Supplementary Material for *Zeroth-Order Online Alternating Direction Method of Multipliers: Convergence Analysis and Applications*

Sijia Liu
University of Michigan
IBM Research, Cambridge

Jie Chen
Northwestern Polytechnical
University, China

Pin-Yu Chen
IBM Research,
Yorktown Heights

Alfred O. Hero
University of Michigan

1 Assumptions and Key Notations

Recall that we consider the regularized loss minimization problem over a time horizon of length T,

$$\begin{array}{ll} \underset{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}{\text{minimize}} & \displaystyle\frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x};\mathbf{w}_t) + \phi(\mathbf{y}) \\ \text{subject to} & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{c}. \end{array} \tag{1}$$

ZOO-ADMM is given by

$$\mathbf{x}_{t+1} = \operatorname*{arg\,min}_{\mathbf{x}\in\mathcal{X}} \left\{ \hat{\mathbf{g}}_t^T\mathbf{x} - \boldsymbol{\lambda}_t^T(\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}_t-\mathbf{c}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}_t-\mathbf{c}\|_2^2 + \frac{1}{2\eta_t}\|\mathbf{x}-\mathbf{x}_t\|_{\mathbf{G}_t}^2 \right\}, \tag{2}$$

$$\mathbf{y}_{t+1} = \operatorname*{arg\,min}_{\mathbf{y}\in\mathcal{Y}} \left\{ \phi(\mathbf{y}) - \boldsymbol{\lambda}_t^T(\mathbf{A}\mathbf{x}_{t+1}+\mathbf{B}\mathbf{y}-\mathbf{c}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x}_{t+1}+\mathbf{B}\mathbf{y}-\mathbf{c}\|_2^2 \right\}, \tag{3}$$

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \rho(\mathbf{A}\mathbf{x}_{t+1}+\mathbf{B}\mathbf{y}_{t+1}-\mathbf{c}), \tag{4}$$

where $\mathbf{G}_t = \alpha\mathbf{I} - \rho\eta_t\mathbf{A}^T\mathbf{A}$.

We first elaborate on our assumptions.

- Assumption A implies that $\|\mathbf{x}-\mathbf{x}'\|_2 \le R$ and $\|\mathbf{y}-\mathbf{y}'\|_2 \le R$ for all $\mathbf{x},\mathbf{x}'\in\mathcal{X}$ and for all $\mathbf{y},\mathbf{y}'\in\mathcal{Y}$.
- Based on Jensen's inequality, Assumptions B implies that $\|\mathbb{E}[\nabla_{\mathbf{x}} f(\mathbf{x};\mathbf{w}_t)]\|_2 \le L_1$.
- Assumption C implies a Lipschitz condition over the gradient $\nabla_{\mathbf{x}} f(\mathbf{x};\mathbf{w}_t)$ with constant $L_g(\mathbf{w}_t)$ (Bubeck et al., 2015; Hazan, 2016). Also based on Jensen's inequality, we have $|\mathbb{E}[L_g(\mathbf{w}_t)]| \le L_g$.

We next introduce key notations used in our analysis. Given the primal-dual variables $\mathbf{x}$, $\mathbf{y}$ and $\boldsymbol{\lambda}$ of problem (1), we define $\mathbf{v} := [\mathbf{x}^T, \mathbf{y}^T, \boldsymbol{\lambda}^T]$, and a primal-dual mapping H

$$H(\mathbf{v}) := \mathbf{C}\mathbf{v} - \begin{bmatrix} 0 \\ 0 \\ \mathbf{c} \end{bmatrix}, \ \mathbf{C} := \begin{bmatrix} 0 & 0 & -\mathbf{A}^T \\ 0 & 0 & -\mathbf{B}^T \\ \mathbf{A} & \mathbf{B} & 0 \end{bmatrix}, \tag{5}$$

where $\mathbf{C}$ is skew symmetric, namely, $\mathbf{C}^T = -\mathbf{C}$. An important property of the affine mapping H is that $\langle \mathbf{v}_1 - \mathbf{v}_2, H(\mathbf{v}_1) - H(\mathbf{v}_2)\rangle = 0$ for every $\mathbf{v}_1$ and $\mathbf{v}_2$. Supposing the sequence $\{\mathbf{v}_t\}$ is generated by an algorithm, we introduce the auxiliary sequence

$$\tilde{\mathbf{v}}_t := [\mathbf{x}_t^T, \mathbf{y}_t^T, \tilde{\boldsymbol{\lambda}}_t^T]^T, \tag{6}$$

where $\tilde{\boldsymbol{\lambda}}_t := \boldsymbol{\lambda}_t - \rho(\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{y}_t - \mathbf{c})$.

2 Proof of Theorem 1

Since the sequences $\{\mathbf{x}_t\}$, $\{\mathbf{y}_t\}$ and $\{\boldsymbol{\lambda}_t\}$ produced from (2)-(4) have the same structure as the ADMM/O-ADMM steps, the property of ADMM given by Theorem 4 of (Suzuki, 2013) is directly applicable to our case, yielding

$$
\begin{aligned}
&\sum_{t=1}^{T}(f_t(\mathbf{x}_t)+\phi(\mathbf{y}_t)) - \sum_{t=1}^{T}(f_t(\mathbf{x})+\phi(\mathbf{y})) + \sum_{t=1}^{T}(\tilde{\mathbf{v}}_t-\mathbf{v})^T H(\tilde{\mathbf{v}}_t) \\
\leq & \frac{\|\mathbf{x}_1-\mathbf{x}\|_{\mathbf{G}_1}^2}{2\eta_1} + \sum_{t=2}^{T}\left(\frac{\|\mathbf{x}_t-\mathbf{x}\|_{\mathbf{G}_t}^2}{2\eta_t} - \frac{\|\mathbf{x}_t-\mathbf{x}\|_{\mathbf{G}_{t-1}}^2}{2\eta_{t-1}}\right) + \langle \boldsymbol{\lambda}, \mathbf{A}(\mathbf{x}_{T+1}-\mathbf{x}_1)\rangle \\
&+ \frac{\rho}{2}\|\mathbf{y}_1-\mathbf{y}\|_{\mathbf{B}^T\mathbf{B}} + \frac{\|\boldsymbol{\lambda}_1-\boldsymbol{\lambda}\|_2^2}{2\rho} - \frac{\|\boldsymbol{\lambda}_{T+1}-\boldsymbol{\lambda}\|_2^2}{2\rho} + \langle \mathbf{B}(\mathbf{y}-\mathbf{y}_{T+1}), \boldsymbol{\lambda}_{T+1}-\boldsymbol{\lambda}\rangle \\
&- \langle \mathbf{B}(\mathbf{y}-\mathbf{y}_1), \boldsymbol{\lambda}_1-\boldsymbol{\lambda}\rangle - \sum_{t=1}^{T}\frac{\|\lambda_t-\lambda_{t+1}\|_2^2}{2\rho} - \sum_{t=1}^{T}\frac{\sigma}{2}\|\mathbf{x}_t-\mathbf{x}\|_2^2 + \sum_{t=1}^{T}\frac{\eta_t}{2}\|\hat{\mathbf{g}}_t\|_{\mathbf{G}_t^{-1}}^2.
\end{aligned} \tag{7}
$$

Here for notational simplicity we have used, and henceforth will continue to use, $f_t(\mathbf{x}_t)$ instead of $f(\mathbf{x}_t;\mathbf{w}_t)$.

In (7), based on $\mathbf{G}_t = \alpha\mathbf{I} - \rho\eta_t\mathbf{A}^T\mathbf{A}$, we have

$$
\frac{\|\mathbf{x}_t-\mathbf{x}\|_{\mathbf{G}_t}^2}{2\eta_t} - \frac{\|\mathbf{x}_t-\mathbf{x}\|_{\mathbf{G}_{t-1}}^2}{2\eta_{t-1}} = \left(\frac{\alpha}{2\eta_t} - \frac{\alpha}{2\eta_{t-1}}\right)\|\mathbf{x}_t-\mathbf{x}\|_2^2,
$$

which yields

$$
\sum_{t=2}^{T}\left(\frac{\|\mathbf{x}_t-\mathbf{x}\|_{\mathbf{G}_t}^2}{2\eta_t} - \frac{\|\mathbf{x}_t-\mathbf{x}\|_{\mathbf{G}_{t-1}}^2}{2\eta_{t-1}}\right) - \sum_{t=1}^{T}\frac{\sigma}{2}\|\mathbf{x}_t-\mathbf{x}\|_2^2 \leq \sum_{t=2}^{T}\max\{\frac{\alpha}{2\eta_t} - \frac{\alpha}{2\eta_{t-1}} - \frac{\sigma}{2}, 0\}R^2. \tag{8}
$$

We also note that the terms $\frac{1}{2\eta_1}\|\mathbf{x}_1-\mathbf{x}\|_{\mathbf{G}_1}^2$, $\langle \boldsymbol{\lambda}, \mathbf{A}(\mathbf{x}_{T+1}-\mathbf{x}_1)\rangle$, $\frac{\rho}{2}\|\mathbf{y}_1-\mathbf{y}\|_{\mathbf{B}^T\mathbf{B}}$, $\frac{1}{2\rho}(\|\boldsymbol{\lambda}_1-\boldsymbol{\lambda}\|_2^2 - \|\boldsymbol{\lambda}_{T+1}-\boldsymbol{\lambda}\|_2^2)$, $\langle \mathbf{B}(\mathbf{y}-\mathbf{y}_{T+1}), \boldsymbol{\lambda}_{T+1}-\boldsymbol{\lambda}\rangle$, and $\langle \mathbf{B}(\mathbf{y}-\mathbf{y}_1), \boldsymbol{\lambda}_1-\boldsymbol{\lambda}\rangle$ are *independent* of time t. In particular, we have

$$
\begin{aligned}
&\|\mathbf{x}_1-\mathbf{x}\|_{\mathbf{G}_1}^2 \leq \alpha R^2,\ \langle \boldsymbol{\lambda}, \mathbf{A}(\mathbf{x}_{T+1}-\mathbf{x}_1)\rangle \leq R\|\boldsymbol{\lambda}\|_2\|\mathbf{A}\|_F, \\
&(\|\boldsymbol{\lambda}_1-\boldsymbol{\lambda}\|_2^2 - \|\boldsymbol{\lambda}_{T+1}-\boldsymbol{\lambda}\|_2^2) \leq \|\boldsymbol{\lambda}\|_2^2,\ \langle \mathbf{B}(\mathbf{y}-\mathbf{y}_1), \boldsymbol{\lambda}-\boldsymbol{\lambda}_1\rangle \leq R\|\mathbf{B}\|_F\|\boldsymbol{\lambda}\|_2,
\end{aligned} \tag{9}
$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, and we have used the facts that $\mathbf{G}_t \preceq \alpha\mathbf{I}$ and $\boldsymbol{\lambda}_1 = \mathbf{0}$.

Based on the optimality condition of $\mathbf{y}_{t+1}$ in (3), we have

$$
\langle \partial\phi(\mathbf{y}_{t+1}) - \mathbf{B}^T\boldsymbol{\lambda}_t + \rho\mathbf{B}^T(\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{y}_{t+1} - \mathbf{c}), \mathbf{y}-\mathbf{y}_{t+1}\rangle \geq 0\ , \forall \mathbf{y} \in \mathcal{Y},
$$

which is equivalent to $\langle \partial\phi(\mathbf{y}_{t+1}) - \mathbf{B}^T\boldsymbol{\lambda}_{t+1}, \mathbf{y}-\mathbf{y}_{t+1}\rangle \geq 0$. And thus, we obtain

$$
\langle \boldsymbol{\lambda}_{t+1}, \mathbf{B}(\mathbf{y}-\mathbf{y}_{t+1})\rangle - \langle \boldsymbol{\lambda}, \mathbf{B}(\mathbf{y}-\mathbf{y}_{t+1})\rangle \leq \langle \partial\phi(\mathbf{y}_{t+1}), \mathbf{y}-\mathbf{y}_{t+1}\rangle - \langle \boldsymbol{\lambda}, \mathbf{B}(\mathbf{y}-\mathbf{y}_{t+1})\rangle,
$$

which yields

$$
\langle \mathbf{B}(\mathbf{y}-\mathbf{y}_{t+1}), \boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}\rangle \leq \langle \mathbf{y}-\mathbf{y}_{t+1}, \partial\phi(\mathbf{y}_{t+1}) - \mathbf{B}^T\boldsymbol{\lambda}\rangle \leq R(L_2 + \|\mathbf{B}^T\boldsymbol{\lambda}\|_2), \tag{10}
$$

where we have used the fact that $\|\partial\phi(\mathbf{y}_{t+1})\|_2 \leq L_2$.

Substituting (8)-(10) into (7), we then obtain

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}(f_t(\mathbf{x}_t)+\phi(\mathbf{y}_t)) - \frac{1}{T}\sum_{t=1}^{T}(f_t(\mathbf{x})+\phi(\mathbf{y})) + \frac{1}{T}\sum_{t=1}^{T}(\tilde{\mathbf{v}}_t-\mathbf{v})^T H(\tilde{\mathbf{v}}_t) \\
&+ \frac{1}{T}\sum_{t=1}^{T}\frac{\|\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\|_2^2}{2\rho} \leq \frac{1}{T}\sum_{t=2}^{T}\max\{\frac{\alpha}{2\eta_t} - \frac{\alpha}{2\eta_{t-1}} - \frac{\sigma}{2}, 0\}R^2 + \frac{1}{T}\sum_{t=1}^{T}\frac{\eta_t}{2}\|\hat{\mathbf{g}}_t\|^2 + \frac{K}{T},
\end{aligned} \tag{11}
$$

where K is a constant term related to α, R, η_1, $\mathbf{A}$, $\mathbf{B}$, $\boldsymbol{\lambda}$, ρ and L_2, $K = \frac{\alpha R^2}{2\eta_1} + R\|\boldsymbol{\lambda}\|_2\|\mathbf{A}\|_F + \frac{1}{2\rho}\|\boldsymbol{\lambda}\|_2^2 + R\|\mathbf{B}\|_F\|\boldsymbol{\lambda}\|_2 + R(L_2 + \|\mathbf{B}^T\boldsymbol{\lambda}\|_2)$, and we have used the fact that $\|\hat{\mathbf{g}}_t\|^2_{\mathbf{G}_t^{-1}} \leq \|\hat{\mathbf{g}}_t\|_2^2$ (due to $\mathbf{G}_t^{-1} \preceq \mathbf{I}$).

Based on (11) we continue to prove Theorem 1. When $\mathbf{B}$ is invertible and $\mathbf{y}_t' = \mathbf{B}^{-1}(\mathbf{c} - \mathbf{A}\mathbf{x}_t)$, we obtain

$$\mathbf{B}(\mathbf{y}_t' - \mathbf{y}_t) = \frac{1}{\rho}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}). \tag{12}$$

Based on the convexity of f and ϕ, we obtain

$$\begin{aligned} f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t') &\leq f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t) + \langle \partial\phi(\mathbf{y}_t'), \mathbf{y}_t' - \mathbf{y}_t \rangle \\ &= f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t) + \frac{1}{\rho}\langle (\mathbf{B}^{-1})^T \partial\phi(\mathbf{y}_t'), \boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1} \rangle, \end{aligned} \tag{13}$$

where the last equality holds due to (12).

Let $(\mathbf{x}^*, \mathbf{y}^*)$ be the optimal solution (implying $\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{y}^* - \mathbf{c} = \mathbf{0}$). For any dual variable $\boldsymbol{\lambda}^*$ and $\tilde{\mathbf{v}}_t = [\mathbf{x}_t^T, \mathbf{y}_t^T, \tilde{\boldsymbol{\lambda}}_t^T]^T$, we have

$$\begin{aligned} (\tilde{\mathbf{v}}_t - \mathbf{v}^*)^T H(\tilde{\mathbf{v}}_t) = H(\mathbf{v}^*)^T(\tilde{\mathbf{v}}_t - \mathbf{v}^*) = \begin{bmatrix} -\mathbf{A}^T\boldsymbol{\lambda}^* \\ -\mathbf{B}^T\boldsymbol{\lambda}^* \\ \mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{y}^* - \mathbf{c} \end{bmatrix}^T \begin{bmatrix} \mathbf{x}_t - \mathbf{x}^* \\ \mathbf{y}_t - \mathbf{y}^* \\ \tilde{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}^* \end{bmatrix} \\ = \langle \boldsymbol{\lambda}^*, \mathbf{c} - \mathbf{A}\mathbf{x}_t - \mathbf{B}\mathbf{y}_t \rangle = \frac{1}{\rho}\langle \boldsymbol{\lambda}^*, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1} \rangle \end{aligned} \tag{14}$$

where $\mathbf{v}^* := [(\mathbf{x}^*)^T, (\mathbf{y}^*)^T, (\boldsymbol{\lambda}^*)^T]^T$, and the affine mapping $H(\cdot)$ is given by (5).

Setting $\boldsymbol{\lambda}^* = (\mathbf{B}^{-1})^T\partial\phi(\mathbf{y}_t')$, based on (13) and (14) we have

$$\begin{aligned} & f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t') - (f_t(\mathbf{x}^*) + \phi(\mathbf{y}^*)) \\ \leq & f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t) + (\tilde{\mathbf{v}}_t - \mathbf{v}^*)^T H(\tilde{\mathbf{v}}_t) - (f_t(\mathbf{x}^*) + \phi(\mathbf{y}^*)). \end{aligned} \tag{15}$$

Combining (11) and (15) yields

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^T (f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t')) - \frac{1}{T}\sum_{t=1}^T (f_t(\mathbf{x}^*) + \phi(\mathbf{y}^*)) + \frac{1}{T}\sum_{t=1}^T \frac{\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t\|_2^2}{2\rho} \\ \leq & \frac{1}{T}\sum_{t=1}^T (f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t)) - \frac{1}{T}\sum_{t=1}^T (f_t(\mathbf{x}^*) + \phi(\mathbf{y}^*)) + \frac{1}{T}\sum_{t=1}^T (\tilde{\mathbf{v}}_t - \mathbf{v}^*)^T H(\tilde{\mathbf{v}}_t) + \frac{1}{T}\sum_{t=1}^T \frac{\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t\|_2^2}{2\rho} \\ \leq & \frac{1}{T}\sum_{t=2}^T \max\{\frac{\alpha}{2\eta_t} - \frac{\alpha}{2\eta_{t-1}} - \frac{\sigma}{2}, 0\}R^2 + \frac{1}{T}\sum_{t=1}^T \frac{\eta_t}{2}\|\hat{\mathbf{g}}_t\|_2^2 + \frac{K}{T}. \end{aligned} \tag{16}$$

Since $\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t = \rho(\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{y}_{t+1} - \mathbf{c})$, from (16) we have

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^T (f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t')) - \frac{1}{T}\sum_{t=1}^T (f_t(\mathbf{x}^*) + \phi(\mathbf{y}^*)) + \frac{\rho}{2T}\sum_{t=1}^T \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{y}_{t+1} - \mathbf{c}\|_2^2 \\ \leq & \frac{1}{T}\sum_{t=2}^T \max\{\frac{\alpha}{2\eta_t} - \frac{\alpha}{2\eta_{t-1}} - \frac{\sigma}{2}, 0\}R^2 + \frac{1}{T}\sum_{t=1}^T \frac{\eta_t}{2}\|\hat{\mathbf{g}}_t\|_2^2 + \frac{K}{T}. \end{aligned} \tag{17}$$

Taking expectations for both sides of (17) with respect to its randomness, we have

$$\begin{aligned} & \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T (f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t')) - \frac{1}{T}\sum_{t=1}^T (f_t(\mathbf{x}^*) + \phi(\mathbf{y}^*))\right] + \mathbb{E}\left[\frac{\rho}{2T}\sum_{t=1}^T \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{y}_{t+1} - \mathbf{c}\|_2^2\right] \\ \leq & \frac{1}{T}\sum_{t=2}^T \max\{\frac{\alpha}{2\eta_t} - \frac{\alpha}{2\eta_{t-1}} - \frac{\sigma}{2}, 0\}R^2 + \frac{1}{T}\sum_{t=1}^T \frac{\eta_t}{2}\mathbb{E}[\|\hat{\mathbf{g}}_t\|_2^2] + \frac{K}{T}. \end{aligned} \tag{18}$$

Based on (Duchi et al., 2015, Lemma 1), the second-order statistics of the gradient estimate $\hat{\mathbf{g}}_t$ is given by

$$\mathbb{E}_{\mathbf{z}_t}[\hat{\mathbf{g}}_t] = \mathbf{g}_t + \beta_t L_g(\mathbf{w}_t)\nu(\mathbf{x}_t, \beta_t), \tag{19}$$

$$\mathbb{E}_{\mathbf{z}_t}[\|\hat{\mathbf{g}}_t\|_2^2] \leq 2s(m)\|\mathbf{g}_t\|_2^2 + \frac{1}{2}\beta_t^2 L_g(\mathbf{w}_t)^2 M(\mu)^2, \tag{20}$$

where $\mathbf{g}_t = \nabla_{\mathbf{x}} f(\mathbf{x}; \mathbf{w}_t)|_{\mathbf{x}=\mathbf{x}_t}$, $\|\nu(\mathbf{x}_t, \beta_t)\|_2 \leq \frac{1}{2}\mathbb{E}_{\mathbf{z}}[\|\mathbf{z}\|_2^3]$, $L_g(\mathbf{w}_t)$ is defined in Assumption C, and $s(m)$ and $M(\mu)$ are introduced in Assumption E. According to (20), we have

$$\begin{aligned}\mathbb{E}[\|\hat{\mathbf{g}}_t\|_2^2] = \mathbb{E}\left[\mathbb{E}_{\mathbf{z}}[\|\hat{\mathbf{g}}_t\|_2^2]\right] \leq & \mathbb{E}\left[2s(m)\|\mathbf{g}_t\|_2^2 + \frac{1}{2}\beta_t^2 L_{g,t}^2 M(\mu)^2\right] \\ \leq & 2s(m)L_1^2 + \frac{1}{2}\beta_t^2 L_g^2 M(\mu)^2,\end{aligned} \tag{21}$$

where for ease of notation, we have replaced $L_g(\mathbf{w}_t)$ with $L_{g,t}$, and the last inequality holds due to Assumptions B and C.

Substituting (21) into (18), the expected average regret can be bounded as

$$\overline{\text{Regret}}_T(\mathbf{x}_t, \mathbf{y}_t', \mathbf{x}^*, \mathbf{y}^*) \leq \frac{1}{T}\sum_{t=2}^T \max\{\frac{\alpha}{2\eta_t} - \frac{\alpha}{2\eta_{t-1}} - \frac{\sigma}{2}, 0\}R^2 + \frac{s(m)L_1^2}{T}\sum_{t=1}^T \eta_t + \frac{M(\mu)^2 L_g^2}{4T}\sum_{t=1}^T \eta_t\beta_t^2 + \frac{K}{T}. \tag{22}$$

On the other hand, when $\mathbf{A}$ is invertible and $\mathbf{x}_t' = \mathbf{A}^{-1}(\mathbf{c} - \mathbf{B}\mathbf{y}_t)$, we obtain

$$\mathbf{A}(\mathbf{x}_t' - \mathbf{x}_t) = \frac{1}{\rho}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}).$$

Based on the convexity of f and ϕ, we obtain

$$\begin{aligned}f_t(\mathbf{x}_t') + \phi(\mathbf{y}_t) \leq & f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t) + \langle \nabla f_t(\mathbf{x}_t'), \mathbf{x}_t' - \mathbf{x}_t\rangle \\ = & f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t) + \frac{1}{\rho}\langle (\mathbf{A}^{-1})^T \nabla f_t(\mathbf{x}_t'), \boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\rangle.\end{aligned} \tag{23}$$

Setting $\boldsymbol{\lambda}^* = (\mathbf{A}^{-1})^T \nabla f_t(\mathbf{x}_t')$, based on (23) and (14) we have

$$\begin{aligned}& f_t(\mathbf{x}_t') + \phi(\mathbf{y}_t) - (f_t(\mathbf{x}^*) + \phi(\mathbf{y}^*)) \\ \leq & f_t(\mathbf{x}_t) + \phi(\mathbf{y}_t) + (\tilde{\mathbf{v}}_t - \mathbf{v}^*)^T H(\tilde{\mathbf{v}}_t) - (f_t(\mathbf{x}^*) + \phi(\mathbf{y}^*)).\end{aligned} \tag{24}$$

Since the right hand side (RHS) of (24) and RHS of (15) are same, we can then mimic the aforementioned procedure to prove that the regret $\overline{\text{Regret}}_T(\mathbf{x}_t', \mathbf{y}_t, \mathbf{x}^*, \mathbf{y}^*)$ obeys the same bounds as (22).

3 Simplification of Regret Bound

Consider terms in right hand side (RHS) of (22) together with $\eta_t = \frac{C_1}{\sqrt{s(m)}\sqrt{t}}$ and $\beta_t = \frac{C_2}{M(\mu)t}$, we have

$$\begin{aligned}& \frac{1}{T}\sum_{t=2}^T \max\{\frac{\alpha}{2\eta_t} - \frac{\alpha}{2\eta_{t-1}} - \frac{\sigma}{2}, 0\}R^2 \leq \frac{1}{T}\sum_{t=2}^T (\frac{\alpha}{2\eta_t} - \frac{\alpha}{2\eta_{t-1}})R^2 \leq \frac{1}{\sqrt{T}}\frac{\alpha R^2\sqrt{s(m)}}{2C_1}, \\ & \frac{s(m)L_1^2}{T}\sum_{t=1}^T \eta_t \leq \frac{2C_1\sqrt{s(m)}L_1^2}{\sqrt{T}}, \\ & \frac{M(\mu)^2 L_g^2}{4T}\sum_{t=1}^T \eta_t\beta_t^2 = \frac{C_1 C_2^2 L_g^2}{4\sqrt{s(m)}T}\sum_{t=1}^T \frac{1}{t^{5/2}} \leq \frac{5C_1C_2^2L_g^2}{12T},\end{aligned} \tag{25}$$

where we have used the facts that $\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$,

$$\sum_{t=1}^T (1/t^a) = 1 + \sum_{t=2}^T (1/t^a) \leq 1 + \int_1^\infty (1/t^a) = a/(a-1), \ \forall a > 1, \tag{26}$$

and we recall that $s(m) = m \geq 1$. Substituting (25) into RHS of (22), we conclude that the expected average regret $\overline{\mathrm{Regret}}_T(\mathbf{x}_t, \mathbf{y}_t', \mathbf{x}^*, \mathbf{y}^*)$ is upper bounded by

$$\frac{1}{\sqrt{T}} \frac{\alpha R^2 \sqrt{s(m)}}{2C_1} + \frac{2C_1 \sqrt{s(m)} L_1^2}{\sqrt{T}} + \frac{5C_1 C_2^2 L_g^2}{12T} + \frac{K}{T}. \quad (27)$$

4 Proof of Corollary 1

Given i.i.d. samples $\{\mathbf{w}_t\}$ drawn from the probability distribution P, from Theorem 1 we have

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}_t;\mathbf{w}_t)+\phi(\mathbf{y}_t')\right)-\frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}^*;\mathbf{w}_t)+\phi(\mathbf{y}^*)\right)\right] \\ \leq &\frac{1}{\sqrt{T}}\frac{\alpha R^2\sqrt{s(m)}}{2C_1}+\frac{2C_1\sqrt{s(m)}L_1^2}{\sqrt{T}}+\frac{5C_1C_2^2L_g^2}{12}\frac{1}{T}+\frac{K}{T}. \end{aligned} \quad (28)$$

Based on $F(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\mathbf{w}}[f(\mathbf{x}; \mathbf{w})] + \phi(\mathbf{y})$, from (28) we have

$$\begin{aligned} &\mathbb{E}\left[F(\bar{\mathbf{x}}_t,\bar{\mathbf{y}}_t)-F(\mathbf{x}^*,\mathbf{y}^*)\right] \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}F(\mathbf{x}_t,\mathbf{y}_t)-F(\mathbf{x}^*,\mathbf{y}^*)\right] \\ =&\mathbb{E}_{\mathbf{z}_{1:T}}\left[\mathbb{E}_{\mathbf{w}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}_t;\mathbf{w}_t)+\phi(\mathbf{y}_t')\right)-\frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}^*;\mathbf{w}_t)+\phi(\mathbf{y}^*)\right)\right]\right] \\ \leq&\frac{1}{\sqrt{T}}\frac{\alpha R^2\sqrt{s(m)}}{2C_1}+\frac{2C_1\sqrt{s(m)}L_1^2}{\sqrt{T}}+\frac{5C_1C_2^2L_g^2}{12}\frac{1}{T}+\frac{K}{T}, \end{aligned} \quad (29)$$

where the first inequality holds due to the convexity of F, and the second equality holds since $\mathbf{x}_t$ and $\mathbf{y}_t$ are implicit functions of i.i.d. random variables $\{\mathbf{w}_k\}_{k=1}^{t-1}$ and $\{\mathbf{z}_k\}_{k=1}^{t-1}$, and $\{\mathbf{w}_t\}$ and $\{\mathbf{z}_t\}$ are independent of each other.

5 Proof of Corollary 2

Substituting $\eta_t = \frac{\alpha}{\sigma t}$ and $\beta_t = \frac{C_2}{M(\mu)t}$ into RHS of (22), we have

$$\begin{aligned} &\frac{1}{T}\sum_{t=2}^{T}\max\{\frac{\alpha}{2\eta_t}-\frac{\alpha}{2\eta_{t-1}}-\frac{\sigma}{2},0\}R^2=0, \quad \frac{s(m)L_1^2}{T}\sum_{t=1}^{T}\eta_t\leq\frac{\alpha s(m)L_1^2\log T}{\sigma T}, \\ &\frac{M(\mu)^2L_g^2}{4T}\sum_{t=1}^{T}\eta_t\beta_t^2=\frac{\alpha C_2^2L_g^2}{4\sigma T}\sum_{t=1}^{T}\frac{1}{t^3}\leq\frac{3\alpha C_2^2L_g^2}{8\sigma T}, \end{aligned} \quad (30)$$

where we have used the facts that $\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \log T$ and (26). Based on (30) and (27), we complete the proof.

6 Proof of Corollary 3

We consider the hybrid minibatch strategy

$$\hat{\mathbf{g}}_t = \frac{1}{q_1 q_2} \sum_{j=1}^{q_1} \sum_{i=1}^{q_2} \frac{f(\mathbf{x}_t + \beta_t \mathbf{z}_{t,j}; \mathbf{w}_{t,i}) - f(\mathbf{x}_t; \mathbf{w}_{t,i})}{\beta_t} \mathbf{z}_{t,j} \quad (31)$$

with $\hat{\mathbf{g}}_{t,ij} := \frac{f(\mathbf{x}_t+\beta_t\mathbf{z}_{t,j};\mathbf{w}_{t,i})-f(\mathbf{x}_t;\mathbf{w}_{t,i})}{\beta_t}\mathbf{z}_{t,j}$. Based on (19) and i.i.d. samples $\{\mathbf{w}_{t,i}\}$ and $\{\mathbf{z}_{t,j}\}$, we have

$$\bar{\mathbf{g}}_t := \mathbb{E}[\hat{\mathbf{g}}_{t,ij}] = \mathbb{E}[\mathbf{g}_t] + \beta_t \mathbb{E}[L_{g,t} \nu(\mathbf{x}_t, \beta_t)], \ \forall i, j. \quad (32)$$

where for ease of notation we have replaced $L_g(\mathbf{w}_t)$ with $L_{g,t}$, $\|\nu(\mathbf{x}_t,\beta_t)\|_2 \le \frac{1}{2}\mathbb{E}[\|\mathbf{z}\|_2^3] \le M(\mu)$ due to Assumption E. From (31), we obtain

$$\begin{aligned}\mathbb{E}[\|\hat{\mathbf{g}}_t\|_2^2] =& \mathbb{E}\left[\left\|\frac{1}{q_1q_2}\sum_{i=1}^{q_1}\sum_{j=1}^{q_2}(\hat{\mathbf{g}}_{t,ij}-\bar{\mathbf{g}}_t)+\bar{\mathbf{g}}_t\right\|_2^2\right] = \|\bar{\mathbf{g}}_t\|_2^2 + \mathbb{E}\left[\left\|\frac{1}{q_1q_2}\sum_{i=1}^{q_1}\sum_{j=1}^{q_2}(\hat{\mathbf{g}}_{t,ij}-\bar{\mathbf{g}}_t)\right\|_2^2\right]\\ =&\|\bar{\mathbf{g}}_t\|_2^2 + \frac{1}{q_1q_2}\mathbb{E}[\|\hat{\mathbf{g}}_{t,11}-\bar{\mathbf{g}}_t\|_2^2] = \|\bar{\mathbf{g}}_t\|^2 + \frac{1}{q_1q_2}\mathbb{E}[\|\hat{\mathbf{g}}_{t,11}\|^2] - \frac{1}{q_1q_2}\|\bar{\mathbf{g}}_t\|^2,\end{aligned} \tag{33}$$

where we have used the fact that $\mathbb{E}[\hat{\mathbf{g}}_{t,ij}] = \mathbb{E}[\hat{\mathbf{g}}_{t,11}]$ for any i and j.

The definition of $\bar{\mathbf{g}}_t$ in (32) yields

$$\begin{aligned}\|\bar{\mathbf{g}}_t\|^2 \le& 2\|\mathbb{E}[\mathbf{g}_t]\|_2^2 + 2\|\beta_t\mathbb{E}[L_{g,t}\nu(\mathbf{x}_t,\beta_t)]\|_2^2\\ \le& 2\mathbb{E}[\|\mathbf{g}_t\|_2^2] + 2\beta_t^2\mathbb{E}[L_{g,t}^2]\mathbb{E}[\|\nu(\mathbf{x}_t,\beta_t)\|_2^2] \le 2\mathbb{E}[\|\mathbf{g}_t\|_2^2] + \frac{1}{2}\beta_t^2L_g^2M(\mu)^2,\end{aligned} \tag{34}$$

where the first inequality holds due to Cauchy-Schwarz inequality, and the second inequality holds due to Jensen's inequality. From (20), we obtain

$$\mathbb{E}[\|\hat{\mathbf{g}}_{t,11}\|^2] \le 2s(m)\mathbb{E}[\|\mathbf{g}_t\|_2^2] + \frac{1}{2}\beta_t^2L_g^2M(\mu)^2. \tag{35}$$

Substituting (34) and (35) into (33), we obtain

$$\mathbb{E}[\|\hat{\mathbf{g}}_t\|_2^2] \le \|\bar{\mathbf{g}}_t\|_2^2 + \frac{1}{q_1q_2}\mathbb{E}[\|\hat{\mathbf{g}}_{t,11}\|_2^2] \le 2(1+\frac{s(m)}{q_1q_2})\mathbb{E}[\|\mathbf{g}_t\|_2^2] + \frac{q_1q_2+1}{2q_1q_2}\beta_t^2L_g^2M(\mu)^2. \tag{36}$$

Similar to proof of Theorem 1, substituting (36) into (18), we obtain

$$\begin{aligned}\overline{\mathrm{Regret}}_T(\mathbf{x}_t,\mathbf{y}_t',\mathbf{x}^*,\mathbf{y}^*) \le& \frac{1}{T}\sum_{t=2}^{T}\max\{\frac{\alpha}{2\eta_t}-\frac{\alpha}{2\eta_{t-1}}-\frac{\sigma}{2},0\}R^2 + \frac{1}{T}\sum_{t=1}^{T}\frac{\eta_t}{2}\mathbb{E}[\|\hat{\mathbf{g}}_t\|_2^2] + \frac{K}{T}\\ \le& \frac{1}{T}\sum_{t=2}^{T}\max\{\frac{\alpha}{2\eta_t}-\frac{\alpha}{2\eta_{t-1}}-\frac{\sigma}{2},0\}R^2 + \frac{(q_1q_2+s(m))L_1^2}{q_1q_2T}\sum_{t=1}^{T}\eta_t\\ &+\frac{(q_1q_2+1)L_g^2M(\mu)^2}{4q_1q_2T}\sum_{t=1}^{T}\eta_t\beta_t^2 + \frac{K}{T}.\end{aligned} \tag{37}$$

Substituting $\eta_t = \frac{C_1}{\sqrt{1+\frac{s(m)}{q_1q_2}}\sqrt{t}}$ and $\beta_t = \frac{C_2}{M(\mu)t}$ into (37), we obtain

$$\begin{aligned}&\overline{\mathrm{Regret}}_T(\mathbf{x}_t,\mathbf{y}_t',\mathbf{x}^*,\mathbf{y}^*)\\ \le&\frac{\alpha R^2}{2C_1}\frac{\sqrt{1+\frac{s(m)}{q_1q_2}}}{\sqrt{T}} + 2C_1L_1^2\frac{\sqrt{1+\frac{s(m)}{q_1q_2}}}{\sqrt{T}} + \frac{5C_1C_2^2L_g^2}{12T}\frac{q_1q_2+1}{q_1q_2\sqrt{1+\frac{s(m)}{q_1q_2}}} + \frac{K}{T}\\ \le&\frac{\alpha R^2}{2C_1}\frac{\sqrt{1+\frac{s(m)}{q_1q_2}}}{\sqrt{T}} + 2C_1L_1^2\frac{\sqrt{1+\frac{s(m)}{q_1q_2}}}{\sqrt{T}} + \frac{5C_1C_2^2L_g^2}{6}\frac{1}{T} + \frac{K}{T},\end{aligned} \tag{38}$$

which then completes the proof.

7 ZOO-ADMM for Sensor Selection

We recall that the sensor selection problem can be cast as

$$\begin{aligned}\underset{\mathbf{x},\mathbf{y}}{\text{minimize}} \quad & \frac{1}{T}\sum_{t=1}^{T}f(\mathbf{x};\mathbf{w}_t) + \mathcal{I}_1(\mathbf{x}) + \mathcal{I}_2(\mathbf{y})\\ \text{subject to} \quad & \mathbf{x}-\mathbf{y}=\mathbf{0},\end{aligned} \tag{39}$$

where $\mathbf{y} \in \mathbb{R}^m$ is an auxiliary variable, $f(\mathbf{x}; \mathbf{w}_t) = -\mathrm{logdet}(\sum_{i=1}^m x_i \mathbf{a}_{i,t} \mathbf{a}_{i,t}^T)$ with $\mathbf{w}_t = \{\mathbf{a}_{i,t}\}_{i=1}^m$, and $\{\mathcal{I}_i\}$ are indicator functions

$$\mathcal{I}_1(\mathbf{x}) = \begin{cases} 0 & \mathbf{0} \leq \mathbf{x} \leq \mathbf{1} \\ \infty & \text{otherwise,} \end{cases} \quad \mathcal{I}_2(\mathbf{y}) = \begin{cases} 0 & \mathbf{1}^T \mathbf{y} = m_0 \\ \infty & \text{otherwise.} \end{cases}$$

Based on (39), two key steps of ZOO-ADMM (2)-(3) are given by

$$\mathbf{x}_{t+1} = \operatorname*{arg\,min}_{\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}} \left\{ \|\mathbf{x} - \mathbf{d}_t\|_2^2 \right\}, \tag{40}$$

$$\mathbf{y}_{t+1} = \operatorname*{arg\,min}_{\mathbf{1}^T \mathbf{y} = m_0} \left\{ \|\mathbf{y} - (\mathbf{x}_{t+1} - (1/\rho)\boldsymbol{\lambda}_t)\|_2^2 \right\}, \tag{41}$$

where $\hat{\mathbf{g}}_t$ is the gradient estimate, and $\mathbf{d}_t := \frac{\eta_t}{\alpha}(-\hat{\mathbf{g}}_t + \boldsymbol{\lambda}_t - \rho \mathbf{x}_t + \rho \mathbf{y}_t) + \mathbf{x}_t$. Sub-problems (40) and (41) yield closed-form solutions as below (Parikh and Boyd, 2014)

$$[\mathbf{x}_{t+1}]_i = \begin{cases} 0 & [\mathbf{d}_t]_i < 0 \\ [\mathbf{d}_t]_i & [\mathbf{d}_t]_i \in [0, 1] \\ 1 & [\mathbf{d}_t]_i > 1, \end{cases} \quad \text{and} \tag{42}$$

$$\mathbf{y}_{t+1} = \mathbf{x}_{t+1} - \frac{1}{\rho}\boldsymbol{\lambda}_t + \frac{m_0 - \mathbf{1}^T (\mathbf{x}_{t+1} - \boldsymbol{\lambda}_t/\rho)}{m} \mathbf{1}_m, \tag{43}$$

where $[\mathbf{x}]_i$ denote the ith entry of $\mathbf{x}$.

8 ZOO-ADMM for Sparse Cox Regression

This sparse regression problem can formulated as

$$\begin{array}{ll} \underset{\mathbf{x}, \mathbf{y}}{\text{minimize}} & \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}; \mathbf{w}_i) + \gamma \|\mathbf{y}\|_1 \\ \text{subject to} & \mathbf{x} - \mathbf{y} = \mathbf{0}, \end{array} \tag{44}$$

where $f(\mathbf{x}; \mathbf{w}_i) = \delta_i \left\{ -\mathbf{a}_i^T \mathbf{x} + \log\left(\sum_{j \in \mathcal{R}_i} e^{\mathbf{a}_j^T \mathbf{x}}\right) \right\}$ with $\mathbf{w}_i = \mathbf{a}_i$. By using the ZOO-ADMM algorithm, we can avoid the gradient calculation for the involved objective function in Cox regression. The two key steps of ZOO-ADMM (2)-(3) at iteration i become

$$\mathbf{x}_{i+1} = \frac{\eta_t}{\alpha}(-\hat{\mathbf{g}}_i + \boldsymbol{\lambda}_i - \rho \mathbf{x}_i + \rho \mathbf{y}_i) + \mathbf{x}_i, \tag{45}$$

$$\mathbf{y}_{i+1} = \operatorname*{arg\,min}_{\mathbf{y}} \left\{ \|\mathbf{y}\|_1 + \frac{\rho}{2\gamma} \|\mathbf{y} - \mathbf{d}_i\|_2^2 \right\}, \tag{46}$$

where $\hat{\mathbf{g}}_i$ is the gradient estimate, $\mathbf{d}_i = (\mathbf{x}_{i+1} - (1/\rho)\boldsymbol{\lambda}_i)$, and the solution of sub-problem (46) is given by the soft-thresholding operator at the point $\mathbf{d}_i$ with parameter ρ/γ (Parikh and Boyd, 2014, Sec. 6)

$$[\mathbf{y}_{i+1}]_k = \begin{cases} (1 - \frac{\gamma}{\rho |[\mathbf{d}_i]_k|})[\mathbf{d}_i]_k & [\mathbf{d}_i]_k > \frac{\gamma}{\rho} \\ 0 & [\mathbf{d}_i]_k \leq \frac{\gamma}{\rho}, \end{cases} \quad k = 1, 2, \ldots, m.$$

References

S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4): 157–325, 2016.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

T. Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *International Conference on Machine Learning*, pages 392–400, 2013.