Supplementary Materials for: High-Dimensional Feature Selection for Sample Efficient Treatment Effect Estimation

7 Optimization algorithm

The proximal gradient algorithm for optimizing our objective (2) is shown in Algorithm 1, where we define (for $\theta \in \mathbb{R}^{p\times q}$)

$$[\text{Prox}_\lambda(\theta)]_{i:} = \theta_{i:} \max\left(0, 1 - \frac{\lambda}{\|\theta_{i:}\|_2}\right), \qquad i = 1, \ldots, p,$$

as the proximal operator for the L-1,2 norm.

Algorithm 1 Proximal gradient descent for (2)

Input: matrices $\hat{\Gamma}^{(j)} = \frac{X_j^T X_j}{n}$, $\hat{\gamma}^{(j)} = \frac{X_j^T y_j}{n}$ for $j = 1, \ldots, q$, regularizer ρ_λ and associated $q'_\lambda(\cdot)$, backtracking constant $c \in (0, 1)$, initial step size ζ_0, norm constraint R, and initial iterate θ_0.
$\theta \leftarrow \theta_0$.
while not converged **do**
  for $j = 1, \ldots, q$ **do**
    Compute the jth gradient $\nabla \bar{\mathcal{L}}_n(\theta_{:j}) = \hat{\Gamma}^{(j)}\theta_{:j} - \hat{\gamma}^{(j)} - \sum_{i=1}^p \theta_{ij} \frac{q'_\lambda(\|\theta_{i:}\|_2)}{\|\theta_{i:}\|_2}$.
  end for
  Line search: Let stepsize ζ_t be the largest element of $\{c^t\zeta_0\}_{t=1,\ldots}$ such that

$$\|\text{Prox}_\lambda(\theta - \zeta_t \nabla \bar{\mathcal{L}}_n(\theta)))\|_{1,2} < R.$$

  $\theta \leftarrow \text{Prox}_\lambda(\theta - \zeta_t \nabla \bar{\mathcal{L}}_n(\theta)))$.
end while
Return estimate θ.

8 Additional Experiments

8.1 Synthetic experiments for $q = 40$

Figure 7 shows results for $q = 40$ following the setup in the main text.

Figure 8 compares our joint sparse approach to the individual sparsity approach in the case where all t values share Gaussian coefficients on each element in S (best for joint sparsity) and in the case where only 1 t value has nonzero coefficients in S (best case for individual sparsity). Our joint approach gains approximately $\sqrt{q}$ factor in sample complexity of 0.9 accuracy when each t value has a (standard Normal) nonzero coefficient for each entry of S, and even in the extreme case where only one value of t has nonzero coefficients, we still do not lose performance relative to the individual sparsity approach.

8.2 Additional real data experiments

Cattaneo2. The main text in Table 1 showed results for regularization parameter chosen via cross validation. We now consider robustness of the effect estimation to misspecification of λ. Table 3 shows results for λ chosen too high (yielding very sparse S with average $|S|$ of 3) and too low (yielding nonsparse S with average $|S|$ of 15). Both estimates perform somewhat worse than the results in the main text, but still better than the nonsparse estimate (again shown in the main text), indicating that our approach still tends to select useful covariates.

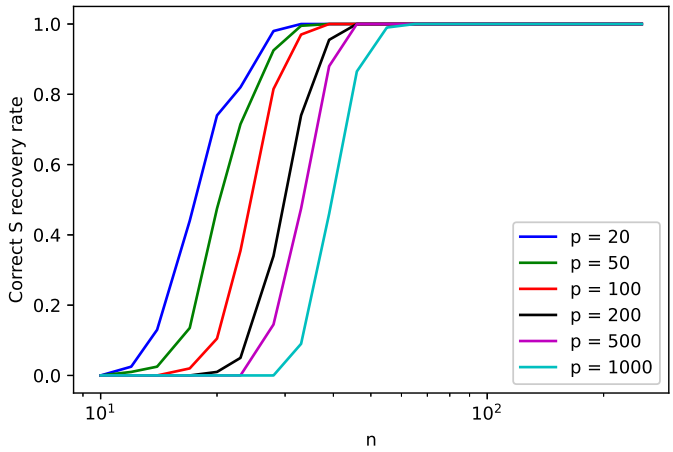

Figure 7: Empirical probability of our joint sparse algorithm (upper) and individual sparsity approach (lower) correctly recovering the sparse set S as a function of n and p, for $q = 40$.

IHDP. For the IHDP data, we know that S is sparse (since the datset is semisynthetic), but we aren't told about the sparsity of the set $X_1 \cup X_2$. To answer this question, we used the doubly robust estimator with covariates selected as those 12 (out of 25) with the largest magnitude coefficients when regressing treatment T versus X. The resulting treatment effect estimate was 5.61, with variance 0.623. This is actually not only worse than our approach, but worse than the nonsparse estimate as well (see main text).

9 Proof of Lemma 1

Proof. Chapter 11 of Pearl, 2009 gives two sufficient conditions for strong ignorability and c-equivalence. If A and A' are two sets of covariates, then if either of

$$\begin{aligned} &\text{(a)} \quad T \perp A'|A, \quad \text{and} \quad Y \perp A|T, A', \\ &\text{(b)} \quad T \perp A|A', \quad \text{and} \quad Y \perp A'|T, A \end{aligned}$$

are satisfied, then A' is c-equivalent to A and we can replace A with A' in the treatment effect estimation.

Let us use the graph in Figure 2 to check the c-equivalence of S to X, using condition (a).

1. $T \perp S|X$ immediately since S is a subset of X.
2. $Y \perp X|T, S$ holds since the graph indicates that T, S form a Markov blanket for Y.

We also verify it for $X_1 \cup X_2$, using condition (b):

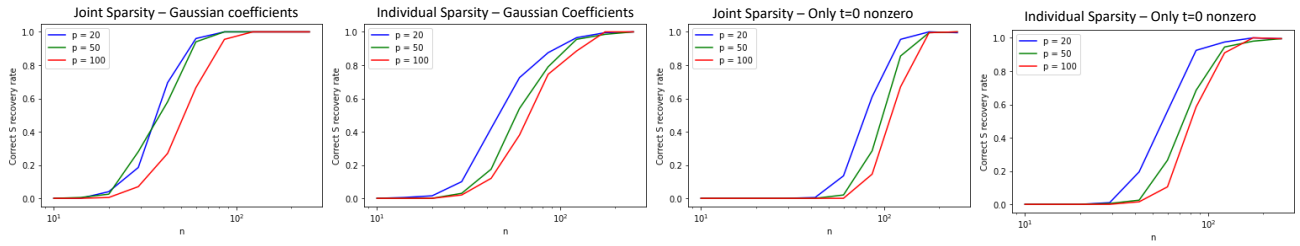

Figure 8: Exact recovery of S $(q, k = 10)$. (a) std Gaussian coefficients. (b) $t = 0$ has unit coefficients, all other t zero.

	Sparse DR Estimate (Ours, too sparse)	Sparse DR Estimate (Ours, less sparse)
Effect of 1 vs. 0	-217.1g(21.2)	-152.2g(24.7)
Effect of 2 vs. 0	-279.4g(17.4)	-195.6g(34.8)
Effect of 3 vs. 0	-302.9g(17.5)	-197.0g(34.0)
Binary effect(>0 vs 0)	-269.1g(14.7)	-194.7g(31.8)

Table 3: Estimated average treatment effects on Cattaneo2 dataset, showing for larger regularization yielding sparser S (on average cardinality of 3) and smaller regularization yielding less sparse S (average cardinality of 15). Compare to Table 1 in the main text. Actions – 0: no smoking, 1: 1-5 cigarettes daily, 2: 6-10 daily, and 3: 11 or more. For binary action effect, the empirical estimated interval is known to be (-250g, -200g). Standard deviations over 20 random data splits are given in parentheses.

1. $T \perp X|(X_1 \cup X_2)$ holds since the graph indicates that $T, X_1 \cup X_2$ form a Markov blanket for T.
2. $Y \perp (X_1 \cup X_2)|T, X$ immediately since $X_1 \cup X_2$ is a subset of X.

□

10 Proof of Lemma 4

We first state the following lemma, which allows us to use the first of the two joint RSC conditions.

Lemma 8. *Suppose $\hat{\theta}$ is a zero subgradient point of the objective (5) supported on S, i.e.*

$$\nabla \mathcal{L}_n(\hat{\theta}_S) + \nabla \rho_\lambda(\hat{\theta}_S) = 0. \tag{18}$$

Let $\tilde{\nu} := \hat{\theta} - \theta^*$. Then $\|\tilde{\nu}\|_F \leq 1$.

Lemma 8 implies that $\|\hat{\theta}_S - \theta^*_S\|_F \leq 1$. Hence the first joint RSC condition (6) applies, so we have

$$\langle \nabla \mathcal{L}(\hat{\theta}_S) - \nabla \mathcal{L}(\theta^*_S), \tilde{\nu} \rangle \geq \alpha_1 \|\tilde{\nu}\|_F^2 - \tau_1 \frac{\log k}{n} \|\tilde{\nu}\|_{1,2}^2. \tag{19}$$

We also have, by the convexity of $\rho_\lambda(\theta) + \mu/2\|\theta\|_F^2$ implied by the μ-amenability of ρ_λ, that

$$\langle \nabla \rho_\lambda(\hat{\theta}_S), \theta^*_S - \hat{\theta_S} \rangle \leq \rho_\lambda(\theta^*_S) - \rho_\lambda(\hat{\theta}_S) + \frac{\mu}{2} \|\tilde{\nu}\|_F^2. \tag{20}$$

We know that since $\hat{\theta}$ is a stationary point, $\langle \mathcal{L}_n(\hat{\theta}_S) + \nabla \rho_\lambda(\hat{\theta}_S), \theta_S - \hat{\theta}_S \rangle \geq 0$ for all feasible θ. Using this fact with (19) and (20) yields

$$\begin{aligned}
&(\alpha_1 - \mu/2)\|\tilde{\nu}\|_F^2 \\
&\leq -\langle \nabla \mathcal{L}_n(\theta^*_S), \tilde{\nu} \rangle + \rho_\lambda(\theta^*_S) - \rho_\lambda(\hat{\theta}_S) + \tau_1 \frac{\log k}{n} \|\tilde{\nu}\|_{1,2}^2 \\
&\leq \rho_\lambda(\theta^*_S) - \rho_\lambda(\hat{\theta}_S) + \left(\|\nabla \mathcal{L}_n(\theta^*_S)\|_{\infty,2} + R\tau_1 \frac{\log k}{n} \right) \|\tilde{\nu}\|_{1,2},
\end{aligned} \tag{21}$$

where we have again applied (24).

Now by (7) and the fact that $\tau_1 = q$ by Lemma 3, we have

$$\begin{aligned}
&\|\nabla \mathcal{L}_n(\theta^*_S)\|_{\infty,2} + R\tau_1 \frac{\log k}{n} \\
&\leq c' \sqrt{\frac{q \log p}{n}} + \sqrt{\frac{R^2 q \log k}{n}} \sqrt{\frac{q \log p}{n}} \\
&\leq \frac{\lambda}{2} + \frac{\lambda}{2} = \lambda,
\end{aligned} \tag{22}$$

where we have used the assumptions that $\lambda \geq c_\ell \sqrt{\frac{q \log p}{n}}$ and $n \geq CR^2 q \log p$ where here we require $c_\ell \geq 2c'$ and $C \geq \frac{1}{4c_\ell^2}$.

Also recall that the definition of (μ, γ) amenability states that the function $\rho_\lambda + \frac{\mu t^2}{2}$ is convex over the real line, $\lim_{t \to 0^+} \rho'_\lambda(t) = \lambda$, and ρ_λ is symmetric about 0. Combining these facts implies that for scalar t, $\lambda|t| \leq \rho_\lambda(t) + \frac{\mu t^2}{2}$. This in turn implies by substitution that $\lambda \|\theta\|_{1,2} \leq \rho_\lambda(\theta) + \frac{\mu \|\theta\|_F^2}{2}$.

We use this fact, the subadditivity of ρ_λ (implied by the condition that $\frac{\rho_\lambda(t)}{t}$ is nonincreasing on $\mathbb{R}^+$), and the inequality (22) to simplify (21) as

$$\begin{aligned}
&(\alpha_1 - \mu/2)\|\tilde{\nu}\|_F^2 \leq \rho_\lambda(\theta_S^*) - \rho_\lambda(\hat{\theta}_S) + \lambda \|\tilde{\nu}\|_{1,2} \\
&\leq \rho_\lambda(\theta_S^*) - \rho_\lambda(\hat{\theta}_S) + \lambda \left(\rho_\lambda(\tilde{\nu})/\lambda + \frac{\mu}{2\lambda} \|\tilde{\nu}\|_F^2 \right) \\
&\leq \rho_\lambda(\theta_S^*) - \rho_\lambda(\hat{\theta}_S) + \lambda \left((\rho_\lambda(\hat{\theta}_S) + \rho_\lambda(\theta_S^*))/\lambda + \frac{\mu}{2\lambda} \|\tilde{\nu}\|_F^2 \right) \\
&= 2\rho_\lambda(\theta_S^*) + \frac{\mu}{2} \|\tilde{\nu}\|_F^2,
\end{aligned}$$

hence $0 \leq (\alpha_1 - \mu)\|\tilde{\nu}\|_F^2 \leq 2\rho_\lambda(\theta_S^*) \leq 2\lambda \|\theta_S^*\|_{1,2} \leq R\lambda$, implying that

$$\|\tilde{\nu}\|_F \leq \sqrt{\frac{R\lambda}{\alpha_1 - \mu}}$$

and thus via a norm inequality

$$\|\tilde{\nu}\|_{1,2} \leq \sqrt{\frac{Rk\lambda}{\alpha_1 - \mu}}.$$

By the triangle inequality we then have

$$\|\hat{\theta}_S\|_{1,2} \leq \|\theta^*\|_{1,2} + \|\hat{\theta}_S - \theta_S^*\|_{1,2} \leq \frac{R}{2} + \sqrt{\frac{Rk\lambda}{\alpha_1 - \mu}} < R.$$

where the last inequality follows by the fact that $R > \frac{4k\lambda}{\alpha_1 - \mu}$ under our assumptions. □

11 Proof of Lemma 3

By the proof of Corollary 1 in [Loh and Wainwright, 2015] and using the fact that our loss function (3) decouples across columns, we have that with probability at least $1 - qc_1 \exp(-cn)$ and $n \geq O(k \log p)$,

$$\langle \nabla \mathcal{L}(\theta + \Delta) - \nabla \mathcal{L}(\theta), \Delta \rangle \geq \frac{1}{2} \min_j (\lambda_{\min}(\Sigma_j)) \|\Delta\|_F^2 - \frac{\log p}{n} \sum_j \|\Delta_{:j}\|_1^2.$$

We require the following lemma.

Lemma 9. *For $A \in \mathbb{R}^{p \times q}$, $\|A\|_{1,2} \geq \frac{1}{\sqrt{q}} \|A^T\|_{2,1}$.*

Proof. We have $\|A\|_{1,2} = \sum_i \|A_{i:}\|_2$ and $\|A^T\|_{2,1} = \sqrt{\sum_j \|A_{:j}\|_1^2}$. Note that

$$\|A\|_{2,1} \leq \sqrt{q} \|A\|_F \leq \sqrt{q} \|A\|_{1,2}.$$

□

Applying Lemma 9, we have

$$\langle \nabla \mathcal{L}(\theta + \Delta) - \nabla \mathcal{L}(\theta), \Delta \rangle \geq \frac{1}{2} \min_j (\lambda_{\min}(\Sigma_j)) \|\Delta\|_F^2 - \frac{q \log p}{n} \|\Delta\|_{1,2}^2,$$

as desired for the $\|\Delta\|_F \leq 1$ case.

If $\|\Delta\|_F \geq 1$, then by the constraint $\|\Delta\|_{1,2} \leq R$ and assumption $n \geq 4R^2 q \log p$ we have

$$\frac{1}{2}\min_j(\lambda_{\min}(\Sigma_j))\|\Delta\|_F^2 - \frac{q\log p}{n}\|\Delta\|_{1,2}^2 \geq \frac{1}{2}\min_j(\lambda_{\min}(\Sigma_j))\|\Delta\|_F - \sqrt{\frac{q\log p}{n}}\|\Delta\|_{1,2}.$$

Moving onto the second part of the lemma, we have (since $\mathcal{L}_n$ is the least squares loss) that

$$\nabla^2\mathcal{L}_n(\theta) = \text{diag}\left(\left\{\frac{X_j^T X_j}{n}\right\}_{j=1}^q\right),$$

where diag indicates the block diagonal matrix formed with the given blocks. Now since the X_j are subgaussian with covariance $\Sigma_x^{(j)}$, we have that (Proposition 2.1 of Vershynin, 2012)

$$|||((1/n)[X_j^T X_j]_{SS}) - ([\Sigma_x^{(j)}]_{SS})|||_2 \leq |||\Sigma_x^{(j)}|||_2\sqrt{\frac{k\log p}{n}}$$

with probability at least $1 - c_1\exp(-c_2\log p)$. Since we have assumed that $\lambda_{\min}([\Sigma_x^{(j)}]_{SS}) > 2\mu$, we therefore have

$$\lambda_{\min}(([X_j^T X_j]_{SS}/n) \geq 2\mu - \mu > \mu$$

for $n > \frac{k\log p|||\Sigma_x^{(j)}|||_2^2}{\mu^2}$.

With the union bound we thus have that the function $\mathcal{L}_n(\theta_S) - \frac{\mu}{2}\|\theta_S\|_F^2$ is strictly convex with probability at least $1 - c_1 q\exp(-c_2\log p)$. By the definition of (μ,γ) amenability, we known that $\rho_\lambda - \frac{\mu}{2}t^2$ is convex. Since the addition of a strictly convex function and a convex function is strictly convex, the lemma results. □

12 Proof of Lemma 8

Suppose $\|\tilde{\nu}\|_F > 1$. Then by joint RSC (6) we have

$$\langle\nabla\mathcal{L}(\hat{\theta}) - \nabla\mathcal{L}(\theta^*), \tilde{\nu}\rangle \geq \alpha_2\|\tilde{\nu}\|_F - \tau_2\sqrt{\frac{\log p}{n}}\|\tilde{\nu}\|_{1,2}.$$

Since $\hat{\theta}$ is a stationary point, $\nabla\mathcal{L}(\hat{\theta}) + \nabla\rho_\lambda(\hat{\theta}) = 0$ and we thus have

$$\langle-\nabla\rho_\lambda(\hat{\theta}) - \nabla\mathcal{L}(\theta^*), \tilde{\nu}\rangle \geq \alpha_2\|\tilde{\nu}\|_F - \tau_2\sqrt{\frac{\log p}{n}}\|\tilde{\nu}\|_{1,2}. \tag{23}$$

Recall that for equal sized matrices A, B

$$\begin{aligned}\langle A, B\rangle &= \sum_i\langle A_{i:}, B_{i:}\rangle \\ &\leq \sum_i\|A_{i:}\|_2\|B_{i:}\|_2 \\ &\leq \left(\max_i\|A_{i:}\|_2\right)\left(\sum_i\|B_{i:}\|_2\right) \\ &= \|A\|_{\infty,2}\|B\|_{1,2},\end{aligned} \tag{24}$$

where for both inequalities we have applied Holder's inequality. We can then write

$$\langle-\nabla\rho_\lambda(\hat{\theta}) - \nabla\mathcal{L}(\theta^*), \tilde{\nu}\rangle \leq \left(\|\nabla\rho_\lambda(\hat{\theta})\|_{\infty,2} + \|\nabla\mathcal{L}(\theta^*)\|_{\infty,2}\right)\|\tilde{\nu}\|_{1,2} \leq (\lambda + \lambda/2)\|\tilde{\nu}\|_{1,2}, \tag{25}$$

where the last inequality follows from the definition of ρ_λ and applying (7) in the main text that yields $\|\nabla\mathcal{L}(\theta^*)\|_{\infty,2} \leq \lambda/2$ when $c_\ell \geq 2c'$.

Combining (25) with (23) yields

$$\begin{aligned}\alpha_2\|\tilde{\nu}\|_F - \tau_2\sqrt{\frac{\log p}{n}}\|\tilde{\nu}\|_{1,2} &\leq 1.5\lambda\|\tilde{\nu}\|_{1,2},\\ \|\tilde{\nu}\|_F &\leq \frac{\|\tilde{\nu}\|_{1,2}}{\alpha_2}\left(1.5\lambda + \tau_2\sqrt{\frac{\log p}{n}}\right)\\ &\leq \frac{2R}{\alpha_2}\left(1.5\lambda + \tau_2\sqrt{\frac{\log p}{n}}\right)\\ &\leq \frac{2R}{\alpha_2}\left(1.5\frac{c_u}{R} + \tau_2\sqrt{\frac{\log p}{n}}\right).\end{aligned}$$

Note that the right hand side is ≤ 1 when c_u is chosen satisfying $c_u \geq \frac{\alpha_2}{6}$ and $n \geq \frac{16}{\alpha_2^2}R^2\tau_2^2\log p$ (since $\tau_2 = \sqrt{q}$, corresponds to having $C \geq \frac{16}{\alpha_2^2}$ in the statement of Theorem 1), yielding a contradiction with our earlier assumption. □

13 Proof of Lemma 5

We have for all $i \in S$

$$\|\hat{\theta}_{i:}\|_2 \geq \|\theta^*_{i:}\|_2 - |\langle\hat{\theta}_{i:} - \theta^*_{i:}, \theta^*_{i:}/\|\theta^*_{i:}\|_2\rangle|.$$

Now by an easy extension of the argument in Appendix D.1.1 of [Loh and Wainwright, 2017], we have that

$$\max_i |\langle\hat{\theta}_{i:} - \theta^*_{i:}, \theta^*_{i:}/\|\theta^*_{i:}\|_2\rangle| \leq c_3\sqrt{\frac{\log p}{n}}$$

with probability at least $1 - c_1\exp(-c_2\min(k, \log p))$. We then have

$$\|\hat{\theta}_{i:}\|_2 \geq \lambda\gamma + c_3\sqrt{\frac{\log p}{n}} - c_3\sqrt{\frac{\log p}{n}} = \lambda\gamma.$$

Recall that by Definition 3 of (μ, γ) amenability, we have that $\rho'_\lambda(t) = 0$ for all $t \geq \gamma\lambda$. □

14 Proof of Lemma 6

Define $\tilde{\nu} = \tilde{\theta} - \hat{\theta}$, where recall $\hat{\theta}$ is the oracle estimate (5). We will show that $\|\tilde{\nu}\|_F \leq 1$. By contradiction, suppose that $\|\tilde{\nu}\|_F > 1$. Then by the RSC condition (6)

$$\langle\nabla\mathcal{L}_n(\tilde{\theta}) - \nabla\mathcal{L}_n(\hat{\theta})\rangle \geq \alpha_2\|\tilde{\nu}\|_F - \tau_2\sqrt{\frac{\log p}{n}}\|\tilde{\nu}\|_{1,2}.$$

Since both $\hat{\theta}$ and $\tilde{\theta}$ are stationary points and $\hat{\theta}$ is an interior local minimum (by Step 2), we have

$$\begin{aligned}\langle\nabla\mathcal{L}_n(\tilde{\theta}) + \nabla\rho_\lambda(\tilde{\theta}), \hat{\theta} - \tilde{\theta}\rangle &\geq 0\\ \nabla\mathcal{L}_n(\hat{\theta}) + \nabla\rho_\lambda(\hat{\theta}) &= 0.\end{aligned}$$

Combining inequalities yields

$$\begin{aligned}\alpha_2\|\tilde{\nu}\|_F - \tau_2\sqrt{\frac{\log p}{n}}\|\tilde{\nu}\|_{1,2} &\leq \langle-\nabla\mathcal{L}_n(\hat{\theta}) + \nabla\rho_\lambda(\tilde{\theta}), \tilde{\nu}\rangle\\ &= \langle\nabla\rho_\lambda(\hat{\theta}) + \nabla\rho_\lambda(\tilde{\theta}), \tilde{\nu}\rangle\\ &\leq (\|\nabla\rho_\lambda(\hat{\theta})\|_{\infty,2} + \|\nabla\rho_\lambda(\tilde{\theta})\|_{\infty,2})\|\tilde{\nu}\|_{1,2},\end{aligned}$$

where we have applied the norm inequality (24). Recall that by (μ, γ)-amenability (see Lemma 8 of [Loh and Wainwright, 2017]) $\|\nabla \rho_\lambda(\theta)\|_{\infty,2} \leq \lambda$ for any θ. Hence we can rearrange and obtain

$$\|\tilde{\nu}\|_F \leq \frac{\|\tilde{\nu}\|_{1,2}}{\alpha_2}\left(2\lambda + \tau_2\sqrt{\frac{\log p}{n}}\right) \leq \frac{2R}{\alpha_2}\left(2\lambda + \tau_2\sqrt{\frac{\log p}{n}}\right)$$

due to the norm constraint on the objective (2). Since we have assumed $\lambda \leq \frac{\alpha_2}{8R}$ and $n \geq \frac{16}{\alpha_2^2}R^2\tau_2^2 \log p$, $\|\tilde{\nu}\|_F \leq 1$ as desired.

We can then apply the appropriate RSC condition from (6) yielding

$$\langle \nabla \mathcal{L}_n(\tilde{\beta}) - \nabla \mathcal{L}_n(\hat{\beta}), \tilde{\nu}\rangle \geq \alpha_1 \|\tilde{\nu}\|_2^2 - \tau_1 \frac{\log p}{n}\|\tilde{\nu}\|_{1,2}^2,$$

and (recalling the definition of $\bar{\mathcal{L}}_n$ from (13))

$$\langle \nabla \bar{\mathcal{L}}_n(\tilde{\beta}) - \nabla \bar{\mathcal{L}}_n(\hat{\beta}), \tilde{\nu}\rangle \geq (\alpha_1 - \mu)\|\tilde{\nu}\|_2^2 - \tau_1 \frac{\log p}{n}\|\tilde{\nu}\|_{1,2}^2. \tag{26}$$

By the first order optimality conditions we have

$$\begin{aligned}
\langle \nabla \bar{\mathcal{L}}_n(\tilde{\theta}), \hat{\theta} - \tilde{\theta}\rangle + \lambda\langle \tilde{z}, \hat{\theta} - \tilde{\theta}\rangle = 0,\\
\langle \nabla \bar{\mathcal{L}}_n(\hat{\theta}), \tilde{\theta} - \hat{\theta}\rangle + \lambda\langle \hat{z}, \tilde{\theta} - \hat{\theta}\rangle = 0,
\end{aligned}$$

where $\tilde{z} \in \partial\|\tilde{\theta}\|_{1,2}$. Combining these and using the definition of subgradient yields

$$\begin{aligned}
&\langle \nabla \bar{\mathcal{L}}_n(\hat{\theta}) - \nabla \bar{\mathcal{L}}_n(\tilde{\theta}), \tilde{\theta} - \hat{\theta}\rangle + \lambda\langle \hat{z}, \tilde{\theta}\rangle - \lambda\|\hat{\theta}\|_{1,2} + \lambda\langle \tilde{z}, \hat{\theta}\rangle - \lambda\|\tilde{\theta}\|_{1,2} \geq 0,\\
&\lambda\|\tilde{\theta}\|_{1,2} - \lambda\langle \hat{z}, \tilde{\theta}\rangle \leq \langle \nabla \bar{\mathcal{L}}_n(\hat{\theta}) - \nabla \bar{\mathcal{L}}_n(\tilde{\theta}), \tilde{\theta} - \hat{\theta}\rangle + \lambda\|\tilde{z}\|_{\infty,2}\|\hat{\theta}\|_{1,2} - \lambda\|\hat{\theta}\|_{1,2},\\
&\lambda\|\tilde{\theta}\|_{1,2} - \lambda\langle \hat{z}, \tilde{\theta}\rangle \leq \langle \nabla \bar{\mathcal{L}}_n(\hat{\theta}) - \nabla \bar{\mathcal{L}}_n(\tilde{\theta}), \tilde{\theta} - \hat{\theta}\rangle,\\
&\lambda\|\tilde{\theta}\|_{1,2} - \lambda\langle \hat{z}, \tilde{\theta}\rangle \leq \tau_1 \frac{\log p}{n}\|\tilde{\nu}\|_{1,2}^2 - (\alpha_1 - \mu)\|\tilde{\nu}\|_F^2,
\end{aligned} \tag{27}$$

where we have used the fact that $\|\tilde{z}\|_{\infty,2} \leq 1$ since $\tilde{\theta}$ is feasible and applied the bound (26).

We also have the following result.

Lemma 10. *If* $\lambda \geq \frac{4R\tau_1 q \log p}{\delta n}$ *and* $\|\hat{z}_{S^c}\|_{\infty,2} \leq 1 - \delta$, *then*

$$\|\tilde{\nu}\|_{1,2} \leq \left(\frac{4}{\delta} + 2\right)\sqrt{k}\|\tilde{\nu}\|_F.$$

Proof. Applying (26) to (27) yields

$$\lambda\langle \hat{z}, \tilde{\theta}\rangle + \lambda\langle \tilde{z}, \hat{\theta}\rangle - \lambda\|\tilde{\theta}\|_{1,2} \geq \langle \nabla \bar{\mathcal{L}}_n(\tilde{\theta}) - \nabla \bar{\mathcal{L}}_n(\hat{\theta}), \tilde{\nu}\rangle \geq (\alpha_1 - \mu)\|\tilde{\nu}\|_2^2 - \tau_1 \frac{\log p}{n}\|\tilde{\nu}\|_{1,2}^2. \tag{28}$$

Recalling that $\hat{\beta}$ is supported on S and $\|\tilde{z}\|_{\infty,2} \leq 1$, we can also write

$$\lambda\langle \tilde{z}, \hat{\theta}\rangle - \lambda\|\tilde{\theta}\|_{1,2} \leq \lambda(\|\hat{\theta}\|_{1,2} - \|\tilde{\theta}_S\|_{1,2} - \|\tilde{\theta}_{S^c}\|_{1,2}) \leq \lambda(\|\tilde{\nu}_S\|_{1,2} - \|\tilde{\nu}_{S^c}\|_{1,2}). \tag{29}$$

Additionally we can use the norm inequality (24) to bound

$$\begin{aligned}
\lambda\langle \hat{z}, \tilde{\nu}\rangle &= \lambda\langle \hat{z}_S, \tilde{\nu}_S\rangle + \lambda\langle \hat{z}_{S^c}, \tilde{\nu}_{S^c}\rangle\\
&\leq \lambda(\|\hat{z}_S\|_{\infty,2}\|\tilde{\nu}_S\|_{1,2} + \|\hat{z}_{S^c}\|_{\infty,2}\|\tilde{\nu}_{S^c}\|_{1,2})\\
&\leq \lambda(\|\tilde{\nu}_S\|_{1,2} + (1-\delta)\|\tilde{\nu}_{S^c}\|_{1,2})
\end{aligned} \tag{30}$$

where we have used the assumption $\|\hat{z}_{S^c}\|_{\infty,2} \leq 1 - \delta$ from the lemma statement.

Combining (28), (29), and (30) yields

$$-\tau_1 \frac{\log p}{n}\|\tilde{\nu}\|_{1,2}^2 \leq (\alpha_1-\mu)\|\tilde{\nu}\|_2^2 - \tau_1 \frac{\log p}{n}\|\tilde{\nu}\|_{1,2}^2 \leq \lambda(2\|\tilde{\nu}_S\|_{1,2} - \delta\|\tilde{\nu}_{S^c}\|_{1,2}).$$

Our assumption on λ implies that $\tau_1 \frac{\log p}{n}\|\tilde{\nu}\|_{1,2} \leq 2R\tau_1 \frac{\log p}{n} \leq \frac{\delta}{2}\lambda$, yielding

$$-\frac{\delta}{2}\lambda\|\tilde{\nu}\|_{1,2} \leq \lambda(2\|\tilde{\nu}_S\|_{1,2} - \delta\|\tilde{\nu}_{S^c}\|_{1,2})$$

or equivalently

$$\frac{\delta}{2}\|\tilde{\nu}_{S^c}\|_{1,2} \leq \left(2+\frac{\delta}{2}\right)\|\tilde{\nu}_S\|_{1,2}.$$

We can then write (using a norm inequality)

$$\|\tilde{\nu}\|_{1,2} = \|\tilde{\nu}_S\|_{1,2} + \|\tilde{\nu}_{S^c}\|_{1,2} \leq \|\tilde{\nu}_S\|_{1,2}\left(1+\frac{4}{\delta}+1\right) \leq \left(2+\frac{4}{\delta}\right)\sqrt{k}\|\tilde{\nu}\|_F.$$

□

Recall we have assumed $\frac{c_u\sqrt{q}}{R} \geq \lambda \geq c_\ell\sqrt{\frac{q\log p}{n}}$, implying for our choices of $\delta = 1/2$ and c_ℓ, c_u

$$\begin{aligned}
\lambda &\geq c_\ell\sqrt{\frac{q\log p}{n}} \\
&= c_\ell\sqrt{\frac{q\log p}{n}}\frac{R}{c_u\sqrt{q}}\frac{c_u\sqrt{q}}{R} \\
&\geq \frac{Rc_\ell^2}{c_u\sqrt{q}}\frac{q\log p}{n} \\
&= \frac{4R\tau_1\sqrt{q}\log p}{\delta n}.
\end{aligned}$$

Thus we can apply Lemma 10 to (27), and have

$$\lambda\|\tilde{\theta}\|_{1,2} - \lambda\langle\hat{z},\tilde{\theta}\rangle \leq \tau_1\frac{k\log p}{n}\left(\frac{4}{\delta}+2\right)^2\|\tilde{\nu}\|_F^2 - (\alpha_1-\mu)\|\tilde{\nu}\|_2^2.$$

If $n \geq \frac{2\tau_1}{\alpha_1-\mu}\left(\frac{4}{\delta}+2\right)^2 k\log p$, $\lambda\|\tilde{\theta}\|_{1,2} - \lambda\langle\hat{z},\tilde{\theta}\rangle \leq 0$. But we know by (24) that $\langle\hat{z},\tilde{\theta}\rangle \leq \|\hat{z}\|_{\infty,2}\|\tilde{\theta}\|_{1,2} \leq \|\tilde{\theta}\|_{1,2}$ which implies $\lambda\|\tilde{\theta}\|_{1,2} - \lambda\langle\hat{z},\tilde{\theta}\rangle \geq 0$. Hence we have $\lambda\|\tilde{\theta}\|_{1,2} - \lambda\langle\hat{z},\tilde{\theta}\rangle = 0$ which implies $\langle\hat{z},\tilde{\theta}\rangle = \|\tilde{\theta}\|_{1,2}$. Our assumption that $\|\hat{z}_{S^c}\|_{1,2} < 1$ (strictly less than 1) implies $\tilde{\theta}_{S^c} = 0$, hence $\tilde{\theta}$ is supported on S. □