease inequality over all blocks $i_k = 1 : N$ and obtain

$$\sum_{i=1}^{N}(F(x_{k,i-1}) - F(x_{k,i})) \geq \eta\sum_{i=1}^{N}\mathcal{G}_i(x_{k,i-1}) + \frac{1}{2}\sum_{i=1}^{N}M_{k,i}\|x_{k+1}^{(i)} - x_k^{(i)}\|^2.$$

Recalling that $x_{k,0} = x_k$ and $x_{k,N} = x_{k+1}$, the sum on the left-hand side is telescopic, and thus

$$\begin{aligned} F(x_k) - F(x_{k+1}) &\geq \eta\sum_{i=1}^{N}\mathcal{G}_i(x_{k,i-1}) + \frac{1}{2}\sum_{i=1}^{N}M_{k,i}\|x_{k+1}^{(i)} - x_k^{(i)}\|^2 \\ &\geq \eta\sum_{i=1}^{N}\mathcal{G}_i(x_{k,i-1}). \end{aligned}$$

Taking again the sum from $k = 0$ to K and using the telescoping property:

$$\eta \sum_{k=0}^{K} \sum_{i=1}^{N} \mathcal{G}_i(x_{k,i-1}) \leq F(x_0) - F(x_{K+1}) \leq F(x_0) - F^*,$$

where the last inequality follows from Assumption 1(ii). Finally, using the nonnegativity of the block-coordinate improvement gaps, we have

$$(K+1) \min_{k=0:K} \sum_{i=1}^{N} \mathcal{G}_i(x_{k,i-1}) \leq \sum_{k=0}^{K} \sum_{i=1}^{N} \mathcal{G}_i(x_{k,i-1}) \leq \frac{F(x_0) - F^*}{\eta},$$

which completes the proof. □

Remark 2 It is worth noting that the rate (9) concerns gaps at different intermediate points, it does not immediately establish the same rate for $\mathcal{G}(x_k)$, where all block gaps are evaluated at one common point. However, the previous theorem guarantees the asymptotic convergence of the block-coordinate improvement gaps evaluated at the intermediate iterates:

$$\lim_{k \to \infty} \mathcal{G}_i(x_{k,i-1}) = 0, \qquad \forall i = 1 : N. \tag{10}$$

Hence, the maximum achievable decrease of the objective function along each visited block vanishes asymptotically.

2.5 Finite termination of RCCM on bounded integer feasible sets

Next, we consider a particular case of problem (1), i.e., minimize a twice differentiable function f over a bounded integer feasible set $\mathcal{D} \subset \mathbb{Z}^n$. For this case a finite termination result follows from the strict decrease given in Lemma 1. Let us assume the feasible set $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_N$, where every $\mathcal{D}_i$ is a nonempty finite subset of $\mathbb{Z}^{n_i}$. Define the block diameters $R_i = \operatorname{diam}(\mathcal{D}_i) = \max_{u,v \in \mathcal{D}_i} \|u - v\|$. The next result provides finite termination and objective-based block optimality for RCCM minimizing a twice differentiable function over a bounded discrete domain. Note that the quadratic regularization and the resulting bound are features of our RCCM analysis when compared to e.g., [19].

Theorem 2 *Let $F = f + 1_{\mathcal{D}}$, with $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_N \subset \mathbb{Z}^n$ finite, Assumption 1 hold and fixed regularization parameters $M_i > 0$. Then, the number T of full iterates with at least one integer coordinate change in RCCM satisfies:*

$$T \leq \left\lfloor \frac{2(F(x_0) - F^*)}{M_{\min}} \right\rfloor. \tag{11}$$

The algorithm RCCM therefore reaches an unchanged full cycle in finite time. Moreover, its terminal point x_T is a fixed point and satisfies for every block i and $x^{(i)} \in \mathcal{D}_i$:

$$F(x_T) \leq F\Big(x^{(i)}, \hat{x}_{T,i}\Big) + \frac{M_i}{2} \|x^{(i)} - x_T^{(i)}\|^2 \ \text{ and } \ \mathcal{G}(x_T) \leq \max_{i=1:N} \frac{M_i R_i^2}{2}. \tag{12}$$

Proof Lemma 1 gives

$$\frac{M_{\min}}{2}\sum_{k=0}^{T-1}\|x_{k+1}-x_k\|^2 \le F(x_0)-F(x_T)\le F(x_0)-F^*.$$

Every changed full cycle satisfies $\|x_{k+1}-x_k\|^2 \ge 1$ because all coordinates are integer-valued. Counting these full iterates gives (11). Moreover, since a full unchanged iterate follows after the last changing iterate, it follows that x_T is a fixed point, i.e. $x_k = x_T$ for all $k > T$. Additionally, since each intermediate subproblem leaves the corresponding block unchanged, optimality of every block subproblem gives (12). Indeed, from subproblem (2) we have:

$$\begin{aligned} F(x_T) &= \min_{x^{(i)}\in\mathcal{D}_i} F\left(x^{(i)},\hat{x}_{T,i}\right)+\frac{M_i}{2}\|x^{(i)}-x_T^{(i)}\|^2 \\ &\le F\left(x^{(i)},\hat{x}_{T,i}\right)+\frac{M_i}{2}\|x^{(i)}-x_T^{(i)}\|^2 \quad \forall x^{(i)}\in\mathcal{D}_i, \end{aligned}$$

which proves the first part of (12). Further, applying the previous inequality at $y_T^{(i)} \in \arg\min_{x^{(i)}\in D_i} F(x^{(i)},\hat{x}_{T,i})$ gives

$$\mathcal{G}_i(x_T) = F(x_T)-F(y_T^{(i)},\hat{x}_{T,i}) \le \frac{M_i}{2}\|y_T^{(i)}-x_T^{(i)}\|^2 \le \frac{M_i R_i^2}{2},$$

proving the second part of (12) □

Note that (12) is a *regularized* best-response condition and does not in general imply the blockwise optimality for F obtained usually under an unregularized block minimization approach as considered in [19]. In what follows, we provide a stronger result for the adaptive strategy (7) of RCCM.

Theorem 3 *Let $F = f + 1_{\mathcal{D}}$, with $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_N \subset \mathbb{Z}^n$ finite, and let Assumption 1 hold. Consider the iterates generated by the RCCM algorithm with adaptive regularization parameters $M_{k,i_k} > 0$ chosen according to* (7) *whenever the corresponding block-coordinate improvement gap is positive. Then, the algorithm reaches an unchanged full cycle after finitely many block updates. Moreover, its terminal point x_T satisfies*

$$\mathcal{G}(x_T) = 0. \tag{13}$$

Equivalently, x_T is a blockwise minimizer of the original objective function:

$$F(x_T) \le F(x^{(i)},\hat{x}_{T,i}), \quad \forall x^{(i)}\in\mathcal{D}_i, \quad i = 1:N. \tag{14}$$

Proof From the descent inequality of Theorem 1, we have

$$F(x_{k,i_k-1}) - F(x_{k,i_k}) \ge \eta\mathcal{G}_{i_k}(x_{k,i_k-1}) + \frac{M_{k,i_k}}{2}\|x_{k+1}^{(i_k)}-x_k^{(i_k)}\|^2.$$

Since $M_{k,i_k} > 0$, whenever the updated block changes, i.e., $x_{k+1}^{(i_k)} \ne x_k^{(i_k)}$, the objective function decreases strictly: $F(x_{k,i_k}) < F(x_{k,i_k-1})$. Moreover, since

$\mathcal{D}$ is finite, the objective function takes only finitely many values on $\mathcal{D}$. Consequently, the algorithm can perform only finitely many block changes and must eventually reach an unchanged full cycle. Let x_T denote the corresponding terminal point. Since the full cycle leaves all blocks unchanged, we have $x_{T,i} = x_T$ for all $i = 1 : N$. Applying the descent inequality to each block of this unchanged cycle yields

$$0 \geq \eta \mathcal{G}_i(x_T), \qquad \forall i = 1 : N.$$

Since $\eta > 0$ and the block-coordinate improvement gaps are nonnegative, it follows

$$\mathcal{G}(x_T) = \max_{i=1:N} \mathcal{G}_i(x_T) = 0.$$

Equivalently, for every block $i = 1 : N$,

$$F(x_T) = \min_{x^{(i)} \in \mathcal{D}_i} F(x^{(i)}, \hat{x}_{T,i}),$$

which proves the blockwise optimality condition. □

Next, we apply our RCCM algorithm on a specific matrix factorization application.

3 Quantized matrix factorization using RCCM algorithm

Quantized matrix factorization (QMF) is a technique that combines matrix factorization with quantization (i.e., elements constrained to bounded integer values) to create low-rank, compressed representations of data, often used for tasks like image compression or efficient training of machine learning models. Traditional compression methods usually represent the compressed data in continuous domains and, therefore they implement quantization steps separately at the end of the compression process. Hence, quantization errors cannot be incorporated into the iterative compression process. Recent work addresses this issue, by introducing quantization constraints into the matrix factorization problem, see e.g. [1, 12].

Let us define the QMF problem. Given a data matrix $Y \in \mathbb{R}^{m \times n}$ (e.g., an image), one approach to tackle the data compression task involves using a low-rank approximation:

$$Y \approx UV^T = \sum_{j=1}^{d} u^{(j)} \left(v^{(j)}\right)^T$$

where $U \in \mathbb{R}^{m \times d}$, $V \in \mathbb{R}^{n \times d}$, with $d \leq \min(n, m)$ is the rank of the approximation, and $u^{(j)}$ and $v^{(j)}$ denote the jth columns of U and V, respectively. The optimization framework assumes minimizing the approximation error quantified via Frobenius norm:

$$\min_{U,V} \frac{1}{2} \|Y - UV^T\|_F^2.$$

This is known as the matrix factorization problem and a globally optimal rank-d solution can be obtained from the truncated singular value decomposition (SVD) [9]. To enhance transmission, computations and storage efficiency in data compression, a quantization procedure is integrated into the matrix factorization problem:

$$\min_{\substack{U\in\mathbb{Z}^{m\times d}_{[\alpha,\beta]},V\in\mathbb{Z}^{n\times d}_{[\alpha,\beta]}\\ \alpha,\beta\in\mathbb{Z}}} \frac{1}{2}\|Y-UV^T\|_F^2+\lambda\log_2(\beta-\alpha+1), \tag{15}$$

where the first term is the distortion, the second is the rate and $\lambda\geq 0$ is a Lagrange multiplier that adjusts the weight between output quality and its size. If $\lambda=0$, then $\alpha<\beta$ are fixed integer numbers; otherwise they are chosen adaptively. Formulation (15) with $\lambda=0$ has been also considered in [1,12]. For simplicity of the exposition we also consider $\lambda=0$ in the sequel, i.e., problem (15) with $\alpha<\beta$ fixed and refer to it as the quantized matrix factorization (QMF) problem:

$$\begin{aligned}&\min_{U\in\mathbb{Z}^{m\times d}_{[\alpha,\beta]},V\in\mathbb{Z}^{n\times d}_{[\alpha,\beta]}} \frac{1}{2}\|Y-UV^T\|_F^2\\ &=\min_{U\in\mathbb{R}^{m\times d},V\in\mathbb{R}^{n\times d}} \frac{1}{2}\|Y-UV^T\|_F^2+1_{\mathbb{Z}^{m\times d}_{[\alpha,\beta]}}(U)+1_{\mathbb{Z}^{n\times d}_{[\alpha,\beta]}}(V).\end{aligned} \tag{16}$$

Constraining U and V to have integer entries in a specified interval $[\alpha,\beta]$, the quantization error is incorporated in the compression process, not at the end of the process. Clearly, the QMF problem (16) is a particular case of the nonconvex composite optimization problem (1) with $x=(U,V)$, twice differentiable nonconvex function

$$f(x)=\frac{1}{2}\|Y-UV^T\|_F^2 \tag{17}$$

and the separable nonconvex function

$$g(x)=1_{\mathbb{Z}^{m\times d}_{[\alpha,\beta]}}(U)+1_{\mathbb{Z}^{n\times d}_{[\alpha,\beta]}}(V). \tag{18}$$

Since the indicator function g has dom$g$ in a compact set, it follows that all the conditions from Assumption 1 hold for QMF problem (16). Notably, for QMF, boundedness of the iterates follows automatically from the integer box constraints. In the QMF problem we fix the block variables $x^{(i)}$ either the columns $u^{(j)}$ or $v^{(j)}$ of the matrices U and V, respectively. Hence, the previous RCCM algorithm can be used to solve the QMF problem (16). After the initialization phase, RCCM algorithm requires that, at iteration k, we fix V_k and we update the columns of U, then proceed analogously by fixing U_{k+1} and updating the columns of V. Moreover, we perform the matrix factor updates columnwise, e.g., examining one column $i_k=1:d$ of U at a time:

$$u^{(i_k)}_{k+1}=\arg\min_{u^{(i_k)}\in\mathbb{R}^m}\frac{1}{2}\|A^{(i_k)}-u^{(i_k)}(v^{(i_k)}_k)^T\|_F^2+1_{\mathbb{Z}^m_{[\alpha,\beta]}}(u^{(i_k)})+\frac{M_u}{2}\|u^{(i_k)}-u^{(i_k)}_k\|^2, \tag{19}$$

where the regularization parameter $M_u \geq 0$ and we denote $A^{(i_k)} = Y - \sum_{j\neq i_k, j=1}^{d} u_{k,i_k-1}^{(j)}(v_k^{(j)})^T$ and $u_{k,i_k-1}^{(j)}$ is the jth column of matrix U_{k,i_k-1}, which comprises a mixture of both the current and the previously updated columns, i.e $U_{k,i_k-1} := [u_{k+1}^{(1)}, \ldots, u_{k+1}^{(i_k-1)}, u_k^{(i_k)}, \ldots, u_k^{(d)}]$, with the convention $U_{k,0} = U_k := [u_k^{(1)}, \ldots, u_k^{(d)}]$. The columns in V are updated solving a similar problem as in (19) with the regularization parameter $M_v \geq 0$. Hence, in RCCM algorithm we need to solve one of the following generic subproblems:

$$\min_{u^{(i)}\in\mathbb{Z}_{[\alpha,\beta]}^m} \frac{1}{2}\|A^{(i)} - u^{(i)}(\bar{v}^{(i)})^T\|_F^2 + \frac{M_u}{2}\|u^{(i)} - \bar{u}^{(i)}\|^2 \tag{20}$$

$$\min_{v^{(i)}\in\mathbb{Z}_{[\alpha,\beta]}^n} \frac{1}{2}\|B^{(i)} - \bar{u}^{(i)}(v^{(i)})^T\|_F^2 + \frac{M_v}{2}\|v^{(i)} - \bar{v}^{(i)}\|^2, \tag{21}$$

where the vectors $\bar{u}^{(i)}, \bar{v}^{(i)}$ and the matrices $A^{(i)}, B^{(i)}$ are fixed. Further, we show how the generic subproblems (20) and (21) can be decomposed into one dimensional optimization problems that have closed form solutions and consequently step (2) in RCCM algorithm can be applied efficiently for solving the QMF problem. Indeed, the objective function in (20) can be rewritten as:

$$\begin{aligned}
&\frac{1}{2}\|A^{(i)} - u^{(i)}(\bar{v}^{(i)})^T\|_F^2 + \frac{M_u}{2}\|u^{(i)} - \bar{u}^{(i)}\|^2 \\
&= \frac{1}{2}\sum_{p=1}^{m}\sum_{q=1}^{n}\left(A_{p,q}^{(i)} - u_p^{(i)}\bar{v}_q^{(i)}\right)^2 + \frac{M_u}{2}\sum_{p=1}^{m}(u_p^{(i)} - \bar{u}_p^{(i)})^2,
\end{aligned}$$

where $A_{p,q}^{(i)}$ denotes the (p,q) element of matrix $A^{(i)}$ and $u_p^{(i)}$ and $\bar{u}_p^{(i)}$ are the pth entries of the vectors $u^{(i)}$ and $\bar{u}_p^{(i)}$, respectively. Similarly, $\bar{v}_q^{(i)}$ is the qth element of vector $\bar{v}^{(i)}$. Since all variables are fixed with the exception of $u_p^{(i)}$, the subproblem (20) can be divided into m one-dimensional quadratic subproblems of the form:

$$\min_{u_p^{(i)}\in\mathbb{Z}_{[\alpha,\beta]}} \frac{1}{2}\sum_{q=1}^{n}\left(A_{p,q}^{(i)} - u_p^{(i)}\bar{v}_q^{(i)}\right)^2 + \frac{M_u}{2}(u_p^{(i)} - \bar{u}_p^{(i)})^2, \tag{22}$$

for $p = 1 : m$, whose solution can be computed explicitly as:

$$\begin{aligned}
(u_p^{(i)})^+ &= \left[\frac{M_u\bar{u}_p^{(i)} + \sum_{q=1}^{n} A_{p,q}^{(i)}\bar{v}_q^{(i)}}{M_u + \sum_{q=1}^{n}(\bar{v}_q^{(i)})^2}\right]_{Z_{[\alpha,\beta]}} \\
&= \text{clamp}\left(\alpha, \beta, \text{round}\left(\frac{M_u\bar{u}_p^{(i)} + \sum_{q=1}^{n} A_{p,q}^{(i)}\bar{v}_q^{(i)}}{M_u + \sum_{q=1}^{n}(\bar{v}_q^{(i)})^2}\right)\right).
\end{aligned} \tag{23}$$

Hence, using the definition of the matrix $A^{(i)}$, the update of the entire column $u^{(i)}$ is as follows:

$$(u^{(i)})^+ = \left[\frac{M_u\bar{u}^{(i)} + A^{(i)}\bar{v}^{(i)}}{M_u + \|\bar{v}^{(i)}\|^2}\right]_{Z_{[\alpha,\beta]}^m} = \left[\frac{M_u\bar{u}^{(i)} + Y\bar{v}^{(i)} - \sum_{s\neq i}(\bar{V}^T\bar{V})_{s,i}\bar{u}^{(s)}}{M_u + \|\bar{v}^{(i)}\|^2}\right]_{Z_{[\alpha,\beta]}^m}.$$

The same arguments apply equally to subproblem (21). In conclusion, the application of Algorithm 1 (RCCM) on QMF problem (16) leads to Algorithm 2. The algorithm in [6] has been used in [1] to solve the QMF problem (16) and called QMF algorithm. Hence, we call Algorithm 1 (RCCM) specialized for the QMF problem (16) as *regularized quantized matrix factorization* (RQMF) algorithm. Since f in (17) is quadratic in U and V, respectively, the main difference between our Algorithm 2 (RQMF) and the proximal alternating linearized minimization algorithm in [6] (which is a projected coordinatewise gradient) when applied to the particular QMF problem (16) consists of the regularization terms for U and V. Hence, for $M_u = M_v = 0$ our algorithm recovers the one in [6] for solving the QMF problem. However, the regularization parameters (M_u, M_v) provide more flexibility to our Algorithm 2, guaranteeing *strict* objective decrease when the iterate changes (see Lemma 1) while the algorithm in [6] ensures only nonincreasing monotonicity (see Theorem 1 in [1]). Additionally, the regularization parameters lead to improved numerical results (see next section).

Algorithm 2 (RQMF)

Input: $Y \in \mathbb{R}^{m\times n}$, M_u, $M_v > 0$, rank d, bounds $[\alpha, \beta]$ and maximum number of full iterations K.
Set $k = 0$. Compute $\bar{U}_0 \in \mathbb{R}^{m\times d}, \bar{V}_0 \in \mathbb{R}^{n\times d}$ using truncated SVD, then project

$$U_0 = [\bar{U}_0]_{\mathbb{Z}^{m\times d}_{[\alpha,\beta]}}, \text{ and } V_0 = [\bar{V}_0]_{\mathbb{Z}^{n\times d}_{[\alpha,\beta]}}$$

while $k < K$ **do**
Compute: $A = [a^{(1)} \cdots a^{(d)}] = YV_k, \quad C = V_k^T V_k$
for $i_k = 1 : d$ **do**

$$1.\ \hat{a}^{(i_k)} = a^{(i_k)} - \left(\sum_{s=1}^{i_k-1} C_{s,i_k} u_{k+1}^{(s)} + \sum_{s=i_k+1}^{d} C_{s,i_k} u_k^{(s)} \right)$$

$$2.\ u_{k+1}^{(i_k)} = \left[\frac{M_u u_k^{(i_k)} + \hat{a}^{(i_k)}}{M_u + \|v_k^{(i_k)}\|^2} \right]_{\mathbb{Z}^m_{[\alpha,\beta]}}$$

end for
Compute: $B = [b^{(1)} \cdots b^{(d)}] = Y^T U_{k+1},\ C = U_{k+1}^T U_{k+1}$
for $i_k = 1 : d$ **do**

$$1.\ \hat{b}^{(i_k)} = b^{(i_k)} - \left(\sum_{s=1}^{i_k-1} C_{s,i_k} v_{k+1}^{(s)} + \sum_{s=i_k+1}^{d} C_{s,i_k} v_k^{(s)} \right)$$

$$2.\ v_{k+1}^{(i_k)} = \left[\frac{M_v v_k^{(i_k)} + \hat{b}^{(i_k)}}{M_v + \|u_{k+1}^{(i_k)}\|^2} \right]_{\mathbb{Z}^n_{[\alpha,\beta]}}$$

end for
if $U_{k+1} = U_k$ and $V_{k+1} = V_k$ **then break**
$k \leftarrow k + 1$
end while

From (23) it is clear that Algorithm 2 has simple iterations (involving only matrix-vector operations). Note that the positive regularization makes our formulas of $u_{k+1}^{(i_k)}$ and $v_{k+1}^{(i_k)}$ well defined even when a factor column is zero.

Remark 3 (i) One can notice that the fourth-order polynomial function f defined in (17) is twice differentiable and dom$g$ of the separable function g defined in (18) is compact. Therefore, all the conditions from Assumption 1 hold for QMF problem (16). Moreover, the iterates generated by (RQMF) $x_k = (U_k, V_k)$ are bounded as they lie in the interval $[\alpha, \beta]$. In conclusion, the sublinear convergence result from Lemma 2 is valid for this particular Algorithm 2 (RQMF) on solving the QMF problem (16).

(ii) Additionally, function f defined in (17) is a fourth-order polynomial, hence semi-algebraic; the interval $[\alpha, \beta]$ intersected with the set of integers is semi-algebraic, hence the indicator functions that define g in (18) are also semi-algebraic. Therefore, the objective function of the QMF problem (16), $F = f + g$, is semi-algebraic, and thus F has the KL property. In conclusion, the convergence results from Lemma 3 are also valid for this particular Algorithm 2 (RQMF) on solving the QMF problem (16).

(iii) However, the QMF objective (16) has discrete domain, so every feasible factorization is a subdifferential critical point and the KL analysis may not be related to local/global optimal points as F_* needs not equal the global optimal value F^* (see Remark 1). Hence, neither a generic first-order stationarity result as given in [1][Proposition 2], [2][Lemma 5] and Lemma 2 nor a KL-based local-rate statement as given in [2][Theorem 11] and Lemma 3 may measure reconstruction quality at an isolated integer point. In conclusion, the result on convergence to stationary points of the proximal alternating linearized minimization algorithm [1, 6] for solving the QMF problem in [1] is vacuous. We therefore provide below a separate block-coordinate improvement gap optimality analysis for QMF problem (16), generalizing and/or improving the results from Sections 2.4 and 2.5.

3.1 A computable discrete column improvement gap and finite termination of RQMF

Denote the reconstruction error $E = Y - UV^T$ and $f(U, V) = \frac{1}{2}\|E\|_F^2$. For $i = 1 : d$ and a row p, set $a_i = \|v^{(i)}\|^2, \quad s_{pi} = (Ev^{(i)})_p$. Changing only the entry (p, i) in U (denoted $u_p^{(i)}$) by a feasible integer displacement h gives the exact expression:

$$f(U + he_p e_i^T, V) - f(U, V) = \tfrac{a_i}{2}h^2 - s_{pi}h. \tag{24}$$

Let $\mathcal{H}_{pi}^U = \{h \in \mathbb{Z} : \alpha \le u_p^{(i)} + h \le \beta\}$ and define the maximum achievable scalar improvement

$$D_{pi}^U = \max_{h \in \mathcal{H}_{pi}^U} \left(s_{pi}h - \tfrac{a_i}{2}h^2\right) \ge 0. \tag{25}$$

Likewise, with $b_i = \|u^{(i)}\|^2$, $t_{qi} = (E^T u^{(i)})_q$, and $\mathcal{H}_{qi}^V = \{h \in \mathbb{Z} : \alpha \le v_q^{(i)} + h \le \beta\}$, define

$$D_{qi}^V = \max_{h \in \mathcal{H}_{qi}^V} \left(t_{qi}h - \tfrac{b_i}{2}h^2\right) \ge 0. \tag{26}$$

Now, the block-coordinate improvement gap from Definition 1 yields for the QMF problem the *column improvement gap*:

$$\mathcal{G}(U,V) = \max_{i=1,\dots,d} \Big\{ f(U,V) - \min_{z \in \mathbb{Z}^m_{[\alpha,\beta]}} f(z, \hat{U}_i; V), \\ f(U,V) - \min_{w \in \mathbb{Z}^n_{[\alpha,\beta]}} f(U; w, \hat{V}_i) \Big\}, \tag{27}$$

where $f(z, \hat{U}_i; V)$ denotes evaluation of f with column $u^{(i)}$ replaced by z (the V expression is analogous). Unlike subdifferential stationarity, $\mathcal{G}$ measures a decrease of the original reconstruction objective function in QMF problem.

Lemma 4 *For every feasible QMF factorization, we have*

$$\mathcal{G}(U,V) = \max \left\{ \max_i \sum_{p=1}^{m} D^U_{pi}, \ \max_i \sum_{q=1}^{n} D^V_{qi} \right\}. \tag{28}$$

In particular, $\mathcal{G}(U,V) = 0$ if and only if (U,V) is the best response with respect to every individual factor column for the objective function f.

Proof For fixed V and the other columns of U, the objective separates over the m entries of $u^{(i)}$, see (22). Equations (24) and (25) show that the best possible decrease from replacing the whole column is $\sum_p D^U_{pi}$. The analogous separation for $v^{(i)}$ gives $\sum_q D^V_{qi}$. Taking the largest decrease over the $2d$ columns of U and V establishes (28). Moreover, each term in (28) is nonnegative because displacement $h = 0$ is feasible, so vanishing of the maximum (i.e., $\mathcal{G}(U,V) = 0$) is equivalent to the minimizers of $\min_{z \in \mathbb{Z}^m_{[\alpha,\beta]}} f(z, \hat{U}_i; V)$ and $\min_{w \in \mathbb{Z}^n_{[\alpha,\beta]}} f(U; w, \hat{V}_i)$ being zero for every column of U and V. This proves the second part of the lemma. □

Formula (28) requires only products with the residual E and scalar bounded-integer quadratic minimizations. If $a_i > 0$, the maximizing displacement in (25) is obtained by projecting $u^{(i)}_p + s_{pi}/a_i$ onto $\mathbb{Z}_{[\alpha,\beta]}$ and subtracting $u^{(i)}_p$; if $a_i = 0$, then $s_{pi} = 0$ and $D^U_{pi} = 0$. The analogous rule applies to D^V_{qi}. The certificate $\mathcal{G}(U,V)$ can therefore be evaluated easily.

Theorem 4 *For the QMF problem* (16) *the algorithm RQMF with fixed regularization parameters $M_u, M_v > 0$ reaches an unchanged full cycle after T full iterates with*

$$T \le \left\lfloor \frac{2(f(x_0) - f^*)}{\min(M_u, M_v)} \right\rfloor. \tag{29}$$

Moreover, the fixed point (U_T, V_T) of RQMF satisfies:

$$\mathcal{G}(U_T, V_T) \le \max \Big\{ \tfrac{M_u}{2} m(\beta - \alpha)^2, \tfrac{M_v}{2} n(\beta - \alpha)^2 \Big\}, \tag{30}$$

and the maximum objective function improvement over any feasible single-entry unit move is at most $\max\{M_u, M_v\}/2$.

Proof Using the same arguments as in the proof of Theorem 2, we get (29). Moreover, at a fixed point, (12) holds for each column of U or V. Replacing that column by an unregularized best response $z \in \mathbb{Z}^m_{[\alpha,\beta]}$ yields (recall that $u_T^{(i)}$ denotes ith column of U_T):

$$f(U_T, V_T) - f(z, \hat{U}_{T,i}; V_T) \leq \tfrac{M_u}{2}\|z - u_T^{(i)}\|^2 \leq \tfrac{M_u m}{2}(\beta - \alpha)^2. \qquad (31)$$

The analogous argument bounds the improvement from replacing one column of V. Taking their maximum proves (30). To prove the last assertion, consider first an arbitrary entry $u_{T,p}^{(i)}$ of the ith column of U_T, and let $h \in \{-1, 1\}$ be such that $u_{T,p}^{(i)} + h \in \mathbb{Z}_{[\alpha,\beta]}$. Define the feasible vector $z \in \mathbb{Z}^m_{[\alpha,\beta]}$ whose jth entry is:

$$z_j = \begin{cases} u_{T,p}^{(i)} + h, & j = p, \\ u_{T,j}^{(i)}, & j \neq p. \end{cases}$$

Thus, z differs from $u_T^{(i)}$ in exactly one entry, and that entry changes by one unit. Consequently, $\|z - u_T^{(i)}\|^2 = h^2 = 1$. Since (U_T, V_T) is a fixed point of RQMF, the optimality condition (31), applied to the ith column of U_T, gives:

$$f(U_T, V_T) - f(z, \hat{U}_{T,i}; V_T) \leq \frac{M_u}{2}\|z - u_T^{(i)}\|^2 = \frac{M_u}{2}.$$

The left-hand side represents the decrease in the objective function f resulting from this single-entry modification. Therefore, any feasible unit change in an entry of U_T can improve f by at most $M_u/2$. An analogous argument applies to the entries of V_T. Combining the two bounds, we conclude that the maximum decrease in the objective function achievable by any feasible single-entry unit move in either factor matrix is at most $\max\{M_u, M_v\}/2$. □

3.2 Discrete column improvement gap optimality for RQMF under rational data

The next result makes explicit how the precision of the data and the regularization parameters determine columnwise discrete optimality. For a given positive integer q the condition $Y \in q^{-1}\mathbb{Z}^{m\times n}$ means that every entry of Y is an integer multiple of $1/q$. For example, integer-valued image data satisfy this condition with $q = 1$. More generally, it holds when the data lie on a fixed rational grid. To the best of our knowledge the next convergence result is the first derived for a coordinate minimization algorithm solving QMF problem (16).

Theorem 5 *Suppose $Y \in q^{-1}\mathbb{Z}^{m\times n}$ for some positive integer q and RQMF uses the fixed regularization parameters*

$$0 < M_u < 1/q, \qquad 0 < M_v < 1/q. \qquad (32)$$

Then RQMF reaches a fixed point after $T \leq \lfloor 2q\big(f(x^0) - f^\star\big)\rfloor$ full iterations and every such fixed point (U_T, V_T) satisfies $\mathcal{G}(U_T, V_T) = 0$.

Proof Finite termination follows from Theorem 2, since the feasible set is finite and $M_u, M_v > 0$. Further we show that any strictly improving feasible integer move decreases f by at least $1/(2q)$. Indeed, take any two feasible factorizations (U, V) and $(\widetilde{U}, \widetilde{V})$, and set $P = UV^T, \widetilde{P} = \widetilde{U}\widetilde{V}^T$. Both P and $\widetilde{P}$ have integer entries and since $Y \in q^{-1}\mathbb{Z}^{m\times n}$, there exists $A \in \mathbb{Z}^{m\times n}$ such that $Y = A/q$. Then, $f(U,V) = \frac{1}{2}\|Y\|_F^2 - \mathrm{Trace}(Y^T P) + \frac{1}{2}\|P\|_F^2$. Therefore,

$$\begin{aligned} f(U,V) - f(\widetilde{U},\widetilde{V}) &= -\mathrm{Trace}(Y^T(P-\widetilde{P})) + \frac{1}{2}\big(\|P\|_F^2 - \|\widetilde{P}\|_F^2\big) \\ &= \frac{q\big(\|P\|_F^2 - \|\widetilde{P}\|_F^2\big) - 2\mathrm{Trace}(A^T(P-\widetilde{P}))}{2q} \in \frac{1}{2q}\mathbb{Z}, \end{aligned}$$

since the numerator is an integer. In conclusion, any strictly positive objective decrease between feasible points is at least $1/(2q)$:

$$f(U,V) - f(\widetilde{U},\widetilde{V}) \geq \frac{1}{2q} \Rightarrow f(U_k,V_k) - f(U_{k+1},V_{k+1}) \geq \frac{1}{2q} 1_{\{(U_{k+1},V_{k+1}) \neq (U_k,V_k)\}}.$$

For any $K \geq 1$, define $T_K = \sum_{k=0}^{K-1} 1_{\{(U_{k+1},V_{k+1}) \neq (U_k,V_k)\}}$. Summing previous inequality, we get

$$\frac{T_K}{2q} \leq \sum_{k=0}^{K-1} (f(U_k,V_k) - f(U_{k+1},V_{k+1})) \leq f(U_0,V_0) - f^\star.$$

The nondecreasing integer sequence T_K is therefore bounded. Its limit T, the total number of full iterations in which at least one integer entry of the factors changes, satisfies the given bound. Let now (U_T, V_T) be any fixed point of RQMF, and set $E = Y - U_T V_T^T$. To prove the second assertion, we first show that no feasible unit change in an entry of U_T can strictly decrease the objective f. Recall from (24) that, for any feasible integer displacement h of the entry $u_{T,p}^{(i)}$, we have (24), i.e.:

$$f(U_T + he_p e_i^T, V_T) - f(U_T, V_T) = \frac{a_i}{2}h^2 - s_{pi}h,$$

where recall that $a_i = \|v_T^{(i)}\|^2$ and $s_{pi} = (Ev_T^{(i)})_p$. Since V_T has integer entries, $a_i \in \mathbb{Z}$. Moreover, $U_T V_T^T$ has integer entries and $Y \in q^{-1}\mathbb{Z}^{m\times n}$, thus $E \in q^{-1}\mathbb{Z}^{m\times n}$ and $s_{pi} \in q^{-1}\mathbb{Z}$. Consequently, for every integer h,

$$f(U_T + he_p e_i^T, V_T) - f(U_T, V_T) \in \frac{1}{2q}\mathbb{Z}. \tag{33}$$

Thus, any strictly improving feasible integer update decreases f by at least $1/(2q)$. On the other hand, because (U_T, V_T) is a fixed point, the block-optimality condition (31), applied to the column $u_T^{(i)}$ with the feasible candidate $u_T^{(i)} + he_p$, yields

$$f(U_T, V_T) - f(U_T + he_p e_i^T, V_T) \leq \frac{M_u}{2}h^2.$$

In particular, for any feasible unit move $h \in \{-1, 1\}$, its improvement is at most $M_u/2 < 1/(2q)$, which contradicts (33), i.e., every strictly positive improvement is at least $1/(2q)$. Hence, no feasible unit change in an entry of U_T can strictly decrease the objective f. We now show that no feasible integer displacement h of larger magnitude than ± 1 can improve f either. Suppose a feasible $h > 1$ satisfies $\frac{a_i}{2}h^2 - s_{pi}h < 0$. As $h > 0$ and $a_i \geq 0$, we obtain $s_{pi} > a_i h/2 \geq a_i/2$. Hence $a_i/2 - s_{pi} < 0$, so the unit move $h = 1$ would also strictly improve f, a contradiction. Note that this unit move is feasible because $\alpha \leq u_{T,p}^{(i)} < u_{T,p}^{(i)} + 1 < u_{T,p}^{(i)} + h \leq \beta$. Similarly, if a feasible $h < -1$ satisfies $\frac{a_i}{2}h^2 - s_{pi}h < 0$, then $s_{pi} < a_i h/2 \leq -a_i/2$, implying $a_i/2 + s_{pi} < 0$. Thus the feasible unit move $h = -1$ would strictly improve f, again a contradiction. Together with the unit-move case, this proves that no feasible change in any individual entry of U_T can decrease f. Since for fixed V_T and the other columns of U_T, the objective is separable over the entries of $u_T^{(i)}$, it follows that no feasible replacement of any entire column of U_T can decrease f. The identical argument applies to V_T, using $b_i = \|u_T^{(i)}\|^2$ and $t_{qi} = (E^T u_T^{(i)})_q$, and the condition $0 < M_v < 1/q$. We conclude that (U_T, V_T) is the best response with respect to each column of both factor matrices for f. Finally, Lemma 4 gives $\mathcal{G}(U_T, V_T) = 0$. □

3.3 Convergence rate in column improvement gap under adaptive regularization for RQMF

The preceding results show that positive regularization ensures strict objective decrease whenever an RQMF iterate changes. However, for fixed regularization parameters, an unchanged column need not be a best response for the original QMF objective. Moreover, Theorem 5 ensures exact columnwise optimality under a rational-grid assumption on data Y and sufficiently small regularization parameters. We now consider an adaptive choice of the regularization parameters in RQMF that removes these restrictions by preventing regularization from suppressing a strictly improving integer column update.

Recall the notation $E = Y - UV^T$, $a_i = \|v^{(i)}\|^2$, $s_{pi} = (Ev^{(i)})_p$, $b_i = \|u^{(i)}\|^2$, and $t_{qi} = (E^T u^{(i)})_q$ introduced in Section 3.1. For a feasible factorization (U, V) and $i = 1 : d$, we define

$$\tau_i^U(U, V) = \max\left\{0, \max_{\substack{p=1:m \\ h \in \mathcal{H}_{pi}^U \cap \{-1,1\}}} (2s_{pi}h - a_i)\right\}, \tag{34}$$

$$\tau_i^V(U, V) = \max\left\{0, \max_{\substack{q=1:n \\ h \in \mathcal{H}_{qi}^V \cap \{-1,1\}}} (2t_{qi}h - b_i)\right\}. \tag{35}$$

An empty inner maximum is interpreted as $-\infty$. These quantities are nonnegative and depend only on the current factorization. The next lemma shows

that they are the exact thresholds above which regularization can prevent the corresponding column from changing.

Lemma 5 *Let (U, V) be a feasible factorization of the QMF problem* (16) *and let $i = 1 : d$. For $M_u \geq 0$, the column $u^{(i)}$ minimizes*

$$\min_{z \in \mathbb{Z}^m_{[\alpha,\beta]}} \left\{ f(z, \hat{U}_i; V) + \frac{M_u}{2} \|z - u^{(i)}\|^2 \right\}$$

if and only if $M_u \geq \tau_i^U(U, V)$. Furthermore, it is the unique minimizer if $M_u > \tau_i^U(U, V)$, whereas it is not a minimizer if $0 \leq M_u < \tau_i^U(U, V)$. The analogous statements hold for $v^{(i)}$, with M_v and $\tau_i^V(U, V)$. In particular,

$$\mathcal{G}(U, V) = 0 \quad \Longleftrightarrow \quad \tau_i^U(U, V) = \tau_i^V(U, V) = 0, \qquad i = 1 : d. \tag{36}$$

Proof We prove the statements for a column of U, since those for a column of V follow analogously. Fix i and let $h \in \mathbb{Z}^m$ be a feasible displacement, i.e., $h_p \in \mathcal{H}_{pi}^U$ for $p = 1 : m$. Since the new residual is $E - h(v^{(i)})^T$, expanding the squared Frobenius norm gives

$$\begin{aligned} f(u^{(i)} + h, \hat{U}_i; V) - f(U, V) &= \frac{1}{2} \left(\|E - h(v^{(i)})^T\|_F^2 - \|E\|_F^2 \right) \\ = \frac{1}{2} \|h(v^{(i)})^T\|_F^2 - \langle E, h(v^{(i)})^T \rangle &= \frac{a_i}{2} \|h\|^2 - h^T E v^{(i)} = \sum_{p=1}^{m} \left(\frac{a_i}{2} h_p^2 - s_{pi} h_p \right). \end{aligned}$$

Consequently, the change in the regularized column objective is

$$\sum_{p=1}^{m} \left(\frac{a_i + M_u}{2} h_p^2 - s_{pi} h_p \right). \tag{37}$$

Since the feasible displacements form the Cartesian product of the sets $\mathcal{H}_{pi}^U$ and (37) is separable, the current column is a minimizer if and only if, for every $p = 1 : m$,

$$\frac{a_i + M_u}{2} h_p^2 - s_{pi} h_p \geq 0 \qquad \forall h_p \in \mathcal{H}_{pi}^U. \tag{38}$$

We show that it is sufficient to check feasible unit displacements. If $h_p \geq 1$ is feasible, then $1 \in \mathcal{H}_{pi}^U$ and

$$\frac{a_i + M_u}{2} h_p^2 - s_{pi} h_p = h_p \left(\frac{a_i + M_u}{2} h_p - s_{pi} \right) \geq h_p \left(\frac{a_i + M_u}{2} - s_{pi} \right).$$

Thus, the condition for the feasible unit move $h_p = 1$ implies (38) for every positive feasible integer displacement. Similarly, if $h_p \leq -1$ is feasible, then $-1 \in \mathcal{H}_{pi}^U$ and

$$\frac{a_i + M_u}{2} h_p^2 - s_{pi} h_p = -h_p \left(-\frac{a_i + M_u}{2} h_p + s_{pi} \right) \geq -h_p \left(\frac{a_i + M_u}{2} + s_{pi} \right).$$

Hence, the condition for the feasible unit move $h_p = -1$ implies (38) for every negative feasible integer displacement. The case $h_p = 0$ gives equality. Therefore, (38) is equivalent to

$$\frac{a_i + M_u}{2} - s_{pi}\sigma \geq 0 \qquad \forall \sigma \in \mathcal{H}^U_{pi} \cap \{-1, 1\}, \quad p = 1 : m.$$

Rearranging, this is equivalent to

$$M_u \geq 2s_{pi}\sigma - a_i \qquad \forall p,\ \forall \sigma \in \mathcal{H}^U_{pi} \cap \{-1, 1\}.$$

Since $M_u \geq 0$, the preceding inequalities hold if and only if $M_u \geq \tau^U_i(U, V)$, proving the first statement. If $M_u > \tau^U_i(U, V)$, every feasible nonzero unit displacement has a strictly positive regularized objective change. The previous inequalities then show that every feasible nonzero scalar displacement also has a strictly positive change. Therefore, the entire column $u^{(i)}$ is the unique minimizer. Conversely, if $M_u < \tau^U_i(U, V)$, there exist p and a feasible signed unit move σ such that $2s_{pi}\sigma - a_i > M_u$. This gives

$$\frac{a_i + M_u}{2} - s_{pi}\sigma < 0,$$

so the unit move strictly improves the regularized objective, and $u^{(i)}$ cannot be a minimizer. Equality $M_u = \tau^U_i(U, V)$ may result in multiple minimizers. Finally, applying the previous equivalence with $M_u = 0$ shows that $\tau^U_i(U, V) = 0$ if and only if $u^{(i)}$ is an unregularized best response. The analogous statement holds for $v^{(i)}$. By Lemma 4, these conditions hold for all columns if and only if $\mathcal{G}(U, V) = 0$, proving (36). □

Based on these thresholds, we introduce an adaptive choice of regularization parameters for RQMF. Let $\theta \in (0, 1)$ and let $\bar{M}_u, \bar{M}_v > 0$ be arbitrary fixed parameters. At each visit to column i, using the current factorization (U, V), we choose

$$M_{u,k,i} = \begin{cases} \theta\tau^U_i(U, V), & \tau^U_i(U, V) > 0 \\ \bar{M}_u, & \tau^U_i(U, V) = 0 \end{cases}, \quad M_{v,k,i} = \begin{cases} \theta\tau^V_i(U, V), & \tau^V_i(U, V) > 0 \\ \bar{M}_v, & \tau^V_i(U, V) = 0. \end{cases} \tag{39}$$

Since the thresholds depend on the current factorization, they must be evaluated before each column update rather than only once per full cycle. The update order remains that of Algorithm 2 (RQMF), with M_u and M_v replaced by the corresponding adaptive parameters, equivalently, the column updates at the current factorization can be expressed as

$$(u^{(i)})^+ = \left[u^{(i)} + \frac{Ev^{(i)}}{a_i + M_{u,k,i}}\right]_{\mathbb{Z}^m_{[\alpha,\beta]}} \quad \text{and} \quad (v^{(i)})^+ = \left[v^{(i)} + \frac{E^T u^{(i)}}{b_i + M_{v,k,i}}\right]_{\mathbb{Z}^n_{[\alpha,\beta]}}.$$

Every adaptive regularization parameter is strictly positive, so the denominators remain positive even when an opposite factor column is zero. The next theorem establishes finite termination of this adaptive RQMF algorithm and exact column improvement gap optimality for arbitrary real data Y.

Theorem 6 *Let $Y \in \mathbb{R}^{m\times n}$ and consider RQMF updates with adaptive regularization parameters* (39)*. Then, the following statements hold.*
(i) At any column visit with threshold $\tau > 0$, the updated column changes and the objective function decreases by at least $(1-\theta)\tau/2$.
(ii) At any column visit with $\tau = 0$, the corresponding column is an unregularized best response and remains unchanged.
(iii) The algorithm performs only finitely many changed column updates and reaches a terminal factorization (U_T, V_T) satisfying

$$\mathcal{G}(U_T, V_T) = 0. \tag{40}$$

In particular, the number of changed column updates is at most $|\mathcal{D}|$, where $\mathcal{D} = \mathbb{Z}^{m\times d}_{[\alpha,\beta]} \times \mathbb{Z}^{n\times d}_{[\alpha,\beta]}$.

Proof We first prove (i) for an update of column $u^{(i)}$, and write $\tau = \tau_i^U(U,V) > 0$. Since the maximum in (34) is taken over a finite set, there exist $p \in \{1,\dots,m\}$ and a feasible $\sigma \in \{-1,1\}$ such that $\tau = 2s_{pi}\sigma - a_i > 0$. Consider the feasible factorization obtained by replacing $u^{(i)}$ with $u^{(i)} + \sigma e_p$. From (24), its unregularized objective change is

$$f(U + \sigma e_p e_i^T, V) - f(U,V) = \frac{a_i}{2}\sigma^2 - s_{pi}\sigma = \frac{a_i}{2} - s_{pi}\sigma = -\frac{\tau}{2}.$$

Let $(u^{(i)})^+$ be an exact minimizer of the regularized column subproblem with $M = M_{u,k,i} = \theta\tau$. Using the feasible candidate $u^{(i)} + \sigma e_p$ in the optimality inequality, we get

$$\begin{aligned} &f((u^{(i)})^+, \hat{U}_i; V) + \frac{M}{2}\|(u^{(i)})^+ - u^{(i)}\|^2 \le f(u^{(i)} + \sigma e_p, \hat{U}_i; V) + \frac{M}{2}\|\sigma e_p\|^2 \\ &= f(U,V) - \frac{\tau}{2} + \frac{M}{2} = f(U,V) - \frac{1-\theta}{2}\tau. \end{aligned} \tag{41}$$

This proves (i) for an update of U. The proof for an update of V is identical, using τ_i^V, b_i, t_{qi}, and a feasible unit displacement of an entry of $v^{(i)}$. To prove (ii), suppose that $\tau_i^U(U,V) = 0$. By Lemma 5, the current column $u^{(i)}$ minimizes the unregularized column objective. Therefore, for every feasible $z \in \mathbb{Z}^m_{[\alpha,\beta]}$, $f(z, \hat{U}_i; V) \ge f(U,V)$ and since the adaptive rule chooses $M_{u,k,i} = \bar{M}_u > 0$ in this case, for every feasible $z \neq u^{(i)}$ we have

$$f(z, \hat{U}_i; V) + \frac{\bar{M}_u}{2}\|z - u^{(i)}\|^2 > f(U,V),$$

while the regularized objective at $z = u^{(i)}$ equals $f(U,V)$. Thus $u^{(i)}$ is the unique regularized minimizer, and the update leaves it unchanged. The argument for a column of V is analogous.

Finally, we prove (iii). Every column visit either leaves the current factorization unchanged (according to (ii)), or strictly decreases f (according to (i)). All factorizations produced by the algorithm belong to the finite set $\mathcal{D}$. Hence two distinct changed visits cannot lead back to the same factorization: otherwise

the strictly decreasing objective would take the same value at two states along a strictly decreasing chain. It follows that there can be at most $|\mathcal{D}|$ changed column updates. After the last changed visit, all subsequent column visits leave the factors unchanged, and a complete unchanged cycle is reached. Let (U_T, V_T) be the factorization at such a cycle. During this cycle the factors remain constant. If any visited column had a positive threshold evaluated at (U_T, V_T), statement (i) would force it to change, a contradiction. Consequently,

$$\tau_i^U(U_T, V_T) = 0, \qquad \tau_i^V(U_T, V_T) = 0, \qquad i = 1 : d.$$

The equivalence (36) yields $\mathcal{G}(U_T, V_T) = 0$, completing the proof of (iii). □

Remark 4 The adaptive parameters (39) remain positive, but they are not required to satisfy the uniform bound $M_u, M_v < 1/q$ from Theorem 5. The conclusion $\mathcal{G}(U_T, V_T) = 0$ also does not imply global optimality. □

4 Numerical experiments on image compression

In this section we present numerical results on image compression using real data. We evaluate our proposed Algorithm 2 (RQMF), against several baselines: Singular Value Decomposition (SVD) based compression [15], standard JPEG [31], and quantization-aware matrix factorization (QMF) algorithm [1]. For RQMF and QMF algorithms we choose $K = 10$ full iterations. For simulations we use three datasets: KODAK (24 lossless images with a resolution of 768×512) [20], CLIC 2024 (32 high-resolution, high-quality images) [28] and ImageNet validation dataset (50.000 images with various resolutions, all higher than 256×256) [8]. Then, the compressed ImageNet dataset is also used in a classification task using 3 well-known convolutional networks AlexNet [22], ResNet50 [16] and MobileNetV2 [29] based on floating point and integer arithmetic representations. All experimental results were implemented in Python on a laptop with Intel Core i9 up to 5.4 GHz (14 physical cores and 20 threads) and 32 GB RAM memory. The code can be found on Github: rqmf-image-compression.

4.1 Preprocessing of color images

We follow the same preprocessing workflow as in [1], see also [15]. More specifically, for color image compression the recommended color transformation is the (YC_BC_R) color space instead of the RGB model, as the human visual system is more sensitive to luminance variations than to chrominance differences [27,31]. Hence, in our experiments we decompose our initial RGB image $X \in \mathbb{Z}_{[0,255]}^{m\times n\times 3}$ in 3 matrices: (i) $Y \in \mathbb{R}_{[0,\,255]}^{m\times n}$ called *luma* which encapsulates the brightness information; (ii) Y_{C_B}, $Y_{C_R} \in \mathbb{R}_{[0,\,255]}^{m\times n}$ correspond to the chrominance components encoding the blue-difference and red-difference signals, respectively. Note

that in this representation, the values are continuous but restricted to the interval $[0, 255]$. Moreover, we downsample each chrominance channel using an average-pooling kernel of size $(2, 2)$, which results in Y_{C_B}, $Y_{C_R} \in \mathbb{R}^{\frac{m}{2} \times \frac{n}{2}}_{[0,\,255]}$. Due to the nature of our matrix decomposition–based compression method, it is beneficial to use patching in the preprocessing step [1, 15]. Hence, each channel is decomposed into non-overlapping 8×8 patches. These patches are then flattened and stacked vertically, resulting in the new matrices $Y \in \mathbb{R}^{\frac{mn}{64} \times 64}_{[0,\,255]}$, Y_{C_B}, $Y_{C_R} \in \mathbb{R}^{\frac{mn}{256} \times 64}_{[0,\,255]}$. When the compression interval $[\alpha, \beta]$ is centered in zero, slightly improved compression results are obtained when these matrices are also centered around zero, e.g., $Y \leftarrow Y - 128$.

4.2 Experimental setup for image quantized compression

For all methods, the integer bounds for the factor matrices were set to $[\alpha, \beta] = [-16, 15]$ and the maximum number of iterations for RQMF and QMF algorithms was set to $K = 10$. We initialize algorithms RQMF and QMF with U_0, V_0 from truncated SVD. The factorization rank, d, was controlled by the quality factor $Q \in [0, 1]$, where $d = \max(\text{round}(Q \times \min(m, n)), 1)$.

Metrics: we evaluate the methods by using standard rate-distortion metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Bits Per Pixel (bpp). The first two metrics assess quality of compressed images, while the last metric measures memory usage.

Parameters choice for RQMF: Note that Algorithm 2 (RQMF) uses 2 regularization parameters M_u and M_v. Our objective is to minimize the final bit rate (bpp) while maintaining a high PSNR and SSIM. This requires a content-aware trade-off: (i) *high complexity images* (i.e., high contrast, high entropy) demand weak regularization (M_u, M_v small) to preserve fine details and energy; (ii) *low complexity images* (i.e., uniform areas) can tolerate strong regularization (M_u, M_v large), which significantly lowers the entropy of the resulting integer factors, thereby maximizing lossless compression efficiency and minimizing bpp. Hence, we set the parameters M_u, M_v depending on the entropy of the image:

$$M_u = M_v = \gamma \cdot \left(1 - \frac{H(Y)}{H_{\max}}\right), \tag{42}$$

where γ is a scaling factor, $H(Y)$ is the Shannon entropy of image Y and $H_{\max}$ is constant and represents the maximum entropy. Given the reduced sensitivity of the human visual system to color errors compared to luminance errors, we employ a differentiated strategy (empirically designed content-adaptive heuristic) for regularization of the luminance and chroma channels: *Luminance (Y)* - the γ is set to a restrictive value ($\gamma = 60$) to strictly preserve structural details; *Chrominance (C_R, C_B)* - a strong, content-adaptive regularization γ

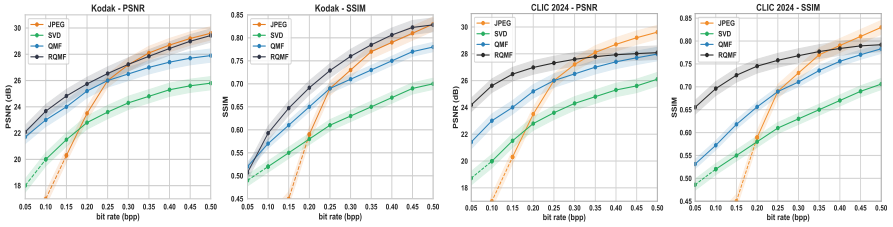

Fig. 1 Rate-distortion performance on the Kodak and CLIC 2024 datasets. The average PSNR and average SSIM for each method are plotted as functions of bit rate. Shaded areas represent standard errors. Dashed lines indicate extrapolated values predicted using LOESS [7] for extremely low bit rates that are otherwise unattainable.

is applied to maximize the compression efficiency (minimize bpp) in these channels ($\gamma = 80$). The term $(1 - H_c/H_{\max})$ quantifies *compresibility*: a high value indicates low information entropy (H_c) and high potential for lossless compression.

4.3 Numerical results for image quantized compression

The numerical results for image quantized compression are reported in Figures 1 and 3. The rate-distortion curves in terms of PSNR and SSIM versus bpp are presented in Figure 1 for our algorithm RQMF and the baselines QMF, SVD and JPEG. As one can see from Figure 1, RQMF consistently achieves superior performance, significantly surpassing SVD, JPEG and QMF at low bit rates. More specifically, on the Kodak images, RQMF achieves consistently stronger PSNR and SSIM performance across all bit rates compared to SVD, JPEG and QMF. At higher bit rates, its PSNR performance is comparable to that of JPEG, while it provides the best overall visual quality (SSIM), as illustrated in the top-right of Figure 1. For high-resolution images, such as CLIC 2024 dataset, RQMF notably outperforms SVD, JPEG and QMF at low bit rates below 0.3 bpp, enabling bit rate savings without introducing excessive visual degradation. Across both datasets, our algorithm RQMF achieves PSNR values above 24 dB and SSIM scores above 0.60 for bpp$>$ 0.1, demonstrating that combination of coordinate minimization and proper regularization yields an algorithm that enjoys convergence guarantees and it is also efficient, robust in preserving visual quality and better in several metrics than an algorithm which does not use regularization as QMF [1].
In terms of quality Figure 2 shows that RQMF maintains better visual fidelity than QMF, JPEG and SVD, effectively preserving color fidelity and intricate details, particularly in low-bit-rate scenarios.
The mean compression and decompression CPU times are reported in Table 1. The compression and decompression speed of RQMF is comparable to QMF, benefiting from the fast matrix multiplication operations, which are significantly faster than the DCT operations required by JPEG.

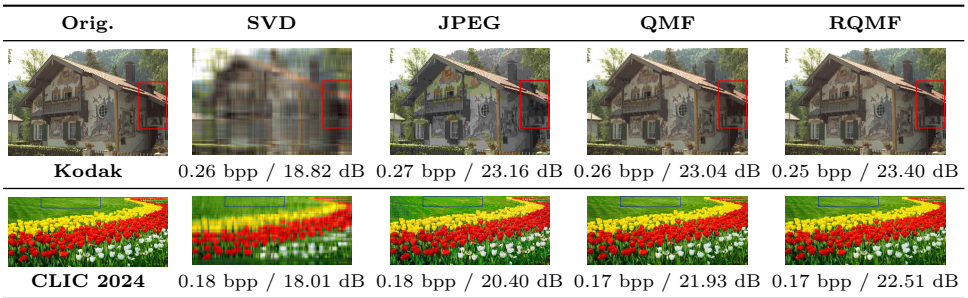

Fig. 2 Qualitative performance comparison on example images from the Kodak and CLIC 2024 datasets. The colored bounding boxes highlight the artifacts introduced by the compression methods.

Table 1 Mean compression and decompression CPU times for the four methods at bit rates of 0.15 bpp and 0.25 bpp, measured on the Kodak and CLIC 2024 datasets.

Dataset	Bit rate	JPEG	SVD	QMF	RQMF
Kodak	0.15 bpp	5.4 ms	1.6 ms	3.2 ms	3.2 ms
	0.25 bpp	5.2 ms	1.5 ms	3.1 ms	3.1 ms
CLIC 2024	0.15 bpp	29.6 ms	7.2 ms	15.2 ms	14.9 ms
	0.25 bpp	26.5 ms	5.9 ms	13.9 ms	13.8 ms

4.4 Ablation studies

We performed ablation studies to assess the influence of the factor bounds $[\alpha, \beta]$, initialization (U_0, V_0), number of iterations (K) and regularization parameters (M_u, M_v) on the compression for algorithm RQMF. The results are provided in Figure 3. In the left figure, the factor range ($[-16, 15]$) provides the best PSNR–bit-rate trade-off for RQMF. Narrower factor bounds such as $[-8, 7]$ reduce bit allocation and improve compression, but they are overly restrictive. The second figure illustrates the impact of the number of (full) iterations K on the compression. Increasing the number of iterations improves PSNR, although the gains eventually are insignificant (around $K = 10$) and no longer justify the additional computational effort. Note that for $K = 0$, RQMF employs no iterations and relies only on the SVD initialization, thus the poor result. In the third figure, we evaluate the impact of the initialization on the performance of RQMF, executing different number of iterations ranging from 10 to 200. It can be observed that initialization based on truncated SVD with $K = 10$ (black line), consistently achieves higher compression quality than random initialization, even after 10 to 200 iterations. Finally, from the right figure one can notice that our algorithm is robust w.r.t. a wide range of the regularization parameters (M_u, M_v) (in simulations we choose $M_u = M_v$ as in (42) and vary γ for luminance and consider $\gamma + 20$ for chrominance).

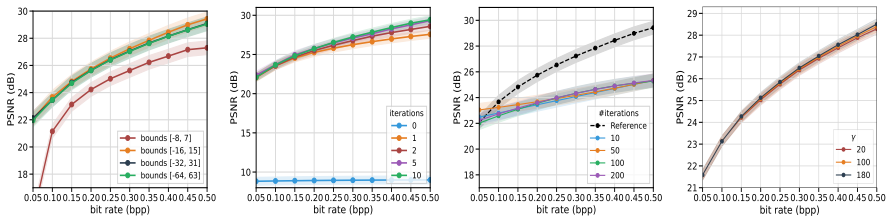

Fig. 3 Sensitivity of RQMF w.r.t. $[\alpha, \beta]$ bounds, number of iterations K, initialization of (U, V) and regularization parameters (M_u, M_v) on Kodak images.

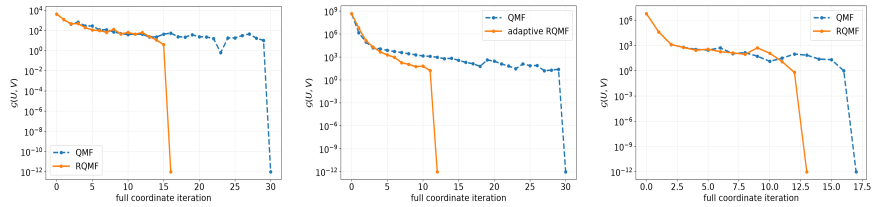

Fig. 4 The column improvement gap along full coordinate iterations for RQMF and QMF on a random real (first 2 plots) and rational (last plot) data matrix Y: RQMF with $M_u = Mv = 0.5$ (left), adaptive RQMF with $\theta = 0.01$ and $\bar{M}_u = \bar{M}_v = 1$ (middle), RQMF with $M_u = M_v = 0.02 < 1/q$ (right).

We also investigate the behaviour of (adaptive) RQMF w.r.t. the column improvement gap. For this purpose we consider some synthetic data, generating a random data matrix Y in two ways: in the first $Y = Y_Z + \epsilon \in \mathbb{R}^{100\times 1000}$, where Y_Z is a random matrix with only integer entries in the interval $[-15, 16]$ and ϵ is a gaussian noise with a standard deviation 0.5; in the second $Y = q^{-1}Y_Z$, where $q = 32$. In both scenarios we consider $d = 16, \alpha = -15, \beta = 16$. In Figure 4, we plot the column improvement gap for the data matrix with real entries: RQMF with fixed $M_u = Mv = 0.5$ (left) and adaptive variant (39) with $\theta = 0.01$ and $\bar{M}_u = \bar{M}_v = 1$ (middle); for a rational data matrix, RQMF with fixed $M_u = M_v = 0.02 < 1/q$. From the plots one can observe that the regularization ensures a smaller number of full coordinate iterations to achieve exact column improvement gap compared to QMF [1].

4.5 Application of image compression on classification

To assess the practical impact of our quantized compression scheme beyond rate-distortion metrics, we evaluate the compressed images in a downstream classification task. Our main motivation comes from embedded and edge devices, which typically rely on low-precision (integer) arithmetic due to limited memory, power and computational budgets. In such scenarios, the entire processing chain, from image preprocessing and storage to neural network inference, benefits from an integer representation of the data. Since the factor matrices produced by our RQMF are constrained to bounded integer values,

the compressed images can be stored, transmitted and reconstructed using `INT8` arithmetic[1], making our method naturally compatible with integer-only inference pipelines. The main benefits of quantization of the entire inference pipeline consist in a memory reduction of almost 4x, 2-4x speedup in computation and power savings proportional to memory savings (see Table II).

Network architectures: We consider three convolutional network (CNN) architectures representing distinct stages and design principles in the evolution of image classification networks. AlexNet serves as an early, parameter-heavy conventional CNN baseline; ResNet50 represents a deeper residual architecture; and MobileNetV2 represents a lightweight model explicitly designed for mobile and embedded inference. This selection enables us to investigate whether the effects of image compression and post-training quantization are consistent across architectures with substantially different capacities, computational structures and sensitivities to low-precision inference.

- *AlexNet* [22]: one of the first deep convolutional networks to achieve breakthrough performance on ImageNet. It consists of 5 convolutional layers followed by 3 fully connected layers ($\approx$ 61M parameters).
- *ResNet50* [16]: a 50-layer deep residual network enabling the training of very deep models ($\approx$ 25.6M parameters).
- *MobileNetV2* [29]: a lightweight architecture explicitly designed for mobile, embedded and edge devices ($\approx$ 3.5M parameters). It replaces standard convolutions with depthwise separable convolutions, drastically reducing the number of operations and parameters. Hence, MobileNetV2 is the most relevant architecture for embedded scenarios.

Evaluation pipeline: We consider ImageNet validation set (50.000 images), where compression is applied to the full-resolution original image, and then we use the standard evaluation preprocessing steps: resize to 256×256, center crop to 224×224, scale to $[0, 1]$ and finally normalized. For inference we consider two numerical representations of each network: (i) the original pretrained model using 32-bit floating point (FP32) arithmetic, and (ii) its post-training quantized (PTQ) counterpart, in which both weights and activations are quantized to 8-bit integers [13,32]. We run these convolutional networks on RQMF, SVD and JPEG compressed images at matched bit rates. The classification performance is measured via the standard Top-1 and Top-5 accuracies.

Post-training quantization of the networks: In order to decrease the storage and compute requirements of the model during inference its parameters are stored as integers with a low number of bits. To obtain the integer counterparts of the networks, we employ post-training quantization (PTQ), which is a family of techniques that convert a pre-trained floating point network to low-precision (integer) arithmetic *without any retraining* [18,21]. Compared to

[1] The factor entries use a 5-bit integer alphabet $[-16, 15]$, which can be stored using `INT8` data types.

quantization-aware training, PTQ is simpler to deploy, does not require access to the original training pipeline, making it the method of choice for embedded and edge scenarios. For the reader's convenience, in what follows we briefly outline the principles and procedures underlying post-training quantization of neural networks based on the works [18, 21]. In *weight and activation quantization*, the network weights and the activations are converted to low precision (e.g., `INT8`), which enables integer-only inference. Our presentation of PTQ below follows [21].

Uniform affine quantization. Let $x \in [x_{\min}, x_{\max}]$ be a real-valued quantity (a weight or an activation) to be mapped onto b-bit integers with $N = 2^b$ representable levels ($N = 256$ for `INT8`). The *uniform affine quantizer* is parameterized by a scale $\Delta > 0$ and an integer zero-point z:

$$\Delta \;=\; \frac{x_{\max} - x_{\min}}{N-1}, \qquad z \;=\; \operatorname{round}\left(-\frac{x_{\min}}{\Delta}\right), \tag{43}$$

and maps a real value x to its integer representation via

$$x_{\text{int}} \;=\; \operatorname{clamp}\Big(0,\, N-1,\, \operatorname{round}\big(\tfrac{x}{\Delta}\big) + z\Big). \tag{44}$$

This zero-point z guarantees that the real value 0 is represented *exactly*, an important property for common operations such as zero padding; for one-sided distributions the range $[x_{\min}, x_{\max}]$ is relaxed to include zero. A frequently used simplification is the *symmetric* quantizer, which fixes $z = 0$ and uses a range symmetric around zero:

$$\Delta \;=\; \frac{\max\big(|x_{\min}|, |x_{\max}|\big)}{N/2 - 1}, \quad x_{\text{int}} \;=\; \operatorname{clamp}\Big(-\tfrac{N}{2},\, \tfrac{N}{2} - 1,\, \operatorname{round}\big(\tfrac{x}{\Delta}\big)\Big),$$

which removes the cross-terms involving the zero-point in integer convolution and thereby simplifies fast Single Instruction Multiple Data implementations at the cost of a slightly less efficient use of integer grid for skewed distributions.

Granularity of quantization. A key design choice in PTQ is the *granularity* of the quantizer. With per-layer (per-tensor) quantization, a single pair (Δ, z) is used for an entire weight tensor $W \in \mathbb{R}^{K \times K \times C_{\text{in}} \times C_{\text{out}}}$, whereas per-channel quantization assigns an individual scale to each convolutional kernel, i.e., in symmetric case to each output channel $c \in \{1, \ldots, C_{\text{out}}\}$:

$$\Delta_c \;=\; \frac{\max\limits_{k,l,m} \big|W(k,l,m;c)\big|}{N/2 - 1}, \tag{45}$$

Since scale is constant within each kernel, both granularities remain compatible with efficient integer dot-product and convolution kernels; activations, in contrast, are quantized per-tensor, as per-channel activation scales complicate inner accumulation of convolutions and matrix multiplication operations.

Range estimation. For weights, the quantizer parameters are obtained directly from the empirical extrema, $x_{\min} = \min_i w_i$ and $x_{\max} = \max_i w_i$ (per tensor

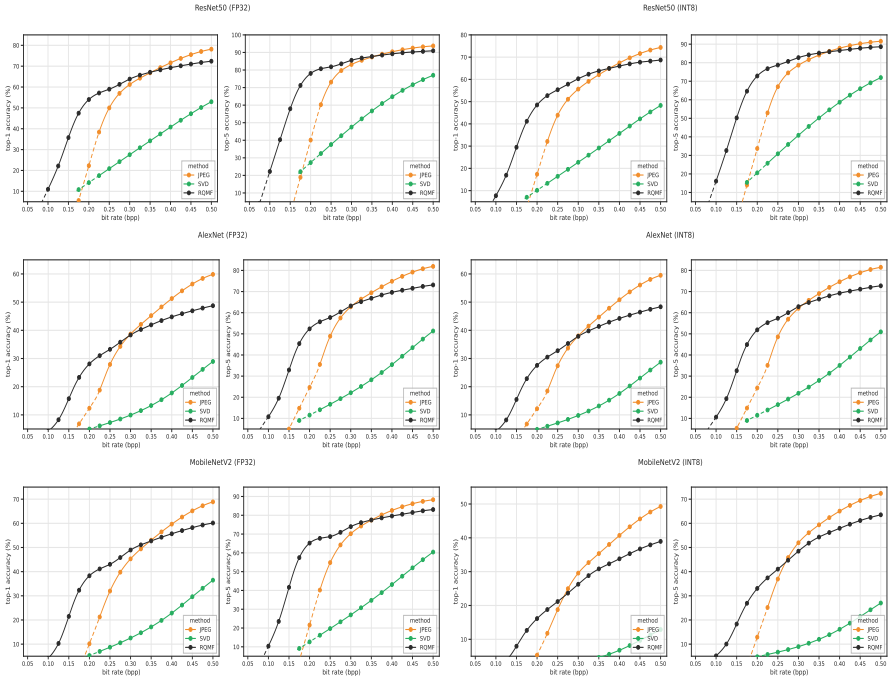

Fig. 5 Top-1 and top-5 accuracies on the ImageNet validation set at several bit rates for ResNet50, AlexNet and MobileNetV2 in both FP32 and INT8 representations.

or per channel). Since activation ranges are input-dependent, they must be estimated in a short calibration step, in which a small set of representative images (typically around one hundred mini-batches) is propagated through the network and the dynamic ranges of each layer are collected as exponential moving averages of the per-batch extrema:

$$x_{\min}^{(t)} = (1-\alpha)\, x_{\min}^{(t-1)} + \alpha \min(x^{(t)}),\ x_{\max}^{(t)} = (1-\alpha)\, x_{\max}^{(t-1)} + \alpha \max(x^{(t)}), \quad (46)$$

with smoothing factor $\alpha \in (0,1)$; more elaborate criteria, such as selecting the clipping range that minimizes the Kullback–Leibler divergence between the original and the quantized distribution, can further improve accuracy [21].

As shown in [21], per-channel quantization of weights combined with per-layer quantization of activations achieves classification accuracies within 2% of the floating point baseline for a wide range of architectures at 8 bits. Larger, over-parameterized networks (e.g., ResNets) are generally more robust to quantization than compact ones (e.g., MobileNet). In this work we adopt the following PTQ recipe: symmetric per-channel quantization of weights (see eq. (45)), asymmetric per-tensor quantization of activations (see eqs. (43)–(44)) and calibration-based range estimation (see eq. (46)) - to obtain the `INT8` counterparts of the networks used in our classification experiments.

Numerical results for classification. Figure 5 and Table 2 report Top-1 and Top-5 accuracies of the three networks (AlexNet, MobileNetV2 and ResNet50), in both `FP32` and `INT8` representations, on the original ImageNet dataset and also compressed with RQMF, JPEG and SVD at different bit rates. First, comparing the two numerical representations, post-training quantization is virtually lossless for AlexNet and ResNet50 (at most a 4-point Top-1 drop), confirming that the quantization error is negligible at the classification level for these architectures. MobileNetV2 is more sensitive to quantization, as expected for compact models built on depthwise separable convolutions, which are notoriously challenging for standard post-training quantization; notably, this gap is already present on uncompressed images and is therefore independent of the image codec. Second, the images compressed with RQMF preserve the semantic content required for classification: RQMF yields higher accuracies than JPEG at low bit rates (below $\sim$0.3 bpp) for all networks and both precisions, consistent with the rate-distortion results of the previous section. Table 2 showcases the performance of the three networks at low bit rate (0.3 bpp) compression on ImageNet. One can notice that the average inference time per image (including memory access, transfer times and preprocessing in order to simulate a real embedded system scenario) has a 2-3x speedup for `INT8` representation of the networks. Moreover, one can observe approximately 4x reduction in memory usage. Overall, these results indicate that the proposed integer-based compression (algorithm RQMF) can be combined with integer-only network inference into a fully quantized end-to-end pipeline suitable for embedded devices, with negligible impact on the classification performance.

5 Conclusions

In this paper we have considered cyclic coordinate minimization with quadratic regularization for solving general (possibly nonconvex) composite optimization problems. We have analyzed its convergence behaviour and derived rates depending on the assumptions on the problem. We have also shown that, in the presence of bounded integer constraints, the first-order stationarity condition is not informative since it is satisfied at every feasible point. For this reason, we have established finite termination and introduced an objective-based optimality measure characterizing blockwise improvements for such particular problems, together with explicit bounds depending on the regularization parameters. Then, we have specialized the algorithm to quantized matrix factorization problems that arise in e.g., image compression. Finally, we have evaluated both the original and quantized compressed images in a classification task using several well-known convolutional networks. The numerical results on image compression and classification using real data have shown the flexibility and efficiency of our algorithm when compared to well-established methods from the literature.

Table 2 Top-1 and top-5 accuracies and average inference time on the ImageNet validation set at a target bit rate of approximately 0.30 bpp.

CNNs	Parameters	Precision	Methods	Top-1 (%)	Top-5 (%)	Time (ms/img)	Disk (MB)
ResNet50	25.6M	FP32	Orig.	80.34	95.13	27.61	97.79
			JPEG	59.57	81.85	14.70	
			SVD	14.20	27.33	21.01	
			RQMF	64.15	85.80	22.34	
		INT8	Orig.	77.36	93.63	13.87	24.95
			JPEG	53.85	77.05	10.45	
			SVD	10.17	20.64	10.19	
			RQMF	60.63	83.01	10.55	
MobileNetV2	3.5M	FP32	Orig.	72.01	90.62	27.86	13.60
			JPEG	43.14	67.92	14.39	
			SVD	5.30	12.68	21.17	
			RQMF	49.34	74.36	22.53	
		INT8	Orig.	52.79	76.39	10.08	3.75
			JPEG	27.81	49.61	6.41	
			SVD	1.75	4.79	8.22	
			RQMF	26.68	49.01	8.60	
AlexNet	61.1M	FP32	Orig.	56.51	79.07	28.53	233.09
			JPEG	36.95	60.78	13.73	
			SVD	4.99	11.58	22.50	
			RQMF	38.69	63.57	23.97	
		INT8	Orig.	56.16	78.87	9.22	58.47
			JPEG	36.30	60.07	5.29	
			SVD	4.87	11.51	7.26	
			RQMF	38.21	63.23	7.69	

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Data Availability The code repository address has been provided in the abstract.

Ethical approval Not applicable.

Consent for publication Not applicable.

A Apendix

Proof of Lemma 2

Proof First, using the optimality condition for $x_{k+1}^{(i_k)}$ and differentiability of f, we get:

$$0 \in \nabla_{(i_k)} f(x_{k,i_k}) + \partial g_{i_k}(x_{k+1}^{(i_k)}) + M_{i_k}\left(x_{k+1}^{(i_k)} - x_k^{(i_k)}\right),$$

that is

$$-\left(\nabla_{(i_k)} f(x_{k,i_k}) + M_{i_k}\left(x_{k+1}^{(i_k)} - x_k^{(i_k)}\right)\right) \in \partial g_{i_k}(x_{k+1}^{(i_k)}).$$

Combining this relation with basic calculus rules for subdifferentiability and Assumption 1 (i), we finally get:

$$\partial F(x_{k+1}) = \nabla f(x_{k+1}) + \partial g(x_{k+1}) = \left[\nabla_{(i)} f(x_{k+1}) + \partial g_i(x_{k+1}^{(i)})\right]_{i=1}^{N}$$
$$\ni \left[\nabla_{(i)} f(x_{k+1}) - \left(\nabla_{(i)} f(x_{k,i}) + M_i\left(x_{k+1}^{(i)} - x_k^{(i)}\right)\right)\right]_{i=1}^{N}.$$

Then, we establishes a relation between the norm of the subdifferential of the objective function and of the difference between consecutive iterates. Indeed,

$$\min_{s_{k+1}\in\partial F(x_{k+1})} \|s_{k+1}\| \le \left\| \left[\nabla_{(i)} f(x_{k+1}) - \nabla_{(i)} f(x_{k,i}) - M_i\left(x_{k+1}^{(i)} - x_k^{(i)}\right)\right]_{i=1}^{N} \right\|$$
$$\le \left\| \left[\nabla_{(i)} f(x_{k+1}) - \nabla_{(i)} f(x_{k,i})\right]_{i=1}^{N} \right\| + \left\| \left[M_i\left(x_{k+1}^{(i)} - x_k^{(i)}\right)\right]_{i=1}^{N} \right\|$$
$$\le \left\| \left[\nabla_{(i)} f(x_{k+1}) - \nabla_{(i)} f(x_{k,i})\right]_{i=1}^{N} \right\| + M_{\max} \|x_{k+1} - x_k\|.$$

Let us now introduce the notation:

$$\bar{\nabla}^2 f(z(1), \cdots, z(d)) = \begin{bmatrix} \nabla_1^2 f(z(1)) \\ \vdots \\ \nabla_d^2 f(z(d)) \end{bmatrix}, \tag{47}$$

with $\nabla_j^2 f(z(j))$ being the jth row of the hessian of f at the point $z(j) \in \mathbb{R}^n$. By the mean value theorem, we have that there exist d vectors $z_k(j) \in [x_{k,i}, x_{k+1}]$, where i is the block index containing the index j, such that:

$$\nabla_j f(x_{k+1}) - \nabla_j f(x_{k,i}) = \nabla_j^2 f(z_k(j))(x_{k+1} - x_{k,i}) \ \ \forall j = 1:d.$$

Since the sequence $(x_k)_{k\ge 0}$ is assumed bounded, there exists a finite positive constant C_f such that $\|\nabla_j^2 f(z_k(j))\| \le C_f$ for all $k \ge 0$ and $j = 1:d$. Hence, from Cauchy-Schwarz inequality we further get:

$$|\nabla_j f(x_{k+1}) - \nabla_j f(x_{k,i})| \le C_f \|x_{k+1} - x_{k,i}\| \le C_f \|x_{k+1} - x_k\| \ \ \forall k \ge 0, j = 1:d.$$

Finally, we get:

$$\min_{s_{k+1}\in\partial F(x_{k+1})} \|s_{k+1}\| \le \left\| \left[\nabla_{(i)} f(x_{k+1}) - \nabla_{(i)} f(x_{k,i})\right]_{i=1}^{N} \right\| + M_{\max} \|x_{k+1} - x_k\|$$
$$\le (C_f\sqrt{n} + M_{\max}) \|x_{k+1} - x_k\|.$$

Denote $C = C_f\sqrt{n} + M_{\max}$ (note that C appears in the rate, our algorithm does not require explicit knowledge of it). Combining now Lemma 1 with these results, we get:

$$F(x_{k+1}) \le F(x_k) - \frac{M_{\min}}{2}\|x_{k+1} - x_k\|^2 \le F(x_k) - \frac{M_{\min}}{2C^2} \min_{s_{k+1}\in\partial F(x_{k+1})} \|s_{k+1}\|^2. \tag{48}$$

Adding this relation from $k = 0$ to $k = K$, and using Assumption 1 (ii) we obtain the convergence rate (3). □

Proof of Lemma 3.

Proof Combining Lemma 1 and Lemma 2, we have:

$$F(x_{k+1}) \le F(x_k) - \frac{M_{\min}}{2C^2} \min_{s_{k+1}\in\partial F(x_{k+1})} |s_{k+1}|^2. \tag{49}$$

Since we have $\text{dist}(0, \partial F(x_{k+1})) := \min_{s_{k+1} \in \partial F(x_{k+1})} |s_{k+1}|$, inequality (49) becomes

$$F(x_{k+1}) \leq F(x_k) - \frac{M_{\min}}{2C^2} \text{dist}(0, \partial F(x_{k+1}))^2. \tag{50}$$

Further, for $k \geq k_0$, we assume the iterate x_k lies in the neighborhood where the KL inequality (4) holds, i.e.:

$$F(x_{k+1}) - F_* \leq \sigma_q \left(\text{dist}(0, \partial F(x_{k+1}))^2\right)^{\frac{q}{2}}.$$

Substituting the above relation into (50), we obtain

$$F(x_{k+1}) \leq F(x_k) - \frac{M_{\min}}{2C^2} \left(\frac{F(x_{k+1}) - F_*}{\sigma_q}\right)^{2/q},$$

which proves (5). □

References

1. Ashtari, P., Behmandpoor, P., Haredasht, F.N., Chen, J.H., Patrinos, P., Huffel, S.V.: Quantization-aware matrix factorization for low bit rate image compression. Information Sciences (2025)
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka–Łojasiewicz inequality. Mathematics of Operations Research **35**(2), 438–457 (2010). DOI 10.1287/moor.1100.0449
3. Beck, A., Tetruashvili, L.: On the convergence of block coordinate descent type methods. SIAM Journal on Optimization **23**(4) (2013)
4. Bertsekas, D.: Nonlinear Programming. Athena Scientific (1999)
5. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM Journal on Optimization **17**(4), 1205–1223 (2007). DOI 10.1137/050644641
6. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming **146** (2014)
7. Cleveland, W.S., Devlin, S.J.: Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American statistical association **83**(403), 596–610 (1988)
8. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). DOI 10.1109/CVPR.2009.5206848
9. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika **1**(3), 211–218 (1936)
10. Fedotov, R., Budge, B., Kavan, L.: Qmf-blend: Quantized matrix factorization for efficient blendshape compression. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers, pp. 1–10 (2025)
11. Fercoq, O., Richtarik, P.: Accelerated, parallel and proximal coordinate descent. SIAM Journal on Optimization **25**(4) (2015)
12. Frein, R.D.: Quantized nonnegative matrix factorization. International Conference on Digital Signal Processing **doi: 10.1109/ICDSP.2014.6900690** (2014)
13. Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M., Keutzer, K.: A survey of quantization methods for efficient neural network inference. In: Low-Power Computer Vision: Improve the Efficiency of Artificial Intelligence. Chapman and Hall/CRC (2022)
14. Gillis, N.: Nonnegative matrix factorization. SIAM (2020)
15. Golpayegani, Z., Bouguila, N.: Patchsvd: A non-uniform svd-based image compression algorithm. arXiv preprint arXiv:2406.05129 (2024)

16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016). DOI 10.1109/CVPR.2016.90
17. Hou, J., Chau, L.P., Magnenat-Thalmann, N., He, Y.: Sparse low-rank matrix approximation for data compression. IEEE Transactions on Circuits and Systems for Video Technology **27** (2015)
18. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2704–2713 (2018)
19. Jäger, S., Schöbel, A.: The blockwise coordinate descent method for integer programs. Mathematical Methods of Operations Research **91**(2), 357–381 (2020). DOI 10.1007/s00186-019-00673-x
20. Kodak: Kodak lossless true color image suite. `https://r0k.us/graphics/kodak/` (1993)
21. Krishnamoorthi, R.: Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv:1806.08342 (2018)
22. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 25 (2012)
23. Necoara, I.: Coordinate gradient descent methods. chapter in Big Data and Computational Intelligence in Networking, Y. Wu et al. (Eds.), Taylor & Francis LLC - CRC Press (2017)
24. Necoara, I., Chorobura, F.: Efficiency of stochastic coordinate proximal gradient methods on nonseparable composite optimization. Mathematics of Operations Research **50**(2) (2024)
25. Necoara, I., Clipici, D.: Efficient parallel coordinate descent algorithm for convex optimization problems with separable constraints: Application to distributed mpc. Journal of Process Control **23**(3) (2013)
26. Necoara, I., Clipici, D.: Parallel random coordinate descent methods for composite minimization: convergence analysis and error bounds. SIAM Journal on Optimization **26**(1) (2016)
27. Netravali, A.: Digital pictures: Representation. Compression, and Standards (1995)
28. Organizers: Challenge on learned image compression. `https://www.compression.cc/` (2024)
29. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
30. Tseng, P., Yun, S.: Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. Journal of Optimization Theory and Applications **140** (2009)
31. Wallace, G.K.: The jpeg still picture compression standard. Communications of the ACM **34**(4), 30–44 (1991)
32. Wang, M., Rasoulinezhad, S., Leong, P.H., So, H.K.H.: Niti: Training integer neural networks using integer-only arithmetic. IEEE Transactions on Parallel and Distributed Systems **33**(11), 3249–3261 (2022)
33. Wright, S.: Coordinate descent algorithms. Mathematical programming **151**(1), 3–34 (2015)