Supplementary Material for Adaptive Relaxed ADMM

Zheng Xu[1]*, Mário A. T. Figueiredo[2], Xiaoming Yuan[3], Christoph Studer [4], Tom Goldstein[1]
[1]Department of Computer Science, University of Maryland, College Park, MD
[2]Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
[4]Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY

In this supplementary material for adaptive relaxed ADMM (ARADMM), we provide details of proofs, implementations and more experimental results. Section 1 provides the details of proofs for the lemmas, theorems, and propositions in Section 3 and Section 4 of the main text. Section 2 provides the implementation details of the applications, datasets and parameter setting. Section 3 provides more experimental results, including a result table of the complete list of all the experiments, convergence curves, visual results of image restoration and face decomposition, and additional sensitivity analysis.

1. Proofs of Lemmas and Theorems

1.1. Proof of Lemma 1

Proof. Using the dual updates (5), VI (28) can be rewritten as

$$\forall v,\ g(v) - g(v_{k+1}) - (Bv - Bv_{k+1})^T \lambda_{k+1} \geq 0. \quad \text{(S1)}$$

Similarly, in the previous iteration,

$$\forall v,\ g(v) - g(v_k) - (Bv - Bv_k)^T \lambda_k \geq 0. \quad \text{(S2)}$$

After letting $v = v_k$ in (S1) and $v = v_{k+1}$ in (S2), we sum the two inequalities together to conclude

$$(Bv_{k+1} - Bv_k)^T (\lambda_{k+1} - \lambda_k) \geq 0. \quad \text{(S3)}$$

□

Lemma S1. *The optimal solution $z^* = (u^*, v^*, \lambda^*)^T$ and sequence $z_k = (u_k, v_k, \lambda_k)^T$ generated by ADMM satisfy*

$$\begin{aligned}&(\tau_k B\Delta v^*_{k+1} + \Delta\lambda^*_{k+1})^T(\tau_k B\Delta v^+_k + \Delta\lambda^+_k)\\ &\geq \frac{1-\gamma_k}{\gamma_k}\|\tau_k B\Delta v^+_k + \Delta\lambda^+_k\|^2 \quad \text{(S4)}\\ &+ \gamma_k((\tau_k B\Delta v^*_k)^T\Delta\lambda^*_k - (\tau_k B\Delta v^*_{k+1})^T\Delta\lambda^*_{k+1}).\end{aligned}$$

*xuzh@cs.umd.edu

1.2. Proof of Lemma S1

Proof. We replace $y = y^*, z = z^*$ in VI (29) and $y = y_{k+1}, z = z_{k+1}$ in VI (26), and sum the two inequalities to get

$$\begin{aligned}&(\Delta z^*_{k+1})^T\Omega(\Delta z^+_k, \tau_k, \gamma_k)\\ &\qquad \geq (\Delta z^*_{k+1})^T(F(z^*) - F(z_{k+1})). \quad \text{(S5)}\end{aligned}$$

From (29), the monotonicity of $F(z)$, and $\Omega(\Delta z^+_k, \tau_k, \gamma_k)$, we have

$$\begin{aligned}&(\tau_k A\Delta u^*_{k+1})^T((\gamma_k - 1)\Delta\lambda^+_k - \tau_k B\Delta v^+_k)\\ &\quad + (\Delta\lambda^*_{k+1})^T(\Delta\lambda^+_k + (1-\gamma_k)\tau_k B\Delta v^+_k) \geq 0. \quad \text{(S6)}\end{aligned}$$

Using the feasibility of optimal solution $Au^* + Bv^* = b$, λ_{k+1} in (5) and $\tilde{u}_{k+1}$ in (3), we have

$$\tau_k A\Delta u^*_{k+1} = \frac{1}{\gamma_k}\Delta\lambda^+_k + \frac{1-\gamma_k}{\gamma_k}\tau_k B\Delta v^+_k - \tau_k B\Delta v^*_{k+1}. \quad \text{(S7)}$$

We now substitute (S7) into (S6) and simplify to get,

$$\begin{aligned}&(\tau_k B\Delta v^*_{k+1} + \Delta\lambda^*_{k+1})^T(\tau_k B\Delta v^+_k + \Delta\lambda^+_k) \geq\\ &\frac{1-\gamma_k}{\gamma_k}\|\tau_k B\Delta v^+_k + \Delta\lambda^+_k\|^2 + \gamma_k(\tau_k B\Delta v^+_k)^T\Delta\lambda^+_k\\ &+ \gamma_k((\tau_k B\Delta v^*_{k+1})^T\Delta\lambda^+_k + (\tau_k B\Delta v^+)^T\Delta\lambda^*_{k+1}). \quad \text{(S8)}\end{aligned}$$

We can use the fact that

$$\Delta\lambda^*_k = \Delta\lambda^*_{k+1} + \Delta\lambda^+_k \text{ and } \Delta v^*_k = \Delta v^*_{k+1} + \Delta v^+_k, \quad \text{(S9)}$$

to get

$$\begin{aligned}&(\tau_k B\Delta v^*_{k+1})^T\Delta\lambda^+_k + (\tau_k B\Delta v^+)^T\Delta\lambda^*_{k+1}\\ &\quad = (\tau_k B\Delta v^*_k)^T\Delta\lambda^*_k - (\tau_k B\Delta v^*_{k+1})^T\Delta\lambda^*_{k+1}\\ &\qquad - (\tau_k B\Delta v^+_k)^T\Delta\lambda^+_k. \quad \text{(S10)}\end{aligned}$$

Finally, we substitute (S10) into (S8) to get (S4). □

1.3. Proof of Lemma 2

Proof. Begin by deriving

$$\begin{aligned}
&\|\tau_k B\Delta v_k^* + \Delta\lambda_k^*\|^2 \\
&= \|(\tau_k B\Delta v_{k+1}^* + \Delta\lambda_{k+1}^*) + (\tau_k B\Delta v_k^+ + \Delta\lambda_k^+)\|^2 && \text{(S11)}\\
&= \|\tau_k B\Delta v_{k+1}^* + \Delta\lambda_{k+1}^*\|^2 + \|\tau_k B\Delta v_k^+ + \Delta\lambda_k^+\|^2 \\
&\quad + 2(\tau_k B\Delta v_{k+1}^* + \Delta\lambda_{k+1}^*)^T(\tau_k B\Delta v_k^+ + \Delta\lambda_k^+) && \text{(S12)}\\
&\geq \|\tau_k B\Delta v_{k+1}^* + \Delta\lambda_{k+1}^*\|^2 + \frac{2-\gamma_k}{\gamma_k}\|\tau_k B\Delta v_k^+ + \Delta\lambda_k^+\|^2 \\
&\quad + 2\gamma_k((\tau_k B\Delta v_k^*)^T\Delta\lambda_k^* - (\tau_k B\Delta v_{k+1}^*)^T\Delta\lambda_{k+1}^*), && \text{(S13)}
\end{aligned}$$

where (S9) is used for (S11), and Lemma S1 is used for (S13). We now have

$$\begin{aligned}
&\frac{2-\gamma_k}{\gamma_k}\|\tau_k B\Delta v_k^+ + \Delta\lambda_k^+\|^2 \\
&\leq \|\tau_k B\Delta v_k^* + \Delta\lambda_k^*\|^2 - \|\tau_k B\Delta v_{k+1}^* + \Delta\lambda_{k+1}^*\|^2 \\
&\quad -2\gamma_k((\tau_k B\Delta v_k^*)^T\Delta\lambda_k^* - (\tau_k B\Delta v_{k+1}^*)^T\Delta\lambda_{k+1}^*) && \text{(S14)}\\
&= \|\tau_k B\Delta v_k^*\|^2 + \|\Delta\lambda_k^*\|^2 - \|\tau_k B\Delta v_{k+1}^*\|^2 - \|\Delta\lambda_{k+1}^*\|^2 \\
&\quad - 2(\gamma_k-1)(\tau_k B\Delta v_k^*)^T\Delta\lambda_k^* \\
&\quad - 2(\gamma_k-1)(-\tau_k B\Delta v_{k+1}^*)^T\Delta\lambda_{k+1}^* && \text{(S15)}\\
&= \gamma_k(\|\tau_k B\Delta v_k^*\|^2 + \|\Delta\lambda_k^*\|^2) \\
&\quad - (2-\gamma_k)(\|\tau_k B\Delta v_{k+1}^*\|^2 + \|\Delta\lambda_{k+1}^*\|^2) \\
&\quad - (\gamma_k-1)\|\tau_k B\Delta v_k^* + \Delta\lambda_k^*\|^2 \\
&\quad - (\gamma_k-1)\|\tau_k B\Delta v_{k+1}^* - \Delta\lambda_{k+1}^*\|^2 && \text{(S16)}\\
&\leq \gamma_k(\|\tau_k B\Delta v_k^*\|^2 + \|\Delta\lambda_k^*\|^2) \\
&\quad - (2-\gamma_k)(\|\tau_k B\Delta v_{k+1}^*\|^2 + \|\Delta\lambda_{k+1}^*\|^2). && \text{(S17)}
\end{aligned}$$

□

1.4. Proof of Theorem 2

Proof. When $\gamma_k < 2$ as in Assumption 2, Lemma 2 suggests

$$\begin{aligned}
&\frac{1}{\gamma_k}\|B\Delta v_k^+ + \frac{1}{\tau_k}\Delta\lambda_k^+\|^2 \\
&\leq \frac{\gamma_k}{2-\gamma_k}(\|B\Delta v_k^*\|^2 + \frac{1}{\tau_k^2}\|\Delta\lambda_k^*\|^2) \\
&\quad - (\|B\Delta v_{k+1}^*\|^2 + \frac{1}{\tau_k^2}\|\Delta\lambda_{k+1}^*\|^2). && \text{(S18)}
\end{aligned}$$

Assumption 2 also suggests

$$\frac{\gamma_k}{(2-\gamma_k)\tau_k^2} \leq \frac{1+\theta_k^2}{\tau_{k-1}^2} \text{ and } \frac{\gamma_k}{2-\gamma_k} \leq (1+\theta_k^2). \quad \text{(S19)}$$

Then (S18) leads to

$$\begin{aligned}
&\frac{1}{\gamma_k}\|\frac{1}{\tau_k}\Delta v_k^+ + B\Delta\lambda_k^+\|^2 \\
&\leq (1+\theta_k^2)(\|B\Delta v_k^*\|^2 + \frac{1}{\tau_{k-1}^2}\|\Delta\lambda_k^*\|^2) \\
&\quad - (\|B\Delta v_{k+1}^*\|^2 + \frac{1}{\tau_k^2}\|\Delta\lambda_{k+1}^*\|^2). && \text{(S20)}
\end{aligned}$$

Accumulating (S20) from $k=0$ to get

$$\begin{aligned}
&\sum_{k=0}^{N}\prod_{t=k+1}^{N}(1+\theta_t^2)\frac{1}{\gamma_k}\|B\Delta v_k^+ + \frac{1}{\tau_k}\Delta\lambda_k^+\|^2 \\
&\quad \leq \prod_{k=1}^{N}(1+\theta_t^2)(\|B\Delta v_0^*\|^2 + \frac{1}{\tau_0^2}\|\Delta\lambda_0^*\|^2). && \text{(S21)}
\end{aligned}$$

Assumption 2 suggests $\prod_{t=1}^{\infty}(1+\theta_t^2) < \infty$, and $\prod_{t=k+1}^{N}(1+\theta_t^2)\frac{1}{\gamma_k} \geq \frac{1}{\gamma_k} > 0.5$. Then (S21) indicates $\sum_{k=0}^{\infty}\|B\Delta v_k^+ + \frac{1}{\tau_k}\Delta\lambda_k^+\|^2 < \infty$. Hence

$$\lim_{k\to\infty}\|B\Delta v_k^+ + \frac{1}{\tau_k}\Delta\lambda_k^+\|^2 = 0. \quad \text{(S22)}$$

Since $(B\Delta v_k^+)^T\Delta\lambda_k^+ \geq 0$ as in Lemma 1,

$$\lim_{k\to\infty}\|\frac{1}{\tau_k}\Delta\lambda_k^+\|^2 \leq \lim_{k\to\infty}\|B\Delta v_k^+ + \frac{1}{\tau_k}\Delta\lambda_k^+\|^2 = 0 \quad \text{(S23)}$$

$$\lim_{k\to\infty}\|B\Delta v_k^+\|^2 \leq \lim_{k\to\infty}\|B\Delta v_k^+ + \frac{1}{\tau_k}\Delta\lambda_k^+\|^2 = 0. \quad \text{(S24)}$$

The residuals r_k, d_k in (6) then satisfy

$$r_k = \frac{1}{\gamma_k\tau_k}\Delta\lambda_{k-1}^+ - \frac{\gamma_k-1}{\gamma_k}B\Delta v_{k-1}^+ \quad \text{(S25)}$$

$$d_k = \tau_k A^T B\Delta v_{k-1}^+. \quad \text{(S26)}$$

We finally have

$$\begin{aligned}
\lim_{k\to\infty}\|r_k\| &\leq \lim_{k\to\infty}\frac{1}{\gamma_k}\|\frac{1}{\tau_k}\Delta\lambda_{k-1}^+\| + \frac{\gamma_k-1}{\gamma_k}\|B\Delta v_{k-1}^+\| \\
&\leq \lim_{k\to\infty}\frac{\sqrt{1+\theta_k^2}}{\gamma_{k-1}}\|\frac{1}{\tau_k}\Delta\lambda_{k-1}^+\| \\
&\quad + \frac{\gamma_k-1}{\gamma_k}\|B\Delta v_{k-1}^+\| = 0, \text{ and} && \text{(S27)}\\
\lim_{k\to\infty}\|d_k\| &\leq \lim_{k\to\infty}|A|\|\tau_k B\Delta v_{k-1}^+\| \\
&\leq \lim_{k\to\infty}(1+\eta_k^2)\tau_k|A|\,\|B\Delta v_{k-1}^+\| = 0. && \text{(S28)}
\end{aligned}$$

□

1.5. Equivalence of relaxed ADMM and relaxed DRS in Section 4.1

Proof. Referring back to the ADMM steps (2)–(5), and defining $\hat{\lambda}_{k+1} = \lambda_k + \tau_k(b - Au_{k+1} - Bv_k)$, the optimality

condition for the minimization of (2) is

$$0 \in \partial h(u_{k+1}) - A^T\lambda_k - \tau_k A^T(b - Au_{k+1} - Bv_k) \quad \text{(S29)}$$
$$= \partial h(u_{k+1}) - A^T\hat{\lambda}_{k+1}, \quad \text{(S30)}$$

which is equivalent to $A^T\hat{\lambda}_{k+1} \in \partial h(u_{k+1})$, thus[1] $u_{k+1} \in \partial h^*(A^T\hat{\lambda}_{k+1})$. A similar argument using the optimality condition for (4) leads to $v_{k+1} \in \partial g^*(B^T\lambda_{k+1})$. Recalling (10), we arrive at

$$Au_{k+1} - b \in \partial\hat{h}(\hat{\lambda}_{k+1}) \text{ and } Bv_{k+1} \in \partial\hat{g}(\lambda_{k+1}). \quad \text{(S31)}$$

Using these identities, we finally have

$$\hat{\lambda}_{k+1} = \lambda_k + \tau_k(b - Au_{k+1} - Bv_k) \quad \text{(S32)}$$
$$\in \lambda_k - \tau_k\big(\partial\hat{h}(\hat{\lambda}_{k+1}) + \partial\hat{g}(\lambda_k)\big) \quad \text{(S33)}$$
$$\lambda_{k+1} = \lambda_k + \tau_k(b - \tilde{u}_{k+1} - Bv_{k+1}) \quad \text{(S34)}$$
$$= \lambda_k + \gamma_k\tau_k(b - Au_{k+1} - Bv_{k+1}) + (1-\gamma_k)\tau_k(Bv_k - Bv_{k+1}) \quad \text{(S35)}$$
$$\in \lambda_k - \tau_k\big(\partial\hat{h}(\hat{\lambda}_{k+1}) + \partial\hat{g}(\lambda_{k+1})\big) + (1-\gamma_k)\tau_k\left(\partial\hat{g}(\lambda_k) - \partial\hat{g}(\lambda_{k+1})\right), \quad \text{(S36)}$$

showing that the sequences $(\lambda_k)_{k\in\mathbb{N}}$ and $(\hat{\lambda}_k)_{k\in\mathbb{N}}$ satisfy the same conditions (11) and (12) as $(\zeta_k)_{k\in\mathbb{N}}$ and $(\hat{\zeta}_k)_{k\in\mathbb{N}}$, thus proving that ADMM for problem (1) is equivalent to DRS for its dual (10). □

1.6. Proof of Proposition 1 in Section 4.2

Proof. Rearrange DRS step (12) to get

$$0 \in \frac{\zeta_{k+1} - \zeta_k}{(1-\gamma)\tau} + \frac{\gamma}{1-\gamma}\partial\hat{h}(\hat{\zeta}_{k+1}) - \partial\hat{g}(\zeta_k) + \frac{1}{1-\gamma}\partial\hat{g}(\zeta_{k+1}). \quad \text{(S37)}$$

Combine DRS step (11) and (S37) to get

$$0 \in \frac{1}{\tau}\Big(\frac{\zeta_{k+1}}{1-\gamma} + \hat{\zeta}_{k+1} - \frac{2-\gamma}{1-\gamma}\zeta_k\Big) + \frac{1}{1-\gamma}\big(\partial\hat{h}(\hat{\zeta}_{k+1}) + \hat{g}(\zeta_{k+1})\big). \quad \text{(S38)}$$

Inserting the linear assumption (13) to DRS step (11), we can explicitly get the update for $\hat{\zeta}_{k+1}$ as

$$\hat{\zeta}_{k+1} = \frac{1-\beta\tau}{1+\alpha\tau}\zeta_k - \frac{a\tau + b\tau}{1+\alpha\tau}, \quad \text{(S39)}$$

where $a \in \Psi$ and $b \in \Phi$. Inserting the linear model (13) into (S38), we get

$$\zeta_{k+1} = \frac{\gamma - 1 - \alpha\tau}{1+\beta\tau}\hat{\zeta}_{k+1} + \frac{2-\gamma}{1+\beta\tau}\zeta_k - \frac{(a+b)\tau}{1+\beta\tau} \quad \text{(S40)}$$
$$= \zeta_k - \gamma\tau\frac{(\alpha+\beta)\zeta_k + (a+b)}{(1+\alpha\tau)(1+\beta\tau)}, \quad \text{(S41)}$$

[1] An important property relating f and f^* is that $y \in \partial f(x)$ if and only if $x \in \partial f^*(y)$ [11].

where the second equality results from using the expression for $\hat{\zeta}_{k+1}$ from (S39).

The residual r_{DR} at ζ_{k+1} is simply the magnitude of the subgradient (corresponding to elements $a \in \Psi$ and $b \in \Phi$) of the objective and is given by

$$r_{DR} = \|(\alpha+\beta)\zeta_{k+1} + (a+b)\| \quad \text{(S42)}$$
$$= \left|1 - \frac{\gamma\tau(\alpha+\beta)}{(1+\alpha\tau)(1+\beta\tau)}\right| \cdot \|(\alpha+\beta)\zeta_k + (a+b)\|, \quad \text{(S43)}$$

where ζ_{k+1} in (S43) was substituted with (S41). The optimal parameters minimize the residual

$$\tau, \gamma = \arg\min_{\tau,\gamma} r_{\text{DR}} = \arg\min_{\tau,\gamma}\left|1 - \frac{\gamma\tau(\alpha+\beta)}{(1+\alpha\tau)(1+\beta\tau)}\right|. \quad \text{(S44)}$$

This residual has optimal value of zero when

$$\gamma_k = 1 + \frac{1+\alpha\beta\tau_k^2}{(\alpha+\beta)\tau_k}.$$

□

2. Implementation details

We provide the implementation details, datasets, and parameter settings for the various applications.

2.1. Proximal operators

We introduce the proximal operators of some functions that are needed to solve the ADMM subproblems.

The proximal of the ℓ_1 norm is called *shrink*, and is defined by

$$\text{shrink}(z,t) = \arg\min_x \|x\|_1 + \frac{1}{2t}\|x-z\|_2^2 = \text{sign}(z)\odot\max\{|z| - t, 0\}, \quad \text{(S45)}$$

where $\text{sign}(\cdot)$ indicates the elementwise sign of a real valued vector, $\odot$ represents the elementwise multiplication, $|\cdot|$ represents the elementwise absolute value.

The proximal of the nuclear norm is the singular value shrinkage operator

$$\text{SVT}(Z,t) = \arg\min_X \|X\|_* + \frac{1}{2t}\|X-Z\|_F^2 = U\,\mathcal{T}(\Lambda,t)\,V^T, \quad \text{(S46)}$$

where $Z = U\Lambda V^T$ is the singular value decomposition, $\mathcal{T}(\Lambda,t)$ is a diagonal matrix with nonnegative singular values obtained by shrinking the diagonal of Λ by parameter t.

The proximal operator of the hinge loss is

$$\text{pxhinge}(z,t) = \arg\min_x \sum_{i=1}^n \max\{1-x_i, 0\} + \frac{1}{2t}\|x-z\|_2^2 = z + \max\{\min\{1-z, t\}, 0\}. \quad \text{(S47)}$$

2.2. Linear regression with elastic net regularization

Elastic net regularization (EN) is a modification of the ℓ_1 (or LASSO) regularizer that helps preserve groups of highly correlated variables [18, 4], and requires solving

$$\min_x \frac{1}{2}\|Dx - c\|_2^2 + \rho_1\|x\|_1 + \frac{\rho_2}{2}\|x\|_2^2, \tag{S48}$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm, $D \in \mathbb{R}^{n\times m}$ is a data matrix, c contains measurements, and x is the regression coefficients. One way to apply ADMM to this problem is to rewrite it as

$$\begin{aligned}\min_{u,v} \frac{1}{2}\|Du - c\|_2^2 + \rho_1\|v\|_1 + \frac{\rho_2}{2}\|v\|_2^2 \\ \text{subject to } u - v = 0.\end{aligned} \tag{S49}$$

Then the ADMM steps are

$$\begin{aligned}
u_{k+1} &= \arg\min_u \frac{1}{2}\|Du - c\|_2^2 + \frac{\tau_k}{2}\|0 - u + v_k + \lambda_k/\tau_k\|_2^2 \\
&= \begin{cases} (D^TD + \tau_k I_m)^{-1}(\tau_k v_k + \lambda_k + D^Tc) & \text{if } n \geq m \\ (I_m - D^T(\tau_k I_n + DD^T)^{-1}D) & \\ \qquad \cdot(v_k + \lambda_k/\tau_k + D^Tc/\tau_k) & \text{if } n < m \end{cases} \\
\tilde{u}_{k+1} &= \gamma_k u_{k+1} + (1-\gamma_k)v_k \\
v_{k+1} &= \arg\min_v \rho_1\|v\|_1 + \frac{\rho_2}{2}\|v\|_2^2 + \frac{\tau_k}{2}\|0 - \tilde{u}_{k+1} + v + \frac{\lambda_k}{\tau_k}\|_2^2 \\
&= \frac{1}{(\tau_k + \rho_2)}\text{shrink}(\tau_k u_{k+1} - \lambda_k, \rho_1) \\
\lambda_{k+1} &= \lambda_k + \tau_k(0 - \tilde{u}_{k+1} + v_{k+1}).
\end{aligned}$$

We provide the details of the synthetic data for EN regularized linear regression. The same synthetic data has been used in [18, 4]. Based on three random normal vectors $\nu_a, \nu_b, \nu_c \in \mathbb{R}^{50}$, the data matrix $D = [d_1 \dots d_{40}] \in \mathbb{R}^{50\times 40}$ is defined as

$$d_i = \begin{cases} \nu_a + e_i, & i = 1, \dots, 5, \\ \nu_b + e_i, & i = 6, \dots, 10, \\ \nu_c + e_i, & i = 11, \dots, 15, \\ \nu_i \in N(0,1), & i = 16, \dots, 40, \end{cases} \tag{S50}$$

where e_i are random normal vectors from $N(0,1)$. The problem is to recover the vector

$$x^* = \begin{cases} 3, & i = 1, \dots, 15, \\ 0, & \text{otherwise} \end{cases} \tag{S51}$$

from measurements with noise $\hat{e} \in N(0, 0.1)$

$$c = Dx^* + \hat{e}. \tag{S52}$$

Moreover, vision benchmark dataset MNIST digital images [8] and CIFAR-10 object images [7], learning benchmark datasets used in [3, 18] , and large scale datasets used in [10] are investigated. Typical parameters $\rho_1 = \rho_2 = 1$ are used in all experiments.

2.3. Low rank least squares (LRLS)

ADMM has been applied to solve the low-rank least squares problem [17, 16]

$$\min_X \frac{1}{2}\|DX - C\|_F^2 + \rho_1\|X\|_* + \frac{\rho_2}{2}\|X\|_F^2, \tag{S53}$$

where $\|\cdot\|_*$ denotes the nuclear norm, $\|\cdot\|_F$ denotes the Frobenius norm, $D \in \mathbb{R}^{n\times m}$ is a data matrix, $C \in \mathbb{R}^{n\times d}$ contains measurements, and $X \in \mathbb{R}^{m\times d}$ is the variable matrix. If we rewrite the problem as

$$\begin{aligned}\min_{U,V} \frac{1}{2}\|DU - C\|_F^2 + \rho_1\|V\|_* + \frac{\rho_2}{2}\|V\|_F^2, \\ \text{subject to } U - V = 0,\end{aligned} \tag{S54}$$

then the ADMM steps are

$$\begin{aligned}
U_{k+1} &= \arg\min_U \|DU - C\|_F^2 + \frac{\tau_k}{2}\|V_k - U + \lambda_k/\tau_k\|_F^2 \\
&= (D^TD + \tau I_d)^{-1}(D^TC + \tau_k V_k + \lambda_k) \\
\tilde{U}_{k+1} &= \gamma_k U_{k+1} + (1-\gamma_k)V_k \\
V_{k+1} &= \arg\min_V \rho_1\|V\|_* + \frac{\rho_2}{2}\|V\|_F^2 + \frac{\tau_k}{2}\|V - U_{k+1} + \frac{\lambda_k}{\tau_k}\|_F^2 \\
&= \frac{1}{(\tau + \rho_2)}\text{SVT}(\tau_k U_{k+1} - \lambda_k, \rho_1) \\
\lambda_{k+1} &= \lambda_k + \tau_k(0 - U_{k+1} + V_{k+1}).
\end{aligned}$$

A synthetic problem is constructed using a Gaussian data matrix $D \in \mathbb{R}^{1000\times 200}$ and a true low-rank solution given by $X = [L\ 0] \in \mathbb{R}^{200\times 500}$, with $L = L_1^TL_2$, and $L_1, L_2 \in \mathbb{R}^{20\times 200}$. We choose $C = DW + 0.1G$, where D, G are random Gaussian matrices. The binary classification problems from [9, 13] are tested by formulating low-rank least squares in [16], where each column of X represents a linear exemplar classifier trained with a positive sample and all negative samples. For MNIST digital images [8] and CIFAR-10 object images [7], we use the first five labels as positive and the last five labels as negative to construct the binary classification problem. $\rho_1 = \rho_2 = 1$ is used for all experiments.

2.4. SVM and QP

The dual of the support vector machine (SVM) learning problem is a QP

$$\min_z \frac{1}{2}z^TQz - e^Tz \text{ subject to } c^Tz = 0 \text{ and } 0 \leq z \leq C,$$

where z is the SVM dual variable, Q is the kernel matrix, c is a vector of labels, e is a vector of ones, and $C > 0$ [2]. We also consider the canonical QP

$$\min_x \frac{1}{2}x^TQx + q^Tx \quad \text{subject to } Dx \leq c, \tag{S55}$$

which could be solved by applying ADMM to

$$\begin{aligned} &\min_{u,v} \frac{1}{2}u^TQu + q^Tu + \iota_{\{z:\, z_i \le c\}}(v) \\ &\text{subject to } \; Du - v = 0. \end{aligned} \tag{S56}$$

Here, ι_S is the characteristic function of the set S; $\iota_S(v) = 0$, if $v \in S$, and $\iota_S(v) = \infty$, otherwise. The steps of ADMM for the canonical QP are

$$\begin{aligned} u_{k+1} &= \arg\min_u \frac{1}{2}u^TQu + q^Tu + \frac{\tau_k}{2}\|0 - Du + v_k + \lambda_k/\tau_k\|_2^2 \\ &= (\tau_k D^TD + Q)^{-1}(D^T(\lambda_k + \tau_k v_k) - q) \\ \tilde{u}_{k+1} &= \gamma_k Du_{k+1} + (1-\gamma_k)v_k \\ v_{k+1} &= \arg\min_v \frac{\tau_k}{2}\|0 - \tilde{u}_{k+1} + v + \lambda_k/\tau_k\|_2^2 \text{ subject to } v \le c \\ &= \min\{\tilde{u}_{k+1} - \lambda_k/\tau_k, c\} \\ \lambda_{k+1} &= \lambda_k + \tau_k(0 - Du_{k+1} + v_{k+1}). \end{aligned}$$

The ADMM steps for general QP are applied to dual SVM by absorbing the linear constraint $c^Tu = 0$ of dual SVM into the linear constraint $Du - v = 0$ using an augmented matrix $\hat{D} = [D^T \;\; c]^T$ and variable $\hat{v} = [v^T \;\; 0]^T$.

The synthetic problem for canonical QP is generated following [4], where $Q \in \mathbb{R}^{500\times 500}$ is a random matrix with condition number approximately 4.5×10^5, and 250 random inequality constrains are used. Binary classification problems from [9, 13] are used for dual SVM, with linear kernel and $C = 1$. The features are centered to have zeros mean and unit variance for each dataset.

2.5. Consensus ℓ_1 regularized logistic regression

ADMM has become an important tool for solving distributed problems [1]. Here, we consider the consensus ℓ_1-regularized logistic regression

$$\begin{aligned} &\min_{x_i,z} \sum_{i=1}^N \sum_{j=1}^{n_i} \log(1 + \exp(-c_j D_j^T x_i)) + \rho\|z\|_1 \\ &\text{subject to } \; x_i - z = 0, i = 1, \ldots, N, \end{aligned} \tag{S57}$$

where $x_i \in \mathbb{R}^m$ represents the local variable on the ith distributed node, z is the global variable, n_i is the number of samples in the ith block, $D_j \in \mathbb{R}^m$ is the jth sample, and $c_j \in \{-1, 1\}$ is the corresponding label. Then ADMM steps are

$$\begin{aligned} u_{i,k+1} &= \arg\min_u \sum_{j=1}^{n_i} \log(1 + \exp(-c_j D_j x_i)) \\ &\qquad + \frac{\tau_k}{2}\|0 - u_i + v_k + \lambda_{i,k}/\tau_k\|_2^2 \\ \tilde{u}_{i,k+1} &= \gamma_k u_{i,k+1} + (1-\gamma_k)v_k \\ v_{k+1} &= \arg\min_v \rho\|v\|_1 + \frac{\tau_k}{2}\sum_{i=1}^N \|0 - \tilde{u}_{i,k+1} + v + \lambda_{i,k}/\tau_k\|_2^2 \\ &= \text{shrink}(\bar{u}_{k+1} - \bar{\lambda}_k/\tau_k, \rho/(N\tau)) \\ &\qquad \text{where } \; \bar{u}_{k+1} = \sum_{i=1}^N \tilde{u}_{i,k+1}/N, \;\; \bar{\lambda}_k = \sum_{i=1}^N \lambda_{i,k}/N \\ \lambda_{i,k+1} &= \lambda_{i,k} + \tau_k(0 - u_{i,k+1} + v_{k+1}). \end{aligned}$$

Subproblem (S62) can be solved with BFGS gradient method.

We apply homogeneous coordinates to absorb the bias term of the linear classifier into variable x_i. The synthetic problem with 1000 samples and 25 features is constructed with two 20-dimensional Gaussian distributions and 5 auxiliary features. For MNIST digital images [8] and CIFAR-10 object images [7], we use the first five labels as positive and the last five labels as negative to construct the binary classification problem. Binary classification problems in [9, 13] are also used to test the effectiveness of the proposed method. The features are centered to have zeros mean and unit variance for each dataset, and $\rho = 1$ is used for all experiments. We split the data equally into two blocks and use a loop to simulate the distributed computing of consensus subproblems.

2.6. Unwrapping SVM

ADMM can also be applied to the primal form of SVM [5],

$$\min_x \frac{1}{2}\|x\|_2^2 + C\sum_{j=1}^n \max\{1 - c_j D_j^T x, 0\}, \tag{S58}$$

where $D_j \in \mathbb{R}^m$ is the jth sample, and $c_j \in \{-1, 1\}$ is the corresponding label. Unwrapped SVM solves the equivalent problem,

$$\begin{aligned} &\min_{u,v} \; C\textstyle\sum_{j=1}^n \max\{1 - u_j, \, 0\} + \frac{1}{2}\|v\|_2^2, \\ &\text{subject to } \; -u + A^Tv = 0 \text{ where} A = [c_i D_i]_{i=1\ldots n}. \end{aligned}$$

with ADMM steps,

$$
\begin{aligned}
u_{k+1} &= \arg\min_u C\sum_{j=1}^{n}\max\{1-u,\,0\} + \frac{\tau_k}{2}\|u - A^T v_k + \lambda_k/\tau_k\|_2^2 \\
&= \text{pxhinge}(A^T v_k - \lambda_k/\tau_k, C/\tau_k), \\
\tilde{u}_{k+1} &= \gamma_k u_{k+1} + (1-\gamma_k)A^T v_k \\
v_{k+1} &= \arg\min_v \frac{1}{2}\|v\|_2^2 + \frac{\tau_k}{2}\|\tilde{u}_{k+1} - A^T v + \lambda_k/\tau_k\|_2^2 \\
&= (I + \tau_k AA^T)^{-1}(A(\lambda_k - \tau_k\tilde{u}_{k+1})) \\
\lambda_{k+1} &= \lambda_k + \tau(\tilde{u}_{k+1} - A^T v_{k+1}).
\end{aligned}
$$

We apply homogeneous coordinates and use the synthetic and benchmark datasets introduced for logistic regression in Section 2.5. $C = 0.01$ is used for all experiments.

2.7. Total variation image restoration (TVIR)

Total variation is often used for image restoration [12, 4],

$$\min_x \frac{1}{2}\|x - c\|_2^2 + \rho\|\nabla x\|_1 \tag{S59}$$

where c represent a given noisy image, ∇ is the gradient linear operator, $\|\cdot\|_2, \|\cdot\|_1$ are the ℓ_2, ℓ_1 norm of vectors. The gradient operator ∇ can be represented as $\nabla = (\mathcal{F}^T L_1 \mathcal{F};\, \mathcal{F}^T L_2 \mathcal{F})$, where $\mathcal{F}$ denotes the discrete Fourier transform, $\mathcal{F}^T$ denotes the inverse Fourier transform, L_1, L_2 represent the gradient operator in the Fourier domain for the first and second dimension of an image, respectively. We solve the equivalent problem

$$\min_{u,v} \frac{1}{2}\|u - c\|_2^2 + \rho\|v\|_1 \quad \text{subject to} \quad \nabla u - v = 0. \tag{S60}$$

using the ADMM steps

$$
\begin{aligned}
u_{k+1} &= \arg\min_u \frac{1}{2}\|u - c\|_2^2 + \frac{\tau_k}{2}\|0 - \nabla u + v_k + \lambda_k/\tau_k\|_2^2 \\
&= (I + \tau\nabla^T\nabla)^{-1}(c + \tau_k\nabla^T(v_k + \lambda_k/\tau_k)) \\
\tilde{u}_{k+1} &= \gamma_k\nabla u_{k+1} + (1-\gamma_k)v_k \\
v_{k+1} &= \arg\min_v \rho\|v\|_1 + \frac{\tau_k}{2}\|0 - \tilde{u}_{k+1} + v + \lambda_k/\tau_k\|^2 \\
&= \text{shrink}(\nabla u_{k+1} - \lambda_k/\tau_k, \rho/\tau_k) \\
\lambda_{k+1} &= \lambda_k + \tau_k(0 - \nabla u_{k+1} + v_{k+1}).
\end{aligned}
$$

We test on three grayscale images, "Babara," "cameraman," and "Lena." Images are scaled with pixel values in the 0 to 255 range and then contaminate with Gaussian white noise of standard deviation 20. $\rho = 10$ is used for all experiments.

2.8. Robust PCA

Robust principal component analysis (RPCA) has broad application in videos and face images [14]. RPCA recovers a low-rank matrix and a sparse matrix by solving

$$\min_{Z,E} \|Z\|_* + \rho\|E\|_1 \;\text{ subject to }\; Z + E = C, \tag{S61}$$

where nuclear norm $\|\cdot\|_*$ is used for the low rank matrix Z, $\|\cdot\|_1$ is used for the sparse error E. ADMM is applied to RPCA by steps,

$$
\begin{aligned}
Z_{k+1} =& \|Z\|_* + \frac{\tau_k}{2}\|C - E_k - Z + \lambda_k/\tau\|_F^2 && \text{(S62)} \\
=& \text{SVT}(C - E_k + \lambda_k/\tau, 1/\tau_k) && \text{(S63)} \\
\tilde{Z}_{k+1} =& \gamma_k Z_{k+1} + (1-\gamma_k)(C - E_k) && \text{(S64)} \\
E_{k+1} =& \rho\|E\|_1 + \frac{\tau_k}{2}\|C - E - \tilde{Z}_{k+1} + \lambda_k/\tau\|_F^2 && \text{(S65)} \\
=& \text{shrink}(C - Z_{k+1} + \lambda_k/\tau_k, \rho/\tau_k) && \text{(S66)} \\
\lambda_{k+1} =& \lambda_k + \tau_k(C - E_{k+1} - Z_{k+1}). && \text{(S67)}
\end{aligned}
$$

Extended Yale B Face dataset is used by applying RPCA for each individual human, and the measurement matrix C is constructed by vectorizing each image as a row of the matrix. $\rho = 0.05$ is selected based on the visual performance of robust PCA decomposition for faces.

3. More experimental results

A complete convergence table for all algorithms and all applications is provided in Table 1. We implement baselinses vanilla ADMM and relaxed ADMM following [1, 4], residual balancing following [1, 6], and adaptive ADMM following [15]. The proposed ARADMM performs best in all test cases.

Fig. 1 shows the convergence curve (relative residual with respect to iterations) for the synthetic problem of EN regularized linear regression and QP. Consider the stop criterion (main text equation (7))

$$
\begin{aligned}
\|r_k\| &\le \epsilon^{tol}\max\{\|Au_k\|, \|Bv_k\|, \|b\|_2\} \\
\|d_k\| &\le \epsilon^{tol}\|A^T\lambda_k\|,
\end{aligned} \tag{S68}
$$

the relative residual is defined as

$$r^{rel} = \max\{\frac{\|r_k\|}{\max\{\|Au_k\|, \|Bv_k\|, \|b\|_2\}}, \frac{\|d_k\|}{\|A^T\lambda_k\|}\} \tag{S69}$$

Fig. 1 clearly shows the proposed ARADMM converges fastest.

TV image restoration successfully recovers the image from noisy observation (Fig. 2). Robust PCA decomposes the original faces into intrinsic images (low rank) and shadings (sparse), as shown in Fig. 3.

The sensitivity to τ_0, γ_0 for the synthetic QP problem is provided in Fig. 4. The analysis is similar to EN regularized linear regression in the main text. Notice in Fig. 4 (right), the curves of RB and relaxed ADMM overlap since the RB method never adjusts τ when $\gamma \in [1.3,\, 2]$.

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. and Trends in Mach. Learning*, 3:1–122, 2011. 5, 6

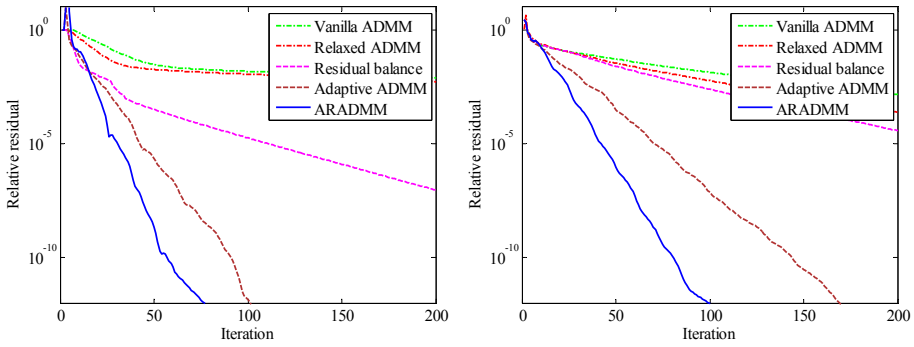

Figure 1. Convergence curves show the relative residual vs iteration number for the synthetic problem of (left) EN regularized linear regression and (right) canonical QP.

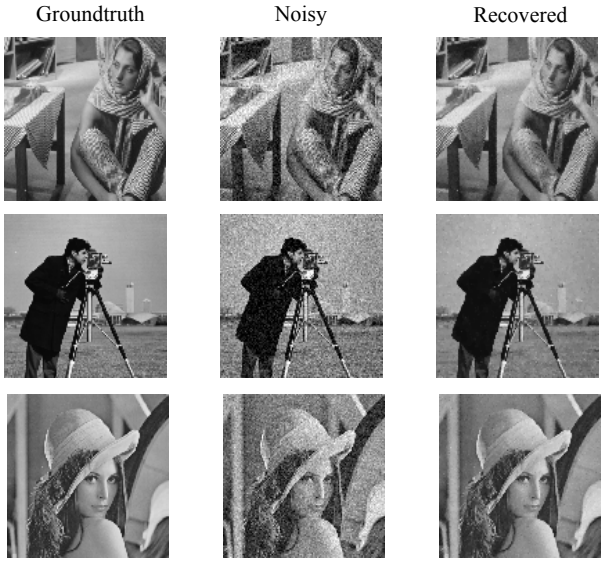

Figure 2. The groundtruth image (left), noisy image (middle) and recovered image by TVIR and ARADMM (right) are shown.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 4

[3] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. 4

[4] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014. 4, 5, 6

[5] T. Goldstein, G. Taylor, K. Barabin, and K. Sayre. Unwrapping ADMM: efficient distributed computing via transpose reduction. In *AISTATS*, 2016. 5

[6] B. He, H. Yang, and S. Wang. Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Jour. Optim. Theory and Appl.*, 106(2):337–356, 2000. 6

[7] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009. 4, 5

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 4, 5

[9] S.-I. Lee, H. Lee, P. Abbeel, and A. Ng. Efficient L1 regularized logistic regression. In *AAAI*, volume 21, page 401, 2006. 4, 5

[10] J. Liu, J. Chen, and J. Ye. Large-scale sparse logistic regression. In *ACM SIGKDD*, pages 547–556, 2009. 4

[11] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970. 3

[12] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. 6

[13] M. Schmidt, G. Fung, and R. Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *ECML*, pages 286–297. Springer, 2007. 4, 5

[14] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in neural information processing systems*, pages 2080–2088, 2009. 6

[15] Z. Xu, M. A. Figueiredo, and T. Goldstein. Adaptive ADMM with spectral penalty parameter selection. *AISTATS*, 2017. 6

[16] Z. Xu, X. Li, K. Yang, and T. Goldstein. Exploiting low-rank structure for discriminative sub-categorization. In *BMVC, Swansea, UK, September 7-10, 2015*, 2015. 4

[17] J. Yang and X. Yuan. Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation*, 82(281):301–329, 2013. 4

[18] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. 4

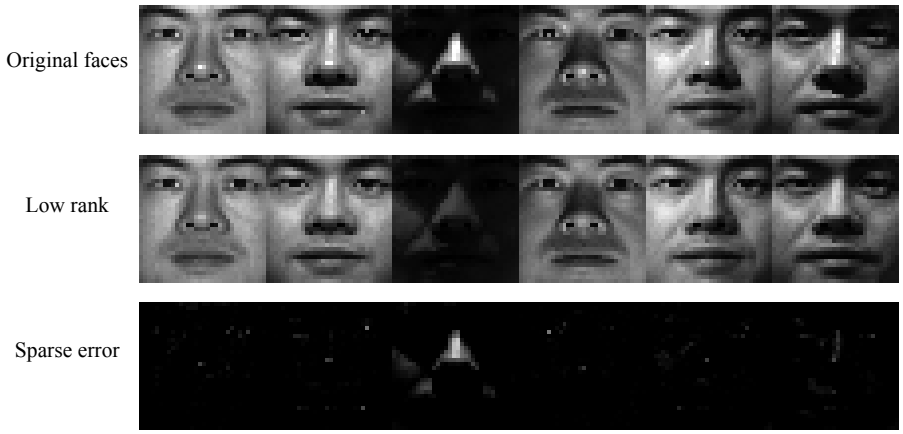

Figure 3. Sample face images of human subject 2 and recovered low rank faces and sparse errors by RPCA and ARADMM. RPCA decomposes the original faces into intrinsic images (low rank) and shadings (sparse).

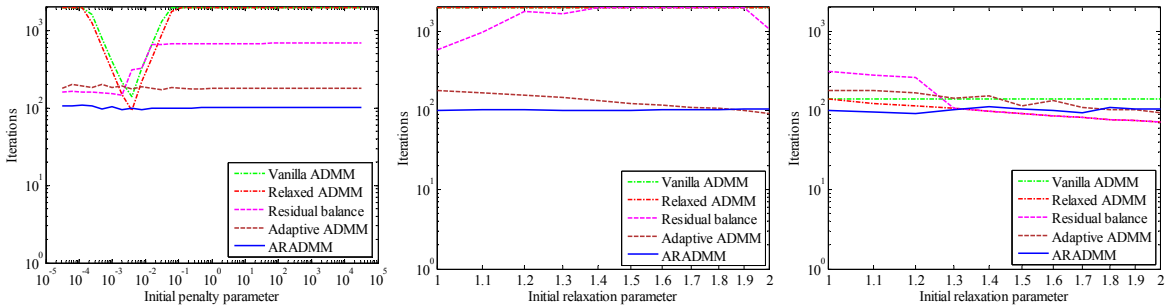

Figure 4. Sensitivity of convergence speed for the synthetic QP problem. (left) Sensitivity to the initial penalty τ_0; (middle) Sensitivity to relaxation γ_0; (right) Sensitivity to relaxation γ_0 when optimal τ_0 is selected by grid search.

Table 1. Iterations (and runtime in seconds) for various applications. Absence of convergence after n iterations is indicated as $n+$.

Application	Dataset	#samples × #features[1]	Vanilla ADMM	Relaxed ADMM	Residual balance	Adaptive ADMM	**Proposed ARADMM**
Elastic net regression	Synthetic	50 × 40	2000+(.642)	2000+(.660)	424(.144)	102(.051)	**70(.026)**
	Boston	506 × 13	2000+(.565)	1533(.429)	71(.024)	29(.011)	**20(.007)**
	Diabetes	768 × 8	762(.199)	503(.194)	45(.016)	21(.009)	**14(.006)**
	Prostate	97 × 8	715(.152)	471(.104)	46(.011)	32(.009)	**17(.005)**
	Servo	130 × 4	309(.110)	202(.039)	54(.012)	26(.007)	**16(.004)**
	MNIST	60000 × 784	1225(29.4)	816(19.9)	94(2.28)	41(.943)	**21(.549)**
	CIFAR10	10000 × 3072	2000+(690)	2000+(697)	556(193)	2000+(669)	**94(31.7)**
	News20	19996 × 1355191	2000+(1.21e4)	2000+(9.16e3)	227(914)	104(391)	**71(287)**
	Rcv1	20242 × 47236	2000+(1.20e3)	1823(802)	196(79.1)	104(35.7)	**64(26.0)**
	Realsim	72309 × 20958	2000+(4.26e3)	2000+(4.33e3)	341(355)	152(125)	**107(88.2)**
Low rank least squares	Synthetic	1000 × 200	2000+(118)	2000+(116)	268(15.1)	26(1.55)	**18(1.04)**
	AC	690 × 14	2000+(3.15)	2000+(3.17)	333(.552)	56(.112)	**44(.076)**
	AH	270 × 13	2000+(1.69)	2000+(1.68)	267(.234)	57(.056)	**40(.042)**
	German	1000 × 24	2000+(4.72)	2000+(4.72)	642(1.52)	130(.334)	**52(.125)**
	Hepatitis	155 × 19	2000+(1.54)	2000+(1.45)	481(.324)	127(.091)	**73(.059)**
	Spect	80 × 22	2000+(1.72)	2000+(1.64)	227(.194)	182(.168)	**76(.072)**
	Spectf	80 × 44	2000+(2.70)	2000+(2.74)	336(.455)	162(.236)	**105(.150)**
	WBC	683 × 10	2000+(3.13)	2000+(3.10)	689(1.11)	61(.107)	**35(.055)**
	MNIST	60000 × 784	200+(1.86e3)	200+(2.08e3)	200+(3.29e3)	200+(3.46e3)	**38(658)**
	CIFAR10	10000 × 3072	200+(7.24e3)	200+(1.33e4)	53(1.60e3)	8(208)	**6(156)**
QP and dual SVM	Synthetic	250 × 500	1224(11.5)	823(7.49)	626(5.93)	170(1.57)	**100(.914)**
	AC	690 × 14	539(7.01)	364(4.63)	78(1.02)	111(1.41)	**68(1.02)**
	AH	270 × 13	347(.764)	266(.585)	103(.227)	92(.194)	**63(.138)**
	German	1000 × 24	2000+(58.8)	2000+(61.8)	1592(45.0)	1393(38.9)	**1238(34.9)**
	Hepatitis	155 × 19	2000+(1.54)	2000+(1.46)	1356(.986)	2000+(1.36)	**774(.486)**
	Spect	80 × 22	2000+(.820)	2000+(.807)	231(.100)	2000+(.873)	**391(.176)**
	Spectf	80 × 44	2000+(.846)	2000+(.777)	169(.070)	175(.086)	**53(.026)**
	WBC	683 × 10	1447(18.1)	972(12.1)	194(2.43)	2000+(25.1)	**102(1.32)**
Consensus logistic regression	Synthetic	1000 × 25	590(9.93)	391(6.97)	70(1.23)	35(.609)	**20(.355)**
	AC	690 × 14	2000+(35.2)	2000+(30.2)	143(2.35)	56(.924)	**41(.727)**
	AH	270 × 13	2000(18.1)	1490(13.9)	79(.930)	32(.391)	**21(.288)**
	German	1000 × 24	2000+(34.3)	2000+(66.6)	151(2.60)	35(.691)	**26(.580)**
	Hepatitis	155 × 19	2000+(30.1)	2000+(25.6)	135(1.86)	71(.857)	**41(.518)**
	Spect	80 × 22	1543(18.0)	1027(12.6)	105(1.13)	49(.481)	**42(.431)**
	Spectf	80 × 44	1005(20.1)	667(14.4)	117(1.98)	145(1.63)	**85(1.07)**
	WBC	683 × 10	934(9.47)	621(6.47)	69(.865)	36(.453)	**23(.320)**
	MNIST	60000 × 784	200+(2.99e3)	200+(3.47e3)	200+(1.37e3)	49(536)	**28(333)**
	CIFAR10	10000 × 3072	200+(593)	200+(2.08e3)	200+(1.54e3)	131(165)	**19(33.7)**
Unwrapping SVM	Synthetic	1000 × 25	2000+(1.13)	1418(.844)	2000+(1.16)	355(.229)	**147(.094)**
	AC	690 × 14	739(.973)	484(.552)	1893(2.17)	723(.861)	**334(.375)**
	AH	270 × 13	286(.146)	230(.131)	765(.417)	453(.263)	**114(.069)**
	German	1000 × 24	753(1.88)	560(1.37)	2000+(4.98)	572(1.44)	**213(.545)**
	Hepatitis	155 × 19	257(.156)	235(.086)	411(.149)	214(.075)	**164(.061)**
	Spect	80 × 22	195(.064)	144(.051)	195(.052)	117(.038)	**107(.037)**
	Spectf	80 × 44	567(.203)	367(.112)	567(.185)	207(.068)	**149(.052)**
	WBC	683 × 10	475(.380)	370(.316)	677(.501)	113(.099)	**74(.058)**
	MNIST	60000 × 784	128(130)	118(111)	163(153)	200+(217)	**67(71.0)**
	CIFAR10	10000 × 3072	200+(512)	200+(532)	200+(516)	89(285)	**57(143)**
Image restoration	Barbara	512 × 512	262(35.0)	175(23.6)	74(10.0)	59(8.67)	**38(5.57)**
	Cameraman	256 × 256	311(8.96)	208(5.89)	82(2.29)	88(2.76)	**35(1.08)**
	Lena	512 × 512	347(46.3)	232(31.3)	94(12.5)	68(9.70)	**39(5.58)**
Robust PCA	FaceSet1	64 × 1024	2000+(41.1)	1507(30.3)	560(11.1)	561(11.9)	**267(5.65)**
	FaceSet2	64 × 1024	2000+(41.1)	2000+(41.4)	263(5.54)	388(9.00)	**188(4.02)**
	FaceSet3	64 × 1024	2000+(39.4)	1843(36.3)	375(7.44)	473(9.89)	**299(6.27)**

AC: Australian Credit; AH: Australian Heart; WBC: Wisconsin Breast Cancer.

[1] #constrains × #unknowns for general QP; width × height for image restoration.