Supplementary Materials for
Approximate Leave-One-Out for Fast Parameter Tuning in High Dimensions

A Proof of Equation 7

In this Section, we prove the primal-dual correspondence in (6) and (7). Recall the form of the primal problem:

$$\min_{\boldsymbol{\beta}} \sum_{j=1}^{n} \ell(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_j) + R(\boldsymbol{\beta}). \tag{31}$$

With a change of variable, we may transform (31) into the following form:

$$\min_{\boldsymbol{\beta},\boldsymbol{\mu}} \sum_{j=1}^{n} \ell(-\mu_j; y_j) + R(\boldsymbol{\beta}), \quad \text{subject to: } \boldsymbol{\mu} = -\boldsymbol{X}\boldsymbol{\beta}.$$

We may further absorb the constraint into the objective function by adding a Lagrangian multiplier $\boldsymbol{\theta} \in \mathbb{R}^n$:

$$\max_{\boldsymbol{\theta}} \min_{\boldsymbol{\beta},\boldsymbol{\mu}} \sum_{j=1}^{n} \ell(-\mu_j; y_j) + R(\boldsymbol{\beta}) - \boldsymbol{\theta}^\top(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mu}). \tag{32}$$

Note that in (32), $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ decoupled from each other and we can optimize over them respectively. Specifically, we have that

$$\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) - \boldsymbol{\theta}^\top \boldsymbol{X}\boldsymbol{\beta} = -\max_{\boldsymbol{\beta}} \left\{ \langle \boldsymbol{\beta}, \boldsymbol{X}^\top \boldsymbol{\theta} \rangle - R(\boldsymbol{\beta}) \right\} = -R^*(\boldsymbol{X}^\top \boldsymbol{\theta}), \tag{33}$$

$$\min_{\mu_j} \ell(-\mu_j; y_j) - \theta_j \mu_j = -\max\{\mu_j \theta_j - \ell(-\mu_j; y_j)\} = -\ell^*(-\theta_j; y_j). \tag{34}$$

We plug (33) and (34) in (32) and obtain that

$$\max_{\boldsymbol{\theta}} \sum_{j=1}^{n} -\ell^*(-\theta_j; y_j) - R^*(\boldsymbol{X}^\top \boldsymbol{\theta}). \tag{35}$$

B Primal Dual Equivalence (Proofs of Theorems 5.1 and 5.2)

In this section we prove the equivalence between the two stated methods in the case where the loss and regularizer are twice differentiable. Let ℓ, ℓ^*, R and R^* be twice differentiable. We construct quadratic surrogates by Taylor extensions. The following lemma plays a key role in our analysis:

Lemma B.1. *Let f be a proper closed convex function, such that both f and f^* are twice differentiable. Then, we have for any $\boldsymbol{x}$ in the domain of f and any $\boldsymbol{u}$ in the domain of f^*:*

$$\nabla^2 f^*(\nabla f(\boldsymbol{x})) = [\nabla^2 f(\boldsymbol{x})]^{-1},$$
$$\nabla^2 f(\nabla f^*(\boldsymbol{u})) = [\nabla^2 f^*(\boldsymbol{u})]^{-1}.$$

Proof. This lemma is a known result in convex optimization. However, since the proof is short and for the sake of completeness we include the proof here. For f a proper closed convex function, we have by Theorem 23.5 of [Rockafellar, 1970] that for all $\boldsymbol{x}, \boldsymbol{x}^*$:

$$\boldsymbol{x}^* \in \partial f(\boldsymbol{x}) \Rightarrow \boldsymbol{x} \in \partial f^*(\boldsymbol{x}^*).$$

In particular, if f and f^* are differentiable, we obtain:

$$\boldsymbol{x} = \nabla f^*(\nabla f(\boldsymbol{x})).$$

Taking derivative in $\boldsymbol{x}$ once more, we obtain that:

$$\boldsymbol{I} = [\nabla^2 f^*(\nabla f(\boldsymbol{x}))][\nabla^2 f(\boldsymbol{x})],$$

which immediately gives:

$$\nabla^2 f^*(\nabla f(\boldsymbol{x})) = [\nabla^2 f(\boldsymbol{x})]^{-1}.$$

The proof of the second part is immediate by applying the existing result to f^*. □

Proof of Theorem 5.1. We have the following expressions for $\tilde{\ell}$ and $\tilde{R}$:

$$\begin{aligned}
\tilde{\ell}(z_j; y_j) =& \frac{1}{2}\ddot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j)(z_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}})^2 + \dot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j)(z_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}) + c, \\
\tilde{R}(\boldsymbol{\beta}) =& \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top [\nabla^2 R(\hat{\boldsymbol{\beta}})](\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + [\nabla R(\hat{\boldsymbol{\beta}})]^\top (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + d,
\end{aligned}$$

where $c, d \in \mathbb{R}$ are constants that do not affect the location of the optimizer. We now compute the convex conjugate of $\tilde{\ell}$ and $\tilde{R}$, and we obtain that:

$$\tilde{\ell}^*(w_j; y_j) = \frac{1}{2}\frac{1}{\ddot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j)}(w_j - \dot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j))^2 + (\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}})(w_j - \dot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j)) + c', \tag{36}$$

$$\tilde{R}^*(\boldsymbol{\mu}) = \frac{1}{2}(\boldsymbol{\mu} - \nabla R(\hat{\boldsymbol{\beta}}))^\top [\nabla^2 R(\hat{\boldsymbol{\beta}})]^{-1}(\boldsymbol{\mu} - \nabla R(\hat{\boldsymbol{\beta}})) + \hat{\boldsymbol{\beta}}^\top (\boldsymbol{\mu} - \nabla R(\hat{\boldsymbol{\beta}})) + d', \tag{37}$$

where again $c', d' \in \mathbb{R}$ are constants.

Now, we wish to relate (36) and (37) to $\tilde{\ell}_D^*$ and $\tilde{R}_D^*$. By substituting the primal-dual correspondence described in (8) of the main text for components of (36) and (37), we obtain that:

$$\tilde{\ell}^*(w_j; y_j) = \frac{1}{2}\frac{1}{\ddot{\ell}(\dot{\ell}^*(-\hat{\theta}_j; y_j); y_j)}(w_j + \hat{\theta}_j)^2 + \dot{\ell}^*(-\hat{\theta}_j; y_j)(w_j + \hat{\theta}_j) + c', \tag{38}$$

$$\begin{aligned}
\tilde{R}^*(\boldsymbol{\mu}) = & \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{X}^\top \hat{\boldsymbol{\theta}})^\top [\nabla^2 R(\nabla R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}}))]^{-1}(\boldsymbol{\mu} - \boldsymbol{X}^\top \hat{\boldsymbol{\theta}}) \\
& + [\nabla R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}})]^\top (\boldsymbol{\mu} - \boldsymbol{X}^\top \hat{\boldsymbol{\theta}}) + d'.
\end{aligned} \tag{39}$$

To conclude, we note that according to Lemma B.1 we have

$$\begin{aligned}
\ddot{\ell}(\dot{\ell}^*(-\hat{\theta}_j; y_j); y_j) &= (\ddot{\ell}^*(-\hat{\theta}_j; y_j))^{-1}, \\
\nabla^2 R(\nabla R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}})) &= [\nabla^2 R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}})]^{-1}.
\end{aligned} \tag{40}$$

Substitute (40) in (38) and (39) we obtain the dual of the quadratic surrogate equals

$$\begin{aligned}
\frac{1}{2}\sum_j \tilde{\ell}^*(-\theta_j; y_j) + \tilde{R}^*(\boldsymbol{X}^\top \theta) = & \frac{1}{2}\sum_j \ddot{\ell}^*(-\hat{\theta}_j; y_j)\Big(- \theta_j + \hat{\theta}_j + \frac{\dot{\ell}^*(-\hat{\theta}_j; y_j)}{\ddot{\ell}^*(-\hat{\theta}_j; y_j)}\Big)^2 \\
& + \frac{1}{2}(\boldsymbol{X}^\top \boldsymbol{\theta} - \boldsymbol{X}^\top \hat{\boldsymbol{\theta}})\nabla^2 R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}})(\boldsymbol{X}^\top \boldsymbol{\theta} - \boldsymbol{X}^\top \hat{\boldsymbol{\theta}}) \\
& + [\nabla R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}})]^\top (\boldsymbol{X}^\top \boldsymbol{\theta} - \boldsymbol{X}^\top \hat{\boldsymbol{\theta}}) + c'.
\end{aligned} \tag{41}$$

We recognize that the formula given in (41) exactly corresponds to the second-order Taylor expansion of (15) in the main paper, which is just the form of $\tilde{\ell}_D^*$ and $\tilde{R}_D^*$. □

Additionally, we show that the augmented dual method solves the surrogate quadratic problem.

Proof of Theorem 5.2. We noted in Section 3.2 of the main text that our dual method as described explicitly approximates the loss by its quadratic expansion at the optimal value. We may thus assume without loss of generality that the loss is given by $\ell(\mu; y) = (\mu - y)^2/2$.

In this case, as stated in Section 3.2, we have that

$$\hat{\boldsymbol{\theta}} = \mathbf{prox}_g(\boldsymbol{y}),$$

where we have defined $g(\boldsymbol{u}) = R^*(\boldsymbol{X}^\top \boldsymbol{u})$. In addition, we note that the augmented observation vector $\boldsymbol{y}_a$ must have its i^{th} observation lie on the leave-i-out regression line by definition, and in particular we have that:

$$[\mathbf{prox}_g(\boldsymbol{y}_a)]_i = 0.$$

This motivated us to solve for $\tilde{y}_i^{/i}$ by linearly expanding $\mathbf{prox}_g$ and considering the intersection of its i^{th} coordinate with 0. Specifically, the desired $\tilde{y}_i^{/i}$ is obtained from the solution of the following linear equation in z:

$$[\mathbf{prox}_g(\boldsymbol{y}) + \boldsymbol{J}_{\mathbf{prox}_g}(\boldsymbol{y})\boldsymbol{e}_i(z - y_i)]_i = 0. \tag{42}$$

where $\boldsymbol{J}_{\mathbf{prox}_g}(\boldsymbol{y})$ denotes the Jacobian matrix of $\mathbf{prox}_g$ at $\boldsymbol{y}$.

We show that if R^* is replaced with its quadratic surrogate $\tilde{R}^*$ as defined in the Theorem 5.1, then:

$$[\mathbf{prox}_{\tilde{g}}(\tilde{\boldsymbol{y}}_a)]_i = 0,$$

where $\tilde{g}(\boldsymbol{u}) = \tilde{R}^*(\boldsymbol{X}^\top \boldsymbol{u})$, and $\tilde{\boldsymbol{y}}_a$ denotes the vector $\boldsymbol{y}$, except with its i^{th} coordinate replaced by the ALO value $\tilde{\boldsymbol{y}}_i^{/i}$. Let us note that as $\tilde{g}$ is quadratic, its proximal map $\mathbf{prox}_{\tilde{g}}$ is linear, and the equation may thus be solved directly by a single Newton's step. As a linear map is characterized by its intercept and slope, compared with (42), it remains to show that:

$$\mathbf{prox}_g(\boldsymbol{y}) = \mathbf{prox}_{\tilde{g}}(\boldsymbol{y}), \tag{43}$$

$$\boldsymbol{J}_{\mathbf{prox}_g}(\boldsymbol{y}) = \boldsymbol{J}_{\mathbf{prox}_{\tilde{g}}}(\boldsymbol{y}). \tag{44}$$

We note that (43) is immediate from the definition of $\tilde{g}$, as both the left and right hand sides are equal to the dual optimal $\hat{\boldsymbol{\theta}}$. In order to show (44), since $\tilde{g}$ is quadratic, we may compute its proximal map exactly. From the previous section, we have that:

$$\tilde{g}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \boldsymbol{X}[\nabla^2 R(\nabla R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}}))]^{-1}\boldsymbol{X}^\top(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + [\nabla R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}})]^\top \boldsymbol{X}^\top(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}),$$

We minimize $\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2 + \tilde{g}(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$ and get

$$\mathbf{prox}_{\tilde{g}}(\boldsymbol{y}) = (\boldsymbol{I} + \boldsymbol{X}[\nabla^2 R(\nabla R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}}))]^{-1}\boldsymbol{X}^\top)^{-1}(\boldsymbol{y} - \boldsymbol{X}\nabla R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}})),$$

Notice the primal dual correspondence implies $\hat{\boldsymbol{\beta}} = \nabla R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}})$. In particular we may compute the Jacobian of $\mathbf{prox}_{\tilde{g}}$ at $\boldsymbol{y}$ as $(\boldsymbol{I} + \boldsymbol{X}[\nabla^2 R(\hat{\boldsymbol{\beta}})]^{-1}\boldsymbol{X}^\top)^{-1}$.

On the other hand, we know that the proximal operator $\mathbf{prox}_g$ is exactly the resolvent of the subgradient ∂g:

$$\mathbf{prox}_g = (I + \partial g)^{-1}$$

and in particular we have that:

$$\mathbf{prox}_g(\boldsymbol{y}) + \nabla g(\mathbf{prox}_g(\boldsymbol{y})) = \boldsymbol{y}.$$

Taking derivative again with respect to $\boldsymbol{y}$ and applying the chain rule, we obtain that:

$$\boldsymbol{J}_{\mathbf{prox}_g}(\boldsymbol{y})(\boldsymbol{I} + \nabla^2 g(\mathbf{prox}_g(\boldsymbol{y}))) = \boldsymbol{I},$$

and hence that:

$$\boldsymbol{J}_{\mathbf{prox}_g}(\boldsymbol{y}) = (\boldsymbol{I} + \nabla^2 g(\mathbf{prox}_g(\boldsymbol{y}))^{-1}.$$

Now, note that we have $\mathbf{prox}_g(\boldsymbol{y}) = \hat{\boldsymbol{\theta}}$, and that:

$$\nabla^2 g(\hat{\boldsymbol{\theta}}) = \boldsymbol{X}[\nabla^2 R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}})]\boldsymbol{X}^\top.$$

We are thus done by Lemma B.1. □

C Proof of Primal Approximation Approach

In this section we prove the results of our primal approach on nonsmooth models presented in Section 4 of the main paper rigorously. Since we use a kernel smoothing strategy, we start with some useful preliminary results on kernel smoothing. We then discuss nonsmooth loss and nonsmooth regularizer respectively.

C.1 Properties of Kernel Smoothing

In the paper, we consider the following smoothing strategy for a convex function $f : \mathbb{R} \to \mathbb{R}$:

$$f_h(z) = \frac{1}{h} \int f(u)\phi((z-u)/h)du \tag{45}$$

We make the following assumption about the kernel ϕ:

Compact support: ϕ has a compact support, i.e., $\text{supp}(\phi) = [-C, C]$ for some $C > 0$;

Normalization: ϕ kernel: $\int \phi(w)dw = 1$, $\phi(0) > 0$; $\phi(x) \geq 0$ for every x;

Symmetry: ϕ is smooth and symmetric around 0 on $\mathbb{R}$.

Let $K := \{v_1, \ldots, v_k\}$ denote the set of zero-order singularities of the function f. Denote by $\dot{f}_-$ and $\dot{f}_+$ the left and right derivative of f. Our next lemma summarizes some of the basic properties of f that may be used in the proofs of Theorem 4.1 and 4.2 of the main text.

Lemma C.1. *The smooth function f_h verifies the following properties:*

1. *$f_h(z) \geq f(z)$ for all $z \in \mathbb{R}$;*
2. *For all $z \in K^C$, for all h small enough:*

$$\dot{f}_h(z) = \frac{1}{h} \int \dot{f}(u)\phi((z-u)/h)du, \quad \ddot{f}_h(z) = \frac{1}{h} \int \ddot{f}(u)\phi((z-u)/h)du.$$

3. *For all $z \in K$:*

$$\lim_{h\to 0} \dot{f}_h(z) = \frac{\dot{f}_-(z) + \dot{f}_+(z)}{2}, \quad \lim_{h\to 0} \ddot{f}_h(z) = +\infty.$$

4. *If f is locally Lipschiz in the sense that, for any $A > 0$, and for any $x, y \in [-A, A]$, we have $|f(x) - f(y)| \leq L_A|x - y|$, where L_A is a constant that only depends on A; then $f_h(z)$ converges to $f(z)$ uniformly on any compact set.*

Proof. For part 1, by the normalization property of ϕ, we can treat ϕ as a probability density. Consider the random variable $U \sim \frac{1}{h}\phi(\frac{z-u}{h})$. From the convexity of f and Jensen's inequality we have

$$f_h(z) = \mathbb{E}f(U) \geq f(\mathbb{E}U) = f(z).$$

For part 2, note that

$$\dot{f}_h(z) = \frac{1}{h^2} \int f(u)\dot{\phi}((z-u)/h)du = \int \dot{f}(u)\frac{1}{h}\phi((z-u)/h)du.$$

A similar computation gives the stated equation for $\ddot{f}_h(z)$.

For part 3, when $z \in K$, we have by compact support of ϕ that as $h \to 0$:

$$\begin{aligned}
\dot{f}_h(z) &= \frac{1}{h^2}\int_{z-hC}^{z} f(u)\dot{\phi}((z-u)/h)du + \frac{1}{h^2}\int_{z}^{z+hC} f(u)\dot{\phi}((z-u)/h)du \\
&= \int_{-C}^{0} \dot{f}(z-hw)\phi(w)dw + \int_{0}^{C} \dot{f}(z-hw)\phi(w)dw \\
&\to \int_{-C}^{0} \dot{f}_+(z)\phi(w)dw + \int_{0}^{C} \dot{f}_-(z)\phi(w)dw \\
&= \frac{\dot{f}_+(z)+\dot{f}_-(z)}{2}.
\end{aligned}$$

A similar computation for the second-order derivative yields:

$$\begin{aligned}
\ddot{f}_h(z) &= \frac{1}{h^3}\int_{z-hC}^{z} f(u)\ddot{\phi}((z-u)/h)du + \frac{1}{h^3}\int_{z}^{z+hC} f(u)\ddot{\phi}((z-u)/h)du \\
&= \frac{1}{h}\phi(0)(\dot{f}_+(z)-\dot{f}_-(z)) + \int_{0}^{C} \ddot{f}(z-hw)\phi(w)dw + \int_{-C}^{0} \ddot{f}(z-hw)\phi(w)dw \\
&\to \infty.
\end{aligned}$$

noticing that $\dot{f}_+(z) > \dot{f}_-(z)$.

For part 4, for any compact set $\mathcal{C}$ which can be covered by a large enough set $[-A, A]$ for some $A > 0$, we have

$$\sup_{z\in\mathcal{C}} |f_h(z) - f(z)| \le \sup_{z\in\mathcal{C}} \int_{-C}^{C} |f(z-hw) - f(z)|\phi(w)dw \le 2hCL_{A+C} \to 0, \quad \text{as } h \to 0$$

□

Having established the basic properties of our kernel smoothing strategy, we apply them to non-smooth loss and non-smooth regularizer respectively.

C.2 Proof of Theorem 4.2: Nonsmooth Separable Regularizer With Smooth Loss

Consider the penalized regression problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{n} \ell(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_j) + \lambda \sum_{l} r(\beta_l). \tag{46}$$

with ℓ and r being twice differentiable and nonsmooth functions respectively. Let r_h be the smoothed version of r constructed as in (45). Define

$$\hat{\boldsymbol{\beta}}_h = \arg\min_{\boldsymbol{\beta}} \sum_{j} \ell(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_j) + \lambda \sum_{l} r_h(\beta_l).$$

As before, let K denote the set of all zero-order singularities of r. We make the following assumptions on the regularizer.

Assumption C.1. *We will need the following assumptions on the problem.*

1. *r is locally Lipschiz in the sense that, for any $A > 0$, and for any $x, y \in [-A, A]$, we have $|r(x) - r(y)| \le L_A|x - y|$, where L_A is a constant that only depends on A;*
2. *$\hat{\boldsymbol{\beta}}$ is the unique minimizer of (46);*

3. When $\hat{\beta}_l = v \in K$, the subgradient $g_r(\hat{\beta}_l)$ of r at $\hat{\beta}_l$ satisfies $g_r(\hat{\beta}_l) \in (\dot{r}_-(v), \dot{r}_+(v))$.

4. r is coercive in the sense that $|r(z)| \to \infty$ as $|z| \to \infty$.

Lemma C.2. *Suppose that Assumption C.1 holds. There exists $M > 0$ that only depends on r, ℓ and λ, such that we have for any $h \leq 1$:*

$$\|\hat{\boldsymbol{\beta}}\|_\infty, \|\hat{\boldsymbol{\beta}}_h\|_\infty < M.$$

Proof. Let $h \leq 1$, then the minimizer of the smoothed version $\hat{\boldsymbol{\beta}}_h$ satifies

$$\begin{aligned}
\lambda \sum_{l=1}^{p} r([\hat{\boldsymbol{\beta}}_h]_l) &\leq \lambda \sum_{l=1}^{p} r_h([\hat{\boldsymbol{\beta}}_h]_l) \\
&\leq \sum_i \ell(y_i; 0) + \lambda p r_h(0) \\
&= \sum_i \ell(y_i; 0) + \lambda p \int_{-C}^{C} r(hw)\phi(w)dw \\
&\leq \sum_i \ell(y_i; 0) + \lambda p \sup_{|w| \leq C} r(w).
\end{aligned}$$

On the other hand, the minimizer $\hat{\boldsymbol{\beta}}$ of the original problem satisfies

$$\lambda \sum_{l=1}^{p} r([\hat{\boldsymbol{\beta}}]_l) \leq \sum_i \ell(y_i; 0) + \lambda p r(0) \leq \sum_i \ell(y_i; 0) + \lambda p \sup_{|w| \leq C} r(w).$$

The convexity and coerciveness of r implies that there exists an M, such that for all $h \leq 1$:

$$\|\hat{\boldsymbol{\beta}}_h\|_\infty \leq M \text{ and } \|\hat{\boldsymbol{\beta}}\|_\infty \leq M.$$

□

Lemma C.3. *Suppose that Assumption C.1 holds. Then the smoothed version converges to the original problem in the sense that:*

$$\|\hat{\boldsymbol{\beta}}_h - \hat{\boldsymbol{\beta}}\|_2 \to 0 \text{ as } h \to 0.$$

Proof. By the local Lipschitz condition of r, we have for any $z \leq M$ and $h \leq 1$:

$$0 \leq r_h(z) - r(z) = \int_{-C}^{C} [r(z - hw) - r(z)]\phi(w)dw \leq 2CL_{M+C}h \tag{47}$$

Let $P_h(\boldsymbol{\beta}) := \sum_j \ell(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_j) + \lambda \sum_l r_h(\beta_l)$ denote the primal objective value. (47) implies that:

$$\sup_{\|\boldsymbol{\beta}\|_\infty \leq M} |P(\boldsymbol{\beta}) - P_h(\boldsymbol{\beta})| \leq 2hpCL_{M+C}$$

By Lemma C.2 $\hat{\boldsymbol{\beta}}_h$ is in a compact set. Hence, any of its subsequence contains a convergent sub-subsequence. Let us abuse the notation and denote by $\hat{\boldsymbol{\beta}}_h$ any of such convergent sub-subsquence, that is, assume that $\hat{\boldsymbol{\beta}}_h \to \hat{\boldsymbol{\beta}}_0$. Along such a sub-subsequence, we have that:

$$P(\hat{\boldsymbol{\beta}}_0) = \lim_{h \to 0} P(\hat{\boldsymbol{\beta}}_h) = \lim_{h \to 0} P_h(\hat{\boldsymbol{\beta}}_h) \leq \lim_{h \to 0} P_h(\hat{\boldsymbol{\beta}}) = \lim_{h \to 0} P(\hat{\boldsymbol{\beta}}).$$

The uniqueness of the minimizer implies $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}$. As the above holds along any convergent sub-subsequence, we have that:

$$\|\hat{\boldsymbol{\beta}}_h - \hat{\boldsymbol{\beta}}\|_2 \to 0 \text{ as } h \to 0.$$

□

Lemma C.4 (Convergence of the subgradients)**.** *Suppose that Assumption C.1 holds. Recall that we use* $R(\boldsymbol{\beta}) = \sum_{l=1}^{p} r(\beta_l)$*. We have that:*

$$\|\nabla R_h(\hat{\boldsymbol{\beta}}_h) - \boldsymbol{g}_R(\hat{\boldsymbol{\beta}})\|_2 \to 0, \qquad \text{as } h \to 0.$$

where $g_R(\hat{\boldsymbol{\beta}})$ *is the subgradient of* R *at* $\hat{\boldsymbol{\beta}}$*.*

Proof. By the first-order optimality conditions and the continuity of ℓ, we have that as $h \to 0$:

$$\|\nabla R_h(\hat{\boldsymbol{\beta}}_h) - \boldsymbol{g}_R(\hat{\boldsymbol{\beta}})\|_2 = \Big\| \sum_j \ell(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j) - \sum_j \ell(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_h; y_j) \Big\|_2 \to 0.$$

□

Lemma C.5 (Convergence of the Hessian)**.** *Suppose that Assumption C.1 holds. We have that as* $h \to 0$*:*

$$\ddot{r}_h(\hat{\beta}_{h,i}) \to \begin{cases} \ddot{r}(\hat{\beta}_i) & \text{if } \hat{\beta}_i \notin K, \\ +\infty & \text{if } \hat{\beta}_i \in K. \end{cases}$$

Proof. Let us first consider the case $\hat{\beta}_i \notin K$. As $\mathbb{R} \setminus K$ is open, there exists $\delta > 0$ such that $[\hat{\beta}_i - \delta, \hat{\beta}_i + \delta] \subset \mathbb{R} \backslash K$. Since $\hat{\beta}_{h,i} \to \hat{\beta}_i$ as $h \to 0$, we have for h small enough that:

$$[\hat{\beta}_{h,i} - hC, \hat{\beta}_{h,i} + hC] \subset [\hat{\beta}_i - \delta, \hat{\beta}_i + \delta] \subset \mathbb{R} \backslash K.$$

Since $\ddot{r}$ is smooth on $[\hat{\beta}_i - \delta, \hat{\beta}_i + \delta]$, by the bounded convergence theorem, we have as $h \to 0$:

$$\ddot{r}_h(\hat{\beta}_{h,i}) = \int_{-C}^{C} \ddot{r}(\hat{\beta}_{h,i} - hw)\phi(w)dw \to \int_{-C}^{C} \ddot{r}(\hat{\beta}_i)\phi(w)dw = \ddot{r}(\hat{\beta}_i)$$

Now, let us consider the case where $\hat{\beta}_i \in K$. By Lemma C.4, we have that $\dot{r}_h(\hat{\beta}_{h,i}) \to g_r(\hat{\beta}_i)$, from which we deduce:

$$|\hat{\beta}_{h,i} - \hat{\beta}_i| < hC.$$

Indeed, if we had $\hat{\beta}_i \geq \hat{\beta}_{h,i} + hC$, notice the assumption on the subgradient $g_r(\hat{\beta}_i)$, this would imply:

$$\dot{r}_h(\hat{\beta}_{h,i}) = \int_{-C}^{C} \dot{r}(\hat{\beta}_{h,i} - hw)\phi(w)dw \leq \dot{r}_-(\hat{\beta}_i) < g_r(\hat{\beta}_i),$$

which is contradictory. The same happens if $\hat{\beta}_i \leq \hat{\beta}_{h,i} - hC$. To conclude, note that as $h \to 0$:

$$\begin{aligned}
\ddot{r}_h(\hat{\beta}_{h,i}) &= \int_{\hat{\beta}_{h,i} - hC}^{\hat{\beta}_i} r(u) \frac{1}{h^3} \ddot{\phi}\Big(\frac{\hat{\beta}_{h,i} - u}{h}\Big) du + \int_{\hat{\beta}_i}^{\hat{\beta}_{h,i} + hC} r(u) \frac{1}{h^3} \ddot{\phi}\Big(\frac{\hat{\beta}_{h,i} - u}{h}\Big) du \\
&= \frac{1}{h} \phi\Big(\frac{\hat{\beta}_{h,i} - \hat{\beta}_i}{h}\Big)(\dot{r}_+(\hat{\beta}_i) - \dot{r}_-(\hat{\beta}_i)) + \int_{\frac{\hat{\beta}_{h,i} - \hat{\beta}_i}{h}}^{C} \ddot{r}(\hat{\beta}_{h,i} - hw)\phi(w)dw \\
&\quad + \int_{-C}^{\frac{\hat{\beta}_{h,i} - \hat{\beta}_i}{h}} \ddot{r}(\hat{\beta}_{h,i} - hw)\phi(w)dw \\
&\to +\infty.
\end{aligned}$$

□

Lemma C.6. *Consider a sequence of matrices* $\boldsymbol{A}_n, n \in \mathbb{N}$*, and let* $\boldsymbol{A}_n = \begin{bmatrix} \boldsymbol{A}_{1n} & \boldsymbol{A}_{2n} \\ \boldsymbol{A}_{3n} & \boldsymbol{A}_{4n} \end{bmatrix}$ *where* $\boldsymbol{A}_{1n}, \boldsymbol{A}_{4n}$ *are invertible for all* n*. Additionally, suppose that* $\boldsymbol{A}_{in} \to \boldsymbol{A}_i, i = 1, 2, 3$*, and* $\boldsymbol{A}_{4n}^{-1} \to \boldsymbol{0}$ *as* $n \to \infty$*. Then we have as* $n \to \infty$ *that:*

$$\boldsymbol{A}_n^{-1} \to \begin{bmatrix} \boldsymbol{A}_1^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

Proof. By the Woodbury matrix identity [Woodbury, 1950], we have

$$\begin{aligned}\boldsymbol{A}_n^{-1} =& \begin{bmatrix} (\boldsymbol{A}_{1n} - \boldsymbol{A}_{2n}\boldsymbol{A}_{4n}^{-1}\boldsymbol{A}_{3n})^{-1} & -(\boldsymbol{A}_{1n} - \boldsymbol{A}_{2n}\boldsymbol{A}_{4n}^{-1}\boldsymbol{A}_{3n})^{-1}\boldsymbol{A}_{2n}\boldsymbol{A}_{4n}^{-1} \\ -\boldsymbol{A}_{4n}^{-1}\boldsymbol{A}_{3n}(\boldsymbol{A}_{1n} - \boldsymbol{A}_{2n}\boldsymbol{A}_{4n}^{-1}\boldsymbol{A}_{3n})^{-1} & \boldsymbol{A}_{4n}^{-1}\boldsymbol{A}_{3n}(\boldsymbol{A}_{1n} - \boldsymbol{A}_{2n}\boldsymbol{A}_{4n}^{-1}\boldsymbol{A}_{3n})^{-1}\boldsymbol{A}_{2n}\boldsymbol{A}_{4n}^{-1} + \boldsymbol{A}_{4n}^{-1} \end{bmatrix} \\ \to& \begin{bmatrix} \boldsymbol{A}_1^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\end{aligned}$$

□

Proof of Theorem 4.2. The proof of Theorem 4.2 is a straightforward corollary of the Lemmas C.3, C.4, C.5 and C.6. □

C.3 Proof of Theorem 4.1: Nonsmooth Loss With Smooth Regularizer

We now consider the case of non-smooth loss. The proof is very similar to the previous section, so we briefly mention the common parts and focus on the differences.

Consider nonsmooth loss ℓ and its smoothed version ℓ_h. R is assumed to be smooth. Let us consider:

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{n} \ell(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_j) + R(\boldsymbol{\beta}),$$
$$P_h(\boldsymbol{\beta}) = \sum_{j=1}^{n} \ell_h(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_j) + R(\boldsymbol{\beta}).$$

Let us still use $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} P(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}_h = \arg\min_{\boldsymbol{\beta}} P_h(\boldsymbol{\beta})$ to denote the optimizers. As before, let $K = \{v_1, \ldots, v_k\}$ denote the zero-order singularities of ℓ, and let $V = \{i : \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} \in K\}$ be the set of indices of observations at such singularities.

Assumption C.2. *We need the following assumptions on ℓ, R and $\hat{\boldsymbol{\beta}}$:*

1. *ℓ is locally Lipschitz, that is, for any $A > 0$, for any $x, y \in [-A, A]$, we have $|\ell(x) - \ell(y)| \leq L_A|x - y|$, where L_A is a constant depends only on A.*
2. *$\lambda_{\min}(\boldsymbol{X}_V \boldsymbol{X}_V^\top) > 0$.*
3. *$\hat{\boldsymbol{\beta}}$ is the unique minimizer.*
4. *Whenever $\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} = v \in K$, the subgradient of ℓ at $\boldsymbol{x}_j^\top \hat{\beta}$, $g_\ell(\boldsymbol{x}^\top \hat{\boldsymbol{\beta}})$ satisfies $g_\ell(\boldsymbol{x}^\top \hat{\boldsymbol{\beta}}) \in (\ell_-(v), \ell_+(v))$.*
5. *R is coercive in the sense that $|R(\boldsymbol{\beta})| \to \infty$ as $\|\boldsymbol{\beta}\| \to \infty$.*

Lemma C.7. *Suppose that Assumption C.2 holds. There exists $M > 0$ that only depends on r, ℓ and λ, such that for all $h \leq 1$, we have:*

$$\|\hat{\boldsymbol{\beta}}\|_\infty \leq M \text{ and } \|\hat{\boldsymbol{\beta}}_h\|_\infty \leq M.$$

Proof. Let $h \leq 1$, then $\hat{\boldsymbol{\beta}}_h$ verifies:

$$\begin{aligned} R(\hat{\boldsymbol{\beta}}_h) \leq& \sum_j \ell_h(0; y_j) + pR(0) \\ =& \sum_j \int_{-C}^{C} \ell(hw; y_j)\phi(w)dw + pR(0) \leq \sum_j \sup_{|w| \leq C} \ell(w; y_i) + pR(0). \end{aligned}$$

Additionally, $\hat{\boldsymbol{\beta}}$ verifies:

$$R(\hat{\boldsymbol{\beta}}) \leq \sum_j \ell(0; y_j) + pR(0) \leq \sum_j \sup_{|w| \leq C} \ell(w; y_i) + pR(0).$$

The convexity and coerciveness of R implies that there exists a M, such that for all $h \leq 1$:

$$\|\hat{\boldsymbol{\beta}}_h\|_2 \leq M \text{ and } \|\hat{\boldsymbol{\beta}}\|_2 \leq M.$$

□

Lemma C.8. *Suppose that Assumption C.2 holds. We have that as $h \to 0$:*

$$\|\hat{\boldsymbol{\beta}}_h - \hat{\boldsymbol{\beta}}\|_2 \to 0.$$

Proof. Let $M_x = \max_i \|\boldsymbol{x}_i\|_2$. By the local Lipschitz condition of ℓ, we have that for any $\|\boldsymbol{\beta}\|_2 \leq M$ and $h \leq 1$ that:

$$\begin{aligned}
0 &\leq \ell_h(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta}) - \ell(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta}) \\
&= \int_{-C}^{C} [\ell(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta} - hw) - \ell(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta})] \phi(w) dw \\
&\leq 2CL_{M_x M + C} h.
\end{aligned}$$

This implies:

$$\sup_{\|\boldsymbol{\beta}\|_2 \leq M} |P(\boldsymbol{\beta}) - P_h(\boldsymbol{\beta})| \leq 2nhCL_{M_x M + C}$$

From Lemma C.7, we know $\hat{\boldsymbol{\beta}}_h$ is in a compact set, thus any of its subsequence contains a convergent sub-subsequence. Again abuse the notation and let $\hat{\boldsymbol{\beta}}_h$ denote this convergent sub-subsequence. Suppose that: $\hat{\boldsymbol{\beta}}_h \to \hat{\boldsymbol{\beta}}_0$. Now we have again:

$$P(\hat{\boldsymbol{\beta}}_0) = \lim_{h \to 0} P(\hat{\boldsymbol{\beta}}_h) = \lim_{h \to 0} P_h(\hat{\boldsymbol{\beta}}_h) \leq \lim_{h \to 0} P_h(\hat{\boldsymbol{\beta}}) = \lim_{h \to 0} P(\hat{\boldsymbol{\beta}}).$$

The uniqueness implies $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}$. As the previous result holds along any sub-subsequence, we deduce that:

$$\|\hat{\boldsymbol{\beta}}_h - \hat{\boldsymbol{\beta}}\|_2 \to 0.$$

□

Lemma C.9 (Convergence of gradients)**.** *Suppose that Assumption C.2 holds. Then, we have that for any j, as $h \to 0$:*

$$\|\dot{\ell}_h(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_h) - g_\ell(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}})\|_2 \to 0.$$

Proof. for $j \notin V$, the result is immediate. For $j \in V$, we have that as $h \to 0$:

$$\Big\| \sum_{j \in V} \boldsymbol{x}_j \dot{\ell}_h(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_h; y_j) - \sum_{j \in V} \boldsymbol{x}_j g_\ell(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j) \Big\|_2 \to 0$$

This implies the desired result by the assumption on $\boldsymbol{X}_{V,\cdot}$. □

Lemma C.10 (Convergence of Hessian)**.** *Suppose that Assumption C.2 holds. Then, we have that for any j, as $h \to 0$:*

$$\ddot{\ell}_h(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_h; y_j) \to \begin{cases} \ddot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j) & \text{if } j \notin V \\ +\infty & \text{if } j \in V \end{cases}$$

Proof. Again, the result follows through a similar argument as in the proof of Lemma C.5 for $j \notin V$. For $j \in V$, we have by Lemma C.9 that as $h \to 0$:

$$\dot{\ell}_h(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_h; y_j) \to g_\ell(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j).$$

Following a similar reasoning as in the proof of Lemma C.5, we have that:

$$|\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_h - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}| < hC.$$

Finally, we note that as $h \to 0$:

$$\ddot{\ell}_h(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_h; y_j) \geq \frac{1}{h}\phi\Big(\frac{\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_h - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}}{h}\Big)(\dot{\ell}_+(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}) - \dot{\ell}_-(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}})) \to +\infty$$

□

Proof of Theorem 4.1. Recall $V = \{i : \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} \in K\}$ and $S = [1:n]\backslash V$. Let $\boldsymbol{H}_h$ be the matrix in ALO for smooth loss and smooth regularizer when using ℓ_h. Let $\boldsymbol{L}_h = \mathrm{diag}[\{\ddot{\ell}_h(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j)\}_j]$, $\boldsymbol{L}_S = \mathrm{diag}[\{\ddot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j)\}_{j\in S}]$. $\boldsymbol{L}_{h,S}$ and $\boldsymbol{L}_{h,V}$ are similarly defined. Recall

$$\boldsymbol{H}_h = \boldsymbol{X}(\lambda\nabla^2 R + \boldsymbol{X}^\top \boldsymbol{L}_h \boldsymbol{X})^{-1}\boldsymbol{X}^\top$$

We then have

$$\begin{aligned}
&(\lambda\nabla^2 R + \boldsymbol{X}^\top \boldsymbol{L}_h \boldsymbol{X})^{-1}\\
=&(\underbrace{\lambda\nabla^2 R + \boldsymbol{X}_{S,\cdot}^\top \boldsymbol{L}_{h,S}\boldsymbol{X}_{S,\cdot}}_{\boldsymbol{Y}_h} + \boldsymbol{X}_{V,\cdot}^\top \boldsymbol{L}_{h,V}\boldsymbol{X}_{V,\cdot})^{-1}\\
=&\boldsymbol{Y}_h^{-1} - \boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top(\boldsymbol{L}_{h,V}^{-1} + \boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}
\end{aligned}$$

As a result, we have

$$\begin{aligned}
&(\lambda\nabla^2 R + \boldsymbol{X}^\top \boldsymbol{L}_h X)^{-1}\boldsymbol{X}_{V,\cdot}^\top\\
=&\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top - \boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top(\boldsymbol{L}_{h,V}^{-1} + \boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top\\
=&\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top(\boldsymbol{I}_p - (\boldsymbol{L}_{h,V}^{-1} + \boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top)\\
=&\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top(\boldsymbol{L}_{h,V}^{-1} + \boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{L}_{h,V}^{-1}
\end{aligned}$$

Similarly we can get

$$\begin{aligned}
\boldsymbol{X}_{V,\cdot}(\lambda\nabla^2 R + \boldsymbol{X}^\top \boldsymbol{L}_h \boldsymbol{X})^{-1} =&\boldsymbol{L}_{h,V}^{-1}(\boldsymbol{L}_{h,V}^{-1} + \boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}\\
\boldsymbol{X}_{V,\cdot}(\lambda\nabla^2 R + \boldsymbol{X}^\top \boldsymbol{L}_h \boldsymbol{X})^{-1}\boldsymbol{X}_{V,\cdot}^\top =&\boldsymbol{L}_{h,V}^{-1} - \boldsymbol{L}_{h,V}^{-1}(\boldsymbol{L}_{h,V}^{-1} + \boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{L}_{h,V}^{-1}
\end{aligned}$$

By Lemma C.10, $\boldsymbol{Y}_h \to \boldsymbol{Y} := \lambda\nabla^2 R + \boldsymbol{X}_{S,\cdot}^\top \boldsymbol{L}_S \boldsymbol{X}_{S,\cdot}$, $\boldsymbol{L}_{h,V}^{-1} \to \boldsymbol{0}$, we have

$$\begin{aligned}
\boldsymbol{H}_{h,S,S}\boldsymbol{L}_{h,S} \to&\boldsymbol{X}_{S,\cdot}(\boldsymbol{Y}^{-1} - \boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top(\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}^{-1})\boldsymbol{X}_{S,\cdot}^\top\boldsymbol{L}_S\\
\boldsymbol{H}_{h,S,V}\boldsymbol{L}_{h,V} \to&\boldsymbol{X}_{S,\cdot}\boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top(\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\\
\boldsymbol{H}_{h,V,S}\boldsymbol{L}_{h,S} \to&\boldsymbol{0}\\
\boldsymbol{H}_{h,V,V}\boldsymbol{L}_{h,V} \to&\boldsymbol{I}_V
\end{aligned}$$

This is not enough, however, noticing that in the final formula of the smooth case, we need $\frac{H_{h,ii}}{1-L_{h,ii}H_{h,ii}}$ but for $i \in V$, $1 - L_{h,ii}H_{h,ii} \to 0$ and $H_{h,ii} \to 0$. So further we have

$$\begin{aligned}
&\boldsymbol{L}_{h,V}(\boldsymbol{I}_V - \boldsymbol{H}_{h,VV}\boldsymbol{L}_{h,V})\\
=&\boldsymbol{L}_{h,V}(\boldsymbol{I}_V - (\boldsymbol{L}_{h,V}^{-1} - \boldsymbol{L}_{h,V}^{-1}(\boldsymbol{L}_{h,V}^{-1} + \boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{L}_{h,V}^{-1})\boldsymbol{L}_{h,V})\\
=&(\boldsymbol{L}_{h,V}^{-1} + \boldsymbol{X}_{V,\cdot}\boldsymbol{Y}_h^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\\
\to&(\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}
\end{aligned}$$

As a result, we have

$$\frac{H_{h,ii}}{1 - L_{h,ii}H_{h,ii}} \to \begin{cases} \frac{\boldsymbol{x}_i^\top(\boldsymbol{Y}^{-1} - \boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top(\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}^{-1})\boldsymbol{x}_i}{1 - \boldsymbol{x}_i(\boldsymbol{Y}^{-1} - \boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top(\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}^{-1})\boldsymbol{x}_i\ddot{\ell}_i} & i \in S \\ \frac{1}{[(\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}]_{ii}} & i \in V \end{cases}$$

For $\dot{\ell}_h(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_h; y_i)$, as $h \to 0$, Lemma C.9 implies the limit value the smooth gradients would converge to. Notice that for $j \in V$, we solve for the subgradient by applying least square formula to the 1st order optimality equation. The final results easily follow. □

D Derivation of the Dual for Generalized LASSO

In this section we derive the dual form of the generalized LASSO stated in the main paper. We recall that for a given matrix $\boldsymbol{D} \in \mathbb{R}^{m\times p}$, the generalized LASSO is given by:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{j=1}^{n}(y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda\|\boldsymbol{D}\boldsymbol{\beta}\|_1.$$

Introduce dummy variables $\boldsymbol{z} \in \mathbb{R}^n$, $\boldsymbol{w} \in \mathbb{R}^m$, and consider the following equivalent constrained optimization problem:

$$\min_{\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{z}\|_2^2 + \lambda\|\boldsymbol{w}\|_1$$
$$\text{subject to: } \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{z} \text{ and } \boldsymbol{D}\boldsymbol{\beta} = \boldsymbol{w}.$$

We may now consider the Lagrangian form of the optimization problem, introducing dual variables $\boldsymbol{\theta} \in \mathbb{R}^n$ and $\boldsymbol{u} \in \mathbb{R}^m$, the dual problem is

$$\begin{aligned}
&\max_{\boldsymbol{\theta},\boldsymbol{u}} \min_{\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{z}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 + \boldsymbol{\theta}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{z}) + \boldsymbol{u}^\top(\boldsymbol{D}\boldsymbol{\beta} - \boldsymbol{w}) \\
&= -\min_{\boldsymbol{\theta},\boldsymbol{u}} \big[\max_{\boldsymbol{z}}\{\boldsymbol{\theta}^\top\boldsymbol{z} - \frac{1}{2}\|\boldsymbol{z}\|_2^2\} + \max_{\boldsymbol{w}}\{\boldsymbol{u}^\top\boldsymbol{w} - \lambda\|\boldsymbol{w}\|_1\} + \max_{\boldsymbol{\beta}}\{\boldsymbol{\theta}^\top\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{\beta}\} - \boldsymbol{\theta}^\top\boldsymbol{y}\big]
\end{aligned}$$

Consider the three subproblems within square brackets respectively, we have

$$\max_{\boldsymbol{z}}\{\boldsymbol{\theta}^\top\boldsymbol{z} - \frac{1}{2}\|\boldsymbol{z}\|_2^2\} = \frac{1}{2}\|\boldsymbol{\theta}\|_2^2,$$
$$\max_{\boldsymbol{w}}\{\boldsymbol{u}^\top\boldsymbol{w} - \lambda\|\boldsymbol{w}\|_1\} = \begin{cases} 0 & \text{if } \|\boldsymbol{u}\|_\infty \le \lambda, \\ \infty & \text{otherwise.} \end{cases}$$

where $\boldsymbol{\theta}^\top\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{\beta}$ is unbounded unless $\boldsymbol{X}^\top\boldsymbol{\theta} = \boldsymbol{D}^\top\boldsymbol{u}$. Finally, we substitute the above results into our Lagrangian dual problem to obtain:

$$\min_{\boldsymbol{\theta},\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 - \boldsymbol{\theta}^\top\boldsymbol{y},$$
$$\text{subject to: } \boldsymbol{D}^\top\boldsymbol{u} = \boldsymbol{X}^\top\boldsymbol{\theta} \text{ and } \|\boldsymbol{u}\|_\infty \le \lambda.$$

which is equivalent to the stated dual problem.

E Proof of Nuclear Norm ALO Formula

In this section, we prove Theorem 6.1. We consider the following matrix sensing formulation

$$\hat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \sum_{j=1}^{n} \ell(\langle \boldsymbol{X}_j, \boldsymbol{B}\rangle; y_j)^2 + \lambda R(\boldsymbol{B}).$$

where R is a unitarily invariant function, which will be explained and studied in more detail in Section E.1. This section is laid out as follows: in Section E.1, we briefly discuss basic properties of unitarily invariant functions; In Section E.2 we do ALO for smooth unitarily invariant penalties; In Section E.3 we prove Theorem 6.1 where nuclear norm is considered.

E.1 Properties of Unitarily Invariant Functions

Let $\boldsymbol{B} \in \mathbb{R}^{p_1\times p_2}$, and consider the SVD of $\boldsymbol{B}$ as $\boldsymbol{B} = \boldsymbol{U}\mathrm{diag}[\boldsymbol{\sigma}]\boldsymbol{V}^\top$ with $\boldsymbol{U} \in \mathbb{R}^{p_1\times p_1}$, $\boldsymbol{V} \in \mathbb{R}^{p_2\times p_2}$. We say that a function $R : \mathbb{R}^{p_1\times p_2} \to \mathbb{R}$ is unitarily invariant if there exists an absolutely symmetric function $f : \mathbb{R}^{\min(p_1,p_2)} \to \mathbb{R}$ such that:

$$R(\boldsymbol{B}) = f(\boldsymbol{\sigma}),$$

where we say that $f : \mathbb{R}^q \to \mathbb{R}$ is absolutely symmetric if for any $\boldsymbol{x} \in \mathbb{R}^q$, any permutation τ and signs $\boldsymbol{\epsilon} \in \{-1, 1\}^q$ we have:

$$f(x_1, \ldots, x_q) = f(\epsilon_1 x_{\tau(1)}, \ldots, \epsilon_q x_{\tau(q)}).$$

The properties of R and f are closely related, and in particular we will make use of the following lemma relating their convexity, smoothness and derivatives, proved in [Lewis, 1995].

Lemma E.1 ([Lewis, 1995])**.** *Let $R(\boldsymbol{B}) = f(\boldsymbol{\sigma})$ with $\boldsymbol{B} = \boldsymbol{U}\mathrm{diag}[\boldsymbol{\sigma}]\boldsymbol{V}^\top$ its SVD. There is an one-to-one correspondence between unitarily invariant matrix functions R and symmetric functions f. Furthermore the convexity and/or differentiability of f are equivalent to the convexity and/or differentiability of R respectively. If R is differentiable, its derivative is given by:*

$$\nabla R(\boldsymbol{B}) = \boldsymbol{U}\mathrm{diag}[\nabla f(\boldsymbol{\sigma})]\boldsymbol{V}^\top \tag{48}$$

When f is not differentiable, a similar result holds with gradient replaced by subdifferentials.

$$\partial R(\boldsymbol{B}) = \boldsymbol{U}\mathrm{diag}[\partial f(\boldsymbol{\sigma})]\boldsymbol{V}^\top \tag{49}$$

Based on this lemma, we know that as long as f is convex and/or smooth, the corresponding matrix function will be convex and/or smooth. This enables us to produce convex and smooth unitarily invariant approximation to non-smooth unitarily invariant matrix regularizers.

In addition to the gradient of the unitarily invariant matrix functions, we also need their Hessians. We show this result in the following Theorem E.1 for a sub-class of unitarily invariant functions.

Theorem E.1. *Consider a unitarily invariant function with form $R(\boldsymbol{B}) = \sum_{j=1}^{\min(p_1,p_2)} f(\sigma_j)$, where f is a smooth function on $\mathbb{R}$ and $\boldsymbol{B} = \boldsymbol{U}\mathrm{diag}[\boldsymbol{\sigma}]\boldsymbol{V}^\top$ is its SVD with $\boldsymbol{U} \in \mathbb{R}^{p_1 \times p_1}$, $\boldsymbol{V} \in \mathbb{R}^{p_2 \times p_2}$. Further assume that all the σ_j's are different from each other and nonzero. Let $p_3 = \min(p_1, p_2)$, $p_4 = \max(p_1, p_2)$. Then the Hessian matrix $\nabla^2 R(\boldsymbol{B}) \in \mathbb{R}^{p_1 p_2 \times p_1 p_2}$ takes the following form*

$$\nabla^2 R(\boldsymbol{B}) = \boldsymbol{Q}\begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{bmatrix}\boldsymbol{Q}^\top \tag{50}$$

where the first block $A_1 \in \mathbb{R}^{p_3 \times p_3}$. A_1 is diagonal with $A_{1,(ss,ss)} = f''(\sigma_s)$, $1 \le s \le p_3$. The second block $A_2 \in \mathbb{R}^{p_3(p_3-1)\times p_3(p_3-1)}$. For $1 \le s \ne t \le p_3$, $A_{2,(st,st)} = A_{2,(ts,ts)} = \frac{\sigma_s f'(\sigma_s) - \sigma_t f'(\sigma_t)}{\sigma_s^2 - \sigma_t^2}$, $A_{2,(st,ts)} = A_{2,(ts,st)} = -\frac{\sigma_s f'(\sigma_t) - \sigma_t f'(\sigma_s)}{\sigma_s^2 - \sigma_t^2}$; The third block $A_3 \in \mathbb{R}^{(p_4-p_3)p_3 \times (p_4-p_3)p_3}$; $A_{3,(st,st)} = \frac{f'(\sigma_t)}{\sigma_t}$ for $1 \le t \le p_3 < s \le p_4$. Except for these specified locations, all other components of A_1, A_2, A_3 are zero. $\boldsymbol{Q}$ is an orthogonal matrix with $\boldsymbol{Q}_{\cdot,st} = \mathrm{vec}(\boldsymbol{u}_s \boldsymbol{v}_t^\top)$ where $\boldsymbol{u}_s$, $\boldsymbol{v}_t$ are the s^{th} column of $\boldsymbol{U}$ and t^{th} column of $\boldsymbol{V}$ respectively. $\mathrm{vec}(\cdot)$ denotes the vectorization operator, which aligns all the components of a matrix into a long vector.

Remark E.1. *Since here we are talking about the Hessian matrix of functions on matrix space, we linearize these matrices and treat them as vectors. It would be helpful if we visualize the correspondence between each blocks in* (50) *and the component indices in the original matrix $\boldsymbol{B}$. Specifically we have Figure 3.*

Proof. First by Lemma E.1, the gradient $\nabla R(\boldsymbol{B})$ takes the following form

$$\nabla R(\boldsymbol{B}) = \boldsymbol{U}\mathrm{diag}[\{f'(\sigma_j)\}_j]\boldsymbol{V}^\top$$

In order to find the differential of $\nabla R(\boldsymbol{B})$, we use the similar techniques and notations described in Lemma IV.2 and Theorem IV.3 in [Candes et al., 2013]. To simplify our derivation, we assume $p_1 \ge p_2$. This does not affect the correctness of our final conclusion.

We characterize the differential of the gradient as a linear form. Specifically, along a certain direction $\boldsymbol{\Delta} \in \mathbb{R}^{p_1 \times p_2}$, by Lemma IV.2 in [Candes et al., 2013], we have

$$d\boldsymbol{U}[\boldsymbol{\Delta}] = \boldsymbol{U}\boldsymbol{\Omega}_{\boldsymbol{U}}[\boldsymbol{\Delta}], \quad d\boldsymbol{V}[\boldsymbol{\Delta}] = \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{V}}[\boldsymbol{\Delta}]^\top, \quad d\sigma_s[\boldsymbol{\Delta}] = [\boldsymbol{U}^\top \boldsymbol{\Delta} \boldsymbol{V}]_{ss} \tag{51}$$

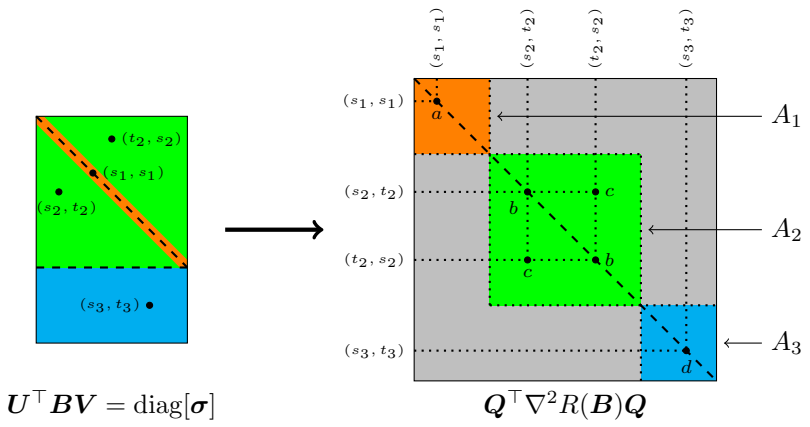

Figure 3: An illustration of the correspondence between the structure of the original matrix and the structure of the Hessian matrix of R. As we have mentioned in Theorem E.1, $a = f''(\sigma_{s_1})$, $b = \frac{\sigma_{s_2} f'(\sigma_{s_2}) - \sigma_{t_2} f'(\sigma_{t_2})}{\sigma_{s_2}^2 - \sigma_{t_2}^2}$, $c = -\frac{\sigma_{s_2} f'(\sigma_{t_2}) - \sigma_{t_2} f'(\sigma_{s_2})}{\sigma_{s_2}^2 - \sigma_{t_2}^2}$; $d = \frac{f'(\sigma_{t_3})}{\sigma_{t_3}}$.

where $\boldsymbol{\Omega}_{\boldsymbol{U}}$ and $\boldsymbol{\Omega}_{\boldsymbol{V}}$ are assymmetric matrices (thus their diagonal values are 0) which can be found by solving the following equation systems:

$$\begin{bmatrix} \boldsymbol{\Omega}_{\boldsymbol{U},st}[\Delta] \\ \boldsymbol{\Omega}_{\boldsymbol{V},st}[\Delta] \end{bmatrix} = -\frac{1}{\sigma_s^2 - \sigma_t^2} \begin{bmatrix} \sigma_t & \sigma_s \\ -\sigma_s & -\sigma_t \end{bmatrix} \begin{bmatrix} (\boldsymbol{U}^\top \boldsymbol{\Delta} \boldsymbol{V})_{st} \\ (\boldsymbol{U}^\top \boldsymbol{\Delta} \boldsymbol{V})_{ts} \end{bmatrix}, \quad \text{if } s \neq t, s \leq p_2 \tag{52}$$

and

$$\boldsymbol{\Omega}_{\boldsymbol{U},st}[\Delta] = \frac{(\boldsymbol{U}^\top \boldsymbol{\Delta} \boldsymbol{V})_{st}}{\sigma_t}, \quad \text{if } s \neq t, s > p_2 \tag{53}$$

The differential of $\nabla R(\boldsymbol{B})$ along a certian direction $\boldsymbol{\Delta}$ can then be calculated through the chain rule as that

$$\begin{aligned} & d\nabla R(\boldsymbol{B})[\boldsymbol{\Delta}] \\ =& d\boldsymbol{U}[\boldsymbol{\Delta}]\mathrm{diag}[\{f'(\sigma_j)\}_j]\boldsymbol{V}^\top + \boldsymbol{U}\mathrm{diag}[\{f''(\sigma_j)d\sigma_j[\boldsymbol{\Delta}]\}_j]\boldsymbol{V}^\top + \boldsymbol{U}\mathrm{diag}[\{f'(\sigma_j)\}_j]d\boldsymbol{V}[\boldsymbol{\Delta}]^\top \\ =& \boldsymbol{U}\big(\boldsymbol{\Omega}_{\boldsymbol{U}}[\boldsymbol{\Delta}]\mathrm{diag}[\{f'(\sigma_j)\}_j] + \mathrm{diag}[\{f''(\sigma_j)d\sigma_j[\boldsymbol{\Delta}]\}_j] + \mathrm{diag}[\{f'(\sigma_j)\}_j]\boldsymbol{\Omega}_{\boldsymbol{V}}[\boldsymbol{\Delta}]\big)\boldsymbol{V}^\top \end{aligned} \tag{54}$$

In the original formula obtained from the primal approach, the Hessian is calculated under the canonical bases [2] $\{\boldsymbol{E}_{st}\}_{s,t}$. In order to simplify the calculation of the Hessian, we instead use the orthonormal bases $\{\boldsymbol{u}_s \boldsymbol{v}_t^\top\}_{s,t}$, and then transform back to $\{\boldsymbol{E}_{st}\}_{s,t}$.

The (kl, st) location of the Hessian matrix under $\{\boldsymbol{u}_s \boldsymbol{v}_t\}_{s,t}$ bases can be calculated by

$$\langle \boldsymbol{u}_k \boldsymbol{v}_l^\top, d\nabla R(\boldsymbol{B})[\boldsymbol{u}_s \boldsymbol{v}_t^\top] \rangle \tag{55}$$

Pluggin equation (54) into (55) we obtain that

$$\begin{aligned} & \langle \boldsymbol{u}_k \boldsymbol{v}_l, d\nabla R(\boldsymbol{B})[\boldsymbol{u}_s \boldsymbol{v}_t^\top] \rangle \\ =& \langle \boldsymbol{E}_{kl}, \boldsymbol{\Omega}_{\boldsymbol{U}}[\boldsymbol{u}_s \boldsymbol{v}_t^\top]\mathrm{diag}[\{f'(\sigma_j)\}_j] + \mathrm{diag}[\{f''(\sigma_j)d\sigma_j[\boldsymbol{u}_s \boldsymbol{v}_t^\top]\}_j] + \mathrm{diag}[\{f'(\sigma_j)\}_j]\boldsymbol{\Omega}_{\boldsymbol{V}}[\boldsymbol{u}_s \boldsymbol{v}_t^\top] \rangle \\ =& \begin{cases} f''(\sigma_t)d\sigma_t[\boldsymbol{u}_t \boldsymbol{v}_t^\top] & s = t = k = l \\ \boldsymbol{\Omega}_{\boldsymbol{U},kl}[\boldsymbol{u}_s \boldsymbol{v}_t^\top]f'(\sigma_l) + f'(\sigma_k)\boldsymbol{\Omega}_{\boldsymbol{V},kl}[\boldsymbol{u}_s \boldsymbol{v}_t^\top] & k \neq l, k \leq p_2 \\ \boldsymbol{\Omega}_{\boldsymbol{U},kl}[\boldsymbol{u}_s \boldsymbol{v}_t^\top]f'(\sigma_l) & 1 \leq l \leq p_2 < k \leq p_1 \end{cases} \end{aligned}$$

[2] $\boldsymbol{E}_{st}$ is defined as a $p_1 \times p_2$ matrix with all of its components being 0 except the (s,t) location being 1.

By (51), we have $d\sigma_j[\boldsymbol{u}_s\boldsymbol{v}_t^\top] = [\boldsymbol{E}_{st}]_{jj} = \delta_{sj}\delta_{tj}$. In addition, $(\boldsymbol{U}^\top\boldsymbol{u}_s\boldsymbol{v}_t^\top\boldsymbol{V}^\top)_{kl} = (\boldsymbol{E}_{st})_{kl} = \delta_{sk}\delta_{tl}$, $(\boldsymbol{U}^\top\boldsymbol{u}_s\boldsymbol{v}_t^\top\boldsymbol{V}^\top)_{lk} = (\boldsymbol{E}_{st})_{lk} = \delta_{sl}\delta_{tk}$. Hence by (52) and (53), we have that

$$\boldsymbol{\Omega}_{\boldsymbol{U},kl}[\boldsymbol{u}_s\boldsymbol{v}_t^\top] = -\frac{\delta_{sk}\delta_{tl}\sigma_l + \delta_{sl}\delta_{tk}\sigma_k}{\sigma_k^2 - \sigma_l^2}, \quad \boldsymbol{\Omega}_{\boldsymbol{V},kl}[\boldsymbol{u}_s\boldsymbol{v}_t^\top] = \frac{\delta_{sk}\delta_{tl}\sigma_k + \delta_{sl}\delta_{tk}\sigma_l}{\sigma_k^2 - \sigma_l^2}, \quad \text{if } s \neq t, s \leq p_2$$

and

$$\boldsymbol{\Omega}_{\boldsymbol{U},kl}[\boldsymbol{u}_s\boldsymbol{v}_t^\top] = \frac{\delta_{sk}\delta_{tl}}{\sigma_l}, \quad \text{if } s \neq t, s > p_2$$

Based on all these, we can obtain that

$$\langle \boldsymbol{u}_k\boldsymbol{v}_l, d\nabla R(\boldsymbol{B})[\boldsymbol{u}_s\boldsymbol{v}_t^\top]\rangle = \begin{cases} f''(\sigma_t) & s = t = k = l \\ \frac{\sigma_s f'(\sigma_s) - \sigma_t f'(\sigma_t)}{\sigma_s^2 - \sigma_t^2} & s \neq t, s \leq p_2, (k,l) = (s,t) \\ -\frac{\sigma_s f'(\sigma_t) - \sigma_t f'(\sigma_s)}{\sigma_s^2 - \sigma_t^2} & s \neq t, s \leq p_2, (k,l) = (t,s) \\ \frac{f'(\sigma_t)}{\sigma_t} & s \neq j, s > p_2, (k,l) = (s,t) \\ 0 & \text{otherwise.} \end{cases}$$

Notice that we obtained the above expressions under the orthonormal bases $\{\boldsymbol{u}_s\boldsymbol{v}_t^\top\}_{s,t}$. In order to get the Hessian form under the canonical bases $\{\boldsymbol{E}_{st}\}_{s,t}$, let $\boldsymbol{Q} \in \mathbb{R}^{p_1p_2 \times p_1p_2}$, with each column $\boldsymbol{Q}_{\cdot,st} = \mathrm{vec}(\boldsymbol{u}_s\boldsymbol{v}_t^\top)$. Denote the matrix form under the canonical bases by $\nabla^2 R(\boldsymbol{B})$ and that under $\{\boldsymbol{u}_s\boldsymbol{v}_t^\top\}_{s,t}$ by $\widetilde{\nabla^2 R(\boldsymbol{B})}$. We then have that

$$\nabla^2 R(\boldsymbol{B}) = \boldsymbol{Q}\widetilde{\nabla^2 R(\boldsymbol{B})}\boldsymbol{Q}^\top$$

This completes our proof. □

E.2 ALO for Smooth Unitarily Invariant Penalties

In this following two sections, we discuss ALO formula for unitarily invariant regularizer R of the form:

$$R(\boldsymbol{B}) = \sum_{j=1}^{\min(p_1,p_2)} r(\sigma_j),$$

where r is a convex and even scalar function. The nuclear norm, Frobenius and numerous other matrix norms all fall in this category. For this section, we consider r as a smooth function. In the next section, we consider the case of the nuclear norm where r is nonsmooth.

Consider the matrix regression problem:

$$\hat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \sum_{j=1}^{n} \ell(\langle \boldsymbol{X}_j, \boldsymbol{B}\rangle; y_j) + \lambda R(\boldsymbol{B}).$$

Let $\hat{\boldsymbol{B}} = \hat{\boldsymbol{U}}\mathrm{diag}[\hat{\boldsymbol{\sigma}}]\hat{\boldsymbol{V}}^\top$. By pluggin the Hessian form we obtained in Theorem E.1 into (20), (21), we have the following ALO formula

$$\langle \boldsymbol{X}_i, \tilde{\boldsymbol{B}}^{/i}\rangle = \langle \boldsymbol{X}_i, \hat{\boldsymbol{B}}\rangle + \frac{H_{ii}}{1 - H_{ii}\ddot{\ell}(\langle \boldsymbol{X}_i, \hat{\boldsymbol{B}}\rangle; y_i)}\dot{\ell}(\langle \boldsymbol{X}_i, \hat{\boldsymbol{B}}\rangle; y_i). \tag{56}$$

where

$$\boldsymbol{H} := \tilde{\boldsymbol{\mathcal{X}}}\left[\tilde{\boldsymbol{\mathcal{X}}}^\top \mathrm{diag}[\ddot{\ell}(\langle \boldsymbol{X}_j, \hat{\boldsymbol{B}}\rangle; y_j)]\tilde{\boldsymbol{\mathcal{X}}} + \lambda\boldsymbol{Q}\boldsymbol{\mathcal{G}}\boldsymbol{Q}^\top\right]^{-1}\tilde{\boldsymbol{\mathcal{X}}}^\top$$

with the matrix $\tilde{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{n \times p_1p_2}$, $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{p_1p_2 \times p_1p_2}$. Each row $\tilde{\boldsymbol{\mathcal{X}}}_{j,\cdot} = \mathrm{vec}(\boldsymbol{X}_j)$. $\boldsymbol{\mathcal{G}}$ is defined by

$$\boldsymbol{\mathcal{G}}_{kl,st} = \begin{cases} r''(\hat{\sigma}_t) & s = t = k = l \\ \frac{\hat{\sigma}_s r'(\hat{\sigma}_s) - \hat{\sigma}_t r'(\hat{\sigma}_t)}{\hat{\sigma}_s^2 - \hat{\sigma}_t^2} & i \neq t, s \leq p_2, (k,l) = (s,t) \\ -\frac{\hat{\sigma}_s r'(\hat{\sigma}_t) - \hat{\sigma}_t r'(\hat{\sigma}_s)}{\hat{\sigma}_s^2 - \hat{\sigma}_t^2} & s \neq t, s \leq p_2, (k,l) = (t,s) \\ \frac{r'(\hat{\sigma}_t)}{\hat{\sigma}_t} & s \neq t, s > p_2, (k,l) = (s,t) \\ 0 & \text{otherwise.} \end{cases} \tag{57}$$

Notice that $[\tilde{\boldsymbol{\mathcal{X}}}\boldsymbol{Q}]_{j,st} = \langle \boldsymbol{X}_j, \hat{\boldsymbol{u}}_s \hat{\boldsymbol{v}}_t^\top \rangle = \hat{\boldsymbol{u}}_s^\top \boldsymbol{X}_j \hat{\boldsymbol{v}}_t$, we have $[\tilde{\boldsymbol{\mathcal{X}}}\boldsymbol{Q}]_{j,\cdot} = \mathrm{vec}(\hat{\boldsymbol{U}}^\top \boldsymbol{X}_j \hat{\boldsymbol{V}})$. Let $\boldsymbol{\mathcal{X}} = \tilde{\boldsymbol{\mathcal{X}}}\boldsymbol{Q}$. This gives us the following nicer form of the $\boldsymbol{H}$ matrix:

$$\boldsymbol{H} := \boldsymbol{\mathcal{X}} \Big[\boldsymbol{\mathcal{X}}^\top \mathrm{diag}[\ddot{\ell}(\langle \boldsymbol{X}_j, \hat{\boldsymbol{B}} \rangle; y_j)] \boldsymbol{\mathcal{X}} + \lambda \boldsymbol{\mathcal{G}} \Big]^{-1} \boldsymbol{\mathcal{X}}^\top$$

E.3 Proof of Theorem 6.1: ALO for Nuclear Norm

For the nuclear norm, we have:

$$\ell(u; y) = \frac{1}{2}(u - y)^2, \quad R(\boldsymbol{B}) = \sum_{j=1}^{\min(p_1, p_2)} \sigma_j$$

Let $P(\boldsymbol{B}) = \frac{1}{2}\sum_{j=1}^n (y_j - \langle \boldsymbol{X}_j, \boldsymbol{B} \rangle)^2 + \lambda \|\boldsymbol{B}\|_*$ denote the primal objective. For the full data optimizer $\hat{\boldsymbol{B}}$ with SVD $\hat{\boldsymbol{B}} = \hat{\boldsymbol{U}} \mathrm{diag}[\hat{\boldsymbol{\sigma}}] \hat{\boldsymbol{V}}$, let $m = \mathrm{rank}(\hat{\boldsymbol{B}})$, the number of nonzero $\hat{\sigma}_j$'s. Furthermore, suppose that we have the following assumption on the full data solution $\hat{\boldsymbol{B}}$.

Assumption E.1. *Let $\hat{\boldsymbol{B}}$ be the full-data minimizer, and let $\hat{\boldsymbol{B}} = \hat{\boldsymbol{U}} \mathrm{diag}[\hat{\boldsymbol{\sigma}}] \hat{\boldsymbol{V}}^\top$ be its SVD.*

1. *$\hat{\boldsymbol{B}}$ is the unique optimizer of the nuclear norm minimization problem,*
2. *For all j such that $\hat{\sigma}_j = 0$, the subgradient $g_r[\hat{\sigma}_j]$ at $\hat{\sigma}_j$ satisfies $g_r[\hat{\sigma}_j] < 1$.*

Since the nuclear norm is nonsmooth, we consider a smoothed version of it. For a matrix and its SVD $\boldsymbol{B} = \boldsymbol{U} \mathrm{diag}[\boldsymbol{\sigma}] \boldsymbol{V}^\top$, and a smoothing parameter $\epsilon > 0$, define the following smoothed version of nuclear norm as

$$R_\epsilon(\boldsymbol{B}) = \sum_{j=1}^{\min(p_1, p_2)} r_\epsilon(\sigma_j), \text{ where } r_\epsilon(x) = \sqrt{x^2 + \epsilon^2}.$$

Let $P_\epsilon(\boldsymbol{B}) = \frac{1}{2}\sum_{j=1}^n (y_j - \langle \boldsymbol{X}_j, \boldsymbol{B} \rangle)^2 + \lambda R_\epsilon(\boldsymbol{B})$ denote the smoothed primal objective, and let $\hat{\boldsymbol{B}}_\epsilon$ be the minimizer of P_ϵ. Note that instead of using the general kernel smoothing strategy we mentioned in the previous section, in this specific case we consider this choice R_ϵ for technical convenience. There are no essential differences between the two smoothing schemes. Finally, let $r(x) = |x|$

Lemma E.1 guarantees the smoothness and convexity of the function R_ϵ. Additionally, r_ϵ verifies several desirable properties:

1. $\dot{r}_\epsilon(x) = \frac{x}{\sqrt{x^2+\epsilon^2}}$, $\ddot{r}_\epsilon(x) = \frac{\epsilon^2}{(x^2+\epsilon^2)^{\frac{3}{2}}}$;
2. $r(x) < r_\epsilon(x) < r(x) + \epsilon$.

In particular, we note that the second property implies that $\sup_x |r(x) - r_\epsilon(x)| \le \epsilon$ and that $\sup_{\boldsymbol{B}} |R(\boldsymbol{B}) - R_\epsilon(\boldsymbol{B})| \le \epsilon \min(p_1, p_2)$.

We now go through a similar strategy as in Appendix C.2 to consider the limit case as $\epsilon \to 0$.

Convergence of the optimizer ($\hat{\boldsymbol{B}}_\epsilon \to \hat{\boldsymbol{B}}$) By definition of $\hat{\boldsymbol{B}}$ as the minimizer of the primal objective, we have that:

$$\lambda \|\hat{\boldsymbol{B}}\|_* \le \frac{1}{2}\sum_j (y_j - \langle \boldsymbol{X}_j, \hat{\boldsymbol{B}} \rangle)^2 + \lambda \|\hat{\boldsymbol{B}}\|_* \le \frac{1}{2}\|\boldsymbol{y}\|_2^2.$$

Similarly, we have that $\hat{\boldsymbol{B}}_\epsilon$ verifies:

$$\begin{aligned}
\lambda \|\hat{\boldsymbol{B}}_\epsilon\|_* &\le \lambda R(\hat{\boldsymbol{B}}_\epsilon) \le \lambda R_\epsilon(\hat{\boldsymbol{B}}_\epsilon) + \lambda \epsilon \min(p_1, p_2) \\
&\le \frac{1}{2}\sum_j (y_j - \langle \boldsymbol{X}_j, \hat{\boldsymbol{B}}_\epsilon \rangle)^2 + \lambda R_\epsilon(\hat{\boldsymbol{B}}_\epsilon) + \lambda \epsilon \min(p_1, p_2) \\
&\le \frac{1}{2}\|\boldsymbol{y}\|_2^2 + \lambda \epsilon \min(p_1, p_2).
\end{aligned}$$

Thus, for all $\epsilon \leq 1$ both $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{B}}_\epsilon$ are contained in a compact set given by $\lambda\|\boldsymbol{B}\|_* \leq \frac{1}{2}\|\boldsymbol{y}\|_2^2 + \lambda \min(p_1, p_2)$.

In particular, any subsequence of $\hat{\boldsymbol{B}}_\epsilon$ contains a convergent sub-subsequence, let us abuse notations and still use $\hat{\boldsymbol{B}}_\epsilon$ for this convergent sub-subsequence. The uniform bound between R and R_ϵ implies that:

$$P(\lim_{\epsilon\to 0} \hat{\boldsymbol{B}}_\epsilon) = \lim_{\epsilon\to 0} P(\hat{\boldsymbol{B}}_\epsilon) = \lim_{\epsilon\to 0} P_\epsilon(\hat{\boldsymbol{B}}_\epsilon) \leq \lim_{\epsilon\to 0} P_\epsilon(\hat{\boldsymbol{B}}) = P(\hat{\boldsymbol{B}}).$$

By the uniqueness of the optimizer $\hat{\boldsymbol{B}}$, we have

$$\lim_{\epsilon\to 0} \hat{\boldsymbol{B}}_\epsilon = \hat{\boldsymbol{B}}.$$

This is true for all such subsequences, which confirms what we want to prove.

Convergence of the gradient ($\nabla R_\epsilon(\hat{\boldsymbol{B}}_\epsilon) \to g_{\|\cdot\|_*}(\hat{\boldsymbol{B}})$) Let $g_{\|\cdot\|_*}$ denote the subgradient of the nuclear norm $\|\cdot\|_*$ in the first order optimality condition of $\hat{\boldsymbol{B}}$. By the continuity of $\dot{\ell}$ and the first order condition, we have:

$$\left\|g_{\|\cdot\|_*}(\hat{\boldsymbol{B}}) - \nabla R_\epsilon(\hat{\boldsymbol{B}}_\epsilon)\right\|_F = \left\|\sum_{j=1}^n \langle \boldsymbol{X}_j, \hat{\boldsymbol{B}} - \hat{\boldsymbol{B}}_\epsilon\rangle \boldsymbol{X}_j\right\|_F \to 0. \tag{58}$$

Let $\hat{\boldsymbol{B}}_\epsilon = \hat{\boldsymbol{U}}_\epsilon \mathrm{diag}[\hat{\boldsymbol{\sigma}}_\epsilon]\hat{\boldsymbol{V}}_\epsilon$ denote the SVD of $\hat{\boldsymbol{B}}_\epsilon$. By Lemma E.1 we have:

$$g_{\|\cdot\|_*}(\hat{\boldsymbol{B}}) = \hat{\boldsymbol{U}}\mathrm{diag}(\{g_r[\hat{\sigma}_j]\}_j)\hat{\boldsymbol{V}}^\top,$$
$$\nabla R_\epsilon(\hat{\boldsymbol{B}}^\epsilon) = \hat{\boldsymbol{U}}_\epsilon \mathrm{diag}(\{\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,j})\}_j)\hat{\boldsymbol{V}}_\epsilon^\top.$$

where $g_r[x] = 1$ if $x > 0$ and $0 \leq g_r[x] \leq 1$ if $x = 0$.

We wish to translate the limit in matrix norm (58) to a limit on their singular values. In order to do this, we use the following lemma from Weyl [Weyl, 1912] or Mirsky [Mirsky, 1960]. We note that our conclusion may follow from either, although we include both for completeness.

Lemma E.2 ([Weyl, 1912],[Mirsky, 1960])**.** *Let A and B be two rectangular matrices of the same shape. Let σ_j denote the j^{th} largest eigenvalue, then we have that for all j:*

$$|\sigma_j(A) - \sigma_j(B)| \leq \|A - B\|_2,$$
$$\sqrt{\sum_j (\sigma_j(A) - \sigma_j(B))^2} \leq \|A - B\|_F.$$

By Lemma E.2, we have that $\hat{\sigma}_{\epsilon,j} \to \hat{\sigma}_j$ and $\frac{\hat{\sigma}_{\epsilon,j}}{\sqrt{\hat{\sigma}_{\epsilon,j}^2 + \epsilon^2}} \to g_r[\hat{\sigma}_j]$ as $\epsilon \to 0$. Additionally, by the assumption $g_r[\hat{\sigma}_j] < 1$ if $\hat{\sigma}_j = 0$, we have that:

$$\frac{\hat{\sigma}_{\epsilon,j}}{\epsilon} \to \begin{cases} +\infty & \text{if } \hat{\sigma}_j > 0, \\ < +\infty & \text{if } \hat{\sigma}_j = 0. \end{cases} \tag{59}$$

This further implies the matrices $\boldsymbol{\mathcal{G}}_\epsilon$ defined as in (57) for R_ϵ satisifies:

$$\lim_{\epsilon\to 0} \boldsymbol{\mathcal{G}}_{\epsilon,kl,ij} = \begin{cases} 0 & s = t = k = l \leq m \\ \infty & s = t = k = l > m \\ \frac{1}{\hat{\sigma}_s + \hat{\sigma}_t} & 1 \leq s \neq t \leq m, (k,l) = (s,t) \\ \frac{1}{\hat{\sigma}_s} & 1 \leq s \leq m < t \leq p_2, (k,l) = (s,t) \\ \frac{1}{\hat{\sigma}_t} & 1 \leq t \leq m < s \leq p_2, (k,l) = (s,t) \\ -\frac{1}{\hat{\sigma}_s + \hat{\sigma}_t} & 1 \leq s \neq t \leq m, (k,l) = (t,s) \\ -\frac{g_r[\hat{\sigma}_t]}{\hat{\sigma}_s} & 1 \leq s \leq m < t \leq p_2, (k,l) = (t,s) \\ -\frac{g_r[\hat{\sigma}_s]}{\hat{\sigma}_t} & 1 \leq t \leq m < s \leq p_2, (k,l) = (t,s) \\ \frac{1}{\hat{\sigma}_t} & 1 \leq t \leq m \leq p_2 < s \leq p_1, (k,l) = (s,t) \\ \infty & m < t \leq p_2 < s \leq p_1, (k,l) = (s,t) \\ 0 & \text{otherwise.} \end{cases} \tag{60}$$

Notice that in (60) we missed a piece of blocks corresponding to $m < s \neq t \leq p_2$, $(k,l) = (s,t)$ or $(k,l) = (t,s)$. We need to process this blocks separately. We will show that the inverse of the corresponding blocks in $\boldsymbol{\mathcal{G}}_\epsilon$ converges to 0. As a result, we can ignore this part according to Lemma C.6.

Each 2×2 sub-matrix within that blocks in $\boldsymbol{\mathcal{G}}_\epsilon$ takes the form

$$\frac{1}{\hat{\sigma}_{\epsilon,s}^2 - \hat{\sigma}_{\epsilon,t}^2} \begin{bmatrix} \hat{\sigma}_{\epsilon,s}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,s}) - \hat{\sigma}_{\epsilon,t}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,t}) & -\hat{\sigma}_{\epsilon,s}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,t}) + \hat{\sigma}_{\epsilon,t}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,s}) \\ -\hat{\sigma}_{\epsilon,s}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,t}) + \hat{\sigma}_{\epsilon,t}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,s}) & \hat{\sigma}_{\epsilon,s}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,s}) - \hat{\sigma}_{\epsilon,t}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,t}) \end{bmatrix}$$

It is easy to verify that the inverse of the above matrix takes the following form

$$\frac{1}{\dot{r}^2(\hat{\sigma}_{\epsilon,s}) - \dot{r}^2(\hat{\sigma}_{\epsilon,t})} \begin{bmatrix} \hat{\sigma}_{\epsilon,s}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,s}) - \hat{\sigma}_{\epsilon,t}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,t}) & \hat{\sigma}_{\epsilon,s}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,t}) - \hat{\sigma}_{\epsilon,t}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,s}) \\ \hat{\sigma}_{\epsilon,s}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,t}) - \hat{\sigma}_{\epsilon,t}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,s}) & \hat{\sigma}_{\epsilon,s}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,s}) - \hat{\sigma}_{\epsilon,t}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,t}) \end{bmatrix} \tag{61}$$

For the two distinct component values in the matrix in (61), we have that

$$\frac{\hat{\sigma}_{\epsilon,s}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,s}) - \hat{\sigma}_{\epsilon,t}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,t})}{\dot{r}^2(\hat{\sigma}_{\epsilon,s}) - \dot{r}^2(\hat{\sigma}_{\epsilon,t})} = \frac{\frac{\hat{\sigma}_{\epsilon,s}^2}{\sqrt{\hat{\sigma}_{\epsilon,s}+\epsilon^2}} - \frac{\hat{\sigma}_{\epsilon,t}^2}{\sqrt{\hat{\sigma}_{\epsilon,t}+\epsilon^2}}}{\frac{\hat{\sigma}_{\epsilon,s}^2}{\hat{\sigma}_{\epsilon,s}+\epsilon^2} - \frac{\hat{\sigma}_{\epsilon,t}^2}{\hat{\sigma}_{\epsilon,t}+\epsilon^2}} = \epsilon \frac{\frac{u_{\epsilon,s}}{\sqrt{1-u_{\epsilon,s}}} - \frac{u_{\epsilon,t}}{\sqrt{1-u_{\epsilon,t}}}}{u_{\epsilon,s} - u_{\epsilon,t}} = \epsilon \frac{1 - \frac{1}{2}\tilde{u}_\epsilon}{(1-\tilde{u}_\epsilon)^{\frac{3}{2}}} \to 0$$

where we did a change of variable $u = \frac{\hat{\sigma}^2}{\hat{\sigma}^2+\epsilon^2}$ and $\tilde{u}_\epsilon$ is a value between $u_{\epsilon,s}$ and $u_{\epsilon,t}$ where we apply Taylor expansion to function $\frac{x}{\sqrt{1-x}}$. The last convergence to 0 is obtained by noticing that $\lim_{\epsilon\to 0} u_{\epsilon,s}, \lim_{\epsilon\to 0} u_{\epsilon,t} \in [0,1)$ due to (59).

Similarly we have the following analysis for the off-diagonal term

$$\frac{\hat{\sigma}_{\epsilon,s}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,t}) - \hat{\sigma}_{\epsilon,t}\dot{r}_\epsilon(\hat{\sigma}_{\epsilon,s})}{\dot{r}^2(\hat{\sigma}_{\epsilon,s}) - \dot{r}^2(\hat{\sigma}_{\epsilon,t})} = \frac{\frac{\hat{\sigma}_{\epsilon,s}\hat{\sigma}_{\epsilon,t}}{\sqrt{\hat{\sigma}_{\epsilon,t}+\epsilon^2}} - \frac{\hat{\sigma}_{\epsilon,s}\hat{\sigma}_{\epsilon,t}}{\sqrt{\hat{\sigma}_{\epsilon,s}+\epsilon^2}}}{\frac{\hat{\sigma}_{\epsilon,s}^2}{\hat{\sigma}_{\epsilon,s}+\epsilon^2} - \frac{\hat{\sigma}_{\epsilon,t}^2}{\hat{\sigma}_{\epsilon,t}+\epsilon^2}} = \frac{\hat{\sigma}_{\epsilon,s}\hat{\sigma}_{\epsilon,t}}{\epsilon} \frac{\sqrt{1-u_{\epsilon,t}} - \sqrt{1-u_{\epsilon,s}}}{u_{\epsilon,s} - u_{\epsilon,t}} = \frac{\hat{\sigma}_{\epsilon,s}\hat{\sigma}_{\epsilon,t}}{\epsilon^2} \frac{\epsilon}{2\sqrt{1-\bar{u}_\epsilon}} \to 0$$

where $\bar{u}_\epsilon$ is a value between $u_{\epsilon,s}$ and $u_{\epsilon,t}$ where we use Taylor expansion to $\sqrt{1-x}$. The last convergence to 0 is obtained based on the same reason as the previous one.

Let $E := \{kl : k \leq m \text{ or } l \leq m\}$, by Lemma C.6, we have

$$\boldsymbol{H}_\epsilon \to \boldsymbol{\mathcal{X}}_{\cdot,E} \Big[\boldsymbol{\mathcal{X}}_{\cdot,E}^\top \boldsymbol{\mathcal{X}}_{\cdot,E} + \lambda \boldsymbol{\mathcal{G}} \Big]^{-1} \boldsymbol{\mathcal{X}}_{\cdot,E}^\top := \boldsymbol{H}$$

where $\boldsymbol{\mathcal{G}}$ is defined in (30).

Finally, we obtain our approximation of leave-i-out prediction by substituting the above formula of $\boldsymbol{H}$ into the general formula (56).

Remark E.2. *Similar to what we did in Figure 3, it is helpful to visualize the structure of $\boldsymbol{\mathcal{G}}$ in correspondence to the blocks of the original matrix. Specifically we have Figure 4.*

F Details of the Numerical Experiments

F.1 Simulated Data

F.1.1 Support Vector Machine

For all SVM simulations the data is generated according to a Gaussian logistic model: the design matrix $\boldsymbol{X}$ is generated as a matrix of i.i.d. $\mathcal{N}(0,1)$; the true parameter $\boldsymbol{\beta}$ is i.i.d. $\mathcal{N}(0,9)$, and each response y_i is generated as an independent Bernoulli with probability p_i given by the following logistic model:

$$\log \frac{p_i}{1-p_i} = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

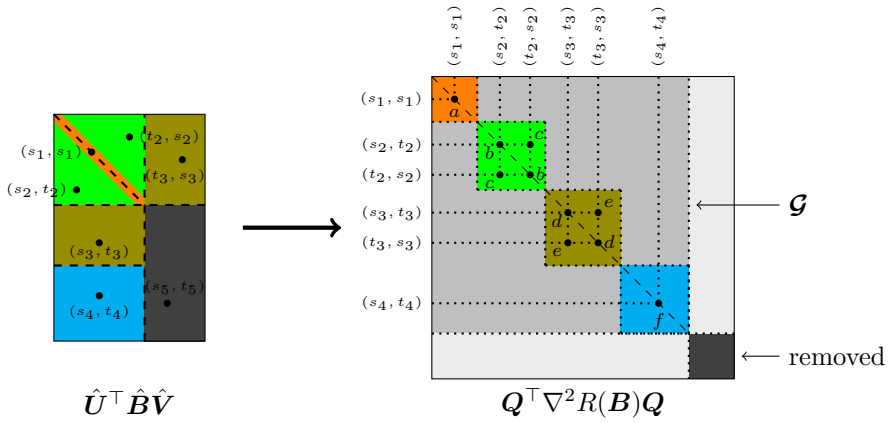

Figure 4: An illustration of the correspondence between the structure of the original matrix and the structure of the $\mathcal{G}$ matrix. As we have mentioned in Theorem E.1, $a = 0$, $b = \frac{1}{\hat{\sigma}_{s_2} + \hat{\sigma}_{t_2}}$, $c = -\frac{1}{\hat{\sigma}_{s_2} + \hat{\sigma}_{t_2}}$, $d = \frac{1}{\hat{\sigma}_{t_3}}$, $e = -\frac{g_r[\hat{\sigma}_{s_3}]}{\hat{\sigma}_{t_3}}$, $f = \frac{1}{\hat{\sigma}_{t_4}}$.

The $n > p$ scenario is generated with $n = 300$ and $p = 80$, and the $n < p$ scenario is generated with $n = 300$ and $p = 600$. We consider a sequence of 40 different values of λ ranging between $e^4 \sim e^{12}$, with their logarithm equally spaced between $[4, 12]$.

The model is fitted using the `sklearn.svm.linearSVC` function in Python package `scikit-learn` [Pedregosa et al., 2011], which is implemented by the `LibSVM` package [Chang and Lin, 2011].

For using the `sklearn.svm.linearSVC`, we set `tolerance`$=10^{-6}$ and `max_iter=10000`. We identify an observation as a support vector if $|1 - y_i \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}| < 10^{-5}$.

F.1.2 Fused LASSO

For all fused LASSO, each component of the design matrix $\boldsymbol{X}$ is generated from i.i.d. $\mathcal{N}(0, 0.05)$. For the true parameter $\boldsymbol{\beta}$, we generated it through the following process: given a number $k < p$, we generate a sparse vector $\boldsymbol{\beta}_0$ with a random sample of k of its components i.i.d. from $\mathcal{N}(0, 1)$. Then we construct a new vector $\boldsymbol{\beta}_1$ as the cumulative sum of $\boldsymbol{\beta}_0$: $\beta_{1,i} = \sum_{j=1}^{i} \beta_{0,j}$; Finally we normalize $\boldsymbol{\beta}_1$ such that it has standard deviation 1. Note that $\boldsymbol{\beta}_1$ is a piecewise constant vector. The response $\boldsymbol{y}$ is generated as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ denotes i.i.d. random gaussian noise from $\mathcal{N}(0, 0.25)$. For our simulation, we use $k = 20$ (so piecewise constant with 20 pieces). The $n > p$ scenario is generated with $n = 200$ and $p = 100$, whereas the $n < p$ scenario is generated with $n = 200$ and $p = 400$.

The model is fitted through a direct translation of the generalized LASSO model into the package `CVX` [Grant and Boyd, 2014]. We use the default tolerance and maximal iteration. We identify the location i such that $\hat{\beta}_{i+1} = \hat{\beta}_i$ by checking if $|\hat{\beta}_{i+1} - \hat{\beta}_i| < 10^{-8}$. For $n > p$, we consider a sequence of 40 tuning parameters from $10^{-2} \sim 10^2$; For $n < p$, we consider a sequence of 30 tuning parameters from $10^{-1} \sim 10$. Both are equally spaced on the log-scale.

F.1.3 Nuclear Norm Minimization

For all nuclear norm simulations the data is generated according to the Gaussian low-rank model; each observation matrix $\boldsymbol{X}_j$ is generated as an i.i.d. $\mathcal{N}(0, 1)$ matrix. The true parameter matrix $\boldsymbol{B}$ is generated as a low rank matrix, by setting $k = 1$ in the following formula

$$\boldsymbol{B} = \sum_{l=1}^{k} \boldsymbol{z}_l \boldsymbol{w}_l^\top,$$

where $\boldsymbol{z}, \boldsymbol{w}$ are independent of each other. $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_{p_1})$, $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_{p_2})$. Hence, the rank of $\boldsymbol{B}$ in our experiments is equal to 1. The response $\boldsymbol{y}$ is generated as $y_j = \langle \boldsymbol{X}_j, \boldsymbol{B} \rangle + \epsilon_j$, where ϵ_j is i.i.d. $\mathcal{N}(0, 0.25)$.

The $n > p$ scenario is generated with $n = 600$, and $\boldsymbol{B} \in \mathbb{R}^{20 \times 20}$ (i.e. $p = 400$). The $n < p$ scenario is generated with $n = 200$, and $\boldsymbol{B} \in \mathbb{R}^{20 \times 20}$ again. For both settings, we consider a sequence of 30 tuning parameters from $5 \times 10^{-1} \sim 5 \times 10$, equally spaced on the log-scale.

The model is fitted using an implementation of a proximal gradient algorithm as described in [Lan et al., 2011], implemented using the Matlab package `TFOCS` [Becker et al., 2011]. The threshold we use to identify singular values with value 0 is $10^{-3} \times \lambda_{\max}(\hat{\boldsymbol{B}})$, where $\lambda_{\max}$ is the maximal singular value of $\hat{\boldsymbol{B}}$.

F.1.4 LASSO Experiment

In our LASSO simulations, we use the setting where $n = 300$, $p = 600$, and the true model is sparse with $k = 60$ non-zeros. These non-zeros are i.i.d. $\mathcal{N}(0, 1)$.

In the misspecification example, the elements of $\boldsymbol{X}$ are i.i.d. $\mathcal{N}(0, 1/k)$. $\boldsymbol{y}$ is generated according to the following non-linear model:

$$y_j = f(\boldsymbol{x}_j^\top \boldsymbol{\beta} + \epsilon_j),$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, 0.25\boldsymbol{I}_n)$, and the function f is given by:

$$f(x) = \begin{cases} \sqrt{x} & \text{if } x \geq 0, \\ -\sqrt{-x} & \text{otherwise.} \end{cases}$$

In the heavy-tailed noise example, the elements of $\boldsymbol{X}$ are i.i.d. $\mathcal{N}(0, 1/k)$. $\boldsymbol{y}$ is generated according to

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where the "heavy-tailed" noise ϵ_j is generated according to a Student-t distribution with three degrees of freedom, and rescaled such that its variance is $\sigma^2 = 0.25$.

In the correlated design example, $\boldsymbol{y}$ is generated according to

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, 0.25\mathbf{I})$, and the "correlated design" $\boldsymbol{X}$ is generated with each row $\boldsymbol{x}_j$ being sampled independently according to a multivariate normal distribution $\boldsymbol{x}_j \sim \mathcal{N}(0, \boldsymbol{C}/k)$, where $\boldsymbol{C}$ is the Toeplitz matrix, given by:

$$\boldsymbol{C} = \begin{pmatrix} \rho & \rho^2 & \dots & \rho^p \\ \rho^2 & \rho & \dots & \rho^{p-1} \\ \vdots & \dots & \ddots & \vdots \\ \rho^p & \rho^{p-1} & \dots & \rho \end{pmatrix}.$$

ρ is set to 0.8 in our experiments. For all settings, we consider a sequence of 25 tuning parameters from $3.16 \times 10^{-3} \sim 3.16 \times 10^{-2}$, equally spaced under log-scale.

All models were solved using the `glmnet` package in Matlab [Qian et al., 2013]. We identify the zero locations of $\hat{\boldsymbol{\beta}}$ by checking $|\beta_j| > 10^{-8}$.

F.1.5 Timing of ALO

For comparing the timing of ALO with that of LOOCV, we consider the LASSO problem with correlated design similar to the one we introduced in Section F.1.4. Specifically, each row of the design matrix has a Toeplitz covariance matrix with $\rho = 0.8$. The true coefficient vector $\boldsymbol{\beta}$ has $\frac{\min(n,p)}{2}$ nonzero components, with each nonzero component of $\boldsymbol{\beta}$ being selected independently from ± 1 with probability 0.5. The noise $\epsilon \sim \mathcal{N}(0, 0.5\boldsymbol{I}_n)$. For each pair of (n, p), we choose a sequence of 50 tuning parameters ranging from λ_0 to $10^{-2.5}\lambda_0$, where $\lambda_0 = \|\boldsymbol{X}^\top \boldsymbol{y}\|_\infty$. Note that for this choice of λ all the regression coefficients are equal to zero.

The timing of one single fit on the full dataset, the ALO risk estimates and the LOOCV risk estimates are reported in Table 1 of the main paper. To obtain the timing of a single fit we run the corresponding function of glmnet along the entire tuning parameter path and record the total time consumed. This process is then repeated for 10 random seeds to obtain the average timing. Every time an estimate is obtained we use our formula to obtain ALO. Hence, the time reported for ALO in Table 1 is again obtained from an average of 10 Monte Carlo samples. To obtain the computation time of LOOCV, we only use 5 random seeds.

As expected, averaged time for LOOCV is close to n times the time required for a single fit. On the other hand, among all the settings we considered in Table 1, ALO takes less than twice the time of a single fit.

F.2 Real-World Data

In this section, we apply our ALO methods to three real-world datasets: Gisette digit recognition [Guyon et al., 2005], the tumor colon tissues gene expression [Alon et al., 1999] and the South Africa heart disease data. All the three datasets have binary response, so we consider classification algorithms. The information of the three datasets is listed in Table 2 below. The column of number of effective features records the number of features after data preprocessing, including removing duplicates and missing columns.

Table 2: Information of the three datasets.

dataset	# samples	# features	# effective features	model used
gisette	6000	5000	4955	SVM
tumor colon	62	2000	1909	logistic + LASSO
heart disease	462	9	9	logistic + LASSO

For gisette, since $n = 6000$ is too large for LOOCV, we randomly subsample 1000 observations and apply linear SVM on it. For the tumor colon tissues and South Africa heart disease dataset, we apply logistic regression with LASSO penalty. The results are shown in Figure 5. The accuracy of ALO is verified on gisette and the heart disease dataset. However, the behavior of ALO is more complicated for the tumor colon tissues dataset. First ALO gives very close estimates to LOOCV for relatively large tuning values, but deviates from LOOCV risk estimates and bends upward after λ decreases to a certain value. Second, we note that the optimal tuning is still correctly captured by ALO.

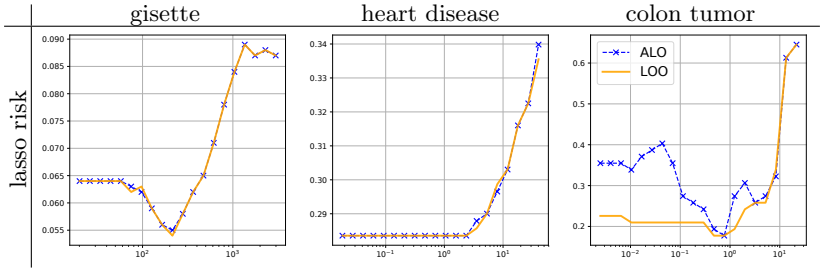

Figure 5: Risk estimates of from ALO versus LOOCV for the three datasets: gisette, South Africa coronary heart disease and colon tumor gene expression. The x-axis is the tuning parameter value λ on log-scale, the y-axis is the risk estimates under 0-1 loss.

There are a few factors which may affect the performance of ALO. First, as implied by the theoretical guarantee on smooth models, the closeness between ALO and LOOCV is a high-dimensional phenomenon, which takes place for relatively large n and p. This requirement of

high-dimensionality is less stringent for $n > p$, when strong convexity of the loss function is to some extent guaranteed, but becomes more significant for $n < p$. On the other hand, from our simulation in Section 7 and the real-data examples in this section, we can see that when $\frac{n}{p}$ is not much smaller than 1 (compared to the $\frac{n}{p}$-ratio in the colon tissue dataset), a few hundreds of observation and features are enough to guarantee the accuracy of ALO risk estimates. Finally, the deviation of ALO estimates tends to happen when the tuning λ becomes smaller than a certain value, typically in the case of $n < p$. As we have mentioned in the main text, for most nonsmooth regularizers, small tuning values induce dense solutions. In most high dimensional datasets, these dense solutions are not favorable in $n < p$ case. From our experiments, this deviation mostly happens after correctly capturing the optimal tuning values.