7 Appendix

7.1 Proof of Theorem 1

Proof. There are two main difficulties in proving the convergence of our algorithm, and none of them is addressed in previous works. First, the Hessian matrix $\mathcal{H}$ is a block-structured matrix as shown in (7), and unfortunately it is low-rank. Second, we have to show the convergence with the active set selection technique.

Let $H(\boldsymbol{\theta}) \in \mathcal{R}^{n\times n}$ be the Hessian $\nabla^2\mathcal{L}(\bar{\boldsymbol{\theta}})$ when $\mathcal{L}(\cdot)$ is strongly convex. When it is not, as discussed in Section 3, we set $H = \nabla^2\mathcal{L}(\bar{\boldsymbol{\theta}}) + \epsilon I$ for a small constant $\epsilon > 0$. Let $\boldsymbol{d} \in \mathbb{R}^{n^2k}$ denotes the minimizer of the quadratic subproblem (8). By definition, we can easily observe that $\sum_{r=1}^{k} \boldsymbol{d}^{(r)} = \boldsymbol{y}$ where

$$\boldsymbol{y}^\star := \operatorname*{argmin}_{\boldsymbol{y}} \boldsymbol{y}^T H(\boldsymbol{\theta})\boldsymbol{y} + \langle \boldsymbol{y}, \nabla\mathcal{L}(\bar{\boldsymbol{\theta}})\rangle + \mathcal{W}(\boldsymbol{y}), \tag{19}$$

where $\mathcal{W}(\boldsymbol{y}) = \min_{\boldsymbol{y}^{(1)},\ldots,\boldsymbol{y}^{(k)}:\sum_{r=1}^{k}\boldsymbol{y}^{(r)}=\boldsymbol{y}} \lambda_r\|\boldsymbol{y}^{(r)}\|_{\mathcal{A}_r}$. Now (19) has a strongly convex Hessian, and thus if $\mathcal{W}(\boldsymbol{y})$ is convex then Theorem 3.1 in [16] can be used to show that the algorithm converges, and thus $F(\sum_{i=1}^{r}\boldsymbol{\theta}^{(r)})$ converges (without active subspace selection).

To show $\mathcal{W}$ is convex, assume we have $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(k)}$ are decomposition of $\boldsymbol{a}$ that attains the minimizer of $\mathcal{W}(\boldsymbol{a})$. By definition we have

$$\begin{aligned}
\mathcal{W}(\alpha\boldsymbol{a} + (1-\alpha)\boldsymbol{b}) &\le \sum_{r=1}^{k} \lambda_r\|\boldsymbol{a}^{(r)} + \boldsymbol{b}^{(r)}\|_{\mathcal{A}_r} \\
&\le \sum_{r=1}^{k} \lambda_r(\alpha\|\boldsymbol{a}^{(r)}\|_{\mathcal{A}_r} + (1-\alpha)\|\boldsymbol{b}^{(r)}\|_{\mathcal{A}_r}) \\
&= \alpha\mathcal{W}(\boldsymbol{a}) + (1-\alpha)\mathcal{W}(\boldsymbol{b}).
\end{aligned}$$

Thus $\mathcal{W}$ is convex, and if we solve each quadratic approximation exactly without active subspace selection, our algorithm converges to the global optimum.

Next we discuss the convergence of our algorithm with active subspace selection. [12] has shown the convergence of active set selection under ℓ_1 regularization, but here the situation is different – we can have infinite many atoms, as in group lasso or nuclear norm cases, so the original proof cannot be applied. To analyze the active subspace selection technique, we will use the convergence proof for Block Coordinate Gradient Descent (BCGD) in [23]. To begin, we give a quick review of the Block Coordinate Gradient Descent (BCGD) algorithm discussed in [23].

BCGD is proposed to solve the composite functions with the following form:

$$\operatorname*{argmin}_{\boldsymbol{x}} F(\boldsymbol{x}) := f(\boldsymbol{x}) + P(\boldsymbol{x}),$$

where $P(\boldsymbol{x})$ is a convex and separable function. Here we consider $\boldsymbol{x}$ to be a p dimensional vector, but in general $\boldsymbol{x}$ can be in any space. At the t-th iteration, BCGD chooses a subset $\mathcal{J}_t$ and compute the descent direction by

$$\boldsymbol{d}_t = \boldsymbol{d}_{H_t}^{\mathcal{J}_t}(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{d}} \left\{ \langle \nabla f(\boldsymbol{x}), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^T H_t \boldsymbol{d} + P(\boldsymbol{x}+\boldsymbol{d}) \mid d_i = 0, \forall i \notin \mathcal{J} \right\} \equiv \operatorname*{argmin}_{\boldsymbol{d}} Q_{H_t}^{\mathcal{J}_t}(\boldsymbol{d}). \tag{20}$$

After computing the descent direction $\boldsymbol{d}_H^{\mathcal{J}_t}$, a line search procedure is used to find the descent direction, where the largest α_t is chosed by searching over $1, 1/2, 1/4, ...$ satisfying

$$F(\boldsymbol{x}_t + \alpha\boldsymbol{d}_t) \le F(\boldsymbol{x}_t) + \alpha_t\sigma\Delta_t, \text{ where } \Delta_t = \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t\rangle + P(\boldsymbol{x}_t + \boldsymbol{d}_t) - P(\boldsymbol{x}_t),$$

and $\sigma \in (0,1)$ is any constant. The line search condition is exactly the same with ours in (10).

It is shown in Theorem 3.1 of [23] that when J_t is selected in a cyclic order covering all the indexes, than BCGD converges to the global optimum for any convex function $f(X)$.

To proof this theorem, we want to show the equivalence between our algorithm and BCGD with one block (BCGD-1block). We first prove the convergence for the case where we only have one regularizer, thus the problem is

$$\operatorname*{argmin}_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_{\mathcal{A}}.$$

The key idea is to carefully define the matrix H_t at each iteration, according the our subspace selection trick. Note that in (20) the matrix H_t can be any positive definite matrix instead of Hessian. At each iteration of Algorithm 1, assume the fixed and free subspace defined in (13) are $\mathcal{S}_{\text{fixed}}$ and $\mathcal{S}_{\text{free}}$. Assume $\dim(\mathcal{S}_{\text{free}}) = q$ and $\dim(\mathcal{S}_{\text{fixed}}) = n - q$, $R = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_q]$ are the orthogonal basis for $\mathcal{S}_{\text{free}}$, then we can construct $\tilde{H}_t$ by

$$\tilde{H}_t = R(R^T H_t R)R^T + R^{\perp}(R^{\perp})^T. \tag{21}$$

It is easy to see that $\tilde{H}_t$ is positive definite if the real Hessian $\nabla^2 f(\boldsymbol{x})$ is positive definite. Using $\tilde{H}_t$ as the Hessian in (20), we first consider $\boldsymbol{d}_{\text{free}}$ to be the minimizer of the following problem:

$$\boldsymbol{d}_{\text{free}} = \underset{\Delta_{\text{free}} \in \mathcal{S}_{\text{free}}}{\operatorname{argmin}} \left\{ \langle \nabla f(\boldsymbol{x}), \Delta_{\text{free}} \rangle + \frac{1}{2} {\Delta_{\text{free}}}^T \tilde{H}_t \Delta_{\text{free}} + P(\boldsymbol{x} + \Delta_{\text{free}}) \right\},$$

which is the quadratic subproblem (20) within the subspace $\mathcal{S}_{\text{free}}$. Next we show that $\boldsymbol{d}_{\text{free}}$ is indeed the optimizer of the whole quadratic subproblem (20) with the original Hessian H_t. To show this, taking derivative of (20) with respect to a $\boldsymbol{a} \in \mathcal{S}_{\text{fixed}}$, the subgradient will be

$$\partial_{\boldsymbol{a}} Q_{\tilde{H}_t}(\boldsymbol{d}_{\text{free}}) = \langle \nabla \mathcal{L}(\boldsymbol{x}), \boldsymbol{a} \rangle + \boldsymbol{a}^T \tilde{H}_t \boldsymbol{d}_{\text{free}} + \partial_{\boldsymbol{a}} P(\boldsymbol{x} + \boldsymbol{d}_{\text{free}}).$$

By the definition of $\tilde{H}_t$ in (21), since $\boldsymbol{a} \in \text{span}(R^{\perp})$ and $\boldsymbol{d}_{\text{free}} \in \text{span})(R)$, we have $\boldsymbol{a}^T \tilde{H}_t \boldsymbol{d}_{\text{free}} = 0$. Also, since $\boldsymbol{a} \in \mathcal{S}_{\text{fixed}}$, $\langle \boldsymbol{x} + \boldsymbol{d}_{\text{free}}, \boldsymbol{a} \rangle = 0$, thus $\partial_{\boldsymbol{a}} P(\boldsymbol{x} + \boldsymbol{d}_{\text{free}}) = [-\lambda, \lambda]$. Also, since $\boldsymbol{a} \in \mathcal{S}_{\text{fixed}}$, $|\langle \nabla f(\boldsymbol{x}), \boldsymbol{a} \rangle| < \lambda$. Therefore, we have

$$0 \in \partial_{\boldsymbol{a}} Q_{\tilde{H}_t}(\boldsymbol{d}_{\text{free}}), \ \ \forall \boldsymbol{a} \in S_{\text{fixed}},$$

also, since $\boldsymbol{d}_{\text{free}}$ is the optimal solution in $\mathcal{S}_{\text{free}}$, the projection of subgradient to $\boldsymbol{a} \in \mathcal{S}_{\text{free}}$ is 0. Therefore $\boldsymbol{d}_{\text{free}}$ is the minimizer of $Q_{\tilde{H}_t}$.

Based on the above discussion, our algorithm that computing the generalized Newton direction in free subspace $\mathcal{S}_{\text{free}}$ is equivalent to another BCGD-1block algorithm with $\tilde{H}_t$ as the approximated Hessian matrix. Therefore based on Theorem 3.1 of [23], our algorithm converges to the global optimum.

When there are more than one set of parameters, i.e., we want to solve (1) with parameter sets $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(k)}$. In this case, the Hessian matrix is a kn by kn matrix. For each parameter set, we will select ${\mathcal{S}_{\text{free}}}^{(i)}$ and ${\mathcal{S}_{\text{fixed}}}^{(i)}$, and only update on ${\mathcal{S}_{\text{free}}}^{(i)}$. To prove the convergence, similar to the above arguments, we can construct the approximate Hessian $\tilde{H}_t$ and show the equivalence between BCGD-block1 with $\tilde{H}_t$ and our algorithm. $\tilde{H}_t$ can be divided into k^2 blocks, each one is a n by n matrix $\tilde{H}_t^{(i,j)}$, where

$$\tilde{H}_t^{(i,j)} = R^{(i)}(R^{(i)})^T H R^{(j)}(R^{(j)})^T + (R^{(i)})^{\perp}((R^{(j)})^{\perp})^T,$$

where $H = \nabla^2 \mathcal{L}(\sum_{i=1}^{k} \boldsymbol{\theta}^{(r)})$, $R^{(i)}$ is the basis of ${\mathcal{S}_{\text{free}}}^{(i)}$, and $(R^{(i)})^{\perp}$ is the basis of ${\mathcal{S}_{\text{fixed}}}^{(i)}$. Assume $\boldsymbol{d}_{\text{free}} \in \mathcal{R}^{nk}$ is the solution of (20) with the constraint that projection to ${\mathcal{S}_{\text{fixed}}}^{(i)}$ is 0 for all i, then it is the solution for both BCGD-1block with $\tilde{H}_t$ and also the update direction produced by Algorithm 1. Also,

$$\boldsymbol{a} \tilde{H}_t \boldsymbol{d}_{\text{free}} = 0 \ \ \forall \boldsymbol{a} \in \{{\mathcal{S}_{\text{fixed}}}^{(1)}, \ldots, {\mathcal{S}_{\text{fixed}}}^{(k)}\},$$

therefore similar to the previous arguments we can show that

$$0 \in \partial_{\boldsymbol{a}} Q_{\tilde{H}_t}(\boldsymbol{d}_{\text{free}}), \ \forall \boldsymbol{a} \in \{{\mathcal{S}_{\text{fixed}}}^{(1)}, \ldots, {\mathcal{S}_{\text{fixed}}}^{(k)}\},$$

which indicates $\boldsymbol{d}_{\text{free}}$ is the minimizer of $Q_{\tilde{H}_t}(\boldsymbol{d})$. Then we can see that the BCGD-1block algorithm with $\tilde{H}_t$ as the Hessian matrix is equivalent to Algorithm 1 when each quadratic subproblem is solved exactly, therefore our algorithm converges to the global optimum of (1) □

7.2 Proof of Lemma 2

Proof. First, since $\mathcal{L}(\cdot)$ is strongly convex,

$$\langle \nabla \mathcal{L}(\bar{\boldsymbol{x}}) - \nabla \mathcal{L}(\bar{\boldsymbol{y}}), \bar{\boldsymbol{x}} - \bar{\boldsymbol{y}} \rangle \geq \eta \|\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}\|^2. \tag{22}$$

Next, by the optimal condition we know

$$-\nabla\mathcal{L}(\bar{\boldsymbol{x}}) \in \partial\|\boldsymbol{x}^{(r)}\|_{\mathcal{A}_r},\ \forall r = 1,\ldots,k$$
$$-\nabla\mathcal{L}(\bar{\boldsymbol{y}}) \in \partial\|\boldsymbol{y}^{(r)}\|_{\mathcal{A}_r},\ \forall r = 1,\ldots,k.$$

By the convexity for each $\|\cdot\|_{\mathcal{A}_r}$,

$$\langle -\nabla\mathcal{L}(\bar{\boldsymbol{x}}) + \nabla\mathcal{L}(\bar{\boldsymbol{y}}), \boldsymbol{x}^{(r)} - \boldsymbol{y}^{(r)}\rangle \geq 0,\ \forall r = 1,..,k.$$

Summing r from 1 to k we get $\langle -\nabla\mathcal{L}(\bar{\boldsymbol{x}}) + \nabla\mathcal{L}(\bar{\boldsymbol{y}}), \bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}\rangle \geq 0$, and combined with (22) we have $\|\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}\| = 0$. □

7.3 Proof of Theorem 2

Proof. Since assumption (16) holds, there exists a positive constant ϵ such that

$$\|\Pi_{(\bar{\mathcal{T}}^{(r)})^\perp}\left(\nabla\mathcal{L}(\bar{\boldsymbol{\theta}}^\star)\right)\|^*_{\mathcal{A}_r} < \lambda_r - \epsilon\ \ \forall r = 1,\ldots,k. \tag{23}$$

We first focus on showing that that $\mathcal{S}_{\text{fixed}}$ will be eventually equal to $(\mathcal{T}^{(r)})^\perp$. Focus on one of the $(\mathcal{T}^{(r)})^\perp$, for any unit vector (in terms of the $\|\cdot\|_{\mathcal{A}_r}$ norm) $\boldsymbol{a} \in (\mathcal{T}^{(r)})^\perp$, we have

$$|\langle \boldsymbol{a}, \nabla\mathcal{L}(\bar{\boldsymbol{\theta}}^\star)\rangle| < \lambda_r - \epsilon. \tag{24}$$

Since the sequence generated by our algorithm converges to the global optimum (as proved in Theorem 1), there exists a T such that

$$\|\nabla\mathcal{L}(\bar{\boldsymbol{\theta}}_t) - \nabla\mathcal{L}(\bar{\boldsymbol{\theta}}^\star)\|^*_{\mathcal{A}_r} < \epsilon,$$

combining with (24) we have

$$\begin{aligned}|\langle \boldsymbol{a}, \nabla\mathcal{L}(\bar{\boldsymbol{\theta}}_t)\rangle| &\leq |\langle \boldsymbol{a}, \nabla\mathcal{L}(\bar{\boldsymbol{\theta}}^*)\rangle| + |\langle \boldsymbol{a}, \nabla\mathcal{L}(\bar{\boldsymbol{\theta}}_t) - \nabla\mathcal{L}(\bar{\boldsymbol{\theta}}^*)\rangle| \\ &< \lambda_r - \epsilon + \|\nabla\mathcal{L}(\bar{\boldsymbol{\theta}}_t) - \nabla\mathcal{L}(\bar{\boldsymbol{\theta}}^*)\|^*_{\mathcal{A}_r} \\ &< \lambda_r \end{aligned} \tag{25}$$

for all $t > T$. Now we consider two cases:

1. If $\langle \boldsymbol{a}, \boldsymbol{\theta}_{t-1}\rangle \neq 0$, then $\boldsymbol{a} \notin {\mathcal{S}_{\text{free}}}^{(r)}$ at the t-th iteration. Since we assume subproblems are exactly solved, and $\boldsymbol{a} \in {\mathcal{S}_{\text{free}}}^{(r)}$, by the optimality condition $|\langle\nabla\mathcal{L}(\bar{\boldsymbol{\theta}}_t)| < \lambda$ implies that $\langle \boldsymbol{\theta}_t, \boldsymbol{a}\rangle = 0$.
2. If $\langle \boldsymbol{a}, \boldsymbol{\theta}_{t-1}\rangle = 0$, combined with (25) we have $\langle \boldsymbol{\theta}_t, \boldsymbol{a}\rangle = 0$.

Therefore, for all $\boldsymbol{a} \in (\mathcal{T}^{(r)})^\perp$, we have $\langle \boldsymbol{\theta}_t, \boldsymbol{a}\rangle = 0$, which implies $(\mathcal{T}^{(r)})^\perp \subset \mathcal{S}_{\text{fixed}}$. On the other hand, by definition $(\mathcal{T}^{(r)}) \cap \mathcal{S}_{\text{fixed}} = \phi$, thus $\mathcal{T}^{(r)} = \mathcal{S}_{\text{free}}$ and $(\mathcal{T}_{(r)})^\perp = \mathcal{S}_{\text{fixed}}$. □

7.4 Proof of Theorem 3

Proof. As shown in Theorem 2, there exists a T such that $\mathcal{S}_{\text{free}} = \text{span}(\{\boldsymbol{a} \mid \langle \boldsymbol{a}, \Theta^\star\rangle\})$ after $t > T$.

Next we show that after finite iterations the line search step size will be 1. For simplicity, let $\bar{\mathcal{L}}(\boldsymbol{\theta}) = \mathcal{L}(\sum_{r=1}^k \boldsymbol{\theta}^{(r)})$ and $\bar{\mathcal{R}}(\boldsymbol{\theta}) = \sum_{r=1}^k \|\boldsymbol{\theta}_r\|_{\mathcal{A}_r}$, and $F(\boldsymbol{\theta}) = \bar{\mathcal{L}}(\boldsymbol{\theta}) + \bar{\mathcal{R}}(\boldsymbol{\theta})$. Since $\nabla^2\mathcal{L}(\bar{\boldsymbol{\theta}})$ is Lipschitz continuous, we have

$$\bar{\mathcal{L}}(\boldsymbol{\theta} + \boldsymbol{d}) \leq \bar{\mathcal{L}}(\boldsymbol{\theta}) + \langle \nabla\bar{\mathcal{L}}(\boldsymbol{\theta}), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^T\nabla^2\bar{\mathcal{L}}(\boldsymbol{\theta})\boldsymbol{d} + \frac{1}{6}\eta\|\boldsymbol{d}\|^3.$$

Plug in into the objective function we have

$$\begin{aligned}F(\boldsymbol{\theta} + \boldsymbol{d}) &\leq \bar{\mathcal{L}}(\boldsymbol{\theta}) + \bar{\mathcal{R}}(\boldsymbol{\theta}) + \langle \nabla\bar{\mathcal{L}}(\boldsymbol{\theta}), \boldsymbol{d}\rangle + \\ &\quad \frac{1}{2}\boldsymbol{d}^T\nabla^2\bar{\mathcal{L}}(\boldsymbol{\theta})\boldsymbol{d} + \frac{1}{6}\eta\|\boldsymbol{d}\|^3 + \bar{\mathcal{R}}(\boldsymbol{\theta} + \boldsymbol{d}) - \bar{\mathcal{R}}(\boldsymbol{\theta}) \\ &\leq F(\boldsymbol{\theta}) + \delta + \frac{1}{2}\boldsymbol{d}^T\nabla^2\bar{\mathcal{L}}(\boldsymbol{\theta})\boldsymbol{d} + \frac{1}{6}\eta\|\boldsymbol{d}\|^3 \end{aligned} \tag{26}$$

To further bound (26), we first show the following Lemma:

Lemma 3. *Let $\boldsymbol{d}^\star$ be the optimal solution of* (20)*, then*

$$\delta = \langle \nabla \bar{\mathcal{L}}(\boldsymbol{\theta}), \boldsymbol{d}^\star \rangle + \bar{\mathcal{R}}(\boldsymbol{\theta} + \boldsymbol{d}^\star) - \bar{\mathcal{R}}(\boldsymbol{\theta}) \leq -(\boldsymbol{d}^\star)^T \nabla^2 \bar{\mathcal{L}}(\boldsymbol{\theta}) \boldsymbol{d}^\star.$$

Proof. Since $\boldsymbol{d}^\star$ is the optimal solution of (20), for any $\alpha < 1$ we have

$$\begin{aligned}\langle \nabla \bar{\mathcal{L}}(\boldsymbol{\theta}) + \boldsymbol{d}^\star \rangle + \frac{1}{2}(\boldsymbol{d}^\star)^T H \boldsymbol{d}^\star + \bar{\mathcal{R}}(\boldsymbol{\theta} + \boldsymbol{d}^s tar) &\leq \langle \nabla \bar{\mathcal{L}}(\boldsymbol{\theta}), \alpha \boldsymbol{d}^\star \rangle + \frac{1}{2}\alpha^2 (\boldsymbol{d}^\star)^T H \boldsymbol{d}^\star + \bar{\mathcal{R}}(\boldsymbol{\theta} + \alpha \boldsymbol{d}^\star) \\ &\leq \alpha \langle \nabla \bar{\mathcal{L}}(\boldsymbol{\theta}), \boldsymbol{d}^\star \rangle + \frac{1}{2}\alpha^2 (\boldsymbol{d}^\star)^T H \boldsymbol{d}^\star + \alpha \bar{\mathcal{R}}(\boldsymbol{\theta} + \boldsymbol{d}) + (1-\alpha)\bar{\mathcal{R}}(\boldsymbol{\theta}),\end{aligned}$$

where we use H to denote the exact Hessian $\nabla^2 \bar{\mathcal{L}}(\boldsymbol{\theta})$ and the second inequality is by the convexity of $\bar{\mathcal{R}}$ since we consider all the atomic norm $\|\cdot\|_{\mathcal{A}}$ to be convex. Therefore

$$(1-\alpha)\left(\langle \nabla \bar{\mathcal{L}}(\boldsymbol{\theta}), \boldsymbol{d}^\star \rangle + \bar{\mathcal{R}}(\boldsymbol{\theta} + \boldsymbol{d}^\star) - \bar{\mathcal{R}}(\boldsymbol{\theta})\right) + \frac{1}{2}(1-\alpha^2)(\boldsymbol{d}^\star)^T H \boldsymbol{d}^\star \leq 0.$$

Dividing both sides by $(1-\alpha)$ we get

$$\delta + \frac{1}{2}(1-\alpha)(\boldsymbol{d}^\star)^T H \boldsymbol{d}^\star \leq 0,$$

therefore $\delta \leq -(\boldsymbol{d}^\star)^T H \boldsymbol{d}^\star$. □

Combining Lemma 3 and (26) we get

$$F(\boldsymbol{\theta} + \boldsymbol{d}^\star) \leq F(\boldsymbol{\theta}) + \frac{\delta}{2} + \frac{1}{6}\eta \|\boldsymbol{d}^\star\|^3.$$

Furthermore, considering $\boldsymbol{\theta}$ in the level set, we can define M to be the largest eigenvalue of Hessians, and thus

$$F(\boldsymbol{\theta} + \boldsymbol{d}^\star) \leq F(\boldsymbol{\theta}) + (\frac{1}{2} - \frac{1}{6}\eta M^2 \|\boldsymbol{d}^\star\|)\delta.$$

Since $\boldsymbol{d} \to 0$ as $\boldsymbol{\theta} \to \boldsymbol{\theta}^\star$, we can find an ϵ-ball around $x^\star$ such that $\frac{1}{2} - \frac{1}{6}\eta M^2 \|\boldsymbol{d}^\star\|) > \sigma$, and $\delta < 0$, thus line search will be satisfied with step size equals to 1.

Based on the above proofs, when $\boldsymbol{\theta}$ is closed enough to $\boldsymbol{\theta}^\star$, $\mathcal{S}_{\text{free}} = \text{span}(\{\boldsymbol{a} \mid \langle \boldsymbol{a}, \boldsymbol{\theta}^\star \rangle\})$ and the step size $\alpha = 1$. Finally we have to explore the structure of dirty model. Since we have k parameter sets $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(k)}$, the Hessian has a block structure presented in (7). Even when $\nabla^2 \mathcal{L}(\boldsymbol{\theta})$ is strongly convex, the rank of the Hessian matrix is at most $O(n)$, while there are totally nk variables or rows. However, our main observation is that when $\nabla^2 \mathcal{L}(\boldsymbol{\theta})$ is positive definite, the Hessian $H \in \mathcal{R}^{pk \times pk}$ has a fixed null space:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}([I, I, \ldots, I][\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(k)}]^T) = \mathcal{L}(E\boldsymbol{\theta}),$$

where $E = [I, I, \ldots, I]$. The null space of H is always the null space of E when $\mathcal{L}$ is strongly convex itself. The following theorem (Theorem 2 in [28] can then be applied:

Lemma 4. *Let $F(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + h(\boldsymbol{\theta})$ where $g(\boldsymbol{\theta})$ has a constant null space $\mathcal{T}^\perp$ and is strongly convex in the subspace $\mathcal{T}$, and has Lipschitz continuous second order derivative $\nabla^2 g(\boldsymbol{\theta})$. If we apply a proximal Newton method (BCGD-block1 with exact Hessian and step size 1) to minimize $F(\boldsymbol{\theta})$, then*

$$\|\boldsymbol{z}_{t+1} - \boldsymbol{z}^\star\| \leq \frac{L_H}{2m}\|\boldsymbol{z}_t - \boldsymbol{z}^\star\|^2,$$

where $\boldsymbol{z}^\star = proj_{\mathcal{T}}(\boldsymbol{\theta}^\star$, $\boldsymbol{z}_t = proj_{\mathcal{T}}(\boldsymbol{z}_t)$, and L_H is the Lipschitz constant for $\nabla^2 g(\boldsymbol{\theta})$.

In our case, $\text{proj}_{\mathcal{T}}([\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k)}]^T) = \sum_{i=1}^k \boldsymbol{\theta}^{(i)}$, therefore we have

$$\|\sum_{i=1}^k \boldsymbol{\theta}_{t+1}^{(i)} - \boldsymbol{\theta}^\star\| \leq C\|\sum_{i=1}^k \boldsymbol{\theta}_t^{(i)} - \boldsymbol{\theta}^\star\|^2,$$

therefore $\bar{\boldsymbol{\theta}} = \sum_{i=1}^k \boldsymbol{\theta}^{(i)}$ has an asymptotic convergence rate. □

7.5 Dual of latent GMRF

The problem (4) can be rewritten by

$$\min_{S,L:L\succeq 0,S-L\succ 0} -\log\det(S-L)+\langle S-L,\Sigma\rangle+\max_{Z:\|Z\|_\infty\le\alpha}\operatorname{trace}(ZS)+\max_{P:\|P\|_2\le\beta}\operatorname{trace}(LP),$$

where $\|Z\|_\infty = \max_{i,j}|Z_{ij}|$ and $\|P\|_2 = \sigma_1(P)$ is the induced two norm. We then interchange min and max to get

$$\begin{aligned}\max_{Z,P:\|Z\|_\infty\le\alpha,\|P\|_2\le\beta}\ \min_{L,S,S-L\succ 0,L\succeq 0} &-\log\det(S-L)+\langle S-L,\Sigma\rangle+\operatorname{trace}(ZS)+\operatorname{trace}(LP)\\ &\equiv g(Z,P,L,S). \end{aligned}\tag{27}$$

Assume we do not have the constraint $L\succeq 0$, then the minimizer will satisfy

$$\begin{aligned}\nabla_L g(Z,P,L,S) &= -(S-L)^{-1}+\Sigma+Z=0\\ \nabla_S g(Z,P,L,S) &= -(S-L)^{-1}-\Sigma+P=0.\end{aligned}$$

Therefore we have

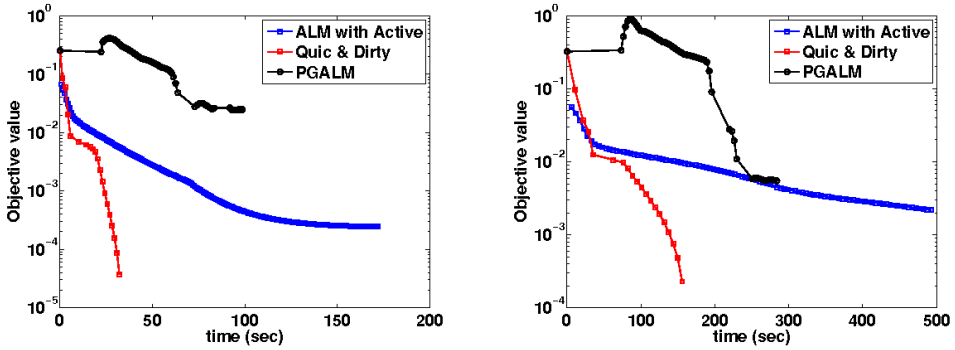
$$Z=-P \ \text{ and } \ S-L=(\Sigma+Z)^{-1}.\tag{28}$$

Combining (28) and (27) we get the dual problem

$$\min_{\Sigma+Z\succeq 0}\log\det(\Sigma+Z)+p \text{ s.t. } \|Z\|_\infty\le\alpha,\|Z\|_2\le\beta.$$

7.6 Alternating Minimization approach for latent GMRF

Another way to solve the latent GMRF problem is to directly applying an alternating minimization scheme to solve (4). The algorithm iteratively fix one of the S, L and update the other. We can still conduct the same active subspace selection technique mentioned in Section 3 in this algorithm. However, this alternating minimization approach can achieve at most linear convergence rate, while our algorithm can achieve super-linear convergence. In this section, we show the detail implementation for this algorithm — ALM(active), and show the comparison results with the Quic & Dirty algorithm (Algorithm 1). The results in Figure 2 shows that our proximal Newton method is faster in terms of the convergence rate.



(a) ER dataset, log scale

(b) Leukemia dataset, log scale

Figure 2: Comparison between QUIC and Alternating Minimization (AM) on gene expression datasets. Note that we implement the active subspace selection approach on both algorithm, so two algorithms have similar speed per iteration. However, we observe that QUIC is much more efficient in terms of final convergence rate.

7.7 Details on the implementation of our multi-task solver

We consider the multi-task learning problem. Assume we have k tasks, each with samples $X^{(r)}\in\mathcal{R}^{d,n_r}$ and labels $\boldsymbol{y}^{(r)}\in\mathcal{R}^{n_r}$. The goal of multi-task learning is to estimate the model $W\in\mathcal{R}^{d\times k}$,

where each column of W, denoted by $\boldsymbol{w}^{(r)}$, is the model for the r-th task. A dirty model has been proposed in [15] to estimate W by $S + B$, where

$$(S, B) = \underset{S,B \in \mathcal{R}^{d\times k}}{\operatorname{argmin}} \|\boldsymbol{y}^{(k)} - X^{(k)}(\boldsymbol{s}^{(k)} + \boldsymbol{b}^{(k)})\|^2 + \lambda_S \|S\|_1 + \lambda_B \|B\|_{1,2}, \tag{29}$$

where $\boldsymbol{s}^{(k)}, \boldsymbol{b}^{(k)}$ are the k-th column of S, B. It was shown in [15] that the combination of sparse and group sparse regularization yields better performance both in theory and in practice.

Instead of considering the squared-loss problem in (29), we further consider optimization problem minimizing the logistic loss:

$$(S, B) = \underset{S,B \in \mathcal{R}^{d\times k}}{\operatorname{argmin}} \sum_{r=1}^{k} \left(\sum_{i=1}^{n_r} \ell_{logistic}(\boldsymbol{y}_i^{(k)}, (\boldsymbol{s}^{(k)} + \boldsymbol{b}^{(k)})^T \boldsymbol{x}_i^{(k)}) \right) + \lambda_S \|S\|_1 + \lambda_B \|B\|_{1,2}, \tag{30}$$

where $\ell_{logistic}(y, a) = \log(1 + e^{-ya})$. This loss function is more suitable for the classification case, as shown in the later experiments.

Let $W = S + B$, and we define $\boldsymbol{w}^{(r)}$ to be the r-th column of W, then the Hessian and gradient of $\mathcal{L}(\cdot)$ can be computed by

$$\nabla_{\boldsymbol{w}^{(r)}} \sum_{i=1}^{n_r} (\sigma(y_i^{(r)} \langle \boldsymbol{w}^{(r)}, \boldsymbol{x}_i^{(r)} \rangle) - 1) y_i^{(r)} \boldsymbol{x}_i^{(r)}, \quad \nabla^2_{\boldsymbol{w}^{(r)}, \boldsymbol{w}^{(r)}} \mathcal{L}(W) = (X^{(r)})^T D^{(r)} X^{(r)},$$

where $\sigma(a) = 1/(1+e^{-a})$ and $D^{(r)}$ is a diagonal matrix with $D_{ii}^{(r)} = \sigma(y_i^{(r)} \langle \boldsymbol{w}^{(r)}, \boldsymbol{x}_i^{(r)} \rangle)$ for all $i = 1, \ldots, n_r$. Note that the Hessian of $\mathcal{L}(W)$ is a block-diagonal matrix, so $\nabla^2_{\boldsymbol{w}^{(r)}, \boldsymbol{w}^{(t)}} \mathcal{L}(W) = 0$. Note also that to form the quadratic approximation at W, we only need to compute $\sigma(y_i^{(r)} \langle \boldsymbol{w}^{(r)}, \boldsymbol{x}_i^{(r)} \rangle)$ for all i, r.

For the sparse-structured parameter component, we select a subset of variables in S to update as in the previous example. For the group-sparse structured component B, we select a subset of "rows" in B to update, following (15). To solve the quadratic approximation subproblem, we again use coordinate descent to minimize with respect to the sparse S component. For the group-sparse component B, we use block coordinate descent, where each time we update variables in one group (one row) using the trust region approach described in [20]. Since $\mathcal{S}_{\text{free}}$ contains only a small subset of blocks, the block coordinate descent can focus on this subset and becomes very efficient.

Algorithm. We first derive the quadratic approximation for the logistic loss function (30). Let $W = S + B$, the gradient for the loss function $\mathcal{L}(W)$ can be written as

$$\nabla_{\boldsymbol{w}^{(r)}} \sum_{i=1}^{n_r} (\sigma(y_i^{(r)} (\boldsymbol{w}^{(r)})^T \boldsymbol{x}_i^{(r)}) - 1) y_i^{(r)} \boldsymbol{x}_i^{(r)},$$

where $\sigma(a) = 1/(1 + e^{-a})$. The Hessian of $\mathcal{L}(W)$ is a block-diagonal matrix, where each $d \times d$ block corresponds to variables in one task, i.e., $\boldsymbol{w}^{(r)}$. Let $H \in \mathcal{R}^{kd \times kd}$ Hessian matrix, each $d \times d$ block can be written as

$$\nabla^2_{\boldsymbol{w}^{(r)}, \boldsymbol{w}^{(r)}} \mathcal{L}(W) = (X^{(r)})^T D^{(r)} X^{(r)},$$

where $D^{(r)}$ is a diagonal matrix with $D_{ii}^{(r)} = \sigma(y_i^{(r)} \langle \boldsymbol{w}^{(r)}, \boldsymbol{x}_i^{(r)} \rangle)$ for all $i = 1, \ldots, n_r$.

Therefore, to form the quadratic approximation of the current solution, the only computation required is to compute $\sigma(y_i^{(r)} \langle \boldsymbol{w}^{(r)}, \boldsymbol{x}_i^{(r)} \rangle)$ for all i, r.

For the lasso part, we select a subset of variables in S to update, according to the subspace selection criterion described 3.2. For the group lasso, we select a subset of "rows" in B to update, as described in 3.2.

For the lasso part, we apply a coordinate descent solver for solving the subproblem. Notice that for the Lasso part each column of S forms a subproblems:

$$\min_{\delta \in \mathbb{R}^d} \frac{1}{2} \delta^T H^{(r)} \delta + \boldsymbol{g}^{(r)} \delta + \|\boldsymbol{s}^{(r)} + \delta\|,$$

where $H^{(r)} = (X^{(r)})^T D^{(r)} X^{(r)}$ and $\boldsymbol{g}^{(r)} = \nabla_{\boldsymbol{s}^{(r)}} \mathcal{L}(W)$. The k subproblems are independent to each other, so we can solve them independently. For each subproblem, we apply the coordinate

descent approach described in [29] to solve it. When update the coordinate δ_i, the key computation is to compute the current gradient $H^{(r)}\delta + \boldsymbol{g}^{(r)}$. Directly computing this is expensive, however, we can maintain $\boldsymbol{p} = D^{(r)}X^{(r)}\delta$ during the updates, and then compute $H^{(r)}\delta = (X^{(r)}_{:,i})^T\boldsymbol{p}$, which only takes $O(\|X^{(r)}_{:,i}\|_0)$ flops.

For solving the group lasso problem, we cannot solve each column independently because the regularization is grouping each row of B. We apply a block coordinate descent method, where each time only one row of B is updated. Let $\delta \in \mathbb{R}^k$ denote the update on the i-th row of B, the subproblem with respect to δ can be written as

$$\frac{1}{2}\sum_{r=1}^{k}\gamma_r(\delta_j)^2 + \boldsymbol{g}^T\delta + \lambda\|\delta + \bar{\boldsymbol{w}}\|, \tag{31}$$

where $\bar{\boldsymbol{w}}$ is the i-th row of W; $\gamma_r = H^{(r)}_{ii}$ and $g_r = \nabla_{W_{ir}}\mathcal{L}(W)$ can be precomputed and will not change during the update.

By taking the gradient of the subproblem (31), we can see that

$$\delta = -\bar{\boldsymbol{w}} + \begin{cases} 0 & \text{if } \|\boldsymbol{g} - \sum_{r=1}^{k}\gamma_r\bar{w}_r^2\| \le \lambda \\ -(\Gamma + \frac{\lambda}{\|\bar{\boldsymbol{w}}+\delta\|}I)^{-1}\boldsymbol{g} & \text{if } \|\boldsymbol{g} - \sum_{r=1}^{k}\gamma_r\bar{w}_r^2\| > \lambda. \end{cases} \tag{32}$$

For the second case, the closed form solution exists when $\Gamma = I$. However, this is no true in general. Instead, we use the iterative trust-region solver proposed in [20] to solve the subproblem, where each iteration of the Newton root finding algorithm only takes $O(k)$ time. Therefore, the computational bottleneck is to compute $\boldsymbol{g}$ in (31). In our case, similar to the Lasso subproblem, we can maintain $\boldsymbol{p}^{(r)} = D^{(r)}X^{(r)}\delta^{(r)}$ for each $r = 1, \ldots, k$ in memory, where $\delta^{(r)}$ is the r-th column of change in W. The gradient can then be computed by $\boldsymbol{g} = H^{(r)}\delta^{(r)} = X^{(r)}\delta^{(r)}$, therefore the time complexity is $O(\bar{n})$ for each coordinate update, where $\bar{n}$ is number of nonzero for each column of $X^{(r)}$.

7.8 Proof of Proposition 1

To prove (a), first we expand the sub-differential

$$\langle \boldsymbol{a}, \partial_{\boldsymbol{\theta}^{(r)}}F(\boldsymbol{\theta})\rangle = \langle \boldsymbol{a}, \partial_{\boldsymbol{\theta}^{(r)}}\mathcal{L}(\bar{\boldsymbol{\theta}}) + \lambda_r\partial_{\boldsymbol{\theta}^{(r)}}\|\boldsymbol{\theta}^{(r)}\|_{\mathcal{A}_r}\rangle = \langle \boldsymbol{a}, G\rangle + \lambda_r\langle \boldsymbol{a}, \rho\rangle \quad \text{for} \quad \rho \in \partial_{\boldsymbol{\theta}^{(r)}}\|\boldsymbol{\theta}^{(r)}\|_{\mathcal{A}_r}$$

and now using the properties of decomposable norms, we calculate

$$\begin{aligned} |\langle \boldsymbol{a}, \rho\rangle| &= |\langle \boldsymbol{a}, \Pi_{\mathcal{T}_r^\perp}\rho\rangle| \\ &\le \|\boldsymbol{a}\|_{\mathcal{A}_r}\|\Pi_{\mathcal{T}_r^\perp}\rho\|^*_{\mathcal{A}_r} \\ &\le 1 \end{aligned}$$

hence $\langle \boldsymbol{a}, G\rangle + \lambda_r\langle \boldsymbol{a}, \rho\rangle \in \langle \boldsymbol{a}, G\rangle - \lambda_r, [\langle \boldsymbol{a}, G\rangle + \lambda_r]$ and the result is shown. In fact, it is not hard to see that *every* element of the set can be written as $\langle \boldsymbol{a}, \rho\rangle$ for some $\rho \in \partial_{\boldsymbol{\theta}^{(r)}}\|\boldsymbol{\theta}^{(r)}\|_{\mathcal{A}_r}$.

To prove (b), note that the optimality condition on σ is that $\sigma^\star$ will satisfy $0 \in \partial_\sigma F(\boldsymbol{\theta} + \sigma^\star\boldsymbol{a})$ and by the chain rule, $\partial_\sigma F(\boldsymbol{\theta} + \sigma\boldsymbol{a}) = \langle \boldsymbol{a}, \partial_{\boldsymbol{\theta}^{(r)}}F(\boldsymbol{\theta} + \sigma\boldsymbol{a})\rangle$. If $\langle \boldsymbol{a}, G\rangle \le \lambda_r$, then by part (a) we see that 0 is in the sub-differential of $\partial_\sigma F(\boldsymbol{\theta})$ and hence $\sigma = 0$ is an optimal point. If F is strongly convex, $\sigma = 0$ is the unique optimal point.

7.9 Proof of Proposition 2

By the definition of proximal operator, in the optimal solution $-G \in \partial_{\boldsymbol{\theta}^{(r)}}\|\text{prox}_{\lambda_r}(G)\|_{\mathcal{A}_r}$. For any $\boldsymbol{a} \in \mathcal{S}^{(r)}_{fixed}$, $\boldsymbol{a} \in \mathcal{T}(\text{prox}^{(r)}_{\lambda_r}(G))^\perp$, thus $|\langle G, \boldsymbol{a}\rangle| < \lambda_r$. Next we consider the projection of gradient to $\mathcal{S}^{(r)}_{fixed}$: let $\rho = \Pi_{S^{(r)}_{fixed}}(G)$, then by the previous statement we know $\|\rho\|_* \le \lambda_r$. Then since $\rho \in \mathcal{T}(\boldsymbol{\theta}^{(r)})^\perp$, we have $\rho \in \lambda_r\partial\|\boldsymbol{\theta}^{(r)}\|_{\mathcal{A}_r}$, therefore constrained to the subspace $\mathcal{S}^{(r)}_{fixed}$, 0 belongs to the sub-gradient, which proves Proposition 2.

7.10 Active Subspace Selection for Group Lasso Regularization

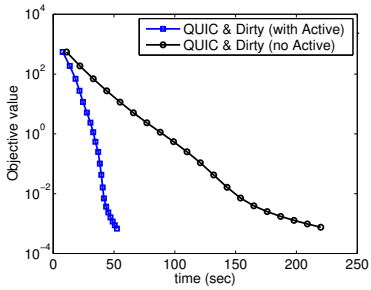


Figure 3: Comparing with/without active subspace selection technique on the RCV1 dataset on the multi-task learning problem with group-lasso regularization and logistic loss. We choose $\lambda = 10^{-3}$ and the final solution only has 1678 nonzero rows, while there are 22283 rows in total.