A. Appendix A: Convergence Analysis

A.1. Proof of Theorem 4.2

Recall primal, dual and Lagrangian forms:

$$P(\boldsymbol{x}) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\phi_i(\langle A_i, \boldsymbol{x}\rangle) + g(\boldsymbol{x}), \tag{20}$$

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \overset{\text{def}}{=} g(\boldsymbol{x}) + \frac{1}{n}\boldsymbol{y}^T A\boldsymbol{x} - \frac{1}{n}\sum_{i=1}^{n}\phi_i^*(y_i) \tag{21}$$

$$D(\boldsymbol{y}) \overset{\text{def}}{=} \min_{x} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \equiv \mathcal{L}(\bar{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \tag{22}$$

where $\bar{\boldsymbol{x}}(\boldsymbol{y}) : \mathbb{R}^n \to \mathbb{R}^d$ is the optimal primal variable with respect to some $\boldsymbol{y}$, namely,

$$\bar{\boldsymbol{x}}(\boldsymbol{y}) = \arg\min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$$

For simplicity, we will use $\bar{\boldsymbol{x}}^{(t)} \overset{\text{def}}{=} \bar{\boldsymbol{x}}(\boldsymbol{y}^{(t)})$ throughout this paper. Similarly, we also use $\bar{\boldsymbol{y}}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^n$ to be the optimal dual variable with respect to some $\boldsymbol{x}$.

Recall with the choice of regularizer of our model, $g(\boldsymbol{x}) = h(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_1$, where $h(\boldsymbol{x}) = \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$ satisfies μ-strong convexity, μ-smooth and separable. The conjugate of loss function(e.g. smooth hinge loss used in our experiments): ϕ^* is γ-strongly convex.

Recall the primal gap defined as $\Delta_p^{(t)} \overset{\text{def}}{=} \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^{(t)}) - D(\boldsymbol{y}^{(t)})$, and dual gap $\Delta_d^{(t)} \overset{\text{def}}{=} D^* - D(\boldsymbol{y}^{(t)})$. In the proof, we will connect the objective change in primal/dual update with the primal/dual gap and show how the sub-optimality: $\Delta^{(t)} = \Delta_p^{(t)} + \Delta_d^{(t)}$ enjoys linear convergence.

Lemma A.1. *(Primal Progress):*

$$\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^{(t)}) \geq \frac{1}{\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|_0 - 1}\Delta_p^{(t)}$$

Proof. This lemma is a direct result by our greedy update rule of our primal variables.

$$\begin{aligned}
& \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}(\bar{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^{(t)}) \\
= & \sum_i \big\{\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}((\bar{x}_i^{(t)} - x_i^{(t)})\boldsymbol{e}_i + \boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\big\} \\
= & \sum_{i \in \text{supp}(\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)})} \big\{\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}((\bar{x}_i^{(t)} - x_i^{(t)})\boldsymbol{e}_i + \boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\big\} \\
\leq & \|\bar{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t)}\|_0 \times \\
& \max_i \big\{\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}((\bar{x}_i^{(t)} - x_i^{(t)})\boldsymbol{e}_i + \boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})\big\} \\
= & \|\bar{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t)}\|_0 \big(\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^{(t)})\big)
\end{aligned}$$

And by adding $\mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^{(t)}) - \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})$ to both sides we finishes the proof. □

Recall $i^{(t)}$ is the selected coordinate to update in dual variable $\boldsymbol{y}^{(t)}$.

Lemma A.2. *(Primal-Dual Progress).*

$$\begin{aligned}
& \Delta_d^{(t)} - \Delta_d^{(t-1)} + \Delta_p^{(t)} - \Delta_p^{(t-1)} \\
\leq & \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) \\
& + \eta(\frac{1}{n}\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\rangle_{i^{(t)}})^2 - \eta(\frac{1}{n}\langle A_{i^{(t)}}, \bar{\boldsymbol{x}}^{(t)}\rangle - \boldsymbol{g})^2
\end{aligned}$$

, where $g \in \frac{1}{n}\partial\phi_{i^{(t)}}^*(\boldsymbol{y}^{(t)})$.

Our goal is to prove that $\Delta_d^{(t)} - \Delta_d^{(t-1)} + \Delta_p^{(t)} - \Delta_p^{(t-1)} \leq -\delta\Delta_p^{(t)} - \delta\Delta_d^{(t)}$ to show linear convergence in sub-optimality. Since $\mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) \leq -\frac{1}{\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|_0}\Delta_p^{(t)}$, this lemma is the middle step that connects to the primal part, and the remaining part represents the dual progress and will be analyzed later.

Proof. The primal and dual gap comes from both primal and dual progresses:

$$\underbrace{\Delta_d^{(t)} - \Delta_d^{(t-1)}}_{\text{dual progress}} + \underbrace{\Delta_p^{(t)} - \Delta_p^{(t-1)}}_{\text{primal progress}}$$

- Dual progress:

  By Danskins' theorem, $-D(\boldsymbol{y})$ is γ-strongly convex. Therefore for any $g \in \partial\phi_{i^{(t)}}^*(\boldsymbol{y}^{(t)})$, we have,

  $$\begin{aligned}
  & \Delta_d^{(t)} - \Delta_d^{(t-1)} = \big(-D(\boldsymbol{y}^{(t)})\big) - \big(-D(\boldsymbol{y}^{(t-1)})\big) \\
  & \leq -(\frac{1}{n}\langle A_{i^{(t)}}, \bar{\boldsymbol{x}}^{(t)}\rangle - g)(y_{i^{(t)}}^{(t)} - y_{i^{(t)}}^{(t-1)}) \\
  & -\frac{\gamma}{2}(y_{i^{(t)}}^{(t)} - y_{i^{(t)}}^{(t-1)})^2
  \end{aligned} \tag{23}$$

- Primal progress:

  Similarly we get,

  $$\begin{aligned}
  & \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^{(t-1)}) \\
  \leq & (\frac{1}{n}\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)}\rangle - g)(y_{i^{(t)}}^{(t)} - y_{i^{(t)}}^{(t-1)}) \\
  & + \frac{\gamma}{2}(y_{i^{(t)}}^{(t-1)} - y_{i^{(t)}}^{(t)})^2
  \end{aligned} \tag{24}$$

  Therefore,

  $$\begin{aligned}
  & \Delta_p^{(t)} - \Delta_p^{(t-1)} \\
  = & \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^{(t-1)}) - (D(\boldsymbol{y}^{(t)}) - D(\boldsymbol{y}^{(t-1)})) \\
  = & \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) + \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^{(t-1)}) \\
  & -(D(\boldsymbol{y}^{(t)}) - D(\boldsymbol{y}^{(t-1)})) \\
  \leq & \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) \\
  & + \frac{1}{n}\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\rangle(y_{i^{(t)}}^{(t)} - y_{i^{(t)}}^{(t-1)})
  \end{aligned}$$

  Here the last inequality comes from inequalities (24) and (23).

  Meanwhile, with the update rule of dual variable:

  $y_{i^{(t)}}^{(t)} \leftarrow \arg\max_\beta \frac{1}{n}\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)}\rangle\beta - \phi_{i^{(t)}}^*(\beta) - \frac{1}{2\eta}(\beta - y_{i^{(t)}}^{(t)})^2$

  Therefore $\exists\ g \in \partial\phi_{i^{(t)}}^*(\boldsymbol{y}^{(t)})$ such that $y_{i^{(t)}}^{(t)} - y_{i^{(t)}}^{(t-1)} = \eta(\frac{1}{n}(\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)}\rangle - g)$. Therefore:

  $$\begin{aligned}
  (23) \quad = & -(\frac{1}{n}\langle A_{i^{(t)}}, \bar{\boldsymbol{x}}^{(t)}\rangle - g)(y_{i^{(t)}}^{(t)} - y_{i^{(t)}}^{(t-1)}) \\
  & -\frac{\gamma}{2}(y_{i^{(t-1)}}^{(t)} - y_{i^{(t)}}^{(t)})^2 \\
  = & \langle\frac{1}{n}A_{i^{(t)}}, \bar{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t)}\rangle(y_{i^{(t)}}^{(t-1)} - y_{i^{(t)}}^{(t)}) \\
  & -(\frac{1}{\eta} + \frac{\gamma}{2})(y_{i^{(t)}}^{(t)} - y_{i^{(t)}}^{(t-1)})^2
  \end{aligned} \tag{25}$$

- Summing together we have:

$$
\begin{aligned}
& \Delta_d^{(t)} - \Delta_d^{(t-1)} + \Delta_p^{(t)} - \Delta_p^{(t-1)} \\
\leq\ & \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) \\
& + \frac{2}{n}\langle A_{i^{(t)}}, \bar{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t)}\rangle (y_{i^{(t)}}^{(t)} - y_{i^{(t)}}^{(t-1)}) \\
& - (\frac{1}{\eta} + \frac{\gamma}{2})(y_{i^{(t)}}^{(t)} - y_{i^{(t)}}^{(t-1)})^2 \\
=\ & \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) \\
& + \frac{2\eta}{n}\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\rangle (\frac{1}{n}\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)}\rangle - g) \\
& - \eta^2 (\frac{1}{\eta} + \frac{\gamma}{2})(\langle \frac{1}{n} A_{i^{(t)}}, \boldsymbol{x}^{(t)}\rangle - g)^2 \\
\leq\ & \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) \\
& + \eta (\frac{1}{n}\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\rangle)^2 - \eta (\frac{1}{n}\langle A_{i^{(t)}}, \bar{\boldsymbol{x}}^{(t)}\rangle - g)^2
\end{aligned}
$$

□

Afterwards, we upper bound the dual progress $(\frac{1}{n}\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\rangle)^2 - (\frac{1}{n}\langle A_{i^{(t)}}, \bar{\boldsymbol{x}}^{(t)}\rangle - g)^2$ by dual gap $\Delta_d^{(t)}$:

Lemma A.3. *(Dual Progress).*

$$
\begin{aligned}
& (\frac{1}{n}\langle A_{i^{(t)}}, x^{(t)} - \bar{x}^{(t)}\rangle)^2 - (\frac{1}{n}\langle A_{i^{(t)}}, \bar{\boldsymbol{x}}^{(t)}\rangle - g)^2 \\
\leq\ & -\frac{\gamma}{2n}\Delta_d^{(t)} + \frac{5R^2}{2n^2}\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|^2 \qquad (26)
\end{aligned}
$$

, where $g \in \frac{1}{n}\partial\phi_{i^{(t)}}^*(y_{i^{(t)}}^{(t)})$.

Proof. For simplicity, we denote $\phi^*(\boldsymbol{y}) = \frac{1}{n}\sum_i \phi_i^*(y_i)$. To begin with,

$$
\begin{aligned}
\Delta_d^{(t)} = D^* - D(\boldsymbol{y}) &\leq \frac{2}{\gamma}\|\frac{1}{n}A\bar{\boldsymbol{x}}^{(t)} - \partial\phi^*(\boldsymbol{y}^{(t)})\|^2 \\
&\leq \frac{2n}{\gamma}\|\frac{1}{n}A\bar{\boldsymbol{x}}^{(t)} - \partial\phi^*(\boldsymbol{y}^{(t)})\|_\infty^2
\end{aligned}
$$

In our algorithm, the greedy choice of $i^{(t)}$ makes sure $\|\frac{1}{n}A\boldsymbol{x}^{(t)} - \partial\phi^*(\boldsymbol{y}^{(t)})\|_{i^{(t)}} = \|\frac{1}{n}A\boldsymbol{x}^{(t)} - \partial\phi^*(\boldsymbol{y}^{(t)})\|_\infty$. However, here we need the relation between $\|\frac{1}{n}A\bar{\boldsymbol{x}}^{(t)} - \partial\phi^*(\boldsymbol{y}^{(t)})\|_{i^{(t)}}$ and $\|\frac{1}{n}A\bar{\boldsymbol{x}}^{(t)} - \partial\phi^*(\boldsymbol{y}^{(t)})\|_\infty$ (assumed to be reached at coordinate i^*). We bridge their gap by $\delta \overset{\text{def}}{=} \frac{1}{n}A(\bar{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t)})$. Since

$$
\begin{aligned}
& -(\frac{1}{n}\langle A_{i^{(t)}}, \bar{\boldsymbol{x}}^{(t)}\rangle - \frac{1}{n}(\phi_{i^{(t)}}^*)'(y_{i^{(t)}}^{(t)}))^2 \\
=\ & -(\frac{1}{n}\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)}\rangle - \frac{1}{n}(\phi_{i^{(t)}}^*)'(y_{i^{(t)}}^{(t)}) + \delta_{i^{(t)}})^2 \\
\leq\ & -\frac{1}{2n^2}\left(\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)}\rangle - (\phi_{i^{(t)}}^*)'(y_{i^{(t)}}^{(t)})\right)^2 + \delta_{i^{(t)}}^2 \\
=\ & -\frac{1}{2}\|\frac{1}{n}A\boldsymbol{x}^{(t)} - \partial\phi^*(\boldsymbol{y}^{(t)})\|_\infty^2 + \delta_{i^{(t)}}^2 \\
\leq\ & -\frac{1}{2}(\frac{1}{n}\langle A_{i^*}, \boldsymbol{x}^{(t)}\rangle - \frac{1}{n}(\phi_{i^*}^*)'(y_{i^*}^{(t)}))^2 + \|\delta\|_\infty^2 \\
=\ & -\frac{1}{2}(\frac{1}{n}\langle A_{i^*}, \bar{\boldsymbol{x}}^{(t)}\rangle - \frac{1}{n}(\phi_{i^*}^*)'(y_{i^*}^{(t)}) - \delta_{i^*})^2 + \|\delta\|_\infty^2 \\
\leq\ & -\frac{1}{4}(\frac{1}{n}\langle A_{i^*}, \bar{\boldsymbol{x}}^{(t)}\rangle - \frac{1}{n}(\phi_{i^*}^*)'(y_{i^*}^{(t)}))^2 + \frac{3}{2}\|\delta\|_\infty^2 \\
=\ & -\frac{1}{4}\|\frac{1}{n}A\bar{\boldsymbol{x}}^{(t)} - \partial\phi^*(\boldsymbol{y}^{(t)})\|_\infty^2 + \frac{3}{2}\|\delta\|_\infty^2 \\
\leq\ & -\frac{\gamma}{2n}\Delta_d^k + \frac{3}{2}\|\delta\|_\infty^2
\end{aligned}
$$

The first inequality follows $-(a+b)^2 = -a^2 - b^2 - 2ab \leq -a^2 - b^2 + \frac{1}{2}a^2 + 2b^2 = -\frac{1}{2}a^2 + b^2$, and replace a by $\frac{1}{n}\langle A_{i^{(t)}}, \boldsymbol{x}^{(t)}\rangle - \frac{1}{n}(\phi_{i^{(t)}}^*)'(y_{i^{(t)}}^{(t)})$ and $b \overset{\text{def}}{=} \delta_{i^{(t)}}$. And similarly for the third inequality.

Meanwhile, since $\|A(\bar{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t)})\|_\infty \leq R\|\bar{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t)}\|$, together we get Lemma A.3. □

Now we have established the connection between the primal and dual progress (change in primal/dual gap) with primal and dual gap, and the only redundant part is $\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|$, but since $\frac{\mu}{2}\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\| \leq \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}(\bar{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^{(t)})$, which could be absorbed in the primal gap. Therefore, back to the main inequality (26):

Proof of Theorem 4.2.

$$
\begin{aligned}
& \Delta_d^{(t)} - \Delta_d^{(t-1)} + \Delta_p^{(t)} - \Delta_p^{(t-1)} \\
\overset{\text{Lemma A.2}}{\leq}\ & \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) \\
& + \langle \frac{1}{n}A\boldsymbol{x}^{(t)} - \nabla\varphi(\boldsymbol{y}^{(t)}), \boldsymbol{y}^{(t)} - \boldsymbol{y}^{(t-1)}\rangle \\
& - 2\langle \frac{1}{n}A\bar{\boldsymbol{x}}^{(t)} - \nabla\varphi(\boldsymbol{y}^{(t)}), \boldsymbol{y}^{(t)} - \boldsymbol{y}^{(t-1)}\rangle \\
\overset{\text{Lemma A.3}}{\leq}\ & \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) - \frac{\eta\gamma}{2n}\Delta_d^{(t)} \\
& + \frac{5\eta R^2}{2n^2}\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|^2 \\
\leq\ & \mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t) - \frac{\eta\gamma}{2n}\Delta_d^{(t)} \\
& + \frac{5\eta R^2}{\mu n^2}\big(\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}(\bar{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^{(t)})\big) \\
=\ & (1 - \frac{5\eta R^2}{\mu n^2})\big(\mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\boldsymbol{x}^t, \boldsymbol{y}^t)\big) - \frac{\eta\gamma}{2n}\Delta_d^{(t)} \\
& + \frac{5\eta R^2}{\mu n^2}\big(\mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\bar{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^t)\big) \\
\overset{\text{Lemma A.1}}{\leq}\ & -(1 - \frac{5\eta R^2}{\mu n^2})\frac{1}{\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|_0 - 1}\Delta_p^{(t)} - \frac{\eta\gamma}{2n}\Delta_d^{(t)} + \\
& \frac{5\eta R^2}{\mu n^2}\big(\mathcal{L}(\boldsymbol{x}^{(t+1)}, \boldsymbol{y}^t) - \mathcal{L}(\bar{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^t)\big) \\
=\ & -\big((1 - \frac{5\eta R^2}{\mu n^2})\frac{1}{\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|_0 - 1} - \frac{5\eta R^2}{\mu n^2}\big)\Delta_p^{(t)} \\
& - \frac{\eta\gamma}{2n}\Delta_d^{(t)}
\end{aligned}
$$

Therefore, we have

$$
\frac{\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|_0}{\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|_0 - 1}(1 - \frac{5\eta R^2}{\mu n^2})\Delta_p^{(t)} + (1 + \frac{\eta\gamma}{2n})\Delta_d^{(t)} \leq \Delta_d^{(t-1)} + \Delta_p^{(t-1)}
$$

i.e. linear convergence. Notice when

$$
\eta^{(t)} \leq \frac{2n^2\mu}{(10R^2 + n\gamma\mu)\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|_0} \qquad (27)
$$

$$
\Delta^{(t)} \leq \frac{1}{1 + \frac{\eta^{(t)}\gamma}{2n}}\Delta^{(t-1)}
$$

Specifically, when inequality holds for (27), and suppose $\|\boldsymbol{x}^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|_0 \leq s$, then it requires $\mathcal{O}(s(\frac{\kappa}{n} + 1)\log\frac{1}{\epsilon})$ iterations to achieve ϵ primal and dual sub-optimality, where $\kappa = \frac{R^2}{\mu\gamma}$. □

B Appendix B: Additional Experimental Results

Finally, we show result for $\lambda = 0.01, 0.1$, and $\mu = 0.01, 0.1, 1$. Here are some comments for results under different parameters.

The winning margin of DGPD is larger on data sets of dense feature matrix than that of sparse feature matrix. One reason for this is, for data of sparse feature matrix, features of higher frequency are more likely to be active than those of lower frequency, and therefore, the feature sub-matrix corresponding to the *active primal variables* are often denser than submatrix matrix corresponding to the inactive ones. This results in a less overall speedup.

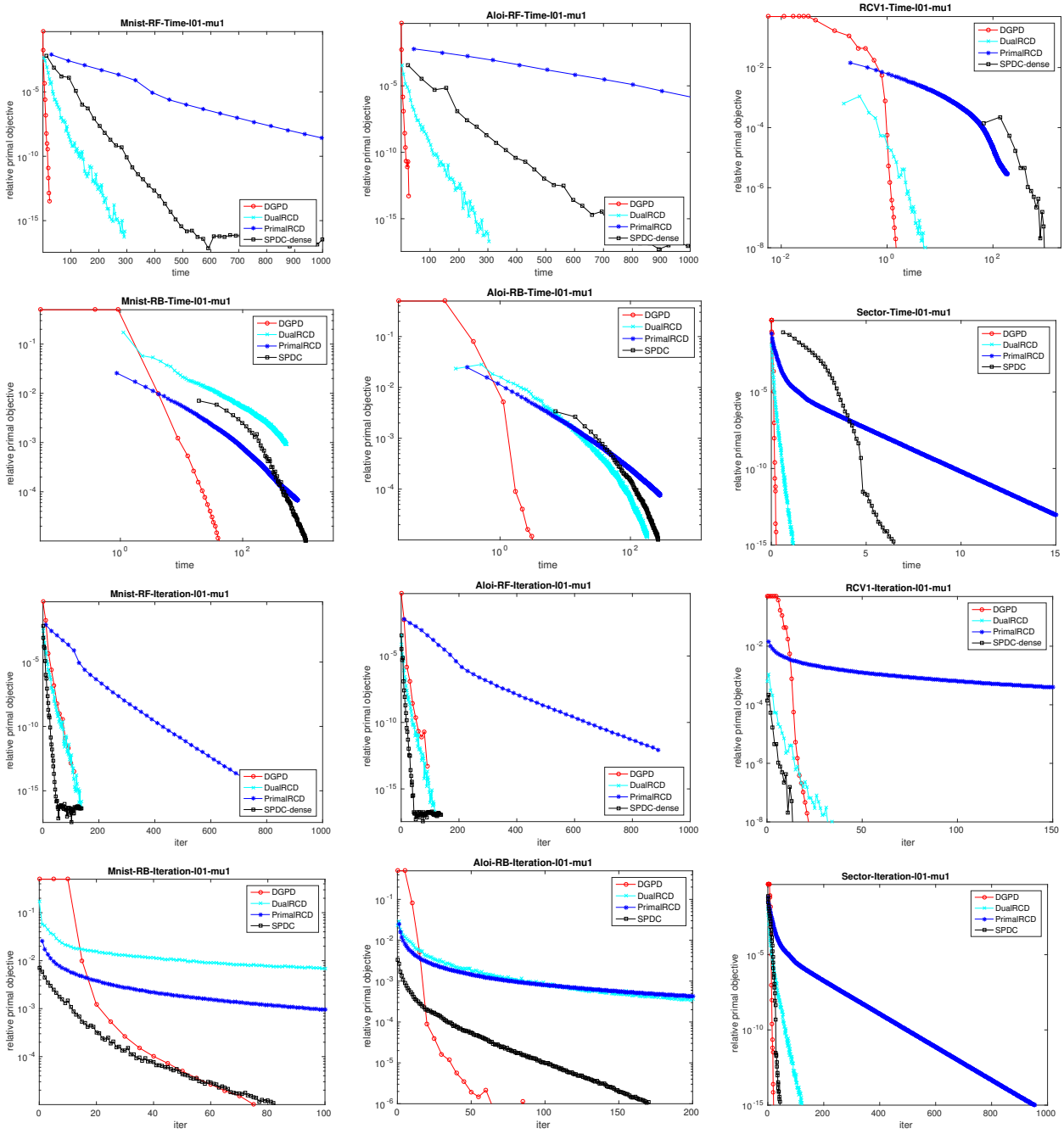

Figure 2. Relative Objective versus Time (the upper 2 rows) and versus # iterations (the lower 2 rows) for $\lambda = 0.1$, $\mu = 1$.

We also observe that in order to achieve the best performance of DGPD, both primal and dual sparsity must hold, and the sparsity is partially controled by the L1/L2 penalty. In particular, when the L1 penalty has too much weight, the primal iterate would become too sparse to yield a reasonable prediction accuracy, which then results in a particularly dense dual iterate due to its non-zero loss on most of the samples. Another example is, when the L2 penalty becomes too large, the classifier would tend to mis-classify many examples in order to gain a large margin, which results in dense dual iterates.

However, in practice such hyperparameter settings are less likely to be chosen due to its inferior prediction performance.

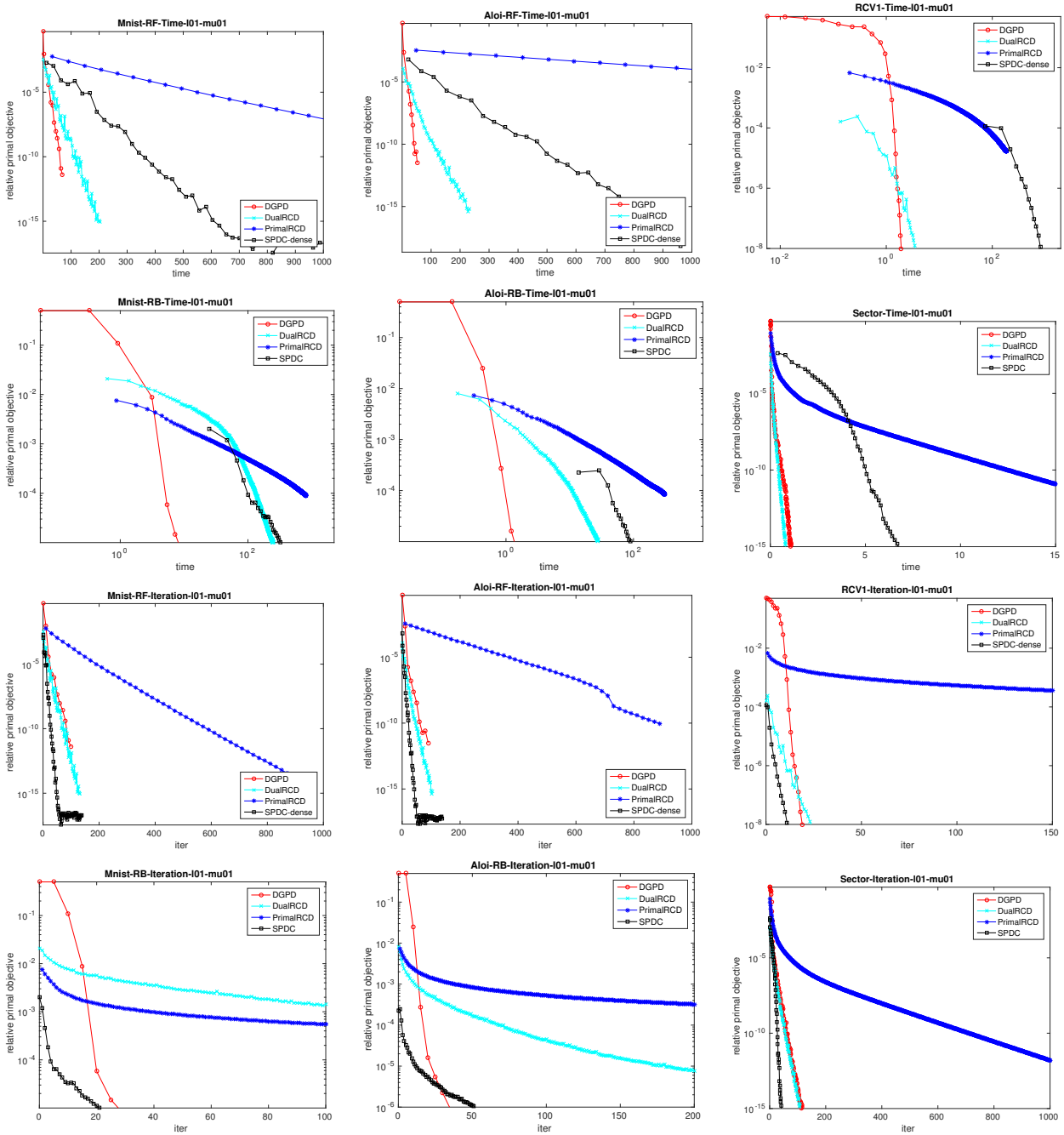

Figure 3. Relative Objective versus Time (the upper 2 rows) and versus # iterations (the lower 2 rows) for $\lambda = 0.1$, $\mu = 0.1$.

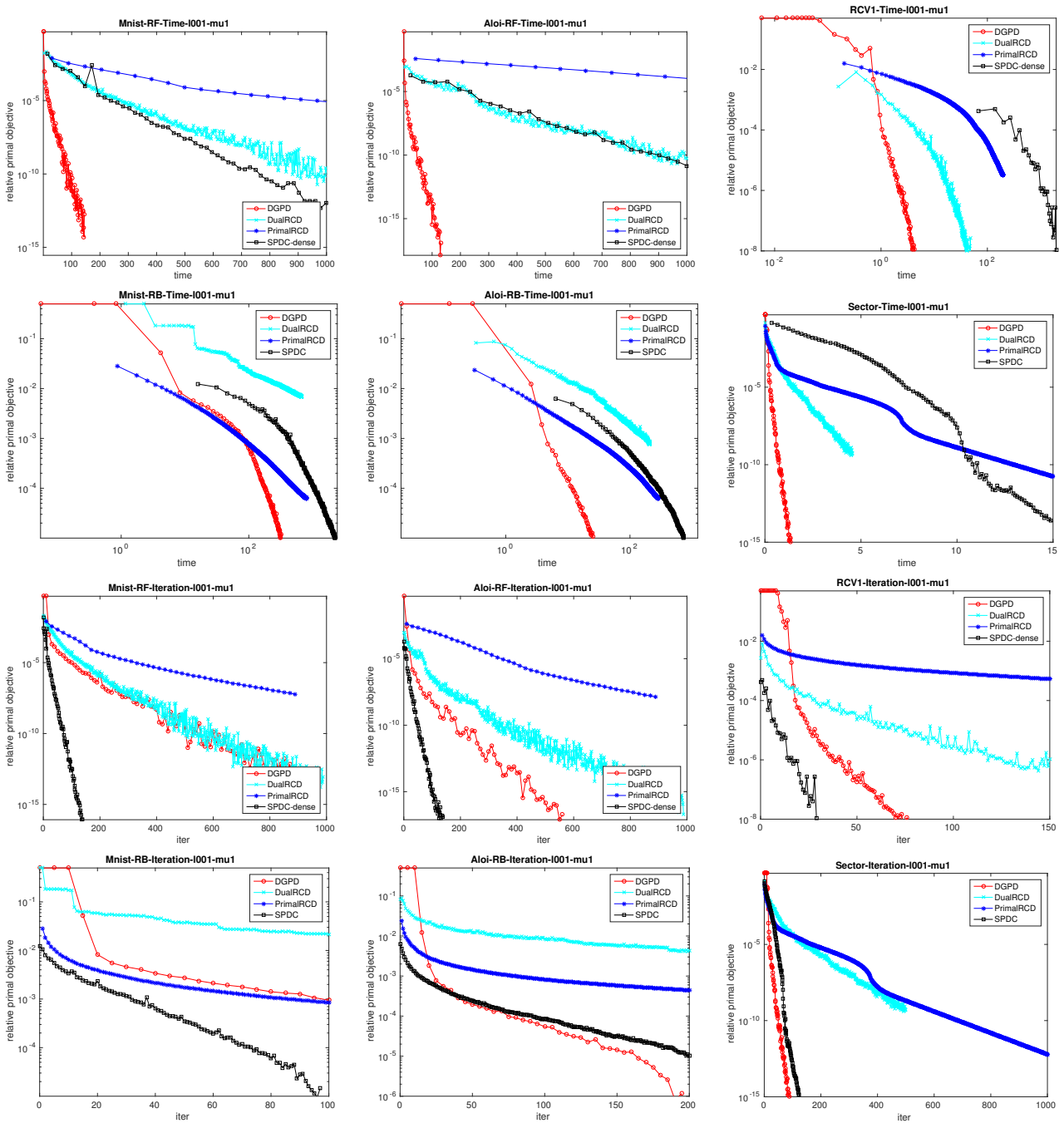

Figure 4. Relative Objective versus Time (the upper 2 rows) and versus # iterations (the lower 2 rows) for $\lambda = 0.01$, $\mu = 1$.

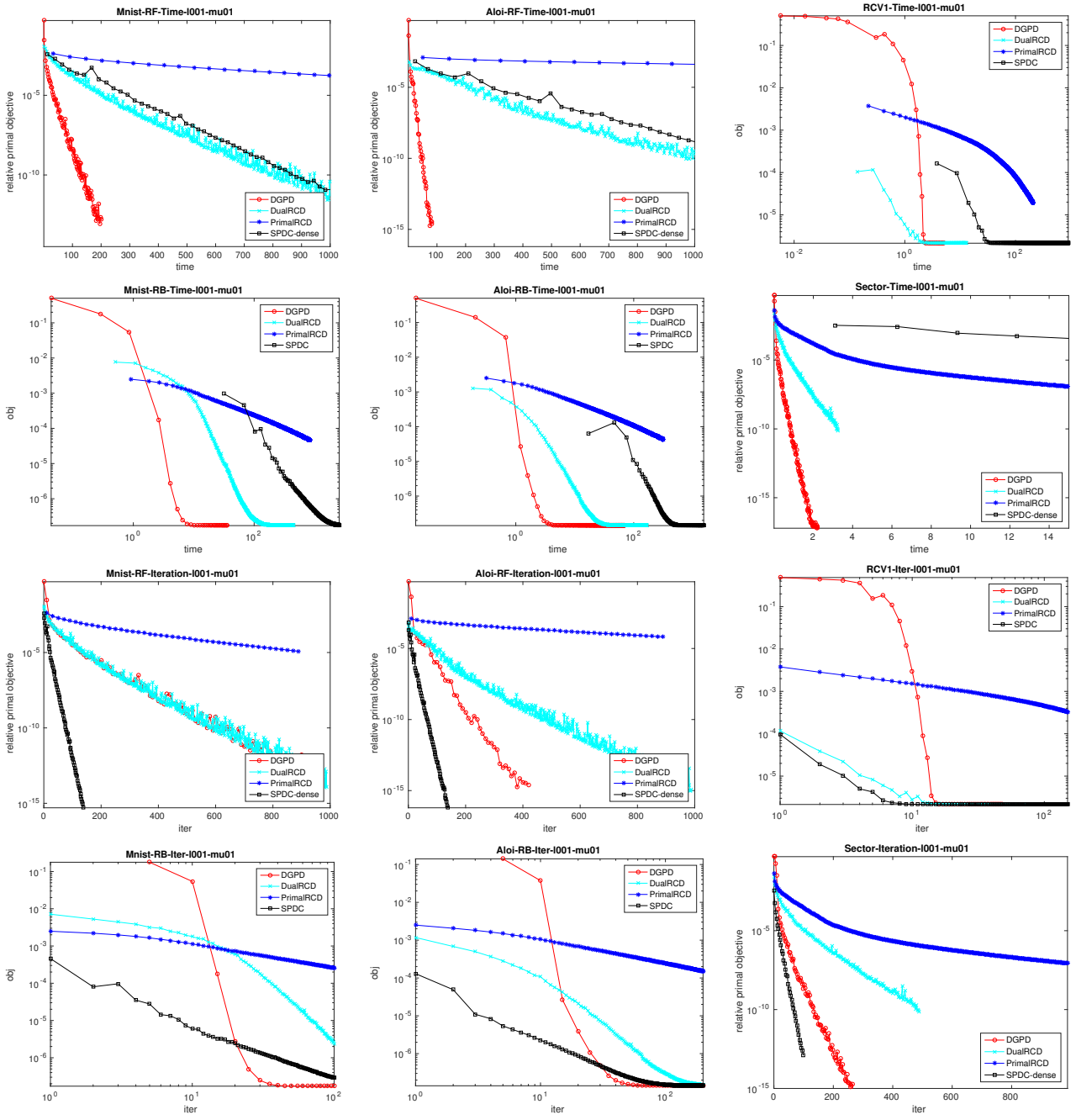

Figure 5. Relative Objective versus Time (the upper 2 rows) and versus # iterations (the lower 2 rows) for $\lambda = 0.01$, $\mu = 0.1$.

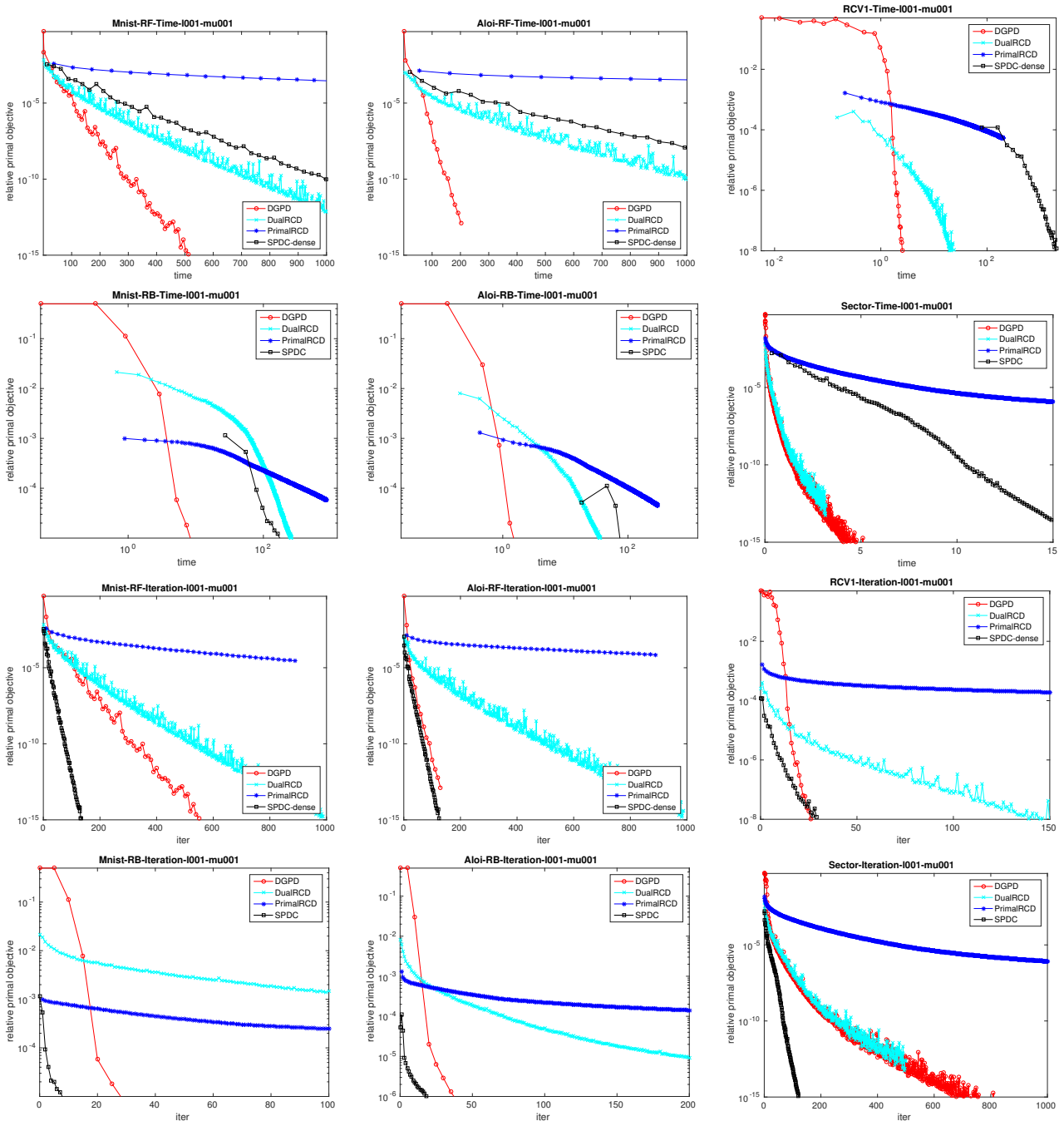

Figure 6. Relative Objective versus Time (the upper 2 rows) and versus # iterations (the lower 2 rows) for $\lambda = 0.01$, $\mu = 0.01$.