Supplementary Material: Stochastic Dual Coordinate Descent with Alternating Direction Multiplier Method

Taiji Suzuki S-TAIJI@IS.TITECH.AC.JP

Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo 152-8552, JAPAN

A. Derivation of the proximal operation for the smoothed hinge loss

By the definition of the smoothed hinge loss, we have that, for $-1 \leq y_i v \leq 0$,

$$
\begin{aligned}
f_i^*(v) &= \sup_{u \in \mathbb{R}} \{uv - f_i(u)\} = \sup_{u \in \mathbb{R}} \left\{ uv - \frac{1}{2}(1 - y_i u)^2 \right\} = \sup_{u \in \mathbb{R}} \left\{ \frac{1}{2}(1 + y_i v)^2 - \frac{1}{2} - \frac{1}{2}(1 + y_i v - y_i u)^2 \right\} \\
&= \frac{1}{2}(1 + y_i v)^2 - \frac{1}{2},
\end{aligned}
$$

and $f_i^*(v) = \infty$ otherwise.

Since

$$
\begin{aligned}
\frac{f_i^*(v)}{C} + \frac{1}{2}(q - v)^2 &= \begin{cases} \frac{1}{2C}(1 + y_i v)^2 - \frac{1}{2C} + \frac{1}{2}(q - v)^2 & (-1 \leq y_i v \leq 0), \\ \infty & (\text{otherwise}), \end{cases} \\
&= \begin{cases} \frac{1+C}{2C}\left(v + \frac{y_i - qC}{1+C}\right)^2 + \frac{v^2 (y_i - qC)^2}{2C(1+C)} + \frac{q^2}{2} & (-1 \leq y_i v \leq 0), \\ \infty & (\text{otherwise}). \end{cases}
\end{aligned}
$$

Thus by minimizing this with respect to v, we have that

$$
\mathrm{prox}(u | f_i^*/C) = \begin{cases} \frac{Cu - y_i}{1+C} & (-1 \leq \frac{Cuy_i - 1}{1+C} \leq 0), \\ -y_i & (-1 > \frac{Cuy_i - 1}{1+C}), \\ 0 & (\text{otherwise}). \end{cases}
$$

B. Proof of the main theorem

In the section, we give the proofs of the theorems in the main body. For notational simplicity, we rewrite the dual problem as follows:

$$
\min_{x \in \mathcal{X}, y \in \mathcal{Y}} \sum_{i=1}^{n} g_i(x_i) + \phi(y), \tag{S-1a}
$$

$$
\text{s.t. } Zx + By = 0, \tag{S-1b}
$$

where $Z \in \mathbb{R}^{p \times n}$, $B \in \mathbb{R}^{p \times d}$. This is equivalent to the dual optimization problem in the main text when $g_i = f_i^*$ and $\phi = n\psi^*(\cdot/n)$ (or equivalently $\phi^* = n\psi$). We write $g(x) = \sum_{i=1}^{n} g_i(x_i)$.

Then we consider the following update rule:

$$y^{(t)} \leftarrow \arg\min_{y} \phi(y) - \langle w^{(t-1)}, Zx^{(t-1)} + By\rangle + \frac{\rho}{2}\|Zx^{(t-1)} + By\|^2 + \frac{1}{2}\|y - y^{(t-1)}\|_Q$$

$$x_i^{(t)} \leftarrow \arg\min_{x_I} \sum_{i\in I} g_i(x_i) - \langle w^{(t-1)}, Z_I x_I + By^{(t)}\rangle + \frac{\rho}{2}\|Z_I x_I + Z_{\setminus I} x_{\setminus I}^{(t-1)} + By^{(t)}\|^2 + \frac{1}{2}\|x_I - x_I^{(t-1)}\|_{G_{ii}}$$

$$w^{(t)} = w^{(t-1)} - \gamma\rho\{n(Zx^{(t)} + By^{(t)}) - (n - n/K)(Zx^{(t-1)} + By^{(t-1)})\}.$$

Assumption 1 can be interpreted as follows. There is an optimal solution (x^*, y^*) and corresponding Lagrange multiplier w^* such that

$$\partial g(x^*) \ni Z^\top w^*, \;\; \partial\phi(y^*) \ni B^\top w^*.$$

We denote by $\nabla f(x)$ an arbitrary element of the subgradient $\partial f(x)$ of a convex function f at x. Moreover, we suppose that each (dual) loss function g_i is v-strongly convex and ϕ is h-smooth:

$$g_i(x_i) - g_i(x_i^*) \geq \langle z_i^\top w^*, x_i - x_i^*\rangle + \frac{v\|x_i - x_i^*\|^2}{2}.$$

We also assume that there exit h and v_ϕ such that, for all y, u and all $y^* \in \mathcal{Y}^*$, there exits $\widehat{y}^* \in \mathcal{Y}^*$ which depends on y and we have

$$\phi(y) - \phi(y^*) \geq \langle B^\top w^*, y - y^*\rangle + \frac{v'_\phi}{2}\|P_{\mathrm{Ker}(B)}(y - y^*)\|^2,$$

$$\phi^*(u) - \phi^*(B^\top w^*) \geq \langle y^*, u - B^\top w^*\rangle + \frac{h'}{2}\|u - B^\top w^*\|^2.$$

Note that the primal and dual are flipped compared with the main text. Once can check that there is a correspondence between v_ψ, h in the main text and v'_ϕ and h' such that $v'_\phi = \frac{v_\psi}{n}$ and $h' = nh$.

Define

$$F(x, y) := \sum_{i=1}^{n} g_i(x_i) + \phi(y) - \langle w^*, Zx + By\rangle \quad (= nF_D(x, y)).$$

By the definition of w^*, one can easily check that

$$F(x, y) - F(x^*, y^*) \geq \frac{nv}{2}\|x - x^*\|^2 \geq 0.$$

We define

$$\begin{aligned} &R'(x, y, w)\\ =&F(x, y) - F(x^*, y^*) + \frac{2}{\rho}\|w^{(t)} - w^*\|^2 + \frac{\rho(1-\gamma)}{2}\|Zx + By\|^2 + \frac{1}{2}\|x - x^*\|^2_{vI_p + H} + \frac{1}{2K}\|y - \mathcal{Y}^*\|^2_Q. \end{aligned}$$

Here again we have that $R' = nR_D$. Let $\hat{n} = n/K$, the expected cardinality of $|I|$, and let $\mathrm{Diag}_{\mathcal{I}}(S)$ be a block diagonal matrix whose $I_k \times I_k$ $(k = 1, \ldots, K)$ diagonal elements are non-zero and given by $(\mathrm{Diag}(S))_{I_k, I_k} = S_{I_k, I_k}$ $(k = 1, \ldots, K)$.

Theorem 2. *Suppose that* $\gamma = \frac{1}{4n}$, $\mathrm{Diag}_{\mathcal{I}}(G) \succ 2\gamma\rho(n - \hat{n})\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)$ *and* $B^\top$ *is injective. Then, under the assumptions, the objective function converges R-linearly:*

$$R'(x^{(t)}, y^{(t)}, w^{(t)}) \leq \left(1 - \frac{\mu}{K}\right)^T R(x^{(0)}, y^{(0)}, w^{(0)}),$$

$$\mathrm{E}[F(x^{(t)}, y^{(t)}) - F(x^*, y^*)] \leq \left(1 - \frac{\mu}{K}\right)^T R(x^{(0)}, y^{(0)}, w^{(0)}),$$

where

$$\mu := \min\left\{\frac{1}{2}\left(\frac{v}{v + \sigma_{\max}(H)}\right), \frac{h'\rho\sigma_{\min}(BB^\top)}{2\max\{1, 4h'\rho, 4h'\sigma_{\max}(Q)\}}, \frac{Kv'_\phi}{4\sigma_{\max}(Q)}, \frac{Kv\sigma_{\min}(BB^\top)}{\sigma_{\max}(Q)(\rho\sigma_{\max}(Z^\top Z) + 4v)}\right\},$$

In particular, we have that

$$\mathrm{E}[\|w^{(t)} - w^*\|^2] \leq \frac{\rho}{2}\left(1 - \frac{\mu}{K}\right)^T R(x^{(0)}, y^{(0)}, w^{(0)}).$$

Theorem 1 in the main text can be obtained using the relation $v'_\phi = \frac{v_\phi}{n}$, $h' = nh$, $F = nF_D$ and $R' = nR_D$. The convergence of the primal objective is obtained by using the following fact: Since g is strongly convex around x^*, we have that

$$\begin{aligned} & g(x) - g(x^*) \geq \langle Z^\top w^*, x - x^* \rangle + \frac{v\|x - x^*\|^2}{2} \ \ (\forall x) \\ \Rightarrow & g^*(u) \leq g^*(Z^\top w^*) + \langle x^*, u - Z^\top w^* \rangle + \frac{\|u - Z^\top w^*\|^2}{2v} \ \ (\forall u). \end{aligned}$$

where we used $Z^\top \lambda^* \in \partial g(x^*)$. Using this, we have that,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n f_i(z_i^\top w^{(t)}) - \frac{1}{n}\sum_{i=1}^n f_i(z_i^\top w^*) &\leq \left\langle Zx^*/n, w^{(t)} - w^* \right\rangle + \frac{\|Z^\top (w^{(t)} - w^*)\|^2}{2nv} \\ &= \left\langle -y^*/n, B^\top (w^{(t)} - w^*) \right\rangle + \frac{\|Z^\top (w^{(t)} - w^*)\|^2}{2nv}, \end{aligned}$$

where we used the relation $Zx^* + By^* = 0$. Moreover, using the relation $\psi(B^\top w) \leq \psi(B^\top w^*) + \langle y^*/n, B^\top (w - w^*) \rangle + l_1\|w - w^*\| + l_2\|w - w^*\|^2$ and the Jensen's inequality $\mathrm{E}[\|w^{(T)} - w^*\|]^2 \leq \mathrm{E}[\|w^{(T)} - w^*\|^2]$, we obtain the assertion.

Proof of Theorem 2.
Step 1 (*Deriving a basic inequality*):

$$\begin{aligned} & g(x^{(t)}) - g(x^{(t-1)}) + \phi(y^{(t)}) - \phi(y^{(t-1)}) \\ = & \sum_{i \in I} g_i(x_i^{(t)}) - \sum_{i \in I} g_i(x_i^{(t-1)}) + \phi(y^{(t)}) - \phi(y^{(t-1)}) \\ = & \sum_{i \in I} g_i(x_i^{(t)}) - \langle w^{(t-1)}, Zx^{(t)} + By^{(t)} \rangle + \frac{\rho}{2}\|Zx^{(t)} + By^{(t)}\|^2 + \frac{1}{2}\|x_I^{(t)} - x_I^{(t-1)}\|_{G_{I,I}}^2 \\ & + \langle w^{(t-1)}, Zx^{(t)} + By^{(t)} \rangle - \frac{\rho}{2}\|Zx^{(t)} + By^{(t)}\|^2 - \frac{1}{2}\|x_I^{(t)} - x_I^{(t-1)}\|_{G_{I,I}}^2 \\ & - \sum_{i \in I} g_i(x_i^{(t-1)}) + \phi(y^{(t)}) - \phi(y^{(t-1)}). \end{aligned} \tag{S-2}$$

Here we define that $\tilde{Z}_I = [Z_{\backslash I} Z_I]$ and $\tilde{x} := \begin{bmatrix} x_{\backslash I}^{(t-1)} \\ x_I \end{bmatrix}$ for a given x_I, and

$$\tilde{g}_I(x_I) := \sum_{i \in I} g_i(x_i) - \left\langle w^{(t-1)}, \tilde{Z}_I\tilde{x} + By^{(t)} \right\rangle + \frac{\rho}{2}\|\tilde{Z}_I\tilde{x} + By^{(t)}\|^2 + \frac{1}{2}\|x_I - x_I^{(t-1)}\|_{G_{I,I}}^2.$$

Then by the update rule of $x^{(t)}$, we have that

$$\tilde{g}_I(x_I^{(t)}) \leq \tilde{g}_I(x_I^*) - \frac{v}{2}\|x_I^{(t)} - x_I^*\|^2 - \frac{\rho}{2}\|Z_I(x_I^{(t)} - x_I^*)\|^2 - \frac{1}{2}\|x_I^{(t)} - x_I^*\|_{G_{I,I}},$$

which implies

$$
\begin{aligned}
&\sum_{i\in I} g_i(x_i^{(t)}) - \left\langle w^{(t-1)}, Zx^{(t)} + By^{(t)} \right\rangle + \frac{\rho}{2}\|Zx^{(t)} + By^{(t)}\|^2 + \frac{1}{2}\|x_I^{(t)} - x_I^{(t-1)}\|_{G_{I,I}}^2 \\
\le & \sum_{i\in I} g_i(x_i^*) - \left\langle w^{(t-1)}, Z_I x_I^* + Z_{\setminus I} x_{\setminus I}^* + By^{(t)} \right\rangle + \frac{\rho}{2}\|Z_I x_I^* + Z_{\setminus I} x_{\setminus I}^{(t-1)} + By^{(t)}\|^2 + \frac{1}{2}\|x_I^* - x_I^{(t-1)}\|_{G_{I,I}}^2 \\
& - \frac{v}{2}\|x_I^{(t)} - x_I^*\|^2 - \frac{\rho}{2}\|Z_I(x_I^{(t)} - x_I^*)\|^2 - \frac{1}{2}\|x_I^{(t)} - x_I^*\|_{G_{I,I}} \\
= & \sum_{i\in I} g_i(x_i^*) - \left\langle w^{(t-1)}, Z_I(x_I^* - x_I^{(t)}) \right\rangle - \left\langle w^{(t-1)}, Zx^{(t)} + By^{(t)} \right\rangle \\
& + \frac{\rho}{2}\|Z_I x_I^* + Z_{\setminus I} x_{\setminus I}^{(t-1)} + By^{(t)}\|^2 - \frac{\rho}{2}\|Zx^{(t)} + By^{(t)}\|^2 + \frac{\rho}{2}\|Zx^{(t)} + By^{(t)}\|^2 + \frac{1}{2}\|x_I^* - x_I^{(t-1)}\|_{G_{I,I}}^2 \\
& - \frac{v}{2}\|x_I^{(t)} - x_I^*\|^2 - \frac{\rho}{2}\|Z_I(x_I^{(t)} - x_I^*)\|^2 - \frac{1}{2}\|x_I^{(t)} - x_I^*\|_{G_{I,I}} \\
= & \sum_{i\in I} g_i(x_i^*) - \left\langle w^{(t-1)}, Z_I(x_I^* - x_I^{(t)}) \right\rangle \\
& - \frac{v}{2}\|x_I^{(t)} - x_I^*\|^2 - \frac{\rho}{2}\|Z_I(x_I^{(t)} - x_I^*)\|^2 - \frac{1}{2}\|x_I^{(t)} - x_I^*\|_{G_{I,I}} \\
& - \rho\langle Z_{\setminus I} x_{\setminus I}^{(t)} + By^{(t)}, Z_I(x_I^{(t)} - x_I^*)\rangle + \frac{\rho}{2}\|Z_I x_I^*\|^2 - \frac{\rho}{2}\|Z_I x_I^{(t)}\|^2 + \frac{1}{2}\|x_I^* - x_I^{(t-1)}\|_{G_{I,I}}^2 \\
& - \left\langle w^{(t-1)}, Zx^{(t)} + By^{(t)} \right\rangle + \frac{\rho}{2}\|Zx^{(t-1)} + By^{(t)}\|^2.
\end{aligned}
$$

Using this, the RHS of Eq. (S-2) can be further bounded by

$$
\begin{aligned}
\text{(RHS)} \le & \sum_{i\in I} g_i(x_i^*) - \sum_{i\in I} g_i(x_i^{(t-1)}) - \langle w^{(t-1)}, Z_I(x_I^* - x_I^{(t)})\rangle \\
& - \frac{v}{2}\|x_I^{(t)} - x_I^*\|^2 - \frac{\rho}{2}\|Z_I(x_I^{(t)} - x_I^*)\|^2 - \frac{1}{2}\|x_I^{(t)} - x_I^*\|_{G_{I,I}} \\
& - \rho\langle Z_{\setminus I} x_{\setminus I}^{(t)} + By^{(t)}, Z_I(x_I^{(t)} - x_I^*)\rangle + \frac{\rho}{2}\|Z_I x_I^*\|^2 - \frac{\rho}{2}\|Z_I x_I^{(t)}\|^2 \\
& + \frac{1}{2}\|x_I^* - x_I^{(t-1)}\|_{G_{I,I}}^2 - \frac{1}{2}\|x_I^{(t)} - x_I^{(t-1)}\|_{G_{I,I}}^2 \\
& + \phi(y^{(t)}) - \phi(y^{(t-1)}).
\end{aligned} \tag{S-3}
$$

Here, we bound the term

$$
-\rho\langle Z_{\setminus i} x_{\setminus i}^{(t)} + By^{(t)}, Z_I(x_I^{(t)} - x_I^*)\rangle + \frac{\rho}{2}\|Z_I x_I^*\|^2 - \frac{\rho}{2}\|Z_I x_I^{(t)}\|^2.
$$

By Lemma 3, the expectation of this term is equivalent to

$$
\begin{aligned}
& \mathrm{E}\left[-\frac{\rho}{n}\langle Zx^{(t-1)} + By^{(t)}, Z(nx^{(t)} - (n-\hat{n})x^{(t-1)} - \hat{n}x^*)\rangle\right] \\
& + \frac{\rho}{2K}\|x^{(t-1)} - x^*\|_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)}^2 - \frac{\rho}{2}\mathrm{E}\left[\|x^{(t)} - x^{(t-1)}\|_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)}^2\right].
\end{aligned}
$$

Note that, for any block diagonal matrix S which satisfies $S_{I_k,I_{k'}} = (S_{i,j})_{(i,j)\in I_k\times I_{k'}} = O$ $(\forall k \ne k')$, we have that

$$
\begin{aligned}
\mathrm{E}[\|x_I^{(t)} - x_I^*\|_{S_{I,I}}^2] &= \mathrm{E}[\|x_I^{(t)} - x_I^{(t-1)} + x_I^{(t-1)} - x_I^*\|_{S_{I,I}}^2] \\
&= \mathrm{E}[\|x_I^{(t)} - x_I^{(t-1)}\|_{S_{I,I}}^2] + \mathrm{E}[2\langle x_I^{(t)} - x_I^{(t-1)}, x_I^{(t-1)} - x_I^*\rangle_{S_{I,I}}] + \mathrm{E}[\|x_I^{(t-1)} - x_I^*\|_{S_{I,I}}^2] \\
&= \mathrm{E}[\|x^{(t)} - x^{(t-1)}\|_S^2] + \mathrm{E}[2\langle x^{(t)} - x^{(t-1)}, x^{(t-1)} - x^*\rangle_S] + \frac{1}{K}\|x^{(t-1)} - x^*\|_S^2 \\
&= \mathrm{E}[\|x^{(t)} - x^*\|_S^2] - \mathrm{E}[\|x^{(t-1)} - x^*\|_S^2] + \frac{1}{K}\|x^{(t-1)} - x^*\|_S^2 \\
&= \mathrm{E}[\|x^{(t)} - x^*\|_S^2] - \left(1 - \frac{1}{K}\right)\mathrm{E}[\|x^{(t-1)} - x^*\|_S^2],
\end{aligned}
$$

where the expectation is taken with respect to the choice of $I \in \{I_1, \dots, I_K\}$. Moreover, for a fixed vector q, we have that

$$
\begin{aligned}
&\mathrm{E}[\langle q_I, x_I^{(t)} - x_I^* \rangle] \\
=&\mathrm{E}[\langle q_I, x_I^{(t)} - x_I^{(t-1)} + x_I^{(t-1)} - x_I^* \rangle] = \mathrm{E}[\langle q, x^{(t)} - x^{(t-1)} \rangle] + \mathrm{E}[\langle q_I, x_I^{(t-1)} - x_I^* \rangle] \\
=&\mathrm{E}[\langle q, x^{(t)} - x^{(t-1)} \rangle] + \mathrm{E}\left[\sum_{k=1}^{K} 1[I = I_k] \langle q_{I_k}, x_{I_k}^{(t-1)} - x_{I_k}^* \rangle\right] \\
=&\mathrm{E}[\langle q, x^{(t)} - x^{(t-1)} \rangle] + \frac{1}{K}\sum_{k=1}^{K} 1[I = I_k] \langle q_{I_k}, x_{I_k}^{(t-1)} - x_{I_k}^* \rangle = \mathrm{E}[\langle q, x^{(t)} - x^{(t-1)} \rangle] + \frac{1}{K}\langle q, x^{(t-1)} - x^* \rangle \\
=&\mathrm{E}\left[\left\langle q, x^{(t)} - \left(1 - \frac{1}{K}\right) x^{(t-1)} - \frac{1}{K} x^* \right\rangle\right].
\end{aligned}
$$

Then, by taking expectation with respect to I and multiplying both sides of the above inequality by n, we have that

$$
\begin{aligned}
&n\mathrm{E}[g(x^{(t)}) + \phi(y^{(t)}) - g(x^{(t-1)}) - \phi(y^{(t-1)})] \\
\leq& g(x^*) - g(x^{(t-1)}) + \mathrm{E}[\langle w^{(t-1)}, Z(nx^{(t)} - (n - \hat{n})x^{(t-1)} - \hat{n}x^*) \rangle] \\
&- \mathrm{E}\left[\frac{nv}{2}\|x^{(t)} - x^*\|^2 + \frac{n\rho}{2}\|x^{(t)} - x^*\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)} + \frac{n}{2}\|x^{(t)} - x^*\|^2_{\mathrm{Diag}_{\mathcal{I}}(G)}\right] \\
&+ \frac{(n-\hat{n})v}{2}\|x^{(t-1)} - x^*\|^2] + \frac{(n-\hat{n})\rho}{2}\|x^{(t-1)} - x^*\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)} + \frac{n-\hat{n}}{2}\|x^{(t-1)} - x^*\|^2_{\mathrm{Diag}_{\mathcal{I}}(G)} \\
&+ \mathrm{E}\left[-\rho\langle Zx^{(t-1)} + By^{(t)}, Z(nx^{(t)} - (n - \hat{n})x^{(t-1)} - \hat{n}x^*) \rangle\right] \\
&+ \frac{\rho\hat{n}}{2}\|x^{(t-1)} - x^*\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)} - \frac{n\rho}{2}\mathrm{E}\left[\|x^{(t)} - x^{(t-1)}\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)}\right] \\
&+ \frac{\hat{n}}{2}\|x^{(t-1)} - x^*\|^2_{\mathrm{Diag}_{\mathcal{I}}(G)} - \frac{n}{2}\mathrm{E}[\|x^{(t)} - x^{(t-1)}\|^2_{\mathrm{Diag}_{\mathcal{I}}(G)}] \\
&+ n\phi(y^{(t)}) - n\phi(y^{(t-1)}).
\end{aligned} \tag{S-4}
$$

Here, note that the last two term $n\phi(y^{(t)}) - n\phi(y^{(t-1)})$ is bounded as

$$
\begin{aligned}
&n\phi(y^{(t)}) - n\phi(y^{(t-1)}) \\
=&\hat{n}(\phi(y^{(t)}) - \phi(y^{(t-1)})) + (n - \hat{n})(\phi(y^{(t)}) - \phi(y^{(t-1)})) \\
\leq&\hat{n}(\phi(y^*) - \phi(y^{(t-1)})) + \left\langle \nabla\phi(y^{(t)}), (n - \hat{n})(y^{(t)} - y^{(t-1)}) + \hat{n}(y^{(t)} - y^*) \right\rangle \\
&- \frac{\hat{n}h'}{2}\|B^\top w^* - \nabla\phi(y^{(t)})\|^2.
\end{aligned}
$$

for arbitrary $y^* \in \mathcal{Y}^*$ where we used Lemma 4 in the last line. Define

$$
\tilde{w}^{(t)} := w^{(t-1)} - \rho(Zx^{(t-1)} + By^{(t)}).
$$

Note that $B^\top \tilde{w}^{(t)} - Q(y^{(t)} - y^{(t-1)}) \in \partial\phi(y^{(t)})$.

Next, adding $\mathrm{E}[n\langle w^*, Z(x^{(t-1)} - x^{(t)}) + B(y^{(t-1)} - y^{(t)}) \rangle]$ to the both sides of Eq. (S-4), we have that

$$
\begin{aligned}
&n\mathrm{E}[F(x^{(t)}, y^{(t)}) - F(x^{(t-1)}, y^{(t-1)})] \\
\leq&\hat{n}(F(x^*, y^*) - F(x^{(t-1)}, y^{(t-1)})) \\
&+ \mathrm{E}[\langle w^{(t-1)} - w^*, Z(nx^{(t)} - (n - \hat{n})x^{(t-1)} - \hat{n}x^*) \rangle] \\
&+ \mathrm{E}[\langle \tilde{w}^{(t)} - w^*, B(ny^{(t)} - (n - \hat{n})y^{(t-1)} - \hat{n}y^*) \rangle] \\
&- \langle Q(y^{(t)} - y^{(t-1)}), ny^{(t)} - (n - \hat{n})y^{(t-1)} - \hat{n}y^* \rangle \\
&- \mathrm{E}\left[\frac{nv}{2}\|x^{(t)} - x^*\|^2 + \frac{n}{2}\|x^{(t)} - x^*\|^2_H\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{(n-\hat{n})v}{2}\|x^{(t-1)} - x^*\|^2 + \frac{n}{2}\|x^{(t-1)} - x^*\|_H^2 \\
&+ \mathrm{E}\left[-\rho\langle Zx^{(t-1)} + By^{(t)}, Z(nx^{(t)} - (n-\hat{n})x^{(t-1)} - \hat{n}x^*)\rangle\right] \\
&- \frac{n}{2}\mathrm{E}\left[\|x^{(t)} - x^{(t-1)}\|_H^2\right] - \frac{\hat{n}h'}{2}\|B^\top w^* - \nabla\phi(y^{(t)})\|^2. \qquad \text{(S-5)}
\end{aligned}
$$

Step 2 (*Rearranging cross terms between* $(x^{(t)}, y^{(t)}, w^{(t)})$ *and* $(x^{(t-1)}, y^{(t-1)}, w^{(t-1)})$):

Now, we define $\hat{x}^{(t)} := nx^{(t)} - (n-\hat{n})x^{(t-1)}$ and $\hat{y}^{(t)} := ny^{(t)} - (n-\hat{n})y^{(t-1)}$. Then by the update rule of $w^{(t)}$, we have that $w^{(t)} = w^{(t-1)} - \gamma\rho(Z\hat{x}^{(t)} + B\hat{y}^{(t)})$. We evaluate the term $\mathrm{E}[\langle w^{(t-1)} - w^*, Z(\hat{x}^{(t)} - \hat{n}x^*)\rangle] + \mathrm{E}[\langle \tilde{w}^{(t)} - w^*, B(\hat{y}^{(t)} - \hat{n}y^*)\rangle]$:

$$
\begin{aligned}
&\langle w^{(t-1)} - w^*, Z(\hat{x}^{(t)} - \hat{n}x^*)\rangle + \langle \tilde{w}^{(t)} - w^*, B(\hat{y}^{(t)} - \hat{n}y^*)\rangle \\
=&\langle w^{(t-1)} - w^*, Z(\hat{x}^{(t)} - \hat{n}x^*)\rangle + \langle w^{(t-1)} - \rho(Zx^{(t-1)} + By^{(t)}) - w^*, B(\hat{y}^{(t)} - \hat{n}y^*)\rangle \\
=&\langle w^{(t)} + \gamma\rho(Z\hat{x}^{(t)} + B\hat{y}^{(t)}) - w^*, Z(\hat{x}^{(t)} - \hat{n}x^*)\rangle \\
&+ \langle w^{(t)} + \gamma\rho(Z\hat{x}^{(t)} + B\hat{y}^{(t)}) - \rho(Zx^{(t-1)} + By^{(t)}) - w^*, B(\hat{y}^{(t)} - \hat{n}y^*)\rangle \\
=&-\frac{1}{\gamma\rho}\langle w^{(t)} - w^*, w^{(t)} - w^{(t-1)}\rangle \\
&+ \gamma\rho\|Z\hat{x}^{(t)} + B\hat{y}^{(t)}\|^2 - \rho\langle Zx^{(t-1)} + By^{(t)}, B(\hat{y}^{(t)} - \hat{n}y^*)\rangle \\
=&-\frac{1}{2\gamma\rho}\left(\|w^{(t)} - w^*\|^2 + \|w^{(t)} - w^{(t-1)}\|^2 - \|w^{(t-1)} - w^*\|^2\right) \\
&+ \gamma\rho\|Z\hat{x}^{(t)} + B\hat{y}^{(t)}\|^2 - \rho\langle Zx^{(t-1)} + By^{(t)}, B(\hat{y}^{(t)} - \hat{n}y^*)\rangle \\
=&\frac{1}{2\gamma\rho}\left(-\|w^{(t)} - w^*\|^2 + \|w^{(t-1)} - w^*\|^2\right) + \frac{\gamma\rho}{2}\|Z\hat{x}^{(t)} + B\hat{y}^{(t)}\|^2 \\
&- \rho\langle Zx^{(t-1)} + By^{(t)}, B(\hat{y}^{(t)} - \hat{n}y^*)\rangle.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\langle w^{(t-1)} - w^*, Z(\hat{x}^{(t)} - \hat{n}x^*) + \langle \tilde{w}^{(t)} - w^*, B(\hat{y}^{(t)} - \hat{n}y^*)\rangle \\
&- \rho\langle Zx^{(t-1)} + By^{(t)}, Z(nx^{(t)} - (n-\hat{n})x^{(t-1)} - \hat{n}x^*)\rangle \\
=&\frac{1}{2\gamma\rho}\left(-\|w^{(t)} - w^*\|^2 + \|w^{(t-1)} - w^*\|^2\right) + \frac{\gamma\rho}{2}\|Z\hat{x}^{(t)} + B\hat{y}^{(t)}\|^2 \\
&- \rho\langle Zx^{(t-1)} + By^{(t)}, Z\hat{x}^{(t)} + B\hat{y}^{(t)}\rangle \\
=&\frac{1}{2\gamma\rho}\left(-\|w^{(t)} - w^*\|^2 + \|w^{(t-1)} - w^*\|^2\right) \\
&+ \frac{\gamma\rho}{2}n^2\|Zx^{(t)} + By^{(t)}\|^2 + \frac{\gamma\rho}{2}(n-\hat{n})^2\|Zx^{(t-1)} + By^{(t-1)}\|^2 \\
&- \gamma\rho n(n-\hat{n})\langle Zx^{(t)} + By^{(t)}, Zx^{(t-1)} + By^{(t-1)}\rangle \\
&- \rho\langle Zx^{(t-1)} + By^{(t)}, Z(nx^{(t)} - (n-\hat{n})x^{(t-1)}) + B(ny^{(t)} - (n-\hat{n})y^{(t-1)})\rangle.
\end{aligned}
$$

Next, we expand the non-squared term:

$$
\begin{aligned}
&-\gamma\rho n(n-\hat{n})\langle Zx^{(t)} + By^{(t)}, Zx^{(t-1)} + By^{(t-1)}\rangle \\
&- \rho\langle Zx^{(t-1)} + By^{(t)}, Z(nx^{(t)} - (n-\hat{n})x^{(t-1)}) + B(ny^{(t)} - (n-\hat{n})y^{(t-1)})\rangle \\
=&-\gamma\rho n(n-\hat{n})\langle Zx^{(t)} - Zx^*, Zx^{(t-1)} - Zx^*\rangle \\
&- \gamma\rho n(n-\hat{n})\langle By^{(t)} - By^*, By^{(t-1)} - By^*\rangle \\
&- \gamma\rho n(n-\hat{n})\langle Zx^{(t)} - Zx^*, By^{(t-1)} - By^*\rangle \\
&- \gamma\rho n(n-\hat{n})\langle By^{(t)} - By^*, Zx^{(t-1)} - Zx^*\rangle
\end{aligned}
$$

$$
\begin{aligned}
&- n\rho\langle Zx^{(t-1)} - Zx^*, Z(x^{(t)} - x^*)\rangle + (n-\hat{n})\rho\|Zx^{(t-1)} - Zx^*\|^2 \\
&+ (n-\hat{n})\rho\langle By^{(t)} - By^*, B(y^{(t-1)} - y^*)\rangle - n\rho\|By^{(t)} - By^*\|^2 \\
&- \rho\langle Zx^{(t-1)} - Zx^*, B(ny^{(t)} - (n-\hat{n})y^{(t-1)} - \hat{n}y^*)\rangle \\
&- \rho\langle B(y^{(t)} - y^*), Z(nx^{(t)} - (n-\hat{n})x^{(t-1)} - \hat{n}x^*)\rangle \\
=& - (\gamma\rho n(n-\hat{n}) + n\rho)\langle Zx^{(t)} - Zx^*, Zx^{(t-1)} - Zx^*\rangle \\
&- (\gamma\rho n(n-\hat{n}) - (n-\hat{n})\rho)\langle By^{(t)} - By^*, By^{(t-1)} - By^*\rangle \\
&- \gamma\rho n(n-\hat{n})\langle Zx^{(t)} - Zx^*, By^{(t-1)} - By^*\rangle \\
&- (\gamma\rho n(n-\hat{n}) + n\rho - (n-\hat{n})\rho)\langle By^{(t)} - By^*, Zx^{(t-1)} - Zx^*\rangle \\
&+ (n-\hat{n})\rho\|Zx^{(t-1)} - Zx^*\|^2 - n\rho\|By^{(t)} - By^*\|^2 \\
&- \rho(n-\hat{n})\langle Zx^{(t-1)} - Zx^*, B(y^* - y^{(t-1)})\rangle \\
&- \rho n\langle B(y^{(t)} - y^*), Z(x^{(t)} - x^*)\rangle. \qquad \text{(S-6)}
\end{aligned}
$$

Using the relation

$$
\begin{aligned}
\langle Zx^{(t)} - Zx^*, By^{(t-1)} - By^*\rangle &= \langle Z(x^{(t)} - x^*), B(y^{(t)} - y^*)\rangle + \langle Z(x^{(t)} - x^*), B(y^{(t-1)} - y^{(t)})\rangle, \\
\langle By^{(t)} - By^*, Zx^{(t-1)} - Zx^*\rangle &= \langle B(y^{(t)} - y^{(t-1)}), Z(x^{(t-1)} - x^*)\rangle + \langle B(y^{(t-1)} - y^*), Z(x^{(t-1)} - x^*)\rangle,
\end{aligned}
$$

the RHS of Eq. (S-6) is equivalent to

$$
\begin{aligned}
&- (\gamma\rho n(n-\hat{n}) + n\rho)\langle Zx^{(t)} - Zx^*, Zx^{(t-1)} - Zx^*\rangle \\
&- (\gamma\rho n(n-\hat{n}) - (n-\hat{n})\rho)\langle By^{(t)} - By^*, By^{(t-1)} - By^*\rangle \\
&+ (n-\hat{n})\rho\|Zx^{(t-1)} - Zx^*\|^2 - n\rho\|By^{(t)} - By^*\|^2 \\
&+ \{-(\gamma\rho n(n-\hat{n}) + \rho\hat{n}) + \rho(n-\hat{n})\}\langle Zx^{(t-1)} - Zx^*, B(y^{(t-1)} - y^*)\rangle \\
&- (\gamma\rho n(n-\hat{n}) + \rho n)\langle B(y^{(t)} - y^*), Z(x^{(t)} - x^*)\rangle \\
&- \gamma\rho n(n-\hat{n})\langle Z(x^{(t)} - x^*), B(y^{(t-1)} - y^{(t)})\rangle \\
&- (\gamma\rho n(n-\hat{n}) + \rho\hat{n})\langle By^{(t)} - By^{(t-1)}, Zx^{(t-1)} - Zx^*\rangle.
\end{aligned}
$$

The last two terms are transformed to

$$
\begin{aligned}
&- \gamma\rho n(n-\hat{n})\langle Z(x^{(t)} - x^*), B(y^{(t-1)} - y^{(t)})\rangle \\
&- (\gamma\rho n(n-\hat{n}) + \rho\hat{n})\langle By^{(t)} - By^{(t-1)}, Zx^{(t-1)} - Zx^*\rangle \\
=&\gamma\rho n(n-\hat{n})\langle Z(x^{(t)} - x^{(t-1)}), B(y^{(t)} - y^{(t-1)})\rangle \\
&- \rho\hat{n}\langle By^{(t)} - By^*, Zx^{(t-1)} - Zx^*\rangle + \rho\hat{n}\langle By^{(t-1)} - By^*, Zx^{(t-1)} - Zx^*\rangle.
\end{aligned}
$$

Thus, the RHS of Eq. (S-6) is further transformed to

$$
\begin{aligned}
&- (\gamma\rho n(n-\hat{n}) + n\rho)\langle Zx^{(t)} - Zx^*, Zx^{(t-1)} - Zx^*\rangle \\
&- (\gamma\rho n(n-\hat{n}) - (n-\hat{n})\rho)\langle By^{(t)} - By^*, By^{(t-1)} - By^*\rangle \\
&+ (n-\hat{n})\rho\|Zx^{(t-1)} - Zx^*\|^2 - n\rho\|By^{(t)} - By^*\|^2 \\
&+ \{-\gamma\rho n(n-\hat{n}) + \rho(n-\hat{n})\}\langle Zx^{(t-1)} - Zx^*, B(y^{(t-1)} - y^*)\rangle \\
&- (\gamma\rho n(n-\hat{n}) + \rho n)\langle B(y^{(t)} - y^*), Z(x^{(t)} - x^*)\rangle \\
&+ \gamma\rho n(n-\hat{n})\langle Z(x^{(t)} - x^{(t-1)}), B(y^{(t)} - y^{(t-1)})\rangle \\
&- \rho\hat{n}\langle By^{(t)} - By^*, Zx^{(t-1)} - Zx^*\rangle.
\end{aligned}
$$

By Lemma 5 and $Zx^* = -By^*$, this is equivalent to

$$
\begin{aligned}
& -\frac{1}{2}(\gamma\rho n(n-\hat{n}) + n\rho)\{\|Zx^{(t)} - Zx^*\|^2 + \|Zx^{(t-1)} - Zx^*\|^2 - \|Zx^{(t)} - Zx^{(t-1)}\|^2\} \\
& -\frac{1}{2}(\gamma\rho n(n-\hat{n}) - (n-\hat{n})\rho)\{\|By^{(t)} - By^*\|^2 + \|By^{(t-1)} - By^*\|^2 - \|By^{(t)} - By^{(t-1)}\|^2\} \\
& + (n-\hat{n})\rho\|Zx^{(t-1)} - Zx^*\|^2 - n\rho\|By^{(t)} - By^*\|^2 \\
& -\frac{1}{2}\{-\gamma\rho n(n-\hat{n}) + \rho(n-\hat{n})\}(\|Zx^{(t-1)} - Zx^*\|^2 + \|B(y^{(t-1)} - y^*)\|^2 - \|Zx^{(t-1)} + By^{(t-1)}\|^2) \\
& + \frac{1}{2}(\gamma\rho n(n-\hat{n}) + \rho n)(\|Zx^{(t)} - Zx^*\|^2 + \|B(y^{(t)} - y^*)\|^2 - \|Zx^{(t)} + By^{(t)}\|^2) \\
& + \gamma\rho n(n-\hat{n})\langle Z(x^{(t)} - x^{(t-1)}), B(y^{(t)} - y^{(t-1)})\rangle \\
& - \rho\hat{n}\langle By^{(t)} - By^*, Zx^{(t-1)} - Zx^*\rangle \\
= & -\frac{\rho\hat{n}}{2}\|Zx^{(t-1)} - Zx^*\|^2 + \frac{1}{2}(\gamma\rho n(n-\hat{n}) + n\rho)\|Zx^{(t)} - Zx^{(t-1)}\|^2 \\
& -\frac{\rho\hat{n}}{2}\|By^{(t)} - By^*\|^2 + \frac{1}{2}(\gamma\rho n(n-\hat{n}) - (n-\hat{n})\rho)\|By^{(t)} - By^{(t-1)}\|^2 \\
& -\frac{1}{2}\{\gamma\rho n(n-\hat{n}) - \rho(n-\hat{n})\}\|Zx^{(t-1)} + By^{(t-1)}\|^2 \\
& -\frac{1}{2}\{\gamma\rho n(n-\hat{n}) + \rho n\}\|Zx^{(t)} + By^{(t)}\|^2 \\
& + \gamma\rho n(n-\hat{n})\langle Z(x^{(t)} - x^{(t-1)}), B(y^{(t)} - y^{(t-1)})\rangle \\
& - \hat{n}\rho\langle By^{(t)} - By^*, Zx^{(t-1)} - Zx^*\rangle \\
= & -\frac{\hat{n}\rho}{2}\|Zx^{(t-1)} + By^{(t)}\|^2 \\
& + \frac{1}{2}(\gamma\rho n(n-\hat{n}) + n\rho)\|Zx^{(t)} - Zx^{(t-1)}\|^2 \\
& + \frac{1}{2}(\gamma\rho n(n-\hat{n}) - (n-\hat{n})\rho)\|By^{(t)} - By^{(t-1)}\|^2 \\
& -\frac{1}{2}\{\gamma\rho n(n-\hat{n}) - \rho(n-\hat{n})\}\|Zx^{(t-1)} + By^{(t-1)}\|^2 \\
& -\frac{1}{2}\{\gamma\rho n(n-\hat{n}) + \rho n\}\|Zx^{(t)} + By^{(t)}\|^2 \\
& + \gamma\rho n(n-\hat{n})\langle Z(x^{(t)} - x^{(t-1)}), B(y^{(t)} - y^{(t-1)})\rangle.
\end{aligned} \tag{S-7}
$$

Since

$$
\begin{aligned}
& \gamma\rho n(n-\hat{n})\langle Z(x^{(t)} - x^{(t-1)}), B(y^{(t)} - y^{(t-1)})\rangle \\
\leq & \frac{\gamma\rho n(n-\hat{n})}{2}\{\|Z(x^{(t)} - x^{(t-1)})\|^2 + \|B(y^{(t)} - y^{(t-1)})\|^2\},
\end{aligned}
$$

the RHS of Eq. (S-7) is bounded by

$$
\begin{aligned}
& -\frac{\hat{n}\rho}{2}\|Zx^{(t-1)} + By^{(t)}\|^2 \\
& + \frac{1}{2}(2\gamma\rho n(n-\hat{n}) + n\rho)\|Zx^{(t)} - Zx^{(t-1)}\|^2 \\
& + \frac{1}{2}(2\gamma\rho n(n-\hat{n}) - (n-\hat{n})\rho)\|By^{(t)} - By^{(t-1)}\|^2 \\
& -\frac{1}{2}\{\gamma\rho n(n-\hat{n}) - \rho(n-\hat{n})\}\|Zx^{(t-1)} + By^{(t-1)}\|^2 - \frac{1}{2}\{\gamma\rho n(n-\hat{n}) + \rho n\}\|Zx^{(t)} + By^{(t)}\|^2.
\end{aligned}
$$

Combining this and Eq. (S-5), and noticing $\|Zx^{(t)} - Zx^{(t-1)}\| = \|Z_I(x_I^{(t)} - x_I^{(t-1)})\| = \|x^{(t)} - x^{(t-1)}\|_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)}$, we obtain

$$
\begin{aligned}
&nE[F(x^{(t)}, y^{(t)}) - F(x^{(t-1)}, y^{(t-1)})] \\
\leq &\hat{n}(F(x^*, y^*) - F(x^{(t-1)}, y^{(t-1)})) \\
&+ \frac{1}{2\gamma\rho}\left(-\|w^{(t)} - w^*\|^2 + \|w^{(t-1)} - w^*\|^2\right) \\
&- \frac{\hat{n}\rho}{2}\|Zx^{(t-1)} + By^{(t)}\|^2 \\
&+ \frac{1}{2}\{\gamma\rho n^2 - \gamma\rho n(n-\hat{n}) - \rho n\}\|Zx^{(t)} + By^{(t)}\|^2 \\
&+ \frac{1}{2}\{\gamma\rho(n-\hat{n})^2 - \gamma\rho n(n-\hat{n}) + \rho(n-\hat{n})\}\|Zx^{(t-1)} + By^{(t-1)}\|^2 \\
&- E\left[\frac{nv}{2}\|x^{(t)} - x^*\|^2 + \frac{n}{2}\|x^{(t)} - x^*\|_H^2\right] \\
&+ \frac{(n-\hat{n})v}{2}\|x^{(t-1)} - x^*\|^2 + \frac{n}{2}\|x^{(t-1)} - x^*\|_H^2 \\
&+ \gamma\rho n(n-\hat{n})E\left[\|x^{(t)} - x^{(t-1)}\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)}\right] - \frac{n}{2}E[\|x^{(t)} - x^{(t-1)}\|^2_{\mathrm{Diag}_{\mathcal{I}}(G)}] \\
&+ (\gamma\rho n(n-\hat{n}) - \frac{(n-\hat{n})\rho}{2})\|B(y^{(t)} - y^{(t-1)})\|^2 \\
&- \langle Q(y^{(t)} - y^{(t-1)}), ny^{(t)} - (n-\hat{n})y^{(t-1)} - \hat{n}y^*\rangle \\
&- \frac{\hat{n}h'}{2}\|B^\top w^* - \nabla\phi(y^{(t)})\|^2.
\end{aligned}
$$

Since we have assumed $\mathrm{Diag}_{\mathcal{I}}(G) \succ 2\gamma\rho(n-\hat{n})\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)$, it holds that

$$
\gamma\rho n(n-\hat{n})E\left[\|x^{(t)} - x^{(t-1)}\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)}\right] - \frac{n}{2}E[\|x^{(t)} - x^{(t-1)}\|^2_{\mathrm{Diag}_{\mathcal{I}}(G)}] \leq 0.
$$

Moreover, we have that

$$
\begin{aligned}
&- \langle Q(y^{(t)} - y^{(t-1)}), ny^{(t)} - (n-\hat{n})y^{(t-1)} - \hat{n}y^*\rangle \\
= &- n\|y^{(t)} - y^{(t-1)}\|_Q^2 + \frac{1}{2}\{\|y^{(t)} - y^{(t-1)}\|_Q^2 + \|y^{(t-1)} - y^*\|_Q^2 - \|y^{(t)} - y^*\|_Q^2\} \\
= &- \left(n - \frac{\hat{n}}{2}\right)\|y^{(t)} - y^{(t-1)}\|_Q^2 + \frac{\hat{n}}{2}\|y^{(t-1)} - y^*\|_Q^2 - \frac{\hat{n}}{2}\|y^{(t)} - y^*\|_Q^2.
\end{aligned}
$$

Finally, we achieve

$$
\begin{aligned}
&nE[F(x^{(t)}, y^{(t)}) - F(x^{(t-1)}, y^{(t-1)})] \\
\leq &\hat{n}(F(x^*, y^*) - F(x^{(t-1)}, y^{(t-1)})) \\
&+ \frac{1}{2\gamma\rho}\left(-\|w^{(t)} - w^*\|^2 + \|w^{(t-1)} - w^*\|^2\right) \\
&- \frac{\rho n(1-\gamma)}{2}\|Zx^{(t)} + By^{(t)}\|^2 + \frac{\rho(n-\hat{n})(1+\gamma)}{2}\|Zx^{(t-1)} + By^{(t-1)}\|^2 \\
&- E\left[\frac{nv}{2}\|x^{(t)} - x^*\|^2 + \frac{n}{2}\|x^{(t)} - x^*\|_H^2\right] \\
&+ \frac{(n-\hat{n})v}{2}\|x^{(t-1)} - x^*\|^2 + \frac{n}{2}\|x^{(t-1)} - x^*\|_H^2 \\
&+ \gamma\rho n(n-\hat{n})E\left[\|x^{(t)} - x^{(t-1)}\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)}\right] - \frac{n}{2}E[\|x^{(t)} - x^{(t-1)}\|^2_{\mathrm{Diag}_{\mathcal{I}}(G)}] \\
&- \frac{\hat{n}\rho}{2}\|Zx^{(t-1)} + By^{(t)}\|^2 \\
&+ (\gamma\rho n(n-\hat{n}) - \frac{(n-\hat{n})\rho}{2})\|B(y^{(t)} - y^{(t-1)})\|^2
\end{aligned}
$$

$$
\begin{aligned}
&-\left(n-\frac{\hat{n}}{2}\right)\|y^{(t)}-y^{(t-1)}\|_Q^2+\frac{\hat{n}}{2}\|y^{(t-1)}-y^*\|_Q^2-\frac{\hat{n}}{2}\|y^{(t)}-y^*\|_Q^2\\
&-\frac{\hat{n}h'}{2}\|B^\top w^*-\nabla\phi(y^{(t)})\|^2. \qquad \text{(S-8)}
\end{aligned}
$$

Note that Eq. (S-8) holds for arbitrary $y^*\in\mathcal{Y}^*$.

Step 3: (*Deriving the assertion*)

(i) Now, since we can take $\nabla\phi(y^{(t)})=B^\top w^{(t-1)}-\rho(Zx^{(t-1)}+By^{(t)})-Q(y^{(t)}-y^{(t-1)})$, it holds that

$$
\|B^\top w^*-\nabla\phi(y^{(t)})\|^2=\|B^\top(w^*-w^{(t-1)})-\rho(Zx^{(t-1)}+By^{(t)})-Q(y^{(t)}-y^{(t-1)})\|^2.
$$

Since $B^\top$ is injective, this gives that

$$
\begin{aligned}
&-\frac{h'}{2}\|B^\top w^*-\nabla\phi(y^{(t)})\|^2\\
\leq&-h'\sigma_{\min}(BB^\top)\|w^*-w^{(t-1)}\|^2+2h'\rho^2\|Zx^{(t-1)}+By^{(t)}\|^2+2h'\|Q(y^{(t)}-y^{(t-1)})\|^2\\
\leq&-h'\sigma_{\min}(BB^\top)\|w^*-w^{(t-1)}\|^2+2h'\rho^2\|Zx^{(t-1)}+By^{(t)}\|^2+2h'\sigma_{\max}(Q)\|y^{(t)}-y^{(t-1)}\|_Q^2.
\end{aligned}
$$

Now, dividing both sides by $\max\{1,4h'\rho,4h'\sigma_{\max}(Q)\}$ (≥ 1), we have

$$
\begin{aligned}
&-\frac{\hat{n}h'}{2}\|B^\top w^*-\nabla\phi(y^{(t)})\|^2\\
\leq&-\frac{\hat{n}h'\sigma_{\min}(BB^\top)}{\max\{1,4h'\rho,4h'\sigma_{\max}(Q)\}}\|w^*-w^{(t-1)}\|^2+\frac{\hat{n}\rho}{2}\|Zx^{(t-1)}+By^{(t)}\|^2+\frac{\hat{n}}{2}\|y^{(t)}-y^{(t-1)}\|_Q^2. \qquad \text{(S-9)}
\end{aligned}
$$

(ii) Next, it holds that, for some $\widehat{y}^*\in\mathcal{Y}^*$,

$$
\frac{1}{2}\left(F(x^*,y^*)-F(x^{(t-1)},y^{(t-1)})\right)\leq-\frac{v'_\phi}{4}\|P_{\mathrm{Ker}(B)}(y^{(t-1)}-\widehat{y}^*)\|^2. \qquad \text{(S-10)}
$$

On the other hand, for arbitrary $a>0$, it follows that

$$
\begin{aligned}
&-\frac{\rho}{8}\|Zx^{(t-1)}+By^{(t-1)}\|^2\\
\leq&-\frac{1}{8}(1-a)\|Z(x^{(t-1)}-x^*)\|^2-\frac{1}{8}(1-a^{-1})\|B(y^{(t-1)}-\widehat{y}^*)\|^2.
\end{aligned}
$$

Thus, setting $a=1+\frac{2v}{\rho\sigma_{\max}(Z^\top Z)}$, we have that

$$
\begin{aligned}
&-\frac{\rho}{8}\|Zx^{(t-1)}+By^{(t-1)}\|^2\\
\leq&\frac{\rho}{8}\frac{2v}{\rho\sigma_{\max}(Z^\top Z)}\sigma_{\max}(Z^\top Z)\|x^{(t-1)}-x^*\|^2\\
&-\frac{\rho}{8}\frac{2v\rho}{\rho\sigma_{\max}(Z^\top Z)+4v}\sigma_{\min}(BB^\top)\|P^\perp_{\mathrm{Ker}(B)}(y^{(t-1)}-\widehat{y}^*)\|^2\\
=&\frac{v}{4}\|x^{(t-1)}-x^*\|^2-\frac{v\rho\sigma_{\min}(BB^\top)}{4(\rho\sigma_{\max}(Z^\top Z)+4v)}\|P^\perp_{\mathrm{Ker}(B)}(y^{(t-1)}-\widehat{y}^*)\|^2. \qquad \text{(S-11)}
\end{aligned}
$$

Combining Eqs. (S-10), (S-11), we have that

$$
\begin{aligned}
&\frac{\hat{n}}{2}\left(F(x^*,y^*)-F(x^{(t-1)},y^{(t-1)})\right)-\frac{\hat{n}\rho}{8n}\|Zx^{(t-1)}+By^{(t-1)}\|^2\\
\leq&\frac{\hat{n}v}{4}\|x^{(t-1)}-x^*\|^2-\frac{\hat{n}}{n}\min\left\{nv'_\phi,\frac{n\rho v\sigma_{\min}(BB^\top)}{\rho\sigma_{\max}(Z^\top Z)+4v}\right\}\frac{\|y^{(t-1)}-\widehat{y}^*\|_Q^2}{4\sigma_{\max}(Q)}\\
\leq&\frac{\hat{n}v}{4}\|x^{(t-1)}-x^*\|^2-\frac{\hat{n}}{n}\min\left\{nv'_\phi,\frac{n\rho v\sigma_{\min}(BB^\top)}{\rho\sigma_{\max}(Z^\top Z)+4v}\right\}\frac{\|y^{(t-1)}-\mathcal{Y}^*\|_Q^2}{4\sigma_{\max}(Q)}. \qquad \text{(S-12)}
\end{aligned}
$$

(iii) By the assumption $\mathrm{Diag}_{\mathcal{I}}(G) \succ 2\gamma\rho(n-\hat{n})\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)$, it holds that

$$\gamma\rho n(n-\hat{n})\mathrm{E}\left[\|x^{(t)}-x^{(t-1)}\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)}\right] - \frac{n}{2}\mathrm{E}[\|x^{(t)}-x^{(t-1)}\|^2_{\mathrm{Diag}_{\mathcal{I}}(G)}] \leq 0. \tag{S-13}$$

(iv) Therefore, if $\gamma = \frac{1}{4n}$, applying Eq. (S-9), Eq. (S-12) and Eq. (S-13) to Eq. (S-8), for

$$\nu = \frac{\hat{n}}{n}\min\left\{\frac{1}{4}\left(\frac{v}{v+\sigma_{\max}(H)}\right), \frac{h'\rho\sigma_{\min}(BB^\top)}{2\max\{1, 4h'\rho, 4h'\sigma_{\max}(Q)\}}, \frac{nv'_\phi/\hat{n}}{4\sigma_{\max}(Q)}, \frac{nv\sigma_{\min}(BB^\top)/\hat{n}}{4\sigma_{\max}(Q)(\rho\sigma_{\max}(Z^\top Z)+4v)}\right\},$$

we have that

$$\begin{aligned}
&\mathrm{E}\Big[F(x^{(t)},y^{(t)}) - F(x^*,y^*) + \frac{1}{2n\gamma\rho}\|w^{(t)}-w^*\|^2 \\
&\quad + \frac{\rho(1-\gamma)}{2}\|Zx^{(t)}+By^{(t)}\|^2 + \frac{1}{2}\|x^{(t)}-x^*\|^2_{vI_p+H} + \frac{\hat{n}}{2n}\|y^{(t)}-y^*\|^2_Q\Big] \\
&\leq (1-\nu)\Big\{F(x^{(t-1)},y^{(t-1)}) - F(x^*,y^*) + \frac{1}{2n\gamma\rho}\|w^{(t-1)}-w^*\|^2 \\
&\quad + \frac{\rho(1-\gamma)}{2}\|Zx^{(t-1)}+By^{(t-1)}\|^2 + \frac{1}{2}\|x^{(t-1)}-x^*\|^2_{vI_p+H} + \frac{\hat{n}}{2n}\|y^{(t-1)}-y^*\|^2_Q\Big\}.
\end{aligned}$$

Setting $\mu := n\nu/\hat{n}$, this gives the assertion.

□

Lemma 3.

$$\begin{aligned}
&\mathrm{E}\left[-\rho\langle Z_{\setminus i}x^{(t)}_{\setminus i} + By^{(t)}, Z_I(x^{(t)}_I - x^*_I)\rangle + \frac{\rho}{2}\|Z_Ix^*_I\|^2 - \frac{\rho}{2}\|Z_Ix^{(t)}_I\|^2\right] \\
\leq &\mathrm{E}\left[-\frac{\rho}{n}\langle Zx^{(t-1)} + By^{(t)}, Z(nx^{(t)} - (n-\hat{n})x^{(t-1)} - x^*)\rangle\right] \\
&+ \frac{\rho}{2n}\|x^{(t-1)}-x^*\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)} - \frac{\rho}{2}\mathrm{E}\left[\|x^{(t)}-x^{(t-1)}\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)}\right].
\end{aligned}$$

Proof.

$$\begin{aligned}
&\rho\langle Z_{\setminus I}x^{(t-1)}_{\setminus I}, Z_I(x^*_I - x^{(t)}_I)\rangle + \rho\langle By^{(t)}, Z_I(x^*_I - x^{(t)}_I)\rangle + \frac{\rho}{2}\|Z_Ix^*_I\|^2 - \frac{\rho}{2}\|Z_Ix^{(t)}_I\|^2 \\
=&\rho\langle Zx^{(t-1)}, Z_I(x^*_I - x^{(t)}_I)\rangle + \rho\langle By^{(t)}, Z_I(x^*_I - x^{(t)}_I)\rangle + \frac{\rho}{2}\|Z_Ix^*_I\|^2 - \frac{\rho}{2}\|Z_Ix^{(t)}_I\|^2 \\
&- \rho\langle Z_Ix^{(t-1)}_i, Z_I(x^*_I - x^{(t)}_I)\rangle \\
=&\rho\langle Zx^{(t-1)}, Z_I(x^*_I - x^{(t-1)}_I + x^{(t-1)}_I - x^{(t)}_I)\rangle + \rho\langle By^{(t)}, Z_I(x^*_I - x^{(t-1)}_I + x^{(t-1)}_I - x^{(t)}_I)\rangle \\
&+ \frac{\rho}{2}\|Z_Ix^*_I\|^2 - \frac{\rho}{2}\|Z_Ix^{(t)}_I\|^2 - \rho\langle Z_Ix^{(t-1)}_I, Z_I(x^*_I - x^{(t)}_I)\rangle \\
=&\rho\langle Zx^{(t-1)} + By^{(t)}, Z_I(x^*_I - x^{(t-1)}_I)\rangle \\
&+ \frac{\rho}{2}\|Z_I(x^{(t-1)}_I - x^*_I)\|^2 - \frac{\rho}{2}\|Z_I(x^{(t)}_I - x^{(t-1)}_I)\|^2 \\
&+ \rho\langle Zx^{(t-1)} + By^{(t)}, Z(x^{(t-1)} - x^{(t)})\rangle.
\end{aligned}$$

The expectation of the RHS is evaluated as

$$\begin{aligned}
&\frac{\rho}{n}\langle Zx^{(t-1)} + By^{(t)}, Z(x^* - x^{(t-1)})\rangle + \frac{\rho}{2n}\|x^{(t-1)} - x^*\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)} \\
&- \frac{\rho}{2}\mathrm{E}[\|Z(x^{(t)} - x^{(t-1)})\|^2] + \rho\mathrm{E}[\langle Zx^{(t-1)} + By^{(t)}, Z(x^{(t-1)} - x^{(t)})\rangle] \\
= &- \frac{\rho}{n}\mathrm{E}[\langle Zx^{(t-1)} + By^{(t)}, Z(nx^{(t)} - (n-\hat{n})x^{(t-1)} - x^*)\rangle] \\
&+ \frac{\rho}{2n}\|x^{(t-1)} - x^*\|^2_{\mathrm{Diag}_{\mathcal{I}}(Z^\top Z)} - \frac{\rho}{2}\mathrm{E}[\|Z(x^{(t)} - x^{(t-1)})\|^2].
\end{aligned}$$

This gives the assertion. □

Lemma 4. *For all $y \in \mathbb{R}^d$ and $y^* \in \mathcal{Y}^*$, we have*

$$\phi(y) - \phi(y^*) \le \langle \nabla\phi(y), y - y^*\rangle - \frac{h'}{2}\|\nabla\phi(y) - \nabla\phi(y^*)\|^2.$$

Proof. By assumption, for all $y^* \in \mathcal{Y}^*$, we have that

$$\begin{aligned}
\phi(y) &= -\phi^*(\nabla\phi(y)) + \langle y, \nabla\phi(y)\rangle \\
&\le -\phi^*(\nabla\phi(y^*)) + \langle y^*, \nabla\phi(y^*) - \nabla\phi(y)\rangle - \frac{h'}{2}\|\nabla\phi(y^*) - \nabla\phi(y)\|^2 + \langle y, \nabla\phi(y)\rangle \\
&= \langle \nabla\phi(y^*), y^*\rangle + \phi(y^*) + \langle y^*, \nabla\phi(y^*) - \nabla\phi(y)\rangle - \frac{h'}{2}\|\nabla\phi(y^*) - \nabla\phi(y)\|^2 + \langle y, \nabla\phi(y)\rangle \\
&= \langle \nabla\phi(y^*), y^*\rangle + \phi(y^*) + \langle y^*, \nabla\phi(y^*) - \nabla\phi(y)\rangle - \frac{h'}{2}\|\nabla\phi(y^*) - \nabla\phi(y)\|^2 + \langle y, \nabla\phi(y)\rangle \\
&= \phi(y^*) + \langle y - y^*, \nabla\phi(y)\rangle - \frac{h'}{2}\|\nabla\phi(y^*) - \nabla\phi(y)\|^2.
\end{aligned}$$

□

C. Auxiliary Lemmas

Lemma 5. *For all symmetric matrix H, we have*

$$(a-b)^\top H(c-b) = \frac{1}{2}\|a-b\|_H^2 - \frac{1}{2}\|a-c\|_H^2 + \frac{1}{2}\|c-b\|_H^2. \tag{S-14}$$

Proof.

$$\begin{aligned}
(a-b)^\top H(c-b) &= \left(a - \frac{c+b}{2} + \frac{c+b}{2} - b\right)^\top H(c-b) \\
&= \left(\frac{a-c}{2} + \frac{a-b}{2}\right)^\top H(c-b) + \left(\frac{c-b}{2}\right)^\top H(c-b) \\
&= \left(\frac{a-c}{2} + \frac{a-b}{2}\right)^\top H\{(a-b) - (a-c)\} + \left(\frac{c-b}{2}\right)^\top H(c-b) \\
&= \frac{1}{2}\|a-b\|_H^2 - \frac{1}{2}\|a-c\|_H^2 + \frac{1}{2}\|c-b\|_H^2.
\end{aligned}$$

□