7 Appendix

7.1 Proof for Theorem 1

Let $g(M)$ be a smooth function such that $\nabla g(M)$ is Lipschitz-continuous with parameter ρ, that is,

$$g(M') - g(M) - \langle \nabla g(M), M' - M \rangle \leq \frac{\rho}{2}\|M' - M\|_F^2.$$

Then $\nabla_j f(c) = z_j^T \nabla g(M) z_j$ is Lipschitz-continuous with parameter γ, a number of order $O(1)$ when $g(.)$ is an empirical risk normalized by N. Let $\mathcal{A}$ be the active set before adding a component $\hat{j}$. Consider the descent amount produced by minimizing $F(c)$ w.r.t. the $c_{\hat{j}}$ given that $0 \in \partial_j F(c)$ for all $j \in \mathcal{A}$ due to the subproblem in the previous iteration. Let $j = \hat{j}$, for any η_j we have

$$\begin{aligned} F(c + \eta_j e_j) - F(c) &\leq \nabla_j f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \\ &\leq \mu \nabla_{j^*} f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \\ &\leq \mu \nabla_{j^*} f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \end{aligned}$$

Minimize w.r.t η_j gives

$$\begin{aligned} &\min_{\eta_j} \; F(c + \eta_j e_j) - F(c) \\ &\leq \min_{\eta_j} \; \mu \nabla_{j^*} f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \\ &\leq \min_{\eta_j} \; \mu \nabla_{j^*} f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \\ &= \min_{\eta_k : k \notin \mathcal{A}} \; \sum_{k \notin \mathcal{A}} \Big(\mu \nabla_k f(c)\eta_k + \lambda|\eta_k| \Big) + \frac{\gamma}{2} \left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 \\ &\leq \min_{\eta_k : k \notin \mathcal{A}} \; \mu \sum_{k \notin \mathcal{A}} \Big(\nabla_k f(c)\eta_k + \lambda|\eta_k| \Big) + \frac{\gamma}{2} \left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 \\ &+ (1 - \mu)\lambda \sum_{k \notin \mathcal{A}} |\eta_k| \end{aligned}$$

where the last equality is justified by Lemma 2 provided later. For $k \in \mathcal{A}$, we have

$$0 = \min_{\eta_k : k \in \mathcal{A}} \; \mu \sum_{k \in \mathcal{A}} \left(\nabla_k f(c)\eta_k + \lambda|c_k + \eta_k| - \lambda|c_k| \right)$$

Combining cases for $k \notin \mathcal{A}$ and $k \in \mathcal{A}$, we can obtain a global estimate of descent amount compared to some optimal solution x^* as follows

$$\begin{aligned} &\min_{\eta_{\hat{j}}} \; F(c + \eta_{\hat{j}} e_{\hat{j}}) - F(c) \\ &\leq \min_{\eta} \; \mu \Big(\langle \nabla f(c), \eta \rangle + \lambda\|c + \eta\|_1 - \lambda\|c\|_1 \Big) \\ &+ \frac{\gamma}{2} \left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 + (1 - \mu)\lambda \sum_{k \notin \mathcal{A}} |\eta_k| \\ &\leq \min_{\eta} \; \mu \Big(F(c + \eta) - F(c) \Big) + \frac{\gamma}{2} \left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 + (1 - \mu)\lambda \sum_{k \notin \mathcal{A}} |\eta_k| \\ &\leq \min_{\alpha \in [0,1]} \; \mu \Big(F(c + \alpha(c^* - c)) - F(c) \Big) + \frac{\alpha\gamma}{2}\|c^*\|_1^2 + \alpha(1 - \mu)\lambda\|c^*\|_1 \\ &\leq \min_{\alpha \in [0,1]} \; -\alpha\mu \Big(F(c) - F(c^*) \Big) + \frac{\alpha^2\gamma}{2}\|c^*\|_1^2 + \alpha(1 - \mu)\lambda\|c^*\|_1. \end{aligned}$$

It means we can always choose an α small enough to guarantee descent if

$$F(c) - F(c^*) > \frac{(1-\mu)}{\mu}\lambda\|c^*\|_1. \tag{20}$$

Then for

$$F(c) - F(c^*) \geq \frac{2(1-\mu)}{\mu}\lambda\|c^*\|_1, \tag{21}$$

we have

$$\begin{aligned}
&\min_{\eta_{\hat{j}}} \; F(c + \eta_{\hat{j}} e_{\hat{j}}) - F(c) \\
&\leq \min_{\alpha\in[0,1]} \; -\frac{\alpha\mu}{2}\Big(F(c) - F(c^*)\Big) + \frac{\alpha^2\gamma}{2}\|c^*\|_1^2.
\end{aligned}$$

Minimizing w.r.t. to α gives the convergence guarantee

$$F(c^t) - F(c^*) \leq \frac{2\gamma\|c^*\|_1^2}{\mu^2}\frac{1}{t}.$$

for any iterate with $F(c^t) - F(c^*) \geq \frac{2(1-\mu)}{\mu}\lambda\|c^*\|_1$.

Lemma 2.

$$\min_{\eta_j} \; \mu\nabla_{j^*} f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \tag{22}$$

$$= \min_{\eta_k : k\notin\mathcal{A}} \; \sum_{k\notin\mathcal{A}}\Big(\mu\nabla_k f(c)\eta_k + \lambda|\eta_k|\Big) + \frac{\gamma}{2}\left(\sum_{k\notin\mathcal{A}}|\eta_k|\right)^2 \tag{23}$$

Proof. The minimization (28) is equivalent to

$$\begin{aligned}
\min_{\eta_k : k\notin\mathcal{A}} \quad & \sum_{k\notin\mathcal{A}}\Big(\mu\nabla_k f(c)\eta_k\Big) \\
s.t. \quad & \left(\sum_{k\notin\mathcal{A}}|\eta_k|\right)^2 \leq C_1 \;,\; \sum_{k\notin\mathcal{A}}|\eta_k| \leq C_2.
\end{aligned}$$

and therefore is equivalent to

$$\begin{aligned}
\min_{\eta_k : k\notin\mathcal{A}} \quad & \mu\sum_{k\notin\mathcal{A}}\nabla_k f(c)\eta_k \\
s.t. \quad & \sum_{k\notin\mathcal{A}}|\eta_k| \leq \min\{\sqrt{C_1}, C_2\}
\end{aligned}$$

which is a linear objective subject to a convex set and thus always has solution that lies on the corner point with only one non-zero coordinate η_{j^*}, which then gives the same minimum as (27). □

7.2 Proof for Lemma 1

Since $supp(\boldsymbol{c}^*) = \mathcal{A}^*$, and $\boldsymbol{c}^*$ is optimal when restricted on the support, we have $\langle\boldsymbol{\eta}, \boldsymbol{c}^*\rangle = 0$ for some $\boldsymbol{\eta} \in \partial F(\boldsymbol{c}^*)$. And since $F(\boldsymbol{c})$ is strongly convex on the support $\mathcal{A}^*$ with parameter β, we have

$$\begin{aligned}
F(0) - F(\boldsymbol{c}^*) &= F(0) - F(\boldsymbol{c}^*) - \langle\boldsymbol{\eta}, 0 - \boldsymbol{c}^*\rangle \\
&\geq \frac{\beta}{2}\|\boldsymbol{c}^* - 0\|_2^2,
\end{aligned}$$

which gives us

$$\|\boldsymbol{c}^*\|_2^2 \leq \frac{2(F(0) - F(\boldsymbol{c}^*))}{\beta}.$$

Combining above with the fact for any $\boldsymbol{c}$, $\|\boldsymbol{c}\|_1^2 \leq \|\boldsymbol{c}\|_0\|\boldsymbol{c}\|_2^2$, we obtain the result.

7.3 Proof for Theorem 2

Lemma 3. *Let $r(W)$ and $r_N(W)$ be the risk* (2) *and the empirical risk respectively, we have*

$$\sup_{W\in\mathbb{R}^{K\times D}:\|W\|_F\le R} |r(W) - r_N(W)| \le \sqrt{\frac{2DK\log(4RKN)}{N} + \frac{1}{N}\log(\frac{1}{\rho})}$$

with probability $1-\rho$.

Proof. Since $\min_{\boldsymbol{z}\in\{0,1\}^N} \frac{1}{2}(y - \boldsymbol{z}^\mathsf{T} W\boldsymbol{x})^2 \le |y|^2 \le 1$ for a given W, by Hoeffding inequality,

$$\begin{aligned} &P\left(|r_N(W) - r(W)| \ge \epsilon\right) \\ &\le \exp(-2N\epsilon^2). \end{aligned}$$

Let $\mathcal{N}(\delta)$ be a δ-covering of the set $\mathcal{W} := \{W \in \mathbb{R}^{K\times D} \mid \|W\|_F \le R\}$ with $|\mathcal{N}(\delta)| \le \left(\frac{4R}{\delta}\right)^{DK}$. Then for any $W \in \mathcal{W}$, we have $\tilde{W} \in \mathcal{N}(\delta)$ with $\|W - \tilde{W}\| \le \delta$. Applying a union bound, we have

$$\begin{aligned} &P\left(\sup_{\tilde{W}\in\mathcal{N}(\delta)} |r_N(\tilde{W}) - r(\tilde{W})| \ge \epsilon\right) \\ &\le \left(\frac{4R}{\delta}\right)^{DK} \exp(-2N\epsilon^2). \end{aligned} \tag{24}$$

Then for $\Delta W := W - \tilde{W}$ satisfying $\|\Delta W\| \le \delta$, we can bound the difference of square loss of W and $\tilde{W}$ by

$$\begin{aligned} &\min_{\boldsymbol{z}\in\{0,1\}^K} \frac{1}{2}(y - \boldsymbol{z}^\mathsf{T} W\boldsymbol{x})^2 - \min_{\boldsymbol{z}\in\{0,1\}^K} \frac{1}{2}(y - \boldsymbol{z}^\mathsf{T} \tilde{W}\boldsymbol{x})^2 \\ &\le \frac{1}{2}(y - \tilde{\boldsymbol{z}}^\mathsf{T} W\boldsymbol{x})^2 - \frac{1}{2}(y - \tilde{\boldsymbol{z}}^\mathsf{T} \tilde{W}\boldsymbol{x})^2 \\ &\le \|\Delta W\|_F\|\tilde{\boldsymbol{z}}\| + 2R\|\tilde{\boldsymbol{z}}\|^2\|\Delta W\|_F \le 3RK\epsilon \end{aligned} \tag{25}$$

where $\tilde{\boldsymbol{z}} = arg\min_{\boldsymbol{z}\in\{0,1\}^K} \frac{1}{2}(y - \boldsymbol{z}^\mathsf{T} W\boldsymbol{x})^2$ and we used the fact that $\|\boldsymbol{x}\| \le 1$ and $|y| \le 1$. By symmetry, we have

$$\left|\min_{\boldsymbol{z}\in\{0,1\}^K} \frac{1}{2}(y - \boldsymbol{z}^\mathsf{T} \tilde{W}\boldsymbol{x})^2 - \min_{\boldsymbol{z}\in\{0,1\}^K} \frac{1}{2}(y - \boldsymbol{z}^\mathsf{T} W\boldsymbol{x})^2\right| \le 3RK\epsilon$$

. Combining (24) with (25), we have

$$\begin{aligned} &\sup_{W\in\mathcal{W}} |r_N(W) - r_N(W)| \\ &\le 6RK\delta + \sqrt{\frac{DK}{2N}\log(\frac{4R}{\delta}) + \frac{1}{2N}\log(\frac{1}{\rho})}. \end{aligned} \tag{26}$$

with probability $1-\rho$. Setting $\delta = 1/(6RK\sqrt{N})$ and apply Jennen's inequality gives the result. □

Then the following gives the proof for Theorem 2.

Proof. Let $\bar{\boldsymbol{z}}_i = arg\min_{\boldsymbol{z}_i\in\{0,1\}^K} (y_i - \boldsymbol{z}_i^\mathsf{T} \bar{W}\boldsymbol{x}_i)^2$ for $i \in [N]$. Denote $\bar{Z}$ as the $N \times K$ matrix stacked from $(\bar{\boldsymbol{z}}_i^\mathsf{T})_{i=1}^N$. Let $\{\bar{\boldsymbol{z}}^k\}_{k=1}^K$ be the columns of $\bar{Z}$ and $\bar{\mathcal{A}}$ be the indexes of atoms in the atomic set (5) that have the same 0-1 patterns to those columns. Denote $\bar{\boldsymbol{c}}$ as the coefficient vector with $\bar{c}_k = 1$ for $k \in \bar{\mathcal{A}}$ and $\bar{c}_k = 0$ for $k \notin \bar{\mathcal{A}}$. By the definition of $F(\boldsymbol{c})$, we have

$$F(\bar{\boldsymbol{c}}) \le r_N(\bar{W}) + \frac{\tau}{2}\|\bar{W}\|_F^2 + \lambda\|\bar{\boldsymbol{c}}\|_1. \tag{27}$$

where $r_N(\bar{W}) := \frac{1}{2N}\sum_{i=1}^{N} \min_{\boldsymbol{z}\in\{0,1\}^K}(y_i - \boldsymbol{z}^\mathsf{T}\bar{W}\boldsymbol{x}_i)^2$ is the empirical risk of $\bar{W}$. Let $\boldsymbol{c}^* := arg\min_{\boldsymbol{c}:supp(\boldsymbol{c})=\bar{\mathcal{A}}} F(\boldsymbol{c})$. We have $F(\boldsymbol{c}^*) \leq F(\bar{\boldsymbol{c}})$. Then from (18),

$$F(\hat{\boldsymbol{c}}) - F(\bar{\boldsymbol{c}}) \leq F(\hat{\boldsymbol{c}}) - F(\boldsymbol{c}^*) \leq \frac{4\gamma K}{\beta\mu^2}\left(\frac{1}{T}\right) + \frac{2(1-\mu)\lambda}{\mu}\sqrt{\frac{2K}{\beta}}. \tag{28}$$

In addition, the risk of $\hat{W}$ satisfies

$$r_N(\hat{W}) + \frac{\tau}{2}\|\hat{W}\|_F^2 + \lambda\|\hat{\boldsymbol{c}}\|_1 \leq F(\hat{\boldsymbol{c}}) \tag{29}$$

by the definition of the empirical risk $r_N(.)$ (since it is minimized w.r.t. the hidden assignments). Combining (27), (28) and (29), we obtain a bound on the difference of empirical risk

$$\begin{aligned} & r_N(\hat{W}) - r_N(\bar{W}) \\ & \leq \underbrace{\frac{\tau}{2}\|\bar{W}\|^2 + \lambda K}_{\text{bias of regularization}} + \underbrace{\frac{4\gamma K}{\beta\mu^2}\left(\frac{1}{T}\right) + \frac{2(1-\mu)\lambda}{\mu}\sqrt{\frac{2K}{\beta}}}_{\text{optimization error}} \end{aligned} \tag{30}$$

The remaining task is to bound the estimation error $|r(W) - r_N(W)|$. Since Algorithm 1 is a descent algorithm w.r.t. $F(\boldsymbol{c})$ and in the beginning $F(0) \leq 1/2$, we have $\|\boldsymbol{c}\|_1 \leq 1/\lambda$ and $\|W\|^2 \leq 1/\tau$ at any iterate. Then we can bound the estimation error by Lemma 3 for $\hat{W}$ belonging to the set $\mathcal{W}(T) := \{\hat{W} \in \mathbb{R}^{T\times D} \mid \|\hat{W}\|_F \leq \sqrt{1/\lambda\tau}\}$, giving

$$|r(\hat{W}) - r_N(\hat{W})| \leq \sqrt{\frac{2DT\log(4TN/\sqrt{\lambda\tau})}{N} + \frac{1}{N}\log(\frac{1}{\rho})}. \tag{31}$$

Combining (30) and (31), and choosing $\lambda = 1/(NK)$, $\tau = 1/(NR^2)$, we obtain $r(\hat{W}) - r(\bar{W}) \leq \epsilon$ with $T \geq \frac{4\gamma}{\mu^2\beta}(\frac{K}{\epsilon})$, and $N \geq \frac{DT}{\epsilon^2}(2\log(\frac{4RKT}{\epsilon}) + \log(\frac{1}{\rho}))$ for any $0 < \epsilon < 1$. □