Supplementary Materials: Alternating Direction Method of Multipliers for Quantization

A On the update rules of ADMM-Q

Consider the following optimization problem mentioned in section 3.2:

$$\min_x \quad f(x) + \mathcal{I}_{\mathcal{A}}(y) \qquad \text{s.t.} \quad x = y.$$

Following the steps of regular ADMM in section 3.1, we have:

$$\mathcal{L}(x, y, \lambda) \triangleq f(x) + \mathcal{I}_{\mathcal{A}}(y) + \langle \lambda, x - y \rangle + \frac{\rho}{2}\|x - y\|_2^2,$$

In regular ADMM, the order of updating variables does not matter for convergence. When we extend its use to quantization, we update y, x and λ in sequence at each iteration, which is convenient for analyzing its convergence.

Primal Update: $\quad y^{r+1} = \arg\min_y \mathcal{L}(x^r, y, \lambda^r), \qquad x^{r+1} = \arg\min_x \mathcal{L}(y, x^{r+1}, \lambda^r)$

Dual Update: $\quad \lambda^{r+1} = \lambda^r + \rho(x^{r+1} - y^{r+1}).$

The update rule of x and λ is clear. We only derive the update rule of y here:

$$\begin{aligned} y^{r+1} &= \arg\min_y \mathcal{L}(x^r, y, \lambda^r) \\ &= \arg\min_y f(x^r) + \mathcal{I}_{\mathcal{A}}(y) + \langle \lambda^r, x^r - y \rangle + \frac{\rho}{2}\|x^r - y\|_2^2 \\ &= \arg\min_y \mathcal{I}_{\mathcal{A}}(y) + \langle \lambda^r, x^r - y \rangle + \frac{\rho}{2}\|x^r - y\|_2^2 \\ &= \arg\min_y \mathcal{I}_{\mathcal{A}}(y) + \langle \lambda^r, x^r - y \rangle + \frac{\rho}{2}\|x^r - y\|_2^2 \\ &= \arg\min_y \mathcal{I}_{\mathcal{A}}(y) + \|y - x^r - \rho^{-1}\lambda^r\|_2^2 \\ &= \mathcal{P}_{\mathcal{A}}(x^r + \rho^{-1}\lambda^r) \end{aligned} \tag{8}$$

B Proofs in Section 3.3

Lemma B.1. *For any $r \geqslant 1$ we have $\lambda^r = -\nabla_x f(x^r)$.*

Proof. based on the algorithm updates and the optimality condition for x^{r+1} we can easily verify that:

$$\nabla_x f(x^{r+1}) + \underbrace{\lambda^r + \rho(x^{r+1} - y^{r+1})}_{\lambda^{r+1}} = 0.$$

□

Lemma 3.4. If $\rho \geqslant L_f$, we have $\mathcal{L}(x^r, y^r, \lambda^r) \geqslant f(y^r) \geqslant f_{\min}, \quad \forall r \geqslant 1$.

Proof. Note that based on Lemma B.1, we have

$$\begin{aligned} \mathcal{L}(x^r, y^r, \lambda^r) &= f(x^r) + \langle \nabla f(x^r), y^r - x^r \rangle + \frac{\rho}{2}\|x^r - y^r\|^2 \\ &\geqslant f(y^r) \geqslant f_{\min} \end{aligned} \tag{9}$$

where the last two inequalities are due to Assumptions 3.2 and 3.1, respectively. □

Lemma 3.5. Define $\sigma(\rho) \triangleq \rho - \mu$. We have

$$\mathcal{L}(x^{r+1}, y^{r+1}, \lambda^{r+1}) - \mathcal{L}(x^r, y^r, \lambda^r) \leqslant \left(\rho^{-1} {L_f}^2 - \frac{\sigma(\rho)}{2}\right) \left\|x^{r+1} - x^r\right\|^2. \tag{10}$$

Proof. Let us re-write (10) as

$$\begin{aligned}\mathcal{L}(x^{r+1}, y^{r+1}, \lambda^{r+1}) - \mathcal{L}(x^r, y^r, \lambda^r) = &\underbrace{\mathcal{L}(x^{r+1}, y^{r+1}, \lambda^{r+1}) - \mathcal{L}(x^{r+1}, y^{r+1}, \lambda^r)}_{(A)} \\ &+ \underbrace{\mathcal{L}(x^{r+1}, y^{r+1}, \lambda^r) - \mathcal{L}(x^r, y^r, \lambda^r)}_{(B)}.\end{aligned}$$

We want to show that $(A) + (B) \leqslant 0$. First of all note that

$$(A) = \langle \lambda^{r+1}, x^{r+1} - y^{r+1} \rangle - \langle \lambda^r, x^{r+1} - y^{r+1} \rangle = \rho^{-1} \left\|\lambda^{r+1} - \lambda^r\right\|^2.$$

By the optimality condition of x^{r+1}, we have:

$$\nabla_x f(x^{r+1}) + \underbrace{\lambda^r + \rho(x^{r+1} - y^{r+1})}_{\lambda^{r+1}} = 0,$$

showing that $\nabla_x f(x^{r+1}) = -\lambda^{r+1}$, or $\nabla_x f(x^r) = -\lambda^r$.

Furthermore, by the lipschitz assumption of $f(\cdot)$, we have $\left\|\nabla_x f(x^{r+1}) - \nabla_x f(x^r)\right\|^2 \leqslant {L_f}^2 \left\|x^{r+1} - x^r\right\|^2$, showing that

$$\left\|\lambda^{r+1} - \lambda^r\right\|^2 \leqslant {L_f}^2 \left\|x^{r+1} - x^r\right\|^2.$$

Therefore,

$$(A) \leqslant \rho^{-1} {L_f}^2 \left\|x^{r+1} - x^r\right\|^2.$$

On the other hand:

$$\begin{aligned}(B) &= \mathcal{L}(x^{r+1}, y^{r+1}, \lambda^r) - \mathcal{L}(x^r, y^r, \lambda^r) \\ &= \mathcal{L}(x^{r+1}, y^{r+1}, \lambda^r) - \mathcal{L}(x^r, y^{r+1}, \lambda^r) + \underbrace{\mathcal{L}(x^r, y^{r+1}, \lambda^r) - \mathcal{L}(x^r, y^r, \lambda^r)}_{\leqslant 0} \\ &\leqslant \mathcal{L}(x^{r+1}, y^{r+1}, \lambda^r) - \mathcal{L}(x^r, y^{r+1}, \lambda^r) \\ &\leqslant -\frac{\sigma(\rho)}{2} \left\|x^{r+1} - x^r\right\|^2,\end{aligned}$$

where $\sigma(\rho)$ is the strong convex modulus of $\mathcal{L}(\cdot, y^{r+1}, \lambda^r)$ (note that $\sigma(\rho) = \rho - \mu$). □

Example B.2. *Consider the optimization problem* $\min_{x \in \mathbb{Z}} \frac{1}{2}(x^2 - x)$. *It is easy to verify that a* ρ*-stationary point does not exist for* $\rho = \frac{1}{2} < L_f = 1$.

Lemma 3.7. Assume $x^\star$ is the global optimal solution to problem (1), then $x^\star$ is ρ-stationary for any $\rho \geqslant L_f$.

Proof. Assume the contrary that $x^\star$ is not ρ-stationary. By Definition 3.6, $x^\star \notin \arg\min_{a \in \mathcal{A}} \frac{\rho}{2} \|a - x^\star + \rho^{-1} \nabla f(x^\star)\|^2$. Expanding the objective and adding the constant term $f(x^\star) - \frac{1}{2\rho} \|\nabla f(x^\star)\|^2$ implies that

$$x^\star \notin \arg\min_{a \in \mathcal{A}} \ \left(\widehat{f}(a; x^\star) := f(x^\star) + \langle \nabla f(x^\star),\ a - x^\star \rangle + \frac{\rho}{2} \|a - x^\star\|^2\right).$$

Since $\rho \geqslant L_f$, we have $f(a) \leqslant \widehat{f}(a; x^\star)$, $\forall\ a$, according to the descent lemma. Moreover, there exists $a^\star \in \arg\min_{a \in \mathcal{A}} \widehat{f}(a; x^\star)$ by the compactness of $\mathcal{A}$. Thus, $f(a^\star) \leqslant \widehat{f}(a^\star; x^\star) < \widehat{f}(x^\star; x^\star) = f(x^\star)$, which contradicts the optimality of $x^\star$. □

Theorem 3.9. Assume $(\bar{x}, \bar{y}, \bar{\lambda})$ is a limit point of the `ADMM-Q` algorithm. Then $\bar{x}$ is a ρ–stationary point of the optimization problem (1).

Proof. Consider a sub-sequence $(x^{r_t}, y^{r_t}, \lambda^{r_t})$, for $t = 0, \cdots$ which converges to $(\bar{x}, \bar{y}, \bar{\lambda})$. First of all due to decrease lemma 3.5 and lower boundedness of augmented Lagrangian, Lemma 3.4, we know that $\lim_{t\to\infty} \|x^{r_t+1} - x^{r_t}\| = 0$. Thus,

$$\lim_{t\to\infty} x^{r_t+1} = \bar{x} \tag{11}$$

Now based on Lemma (B.1), we also know that

$$\bar{\lambda} = \lim_{t\to\infty} \lambda^{r_t} = \lim_{t\to\infty} \nabla f(x^{r_t}) = \nabla f(\bar{x}) \tag{12}$$

$$\lim_{t\to\infty} \lambda^{r_t+1} = \lim_{t\to\infty} \nabla f(x^{r_t+1}) = \nabla f(\bar{x}) \tag{13}$$

Thus, $\lim_{t\to\infty} \lambda^{r_t+1} = \bar{\lambda}$.

Also, as $\mathcal{A}$ is finite, there exists a large enough T, such that $y^{r_t} = \bar{y}$ for $t \geqslant T$. Again due to the fact that $\mathcal{A}$ is finite, we can re-fine the sub-sequence such that $y^{r_t+1} = \hat{y}$. Thus, without loss of generality assume that these two conditions hold, i.e. $y^{r_t} = \bar{y}$ and $y^{r_t+1} = \hat{y}$ for all t for an appropriately refined sub-sequence. This means that

$$\hat{y} \in \arg\min_a \|a - (x^{r_t} + \rho^{-1}\lambda^{r_t})\| \tag{14}$$

Moreover, $\lambda^{r_t+1} = \lambda^{r_t} + \rho(\hat{y} - x^{r_t})$. Taking the $\lim_{t\to\infty}$ from both sides, we get

$$\hat{y} = \bar{x}. \tag{15}$$

Combining the above with (14) we can easily see that

$$\|\bar{x} - (x^{r_t} + \rho^{-1}\lambda^{r_t})\| \leqslant \|a_i - (x^{r_t} + \rho^{-1}\lambda^{r_t})\|,\ i = 0, \cdots, N \tag{16}$$

Taking the limits $\lim_{t\to\infty}$ from both hand sides of the inequality for all the points a_i we have

$$\|\bar{x} - (\bar{x} + \rho^{-1}\bar{\lambda})\| \leqslant \|a_i - (\bar{x} + \rho^{-1}\bar{\lambda})\|,\ i = 0, \cdots, N. \tag{17}$$

Thus,

$$\bar{x} \in \arg\min_{a\in\mathcal{A}} \|a - (\bar{x} - \rho^{-1}\nabla f(\bar{x}))\|, \tag{18}$$

where we used the fact that $\bar{\lambda} = -\nabla f(\bar{x})$. □

C Convergence Analysis for `I-ADMM-Q`

In order to prove the main convergence results, we need a few definitions and helper lemmas. Throughout this section we re-state all the theoretical results and prove them in the order we need them. For a reference of the steps in the algorithm see Algorithm 2.

First, let us define:

$$e^r = \nabla_x \mathcal{L}(x^r, y^r, \lambda^{r-1}) = \nabla f(x^r) + \lambda^{r-1} + \rho(x^r - y^r) = \nabla f(x^r) + \lambda^r \tag{19}$$

Lemma C.1. *Due to $\sigma(\rho)$-strong convexity and $(L_f + \rho)$-smoothness of $\mathcal{L}(\cdot, y^r, \lambda^{r-1})$, we know that*

$$\sigma(\rho)\|x^r - x^r_\star\| \leqslant \|e^r\| \leqslant (\rho + L_f)\|x^r - x^r_\star\| \tag{20}$$

Moreover, due to strong convexity we also know that:

$$\langle e^r, x^r - x^r_\star\rangle \geqslant \sigma(\rho)\|x^r - x^r_\star\|^2 \tag{21}$$

Lemma C.2. *If $\rho \geqslant L_f$ and we also assume that the iterates x^r stay bounded. Then there exists a non-negative number $\bar{D}$ s.t. $\|x^r - y^r\| \leqslant \bar{D}$. With this definition,*

$$\mathcal{L}(x^r, y^r, \lambda^r) \geqslant f_{\min} - \gamma(\rho + L_f)\bar{D}^2 \tag{22}$$

Proof. Note that

$$\mathcal{L}(x^r, y^r, \lambda^r) = f(x^r) + \langle \lambda^r, x^r - y^r \rangle + \frac{\rho}{2}\|x^r - y^r\|^2 \tag{23}$$

$$= \underbrace{f(x^r) + \langle \nabla f(x^r), y^r - x^r \rangle + \frac{\rho}{2}\|x^r - y^r\|^2}_{\geqslant f(y^r)} + \langle e^r, x^r - y^r \rangle \tag{24}$$

$$\geqslant f(y^r) - \|e^r\|\|x^r - y^r\| \tag{25}$$

$$\geqslant f_{\min} - \gamma(\rho + L_f)\bar{D}^2 \tag{26}$$

where the last inequality is due to the assumptions and Lemma C.1. □

Now let us prove sufficient decrease on $\mathcal{L}$ in each iteration.

Lemma C.3. *Let the assumptions of Lemma C.2 be true. Also, define*

$$\alpha = \left(\frac{2L_f^2}{\rho} + \frac{4(\rho + L_f)^2\gamma^2}{\rho} + \frac{\gamma^2(\rho + L_f)}{2} - \frac{(1-\gamma)^2\sigma(\rho)}{2} \right) \tag{27}$$

and $\beta = \frac{4(\rho+L_f)^2\gamma^2}{\rho}$. *Note that* $\sigma(\rho) = \rho - \mu \geqslant 0$. *Furthermore, assume that the parameters* ρ *and* γ *are chosen such that* $\alpha + \beta < 0$. *Then, have*

$$\lim_{r\to\infty} \|x^{r+1} - x^r\| = 0. \tag{28}$$

Proof. Let us re-write (10) as

$$\mathcal{L}(x^{r+1}, y^{r+1}, \lambda^{r+1}) - \mathcal{L}(x^r, y^r, \lambda^r) = \underbrace{\mathcal{L}(x^{r+1}, y^{r+1}, \lambda^{r+1}) - \mathcal{L}(x^{r+1}, y^{r+1}, \lambda^r)}_{(A)} + \underbrace{\mathcal{L}(x^{r+1}, y^{r+1}, \lambda^r) - \mathcal{L}(x^r, y^r, \lambda^r)}_{(B)}.$$

We want to show that $(A) + (B) \leqslant 0$.

$$(A) = \langle \lambda^{r+1}, x^{r+1} - y^{r+1} \rangle - \langle \lambda^r, x^{r+1} - y^{r+1} \rangle = \rho^{-1} \left\| \lambda^{r+1} - \lambda^r \right\|^2.$$

Using our definitions, we have

$$(A) = \rho^{-1}\|\lambda^{r+1} - \lambda^r\|^2 \tag{29}$$

$$= \rho^{-1}\|\nabla f(x^{r+1}) - \nabla f(x^r) + e^r - e^{r+1}\|^2 \tag{30}$$

$$\leqslant \frac{2}{\rho}\left(\|\nabla f(x^{r+1}) - \nabla f(x^r)\|^2 + \|e^{r+1} - e^r\|^2 \right) \tag{31}$$

$$\leqslant \frac{2}{\rho}\left(L_f^2\|x^{r+1} - x^r\|^2 + 2\|e^r\|^2 + 2\|e^{r+1}\|^2 \right) \tag{32}$$

$$\leqslant \frac{2}{\rho}\left(L_f^2\|x^{r+1} - x^r\|^2 + 2(\rho + L_f)^2\gamma^2\left(\|x^{r+1} - x^r\|^2 + \|x^r - x^{r-1}\|^2 \right) \right), \tag{33}$$

where the last inequality is due to Lemma C.1 and the way x^r is chosen in Algorithm 2.

On the other hand:

$$(B) = \mathcal{L}(x^{r+1}, y^{r+1}, \lambda^r) - \mathcal{L}(x^r, y^r, \lambda^r)$$

$$= \mathcal{L}(x^{r+1}, y^{r+1}, \lambda^r) - \mathcal{L}(x^r, y^{r+1}, \lambda^r) + \underbrace{\mathcal{L}(x^r, y^{r+1}, \lambda^r) - \mathcal{L}(x^r, y^r, \lambda^r)}_{\leqslant 0 \text{ (due to update of } y)}$$

$$\leqslant \mathcal{L}(x^{r+1}, y^{r+1}, \lambda^r) - \mathcal{L}(x^r, y^{r+1}, \lambda^r)$$

$$= \underbrace{\mathcal{L}(x^{r+1}, y^{r+1}, \lambda^r) - \mathcal{L}(x_\star^{r+1}, y^{r+1}, \lambda^r)}_{\leqslant \frac{L_f+\rho}{2}\|x^{r+1} - x_\star^{r+1}\|^2} + \underbrace{\mathcal{L}(x_\star^{r+1}, y^{r+1}, \lambda^r) - \mathcal{L}(x^r, y^{r+1}, \lambda^r)}_{\leqslant -\frac{\sigma(\rho)}{2}\|x_\star^{r+1} - x^r\|^2}$$

$$\leqslant \frac{L_f + \rho}{2}\|x^{r+1} - x_\star^{r+1}\|^2 - \frac{\sigma(\rho)}{2}\|x_\star^{r+1} - x^r\|^2,$$

Now note that $\|x^r - x_\star^{r+1}\| \geqslant (1-\gamma)\|x^{r+1} - x^r\|$ and $\|x^{r+1} - x_\star^{r+1}\| \leqslant \gamma\|x^{r+1} - x^r\|$ because of the update rules of Algorithm 2. Plugging in these, we get

$$(B) \leqslant \left(\frac{\gamma^2(\rho + L_f)}{2} - \frac{(1-\gamma)^2\sigma(\rho)}{2}\right)\|x^{r+1} - x^r\|^2 \tag{34}$$

Now combining the inequalities for (A) and (B), we have

$$\mathcal{L}(x^{r+1}, y^{r+1}, \lambda^{r+1}) - \mathcal{L}(x^r, y^r, \lambda^r) \tag{35}$$

$$\leqslant \underbrace{\left(\frac{2L_f^2}{\rho} + \frac{4(\rho + L_f)^2\gamma^2}{\rho} + \frac{\gamma^2(\rho + L_f)}{2} - \frac{(1-\gamma)^2\sigma(\rho)}{2}\right)}_{\alpha}\|x^{r+1} - x^r\|^2 + \underbrace{\frac{4(\rho + L_f)^2\gamma^2}{\rho}}_{\beta}\|x^r - x^{r-1}\|^2 \tag{36}$$

Now for any T:

$$f_{\min} - \gamma(\rho + L_f)\bar{D}^2 \leqslant \mathcal{L}(x^{T+1}, y^{T+1}, \lambda^{T+1}) \tag{37}$$

$$= \mathcal{L}(x^0, y^0, \lambda^0) + \sum_{r=0}^{T} \mathcal{L}(x^{r+1}, y^{r+1}, \lambda^{r+1}) - \mathcal{L}(x^r, y^r, \lambda^r) \tag{38}$$

$$\leqslant (\alpha + \beta)\sum_{r=0}^{T-1}\|x^{r+1} - x^r\|^2 \;+\; \alpha\|x^{T+1} - x^T\|^2 + \mathcal{L}(x^0, y^0, \lambda^0) \tag{39}$$

$$\leqslant (\alpha + \beta)\sum_{r=0}^{T}\|x^{r+1} - x^r\|^2 + \mathcal{L}(x^0, y^0, \lambda^0), \tag{40}$$

where the last inequality is due to the fact the $\beta \geqslant 0$. Now if the parameters are chosen appropriately such that $\alpha + \beta < 0$, then the right hand side of the above inequality is decreasing as T increases, while the left hand side is constant. Therefore, we have $\lim_{T\to\infty}\sum_{r=0}^{T}\|x^{r+1} - x^r\|^2 < \infty$. Thus, $\lim_{r\to\infty}\|x^{r+1} - x^r\| = 0$. □

Theorem C.4. *Assume that all the assumptions of Lemma C.3 is satisfied. Then, For any limit point $(\bar{x}, \bar{y}, \bar{\lambda})$ of the Algorithm 2, $\bar{x}$ is a stationary solution of the problem.*

Proof. Consider a sub-sequence $(x^{r_t}, y^{r_t}, \lambda^{r_t})$, for $t = 0, \cdots$ which converges to $(\bar{x}, \bar{y}, \bar{\lambda})$. First of all due to Lemma C.3, we know that $\lim_{t\to\infty}\|x^{r_t+1} - x^{r_t}\| = 0$ and $\lim_{t\to\infty}\|x^{r_t-1} - x^{r_t}\| = 0$. Thus,

$$\lim_{t\to\infty} x^{r_t+1} = \bar{x} \;\;\&\;\; \lim_{t\to\infty} x^{r_t-1} = \bar{x} \tag{41}$$

Moreover, due to the updates of the algorithm

$$\lim_{t\to\infty}\|x^{r_t+1} - x_\star^{r_t+1}\| \leqslant \lim_{t\to\infty}\gamma\|x^{r_t+1} - x^{r_t}\| = 0 \;\;\&\;\; \lim_{t\to\infty}\|x^{r_t} - x_\star^{r_t}\| \leqslant \lim_{t\to\infty}\gamma\|x^{r_t} - x^{r_t-1}\| = 0 \tag{42}$$

Thus, $\lim_{t\to\infty} e^{r_t} = \lim_{t\to\infty} e^{r_t+1} = 0$, which means

$$\bar{\lambda} = \lim_{t\to\infty}\lambda^{r_t} = -\lim_{t\to\infty}(\nabla f(x^{r_t}) - e^{r_t}) = -\nabla f(\bar{x}) \tag{43}$$

$$\lim_{t\to\infty}\lambda^{r_t+1} = -\lim_{t\to\infty}(\nabla f(x^{r_t+1}) - e^{r_{t+1}}) = -\nabla f(\bar{x}) \tag{44}$$

Thus, $\lim_{t\to\infty}\lambda^{r_t+1} = \bar{\lambda}$.

Also, as $\mathcal{A}$ is finite, there exists a large enough T, such that $y^{r_t} = \bar{y}$ for $t \geqslant T$. Again due to the fact that $\mathcal{A}$ is finite, we can re-fine the sub-sequence such that $y^{r_t+1} = \hat{y}$. Thus, without loss of generality assume that these two conditions hold, i.e. $y^{r_t} = \bar{y}$ and $y^{r_t+1} = \hat{y}$ for all t for an appropriately refined sub-sequence. This means that

$$\hat{y} \in \arg\min_a \|a - (x^{r_t} + \rho^{-1}\lambda^{r_t})\| \tag{45}$$

Moreover, $\lambda^{r_t+1} = \lambda^{r_t} + \rho(x^{r_t+1} - \hat{y})$. Taking the $\lim_{t\to\infty}$ from both sides, we get

$$\hat{y} = \bar{x}. \tag{46}$$

Combining the above with (45) we can easily see that

$$\|\bar{x} - (x^{r_t} + \rho^{-1}\lambda^{r_t})\| \leqslant \|a_i - (x^{r_t} + \rho^{-1}\lambda^{r_t})\|,\ i = 0, \cdots, N \tag{47}$$

Taking the limits $\lim_{t\to\infty}$ from both hand sides of the inequality for all the points a_i we have

$$\|\bar{x} - (\bar{x} + \rho^{-1}\bar{\lambda})\| \leqslant \|a_i - (\bar{x} + \rho^{-1}\bar{\lambda})\|,\ i = 0, \cdots, N. \tag{48}$$

Thus,

$$\bar{x} \in \arg\min_{a\in\mathcal{A}} \|a - (\bar{x} - \rho^{-1}\nabla f(\bar{x}))\|, \tag{49}$$

where we used the fact that $\bar{\lambda} = -\nabla f(\bar{x})$. □

D Convergence Analysis of PGD Algorithm

In this short section, we show that the convergence behavior of Projected Gradient Descent (PGD) algorithm can also be analyzed using Definition 3.6. Each iteration of PGD is gradient descent followed by a projection to the discrete set $\mathcal{A}$. More precisely, PGD update rule is given by

$$x^{r+1} \in \mathcal{P}_{\mathcal{A}}(x^r - \rho^{-1}\nabla_x f(x^r)) \tag{50}$$

Lemma D.1. *Consider the PGD algorithm with the update rule $x^{r+1} \in \mathcal{P}_{\mathcal{A}}(x^r - \rho^{-1}\nabla_x f(x^r))$ with $\rho \geqslant L_f$. Then, for any $r \geqslant 1$ we have $f(x^r) \geqslant f(x^{r+1}) \geqslant f_{min}$.*

Proof. By the update rule of PGD algorithm, we have:

$$\begin{aligned} x^{r+1} &\in \arg\min_{a\in\mathcal{A}} \|a - x^r + \rho^{-1}\nabla f(x^r)\|^2 \\ &\in \arg\min_{a\in\mathcal{A}} f(a; x^r) := f(x^r) + \langle \nabla f(x^r),\ a - x^r\rangle + \frac{\rho}{2}\|a - x^r\|^2. \end{aligned}$$

Since $\rho \geqslant L_f$, we have $f(a; x^r) \geqslant f(a),\ \forall a$. Hence $f(x^r) = f(x^r; x^r) \geqslant f(x^{r+1}; x^r) \geqslant f(x^{r+1})$. □

Theorem D.2. *Assume that f satisfies Assumptions 3.1, 3.2 and 3.3. Assume further that ρ is chosen large enough so that $\rho \geqslant L_f$. Let $\bar{x}$ be a limit point of the PGD algorithm. Then $\bar{x}$ is a ρ–stationary point of the optimization problem* (1).

Proof. By Lemma D.1 and compactness of $\mathcal{A}$, we know the sequence $f(x^r)$ is bounded and monotone, and hence convergent, i.e. $\lim_{r\to\infty} f(x^r) = \bar{f}$. On the other hand, the continuity of $f(\cdot)$ implies that:

$$\exists\{x^{r_t}\} \quad \text{s.t.} \quad \lim_{t\to\infty} x^{r_t} = \bar{x} \in \mathcal{A},\ \lim_{t\to\infty} f(x^{r_t}) = f(\bar{x}).$$

Hence, $\lim_{r\to\infty} f(x^r) = f(\bar{x})$. Moreover, for any fixed $a \in \mathcal{A}$, we have

$$f(x^{r_t+1}) \leqslant f(x^{r_t}) + \langle \nabla f(x_t^r),\ a - x^r\rangle + \frac{\rho}{2}\|a - x^{r_t}\|^2.$$

Letting $t \to \infty$, we obtain:

$$f(\bar{x}) \leqslant f(\bar{x}) + \langle \nabla f(\bar{x}),\ a - \bar{x}\rangle + \frac{\rho}{2}\|a - \bar{x}\|^2,$$

which in turn implies that:

$$\bar{x} \in \arg\min_{a\in\mathcal{A}} f(\bar{x}) + \langle \nabla f(\bar{x}),\ a - \bar{x}\rangle + \frac{\rho}{2}\|a - \bar{x}\|^2$$

or equivalently,

$$\bar{x} \in \arg\min_{a\in\mathcal{A}} \|a - \bar{x} + \rho^{-1}\nabla f(\bar{x})\|^2.$$

Hence, $\bar{x}$ is a ρ-stationary point. □

E On the update rules of ADMM-S and its behavior

The update rules of x and λ variables are similar to the ADMM-Q algorithm. Here we only present the y update rule. Let us define $\beta' = \beta\rho^{-1}$, $z^{r+1} = x^r + \rho^{-1}\lambda^r$ and $\widetilde{z}^{r+1} = \mathcal{P}_{\mathcal{A}}(z^{r+1})$. Following the steps of regular ADMM, the update rule of y can be written as

$$\begin{aligned} y^{r+1} &= \arg\min_y \ \mathcal{L}(x^r, y, \lambda^r) \\ &= \arg\min_y \ f(x^r) + \langle \lambda^r, x^r - y\rangle + \frac{\rho}{2}\|x^r - y\|_2^2 + \beta\mathcal{S}_{\mathcal{A}}(y) \\ &= \arg\min_y \ \frac{1}{2}\|y - x^r - \rho^{-1}\lambda^r\|^2 + \beta\rho^{-1}\mathcal{S}_{\mathcal{A}}(y) \\ &= \arg\min_y \ \frac{1}{2}\|y - z^{r+1}\|^2 + \beta'\|y - \mathcal{P}_{\mathcal{A}}(z^{r+1})\|_2 \\ &= \arg\min_y \ \frac{1}{2}\|y - z^{r+1}\|^2 + \beta'\|y - \widetilde{z}^{r+1}\|_2 \end{aligned} \tag{51}$$

If $y^{r+1} \neq \widetilde{z}^{r+1}$, then we can take the derivative of the above function and set it to 0 to get the update rule of y:

$$\begin{aligned} &(y^{r+1} - z^{r+1}) + \beta' \frac{y^{r+1} - \widetilde{z}^{r+1}}{\|y^{r+1} - \widetilde{z}^{r+1}\|_2} = 0 \\ \Longrightarrow\ & y^{r+1} = z^{r+1} + \beta' \frac{\widetilde{z}^{r+1} - z^{r+1}}{\|\widetilde{z}^{r+1} - z^{r+1}\|_2} \end{aligned} \tag{52}$$

Now we need to find out when the solution is $y^{r+1} = \widetilde{z}^{r+1}$ and when it is given by equation (52). Using the sub-gradient of the function $\|y - \widetilde{z}^{r+1}\|_2$ at the point $y = \widetilde{z}^{r+1}$, we obtain that

$$y^{r+1} = \widetilde{z}^{r+1} \quad \text{if} \quad \|\widetilde{z}^{r+1} - z^{r+1}\|_2 \leqslant \beta'$$

Combining this equation with (52), we obtain the following update rule for y:

$$y^{r+1} = \begin{cases} z^{r+1} + \dfrac{\beta'(\widetilde{z}^{r+1} - z^{r+1})}{\|\widetilde{z}^{r+1} - z^{r+1}\|_2} & , \ \beta' \leqslant \|\widetilde{z}^{r+1} - z^{r+1}\|_2 \\ \widetilde{z}^{r+1} & , \ \beta' > \|\widetilde{z}^{r+1} - z^{r+1}\|_2 \end{cases}$$

Notice that this update rule would keep y^{r+1} very close to the set $\mathcal{A}$, especially when β is large. In fact in the extreme case where β is large enough, i.e. when $\beta' = \frac{\beta}{\rho} \geqslant \sup_z \|z - \mathcal{P}_{\mathcal{A}}(z)\|$, the update rule of y in ADMM-S coincide with the update rule of y in ADMM-Q algorithm. Obviously due to the fact that y^r is not in $\mathcal{A}$, we cannot expect the ADMM-S to converge to a stationary solution defined in Definition 3.6. But in what follows we show that under assumptions similar to what we used for ADMM-Q , we can actually show that the Lagrangian function converges in ADMM-S.

Most of the proofs follow the same steps as in the convergence analysis of ADMM-Q. Thus, they are mostly omitted and we only focus on the overall steps and the results here. First of all it is easy to verify that the result of Lemma B.1 is also true for ADMM-S, i.e. $\lambda^r = -\nabla_x f(x^r)$. Moreover, Let us assume that the y^r iterates stay bounded, i.e. $y^r \in \mathcal{A}'$, where $\mathcal{A}'$ is a compact set. Note that this is a reasonable assumption due to the proximity of y^r to the bounded set $\mathcal{A}$. As f is continuous, we can assume there exists a $f_{\min}$ such that $f(y) \geqslant f_{\min}$ for all $y \in \mathcal{A}'$. Under these assumptions we have the following lemma, which states that the Lagrangian function is lower bounded.

Lemma E.1. *If $\rho \geqslant L_f$, we have $\mathcal{L}(x^r, y^r, \lambda^r) \geqslant f(y^r) \geqslant f_{\min}, \quad \forall r \geqslant 1$.*

The proof is similar to the proof of Lemma 3.4 and is omitted. Moreover, we have the following result which is similar to Lemma 3.5 for ADMM-Q.

Lemma E.2. *Define $\sigma(\rho) \triangleq \rho - \mu$. We have*

$$\mathcal{L}(x^{r+1}, y^{r+1}, \lambda^{r+1}) - \mathcal{L}(x^r, y^r, \lambda^r) \leqslant (\rho^{-1}L_f^2 - \frac{\sigma(\rho)}{2}) \left\|x^{r+1} - x^r\right\|^2. \tag{53}$$

Parameter Pairs		
v	d	$\sigma_{\widetilde{q}}^2$
8	8	30
8	16	30
8	32	30
8	64	30
8	16	10
8	16	50
8	16	70

Table 4: Parameter pairs used in the experiment

The proof of this lemma also follows the same arguments provided in the proof of Lemma 3.5. Based on these two lemmas, we have that augmented Lagrangian function is decreasing and lower bounded when ρ is chosen appropriately. Thus, it has to converge:

Proposition E.3. *If ρ is chosen such that $\rho^{-1}L_f^2 - \frac{\sigma(\rho)}{2} < 0$, then $\mathcal{L}(x^r, y^r, \lambda^r)$ is decreasing and lower bounded. Thus, it converges.*

F Simulations on Convex Quadratic Case

Recall in section 7, we solve the following problem:

$$\min_x \ \frac{1}{2}x^\top Q x + b^\top x \quad \text{s.t. } x \in \mathcal{A} \triangleq v\mathbb{Z}^d, \tag{54}$$

for some given $Q \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$, and $v \in \mathbb{Z}^+$. We generate matrix Q via the rule $Q = \widetilde{Q}^\top\widetilde{Q} + \widetilde{q}\widetilde{q}^\top$, where $\widetilde{Q}_{ij} \sim N(0,1)$, $\widetilde{q}_i \sim N(0,\sigma_{\widetilde{q}}^2)$, $1 \leqslant i,j \leqslant d$. We follow the same procedure as discussed in section 7; see Table 5 for the hyper-parameters used in `ADMM-Q`, `ADMM-S` and `ADMM-R`. We report the results for the following combinations of v, d and $\sigma_{\widetilde{q}}^2$ as seen in Table 4.

Results. Most of the observations in section 7 carry over here regardless of the values of d and $\sigma_{\widetilde{q}}^2$. More precisely, `ADMM-Q` outperforms PGD and GD+Proj with large margins. Both `ADMM-S` and `ADMM-R` not only have better median final objective values, but also smaller variance as compared with `ADMM-Q`. More importantly, the median tends to overlap with the 25% quantile, see Figure 7. It means the objective of at least 25 runs are exactly the same as the minimal objective over 50 runs. We also observe that `ADMM-S` or `ADMM-R` is not always better than `ADMM-Q`. As we conduct more experiments, we observed cases that `ADMM-S` yields large objective value; see, e.g., instance 3 in Figure 4, and compare with Figure 1. Having said that, we observe that `ADMM-S` and `ADMM-R` outperform `ADMM-Q` in most instances.

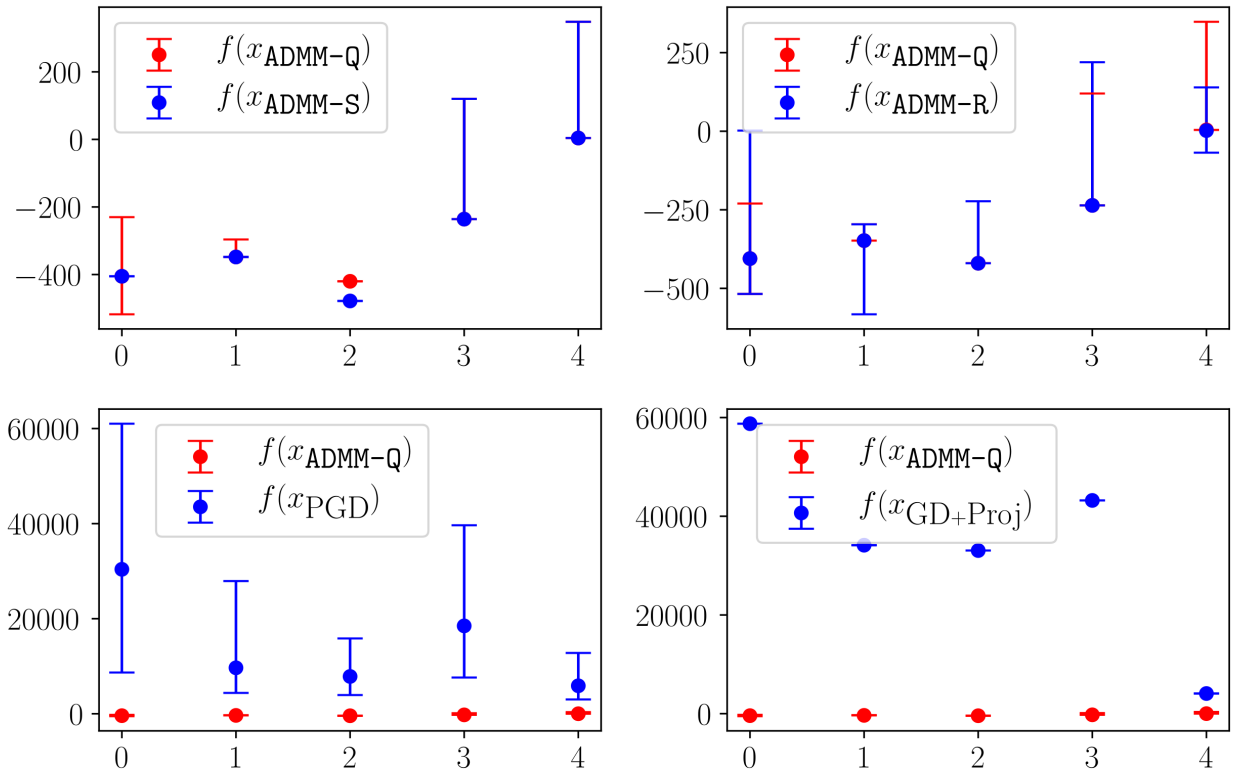

Figure 3: Performance of `ADMM-Q`, `ADMM-S`, `ADMM-R` and PGD on different problem instances with $d = 8$, $\sigma_{\tilde{q}}^2 = 30$

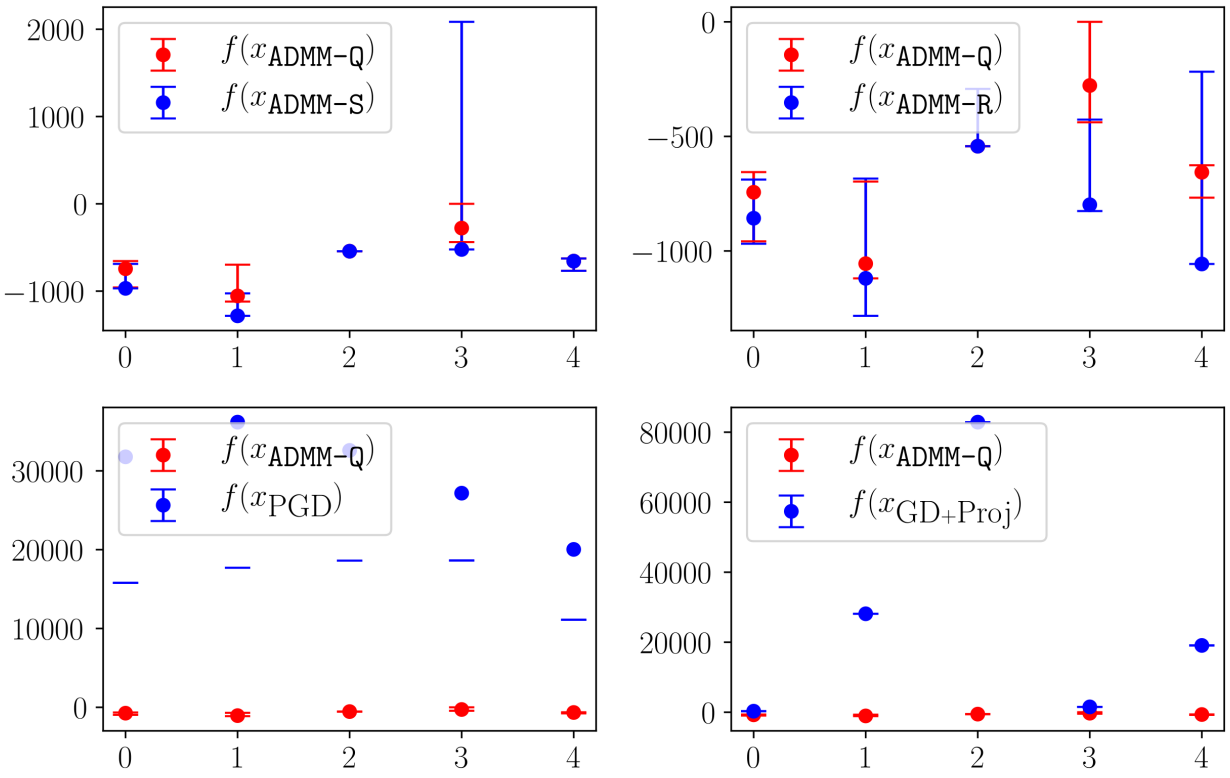

Figure 4: Performance of `ADMM-Q`, `ADMM-S`, `ADMM-R` and PGD on different problem instances with $d = 16$, $\sigma_{\tilde{q}}^2 = 30$, note the difference compared with Figure 1

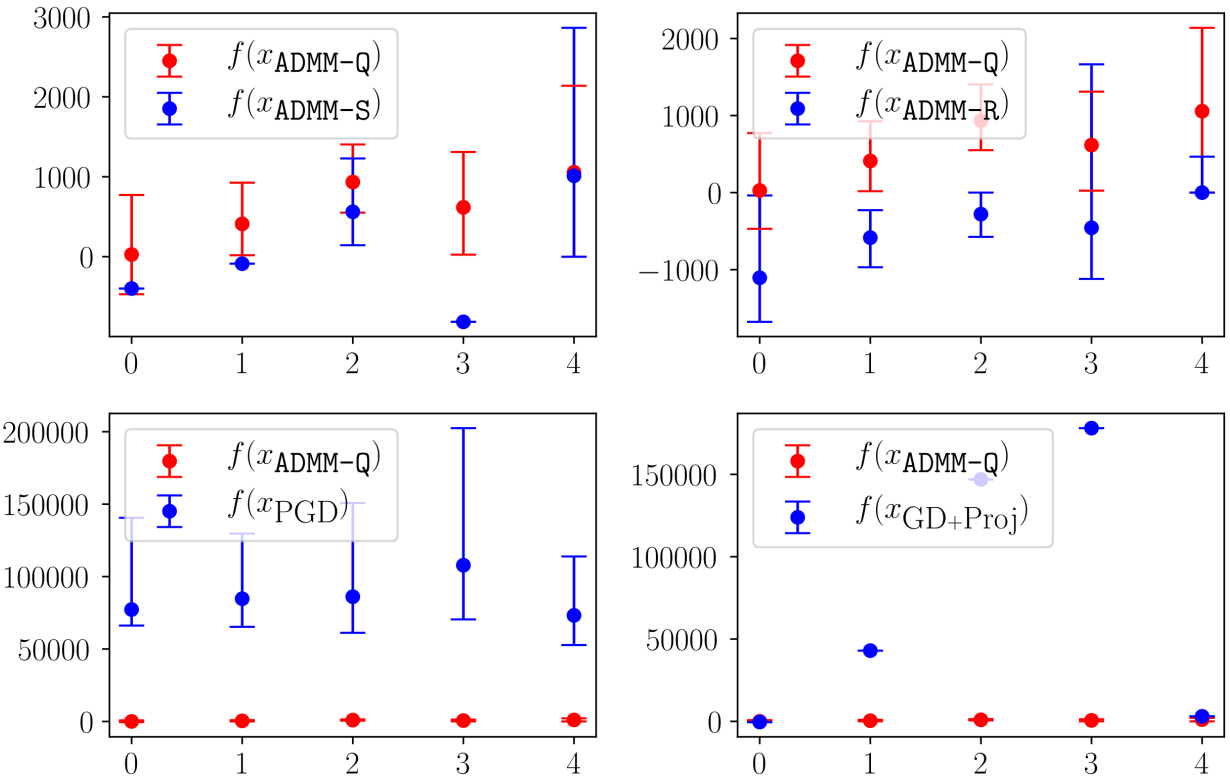

Figure 5: Performance of `ADMM-Q`, `ADMM-S`, `ADMM-R` and PGD on different problem instances with $d = 32$, $\sigma_{\tilde{q}}^2 = 30$

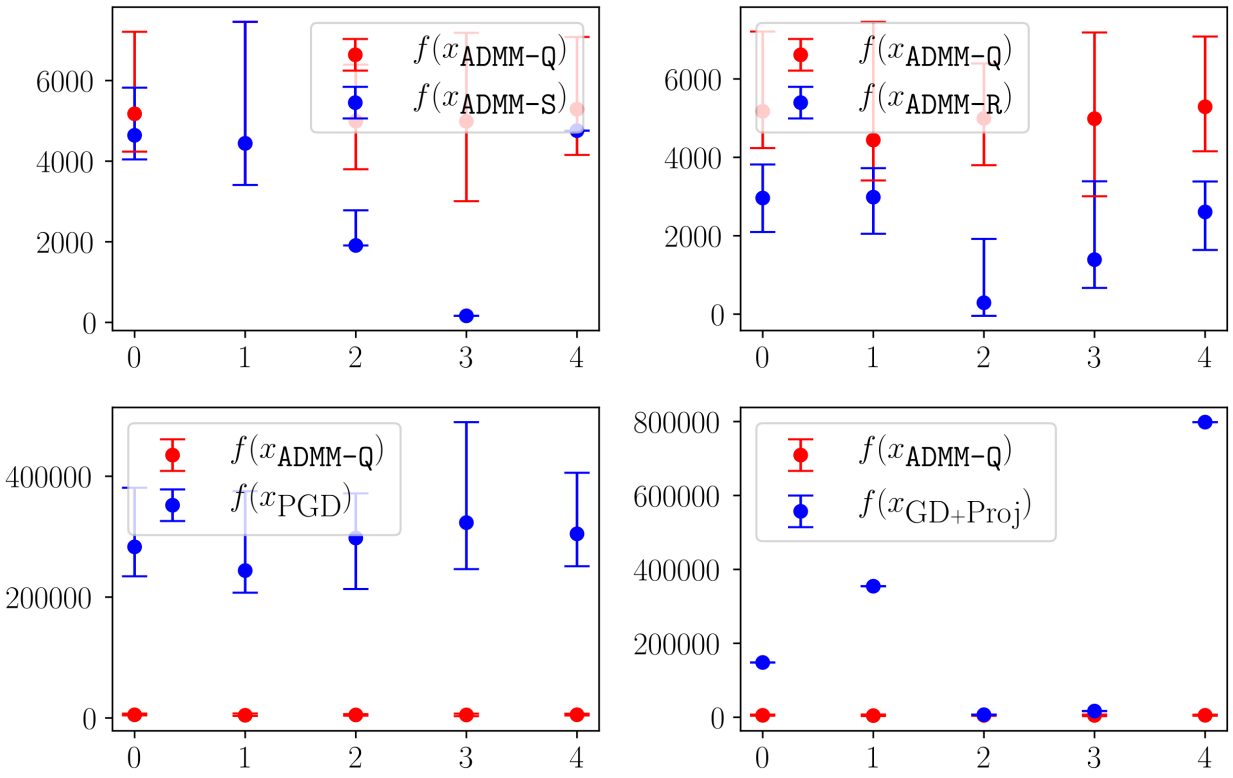

Figure 6: Performance of `ADMM-Q`, `ADMM-S`, `ADMM-R` and PGD on different problem instances with $d = 64$, $\sigma_{\tilde{q}}^2 = 30$

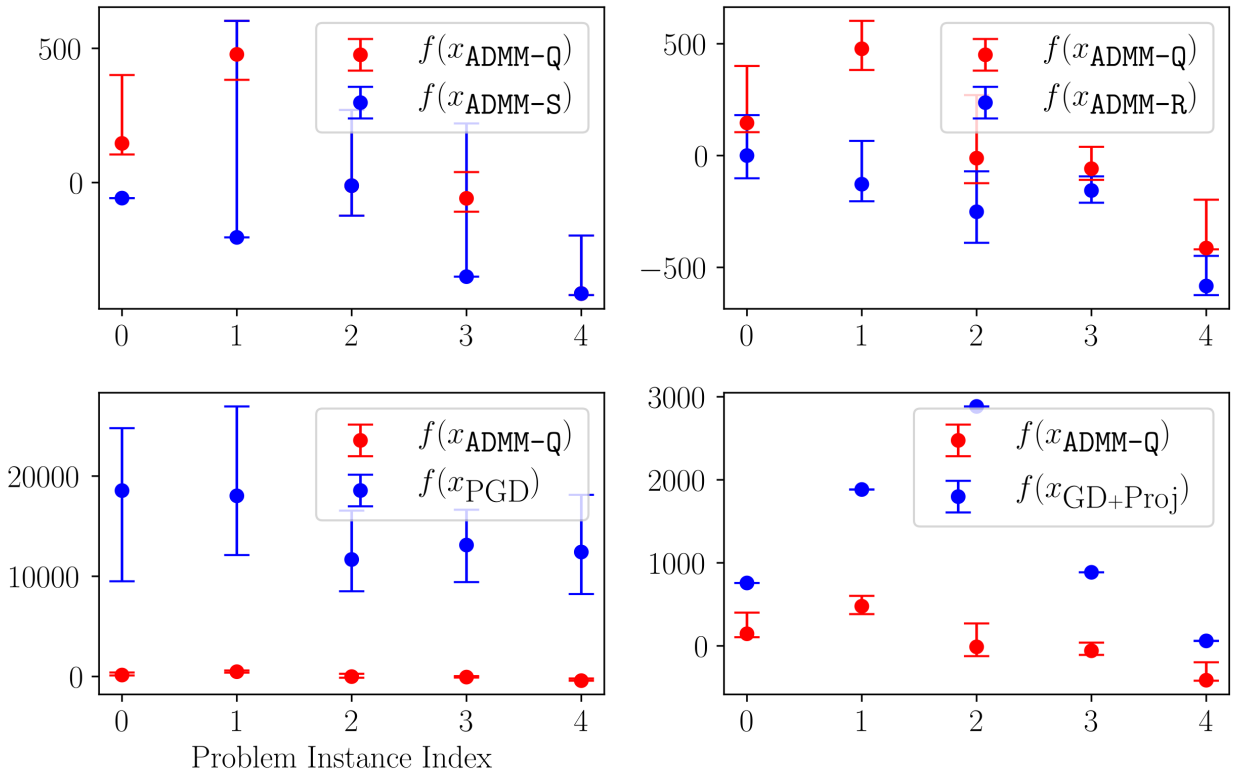

Figure 7: Performance of `ADMM-Q`, `ADMM-S`, `ADMM-R` and PGD on different problem instances with $d = 16$, $\sigma_{\tilde{q}}^2 = 10$

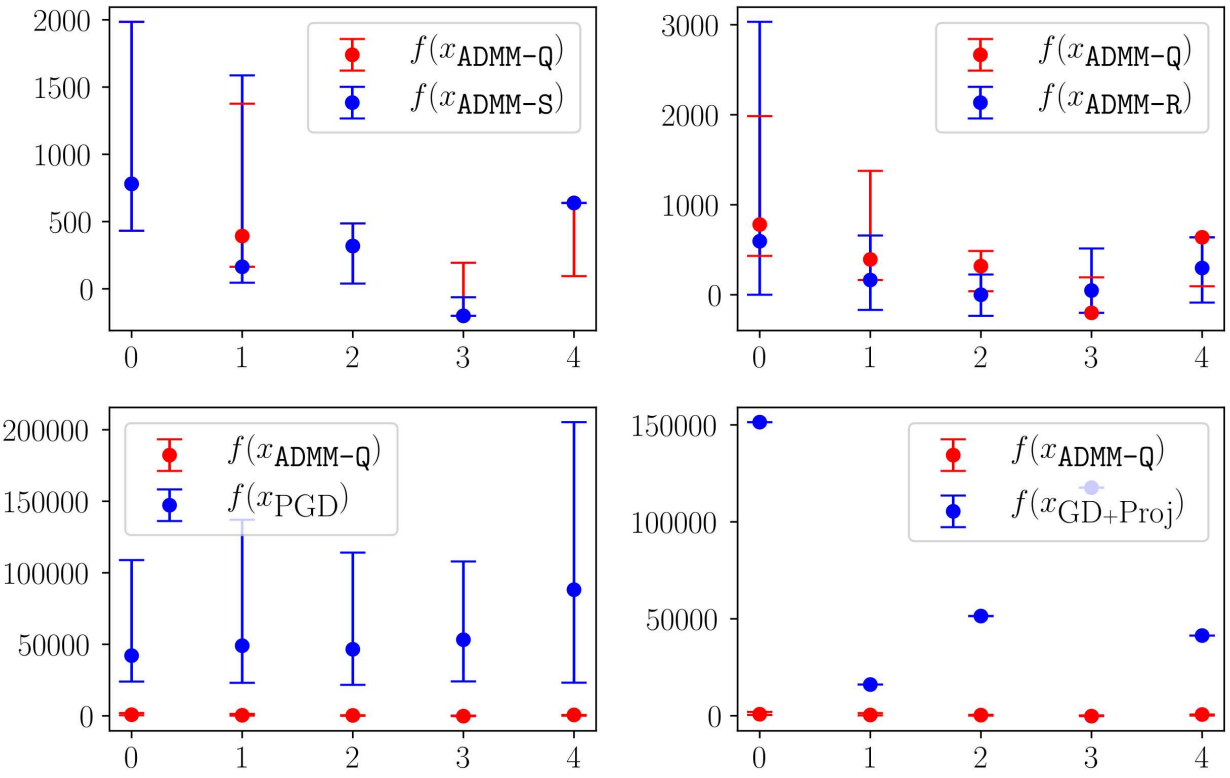

Figure 8: Performance of `ADMM-Q`, `ADMM-S`, `ADMM-R` and PGD on different problem instances with $d = 16$, $\sigma_{\tilde{q}}^2 = 50$

Algorithm	Hyper-parameters	
`ADMM-Q`	None	$\rho = 10^{-k}, k \in \mathbb{Z}, -6 \leqslant k \leqslant 2$
`ADMM-S`	$\beta = 10^{-5}, 10^{-4.5}, 10^{4}, \ldots, 10^{4.5}, 10^{5}$	$\rho = 10^{-k}, k \in \mathbb{Z}, -6 \leqslant k \leqslant 2$
`ADMM-R`	$p_i^r = 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99$	$\rho = 10^{-k}, k \in \mathbb{Z}, -6 \leqslant k \leqslant 2$

Table 5: Hyper-parameters used for `ADMM-Q`, `ADMM-S` and `ADMM-R`

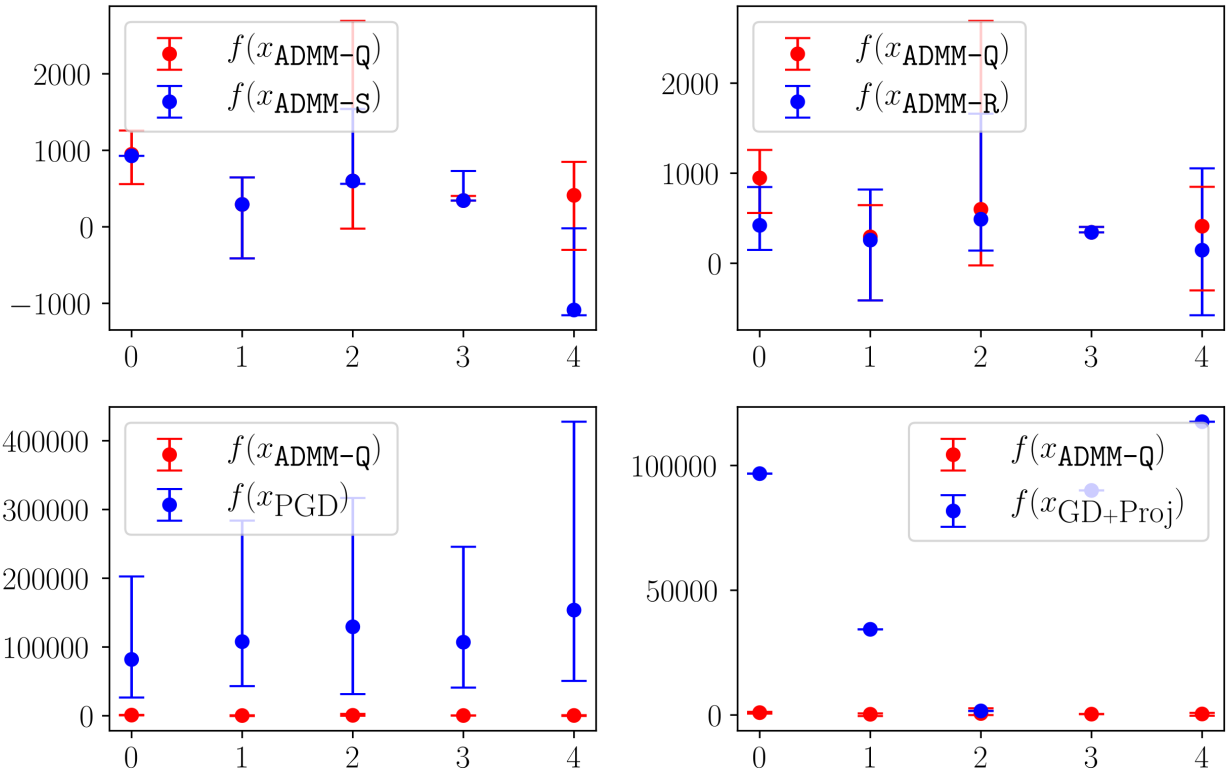

Figure 9: Performance of `ADMM-Q`, `ADMM-S`, `ADMM-R` and PGD on different problem instances with $d = 16$, $\sigma_{\tilde{q}}^2 = 70$

G Simulations on Neural Networks

Table 6 shows the performance of different algorithms on MNIST dataset. It suggests pre-training (with non-binarized weights) further improves the performance of ADMM-based methods. It is worth mentioning that, with pre-training, `ADMM-Q` and its variants and even PGD algorithm are converging extremely fast, sometimes within as few as 3 epochs. While in the presence of pre-training, and PGD and ADMM-based algorithms all work reasonably well, PGD is much more sensitive to initialization. In particular, omitting the pre-training phase drops the performance of PGD much more than the performance of ADMM-based methods. Table 7 shows the results of the experiments on CIFAR-10 dataset. The observation is consistent with that from the MNIST dataset. Pre-training significantly improves the performance of the binarized models including both ADMM-based and PGD. Binarized models trained by ADMM-based algorithms with pre-training have comparable performance with the full precision model.

Algorithm	Accuracy
BinaryConnect Courbariaux et al. [2015]	98.71%
Full Precision	$98.87 \pm 0.04\%$
GD+Proj	$74.92 \pm 4.83\%$
PGD	$92.73 \pm 0.23\%$
`ADMM-Q`	$98.21 \pm 0.16\%$
`ADMM-R`	$97.78 \pm 0.23\%$
`ADMM-S`	$98.21 \pm 0.07\%$
PGD with pre-training	$98.55 \pm 0.05\%$
`ADMM-Q` with pre-training	$98.55 \pm 0.04\%$
`ADMM-R` with pre-training	$98.61 \pm 0.06\%$
`ADMM-S` with pre-training	$98.57 \pm 0.04\%$

Table 6: Testing accuracies for MNIST dataset

Algorithm	Accuracy
Progressive DNN Ye et al. [2019]	93.53%
Full Precision	93.06%
GD+Proj	9.86%
PGD	63.53%
`ADMM-Q`	81.18%
`ADMM-R`	84.87%
`ADMM-S`	84.72%
PGD with pre-training	90.47%
`ADMM-Q` with pre-training	90.42%
`ADMM-R` with pre-training	90.46%
`ADMM-S` with pre-training	90.42%

Table 7: Testing accuracies for CIFAR-10 dataset

Layer Type	Shape
Dropout	0.2
Fully Connected + BatchNorm + ReLU	4096
Dropout	0.5
Fully Connected + BatchNorm + ReLU	4096
Dropout	0.5
Fully Connected + BatchNorm + ReLU	4096
Dropout	0.5
Fully Connected + BatchNorm	10

Table 8: Model architecture for MNIST dataset.

Algorithm		Parameter		
GD (+ Proj)		Learning rate	10^{-2}	10^{-3}
		Epoch	80	40
		Batch-size	512	512
PGD		Learning rate	10^{-2}	10^{-3}
		Epoch	80	40
		Batch-size	512	512
`ADMM-Q`		Learning rate	10^{-2}	10^{-3}
		Epoch	80	40
		Batch-size	512	512
		ρ	10^{-5}	
`ADMM-R`		Learning rate	10^{-2}	10^{-3}
		Epoch	80	40
		Batch-size	512	512
		ρ	10^{-5}	
		p_i^r	0.99	
`ADMM-S`		Learning rate	10^{-2}	10^{-3}
		Epoch	80	40
		Batch-size	512	512
		ρ	10^{-5}	
		β	10^3	
PGD with pre-training	Pre-training	Learning rate	10^{-2}	10^{-3}
		Epoch	20	20
		Batch-size	512	512
	Binariztion	Learning rate	10^{-2}	10^{-3}
		Epoch	20	20
		Batch-size	512	512
`ADMM-Q` with pre-training	Pre-training	Learning rate	10^{-2}	10^{-3}
		Epoch	20	20
		Batch-size	512	512
	Binariztion	Learning rate	10^{-2}	10^{-3}
		Epoch	20	20
		Batch-size	512	512
		ρ	10^{-3}	
`ADMM-R` with pre-training	Pre-training	Learning rate	10^{-2}	10^{-3}
		Epoch	20	20
		Batch-size	512	512
	Binariztion	Learning rate	10^{-2}	10^{-3}
		Epoch	20	20
		Batch-size	512	512
		ρ	10^{-3}	
		p_i^r	0.3	
`ADMM-S` with pre-training	Pre-training	Learning rate	10^{-2}	10^{-3}
		Epoch	20	20
		Batch-size	512	512
	Binariztion	Learning rate	10^{-2}	10^{-3}
		Epoch	20	20
		Batch-size	512	512
		ρ	10^{-3}	
		β	10^3	

Table 9: Training parameters for MNIST dataset.

Algorithm		Parameter					
GD (+ Proj)		Learning rate	10^{-2}	10^{-3}	10^{-4}		
		Epoch	100	100	100		
		Batch-size	512	512	512		
PGD		Learning rate	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-4}
		Epoch	200	200	200	200	400
		Batch-size	512	512	512	512	512
`ADMM-Q`		Learning rate	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-4}
		Epoch	200	200	200	200	400
		Batch-size	512	512	512	512	512
		ρ	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-2}
`ADMM-R`		Learning rate	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-4}
		Epoch	200	200	200	200	400
		Batch-size	512	512	512	512	512
		ρ	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-2}
		p_i^r	0.975				
`ADMM-S`		Learning rate	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-4}
		Epoch	200	200	200	200	400
		Batch-size	512	512	512	512	512
		ρ	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-2}
		β	0.05ρ				
PGD with pre-training	Pre-training	Learning rate	10^{-2}	10^{-3}	10^{-4}		
		Epoch	100	100	100		
		Batch-size	512	512	512		
	Binariztion	Learning rate	10^{-3}	10^{-4}	10^{-5}		
		Epoch	250	250	250		
		Batch-size	512	512	512		
`ADMM-Q` with pre-training	Pre-training	Learning rate	10^{-2}	10^{-3}	10^{-4}		
		Epoch	100	100	100		
		Batch-size	512	512	512		
	Binariztion	Learning rate	10^{-3}	10^{-4}	10^{-5}		
		Epoch	250	250	250		
		Batch-size	512	512	512		
		ρ	0.05ρ				
`ADMM-R` with pre-training	Pre-training	Learning rate	10^{-2}	10^{-3}	10^{-4}		
		Epoch	100	100	100		
		Batch-size	512	512	512		
	Binariztion	Learning rate	10^{-3}	10^{-4}	10^{-5}		
		Epoch	250	250	250		
		Batch-size	512	512	512		
		ρ	10^{-2}				
		p_i^r	0.975				
`ADMM-R` with pre-training	Pre-training	Learning rate	10^{-2}	10^{-3}	10^{-4}		
		Epoch	100	100	100		
		Batch-size	512	512	512		
	Binariztion	Learning rate	10^{-3}	10^{-4}	10^{-5}		
		Epoch	250	250	250		
		Batch-size	512	512	512		
		ρ	10^{-2}				
		β	0.02ρ				

Table 10: Training parameters for CIFAR-10 dataset.

H Link to the Code

Codes are available at `https://github.com/optimization-for-data-driven-science/ADMM-Q`.