A. Preliminaries on Convex Envelopes and Subdifferentials

In this section we review some basic facts about convex envelopes and their sub-differentials that we will use in our theory. Throughout the section we will assume that any infimum is attained. This is true for example if the function is lower semi continuous with bounded level sets, which is the case for our objective function $f(\boldsymbol{x}) = g(\text{card}(\boldsymbol{x})) + \|\boldsymbol{x}\|^2$. In addition the quadratic term grows faster than any linear term of the type $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ and therefore this is also true when we add linear terms.

The convex envelope f^{**} of a function f is the largest convex function that fulfills $f^{**} \leq f$. For a convex function we should have $f^{**}(\sum_i \lambda_i \boldsymbol{x}^j) \leq \sum_j \lambda_j f^{**}(\boldsymbol{x}^j)$, $0 \leq \lambda_j$, $\sum_j \lambda_j = 1$. At a point $\boldsymbol{x} = \sum_j \lambda_j \boldsymbol{x}^j$ where $f(\boldsymbol{x}) > \sum_j \lambda_i f(\boldsymbol{x}^j)$ we compute the value $f^{**}(\boldsymbol{x})$ by minimizing over convex combinations of points using $f^{**}(x) =$

$$\min \left\{ \sum_{j=1}^{d+1} \lambda_j f(\boldsymbol{x}^j); \quad \sum_{j=1}^{d+1} \lambda_j \boldsymbol{x}^j = \boldsymbol{x}, \quad \sum_{j=1}^{d+1} \lambda_j = 1, \quad \lambda_j > 0 \right\}. \tag{20}$$

It can be shown (using Caratheodory's Theorem) that it is enough to consider combinations of $d+1$ points if $\boldsymbol{x} \in \mathbb{R}^d$. Figure 4 shows one example of convex envelope. Here the two functions f^{**} and f coincide at $x = 0$ and $|x| > 1$. If $x \in (0, 1)$ where the functions differ and the value of f^{**} is computed using the convex combination $(1-x)f(0) + xf(1)$. Note that when f and f^{**} differs the function f^{**} will be affine in some direction.

An alternative way of computing f^{**} is using supporting hyperplanes and the conjugate function

$$f^*(\boldsymbol{y}) = \max_{\boldsymbol{x}} \langle \boldsymbol{x}, \boldsymbol{y}\rangle - f(\boldsymbol{x}). \tag{21}$$

From the definition it is clear that

$$f^*(\boldsymbol{y}) \geq \langle \boldsymbol{x}, \boldsymbol{y}\rangle - f(\boldsymbol{x}), \tag{22}$$

for all $\boldsymbol{x}$. Rearranging terms we get

$$f(\boldsymbol{x}) \geq \langle \boldsymbol{x}, \boldsymbol{y}\rangle - f^*(\boldsymbol{y}), \tag{23}$$

which is a an affine function in $\boldsymbol{x}$ and therefore a supporting hyperplane to f. Figure 4 shows three supporting hyperplanes for f. Note that these touch f^{**} in (at least) one point. For each $\boldsymbol{x}$ we can find a hyperplane that touches $f^{**}(\boldsymbol{x})$ which means that

$$f^{**}(\boldsymbol{x}) = \max_{\boldsymbol{y}} \langle \boldsymbol{x}, \boldsymbol{y}\rangle - f^*(\boldsymbol{y}), \tag{24}$$

that is, the convex envelope is the conjugate of the conjugate function.

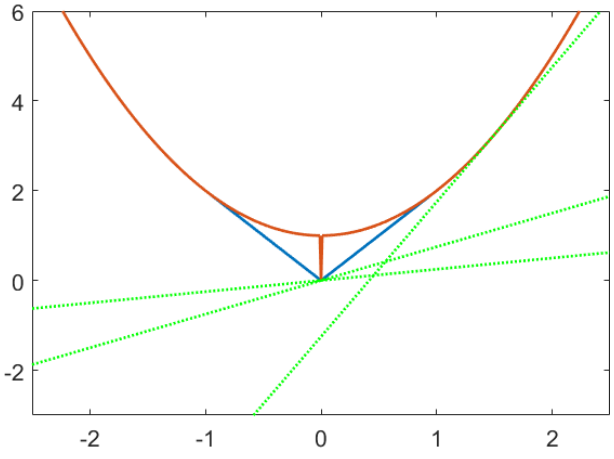

Figure 4: An example of convex envelope. Here $f^{**}(x) = \mu - \max(\sqrt{\mu} - |x|, 0)^2 + x^2$ (blue curve) and $f(x) = \mu \text{card}(x) + x^2$ (orange curve). Green dotted lines are supporting hyperplanes of the form $xy - f^*(y)$ for tree different y.

For a convex function f^{**} the set of sub-gradients $\partial f^{**}(\boldsymbol{x})$ at a point $\boldsymbol{x}$ is defined as all vectors $\boldsymbol{y}$ such that

$$f^{**}(\boldsymbol{x}') - f^{**}(\boldsymbol{x}) \geq \langle \boldsymbol{y}, \boldsymbol{x}' - \boldsymbol{x}\rangle, \quad \forall \boldsymbol{x}'. \tag{25}$$

or equivalently

$$\langle \boldsymbol{y}, \boldsymbol{x}\rangle - f^{**}(\boldsymbol{x}) \geq \langle \boldsymbol{y}, \boldsymbol{x}'\rangle - f^{**}(\boldsymbol{x}'), \quad \forall \boldsymbol{x}'. \tag{26}$$

Since we clearly have equality when $\boldsymbol{x}' = \boldsymbol{x}$ this means that $\langle \boldsymbol{y}, \boldsymbol{x}\rangle - f^{**}(\boldsymbol{x}) = \max_{\boldsymbol{x}'} \langle \boldsymbol{y}, \boldsymbol{x}'\rangle - f^{**}(\boldsymbol{x}') = f^{***}(\boldsymbol{x}) = f^*(\boldsymbol{x})$. Rearranging terms shows that

$$\langle \boldsymbol{y}, \boldsymbol{x}\rangle - f^*(\boldsymbol{y}) = f^{**}(\boldsymbol{x}) = \max_{\boldsymbol{y}'} \langle \boldsymbol{y}', \boldsymbol{x}\rangle - f^*(\boldsymbol{y}'). \tag{27}$$

Thus the set of sub-gradients at a point $\boldsymbol{x}$ are all the vectors $\boldsymbol{y}$ that achieves the maximal value in the second conjugation. In points where f^{**} is non-differentiable the function has several sub-gradients. In a differentiable point the only sub-gradient is the standard gradient.

The following result does not appear to be standard but is crucial for our main theorem. Therefore we state it somewhat more formally below.

Lemma A.1. *Suppose that for a point $\boldsymbol{x}$ we have $f(\boldsymbol{x}) > f^{**}(\boldsymbol{x})$. Then there is a set of points $\{\boldsymbol{x}^j\}$ such that*

$$\boldsymbol{x} = \sum_j \lambda_j \boldsymbol{x}^j, \quad 0 \leq \lambda_j \leq 1, \quad \sum_j \lambda_j = 1, \tag{28}$$

$f^{**}(\boldsymbol{x}^j) = f(\boldsymbol{x}^j)$ *and*

$$f^{**}(\boldsymbol{x}) = \sum_j \lambda_j f(\boldsymbol{x}^j). \tag{29}$$

In addition $\partial f^{**}(\boldsymbol{x}) \subset \bigcap_j \partial f^{**}(\boldsymbol{x}^j)$.

Proof. Consider the convex combination $\boldsymbol{x} = \sum_j \boldsymbol{x}^j$ that solves the minimization in (20).

We have that $f(\boldsymbol{x}^j) \geq f^{**}(\boldsymbol{x}^j)$. Assume further that $f(\boldsymbol{x}^j) > f^{**}(\boldsymbol{x}^j)$ for some j. Then we have

$$f^{**}(\sum_j \lambda_j \boldsymbol{x}^j) = f^{**}(\boldsymbol{x}) = \sum_j \lambda_j f(\boldsymbol{x}^j) > \sum_j \lambda_j f^{**}(\boldsymbol{x}^j), \tag{30}$$

which contradicts the convexity of f^{**}. Therefore $f(\boldsymbol{x}^j) = f^{**}(\boldsymbol{x}^j)$ for all j.

Now consider a subgradient $\boldsymbol{y} \in \partial f^{**}(\boldsymbol{x})$. By definition we have that

$$f^{**}(\boldsymbol{x}') \geq f^{**}(\boldsymbol{x}) + \langle \boldsymbol{y}, \boldsymbol{x}' - \boldsymbol{x} \rangle. \tag{31}$$

Now assume that

$$f^{**}(\boldsymbol{x}^j) > f^{**}(\boldsymbol{x}) + \langle \boldsymbol{y}, \boldsymbol{x}^j - \boldsymbol{x} \rangle, \tag{32}$$

for some j. Then we have

$$\underbrace{\sum_j \lambda_j f^{**}(\boldsymbol{x}^j)}_{=f^{**}(\boldsymbol{x})} > \underbrace{\sum_j \lambda_j f^{**}(\boldsymbol{x})}_{=f^{**}(\boldsymbol{x})} + \underbrace{\sum_j \lambda_j \langle \boldsymbol{y}, \boldsymbol{x}^j - \boldsymbol{x} \rangle}_{=0}, \tag{33}$$

which shows that we must have

$$f^{**}(\boldsymbol{x}^j) = f^{**}(\boldsymbol{x}) + \langle \boldsymbol{y}, \boldsymbol{x}^j - \boldsymbol{x} \rangle. \tag{34}$$

This gives us

$$f^{**}(\boldsymbol{x}^j) + \langle \boldsymbol{y}, \boldsymbol{x}' - \boldsymbol{x}^j \rangle = f^{**}(\boldsymbol{x}) + \langle \boldsymbol{y}, \boldsymbol{x}' - \boldsymbol{x} \rangle \leq f^{**}(\boldsymbol{x}'), \tag{35}$$

which shows that $\boldsymbol{y} \in \partial f^{**}(\boldsymbol{x}^j)$ for all j. □

A.1. The Conjugate of f

We now consider our class of functions $f(\boldsymbol{x}) = g(\text{card}(\tilde{\boldsymbol{x}})) + \|\tilde{\boldsymbol{x}}\|^2$. Consider the conjugate (21). Since $f(\boldsymbol{x})$ only depends on $\tilde{\boldsymbol{x}}$ and not the signs or ordering of the elements it is clear that the elements of $\boldsymbol{x}$ should have the same sign and ordering as those in $\boldsymbol{y}$ to maximize the term $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$. Therefore we get

$$f^*(\boldsymbol{y}) = \max_{\tilde{\boldsymbol{x}}} \langle \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \rangle - g(\text{card}(\tilde{\boldsymbol{x}})) - \|\tilde{\boldsymbol{x}}\|^2. \tag{36}$$

This can equivalently be written

$$\max_k \max_{\|\tilde{\boldsymbol{x}}\|_0 = k} \langle \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \rangle - \sum_{i=1}^{k} \left(g_i + \tilde{x}_i^2 \right). \tag{37}$$

Completing squares gives

$$\max_k \max_{\|\tilde{\boldsymbol{x}}\|_0 = k} -\|\tilde{\boldsymbol{x}} - \frac{1}{2}\tilde{\boldsymbol{y}}\|^2 + \frac{1}{4}\|\tilde{\boldsymbol{y}}\|^2 - \sum_{i=1}^{k} g_i. \tag{38}$$

It is clear that the inner maximization is solved by letting $\tilde{x}_i = \frac{\tilde{y}_i}{2}$ if $i \leq k$ and $\tilde{x}_i = 0$ otherwise. After some simple manipulations this gives the conjugate function

$$f^*(\boldsymbol{y}) = \sum_{i=1}^{n} \max(\frac{1}{4}\tilde{y}_i^2 - g_i, 0). \tag{39}$$

Not that the computations for the matrix case are close to identical. In this case we maximize the scalar product $\langle X, Y \rangle$ when X and Y SVDs with the same U and V matrices (von Neumann's trace theorem), in this case $\langle X, Y \rangle = \langle \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \rangle$.

A.2. The biconjugate of f

Taking the conjugate once more gives

$$f^{**}(\boldsymbol{x}) = \max_{\boldsymbol{y}} \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \sum_{i=1}^{n} \max(\frac{1}{4}\tilde{y}_i^2 - g_i, 0). \tag{40}$$

Again the second term only depends on the elements of $\tilde{\boldsymbol{y}}$ and therefore

$$f^{**}(\boldsymbol{x}) = \max_{\tilde{\boldsymbol{y}}} \langle \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \rangle - \sum_{i=1}^{n} \max(\frac{1}{4}\tilde{y}_i^2 - g_i, 0). \tag{41}$$

For ease of notation we let $\tilde{\boldsymbol{y}} = 2\tilde{\boldsymbol{z}}$ which gives

$$f^{**}(\boldsymbol{x}) = \max_{\tilde{\boldsymbol{z}}} 2\langle \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{z}} \rangle - \sum_{i=1}^{n} \max(\tilde{z}_i^2 - g_i, 0). \tag{42}$$

The maximization over $\tilde{\boldsymbol{z}}$ does in general not have any closed form solution but has to be evaluated numerically. Note however that it is a concave maximization problem that we can solve efficiently. One exception where we can find $\tilde{\boldsymbol{z}}$ is for points $\boldsymbol{x}$ where $f^{**}(\boldsymbol{x}) = f(\boldsymbol{x})$. In what follows we will derive some properties of the maximizing $\boldsymbol{z}$ that simplifies the optimization.

We first consider the elements of $\tilde{\boldsymbol{z}}$ independently without regard for their ordering. Each one has an objective function of the form

$$c_i(\tilde{z}_i) := 2\tilde{x}_i\tilde{z}_i - \max(\tilde{z}_i^2 - g_i, 0). \tag{43}$$

Figure 5 shows $c_i(\tilde{z}_i)$ for different values of $\tilde{x}_i$. When $\tilde{x}_i \geq \sqrt{g_i} \geq 0$ there is a unique maximizing point in $\tilde{z}_i = \tilde{x}_i$. If $0 < \tilde{x}_i \leq \sqrt{g_i}$ the maximizing point is $\tilde{z}_i = \sqrt{g_i}$. In the last case where $0 = \tilde{x}_i \leq \sqrt{g_i}$ any

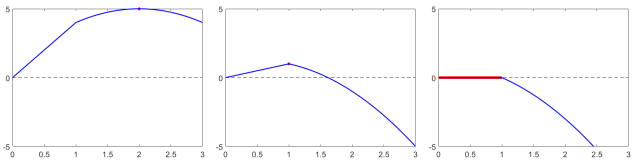

Figure 5: The objective function (43) for $g_i = 1$ and $x_i = 0, \frac{1}{2}$ and 2. The maximizing points are shown in red in each case.

$\tilde{z}_i \in [0, \sqrt{g_i}]$ is a maximizer. Suppose that we select k such that

$$\tilde{x}_i \geq \sqrt{g_i} \quad i \leq k \tag{44}$$
$$\tilde{x}_i < \sqrt{g_i} \quad i > k. \tag{45}$$
$$\tag{46}$$

Then the unconstrained minimizers u_i^* of (43) can be written

$$u_i^* \in \begin{cases} \tilde{x}_i & i \leq k \\ \sqrt{g_i} & i > k, \tilde{x}_i \neq 0 \\ [0, \sqrt{g_i}] & i > k, \tilde{x}_i = 0 \end{cases} \tag{47}$$

Before we proceed any further we note that if $\tilde{x}_i \notin (0, \sqrt{g_i})$ the second case will not occur. We can then select the elements of $\tilde{\boldsymbol{z}}$ so that each $\tilde{z}_i$ maximizes $c_i(\tilde{z}_i)$ without violating the ordering constraint.

Lemma A.2. *If $\tilde{x}_i \notin (0, \sqrt{g_i})$ then the vectors maximizing (42) are given by*

$$\tilde{z}_i^* = \begin{cases} \tilde{x}_i & i \leq k \\ s_i & i > k \end{cases}, \tag{48}$$

where s_i, $i = k+1, ..., n$ is non-increasing and $s_i \in [0, \min(\tilde{x}_k, \sqrt{g_{k+1}})]$. In this case we also have

$$f^{**}(\boldsymbol{x}) = \sum_{i=1}^{k} c_i(\tilde{x}_i) = \sum_{i=1}^{k} 2\tilde{x}_i\tilde{z}_i - \max(\tilde{x}_i^2 - g_i, 0)$$
$$= \sum_{i=1}^{k} g_i + \tilde{x}_i^2 = f(\boldsymbol{x}). \tag{49}$$

Before we proceed to the general case we note that if $\tilde{x}_i \in (0, \sqrt{g_i})$ for some i then

$$c_i(u_i^*) = \begin{cases} \tilde{x}_i^2 + g_i & i \leq k \\ 2\tilde{x}_i\sqrt{g_i} & i > k, \tilde{x}_i \neq 0 \\ 0 & i > k, \tilde{x}_i = 0 \end{cases}. \tag{50}$$

Since $2\tilde{x}_i\sqrt{g_i} < \tilde{x}_i^2 + g_i$ if $x_i < \sqrt{g_i}$ it is clear that this implies that

$$f^{**}(\boldsymbol{x}) < \sum_{i=1}^{\text{card}(\boldsymbol{x})} g_i + \tilde{x}_i^2 = f(\boldsymbol{x}). \tag{51}$$

For the general case the unconstrained minimizers are not non-increasing. To handle this we consider the best value $\tilde{z}_i^*$ given values for its neighbors $\tilde{z}_{i-1}^*$ and $\tilde{z}_{i+1}^*$. It is clear from the figures above that if the unconstrained minimizer u_i^* is unique then the best choice is of $\tilde{z}_i^*$ is

$$\tilde{z}_i^* = \begin{cases} \tilde{z}_{i+1}^* & u_i^* \leq \tilde{z}_{i+1}^* \\ u_i^* & u_i^* \in [\tilde{z}_{i+1}^*, \tilde{z}_{i-1}^*] \\ \tilde{z}_{i-1}^* & u_i^* \geq \tilde{z}_{i-1}^* \end{cases}. \tag{52}$$

Here we have adopted the convention that $\tilde{z}_0 = \infty$ and $\tilde{z}_{n+1} = 0$. In the non-unique case we similarly have that $\tilde{z}_i^* \in [0, \sqrt{g_i}] \cap [\tilde{z}_{i+1}^*, \tilde{z}_{i-1}^*]$ if this intersection is non-empty or $\tilde{z}_i^* = \tilde{z}_{i+1}^*$.

Lemma A.3. *Suppose that $\{u_i^*\}$ is not monotone for all i such that $\tilde{x}_i = 0$. Let p be defined so that the sequence $\{u_i^*\}$ is non-increasing for $i \leq p$ and non-decreasing for $i > p$, whenever $\tilde{x}_i \neq 0$. The constrained maximizers $\tilde{z}_i^*$ will then fulfill*

$$\tilde{z}_i^* \in \begin{cases} \max(u_i^*, \tilde{z}_{i+1}^*) & i \leq p \\ \tilde{z}_{i+1}^* & i > p, \tilde{x}_i \neq 0 \\ [0, \min(u_i^*, \tilde{z}_{i-1}^*)] & i > p, \tilde{x}_i = 0 \end{cases}. \tag{53}$$

Proof. We first consider $i > k$ with $\tilde{x}_i = 0$. Since $\tilde{x}_i$ is non-increasing it is clear we can make $c_i(\tilde{z}_i) = ... = c_n(\tilde{z}_n) = 0$ by letting $\tilde{z}_i = ... = \tilde{z}_n = 0$, regardless of what $\tilde{z}_{i-1}$ is. Any optimal solution therefore has to have $c_i(\tilde{z}_i) = ... = c_n(\tilde{z}_n) = 0$, which is achieved when $\tilde{z}_i^* \in [0, \min(u_i^*, \tilde{z}_{i-1}^*)]$.

Since $\tilde{z}_0^* = \infty$ we have that $u_1^* \notin [\tilde{z}_0^*, \tilde{z}_2^*]$ if and only if $\tilde{z}_2^* > u_1^*$. Therefore it is clear that $\tilde{z}_1^* = \max(u_1^*, \tilde{z}_2^*)$. Now suppose $\tilde{z}_{i-1}^* = \max(u_{i-1}^*, \tilde{z}_i^*)$ for $i \leq p$ then $\tilde{z}_{i-1}^* \geq u_{i-1}^* \geq u_i^*$, which means that either $u_i^* \in [\tilde{z}_{i+1}^*, \tilde{z}_{i-1}^*]$, in which case $\tilde{z}_i^* = u_i^*$, or $u_i^* \leq \tilde{z}_{i+1}^*$ which gives $\tilde{z}_i^* = \tilde{z}_{i+1}^*$. This proves the first case in (53).

Now suppose that $i \geq p+1$ has $\tilde{x}_i \neq 0$. If $\tilde{z}_i^* \leq u_i^*$ then $\tilde{z}_i^* \leq u_{i+1}^*$ since $\{u_i^*\}$ is not decreasing and therefore $\tilde{z}_{i+1}^* = \tilde{z}_i^*$ according to (53). If $\tilde{z}_i^* > u_i^*$ we again have $\tilde{z}_{i+1}^* = \tilde{z}_i^*$ according to (53). □

Corollary A.4. *The constrained minimizers $\tilde{z}_i^*$ can be written*

$$\tilde{z}_i^* = \begin{cases} \max(u_i^*, s) & i \leq p \\ s & i > p, \tilde{x}_i \neq 0 \\ s_i & i > p, \tilde{x}_i = 0 \end{cases}. \tag{54}$$

Here $u_p^* \leq s$, $\{s_i\}$, $i = k+1, ..., n$ *is non-increasing and* $s_i \in [0, \min(\sqrt{g_i}, s)]$.

Proof. The first two cases in (54) are fairly obvious. First it is clear that $s := \tilde{z}_p^* = \tilde{z}_i^*$, for all $i > p$ with $\tilde{x}_i \neq 0$, which is the middle case in (54). Next we see that $\tilde{z}_{p-1}^* = \max(u_{p-1}^*, \tilde{z}_k^*) = \max(u_{p-1}^*, \max(u_p^*, s)) = \max(u_{k-1}^*, s)$ since $u_{p-1}^* \geq u_p^*$. Repeating the same argument again shows the first case in (54).

Finally we note that s_i non increasing and $s_i \in [0, \min(\sqrt{g_i}, s)]$ implies that $s_i \leq s$ and therefore also that $s_i \in [0, \min(\sqrt{g_i}, s)] \cap [0, s_{i-1}] = [0, \min(\sqrt{g_i}, s_{i-1})]$, which shows the third case of (54).

To see that $u_p^* \leq s$ we note that all residuals c_i are non-decreasing with s when $s < u^*$. □

B. Proof of Theorem 3.1

In this section we give the proof of Theorem 3.1 which shows that "fixed cardinality/rank" solutions are stationary in our relaxation (7). The proofs for vector and matrix cases are somewhat different and therefore we treat them separately.

B.1. The vector case

Proof of Theorem 3.1. The objective function of (7) can be written $f^{**}(\boldsymbol{x}) + h(\boldsymbol{x})$ where $h(\boldsymbol{x}) = -\|x\|^2 + \|Ax - b\|^2$. A stationary point therefore fulfills $-\nabla h(x) \in \partial f^{**}(x)$. We have $\nabla h(\boldsymbol{x}) = -2\boldsymbol{x} + 2A^T(A\boldsymbol{x} - \boldsymbol{b})$ which yields

$$-A^T(A\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{z} - \boldsymbol{x}, \tag{55}$$

where $2\boldsymbol{z} \in \partial f^{**}(\boldsymbol{x})$. Now suppose that $\boldsymbol{o}$ fulfills the requirements of the theorem and let S be the set of nonzero elements of $\boldsymbol{o}$. The sub-differential $\partial f^{**}(\boldsymbol{o})$ consists of the maximizing $\boldsymbol{z}$-vectors given in Lemma A.2. The the vector $\boldsymbol{z} - \boldsymbol{o}$ is zero for every element in S. To see that the same is true for $A^T(A\boldsymbol{o} - \boldsymbol{b})$ we note that

$$\boldsymbol{o} = \underset{\text{supp}(\boldsymbol{x})=S}{\arg\min} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 = \underset{\text{supp}(\boldsymbol{x})=S}{\arg\min} \|A_S\boldsymbol{x} - \boldsymbol{b}\|^2, \tag{56}$$

where A_S is constructed by taking A and setting the columns no in S to zero. Therefore the normal equations $A_S^T(A_S\boldsymbol{o} - \boldsymbol{b}) = A_S^T(A\boldsymbol{o} - \boldsymbol{b}) = 0$ hold which shows that the elements of $A^T(A\boldsymbol{o} - \boldsymbol{b})$ that are in S all vanish.

It now remains to show that the elements in the complement of S are smaller than $\min\{\tilde{x}_k, \sqrt{g_{k+1}}\}$. This is however clear since by assumption

$$\|A^T(A^T\boldsymbol{o} - b)\| \leq \|A\|\|\boldsymbol{\epsilon}\| \leq \min\{\tilde{x}_k, \sqrt{g_{k+1}}\}. \tag{57}$$

We remark that estimating the size of the elements by the vector norm is a simple but very crude estimation and the result is therefore likely to hold under much more generous conditions. □

B.2. The matrix case

Proof of Theorem 3.1. Similar to the vector case we need to show that

$$-\mathcal{A}^*(\mathcal{A}O - \boldsymbol{b}) = Z - O, \tag{58}$$

where $2Z \in \partial f^{**}(O)$. The matrix Z is in the sub differential of $f^{**}(O)$ if we can find orthogonal matrices U and V such that $O = UD_{\tilde{\boldsymbol{o}}}V^T$ and $Z = UD_{\tilde{\boldsymbol{z}}}V^T$. Here $\tilde{\boldsymbol{o}}$ and $\tilde{\boldsymbol{z}}$ are the singular values of O and Z respectively. The matrices $D_{\tilde{\boldsymbol{o}}}$ and $D_{\tilde{\boldsymbol{z}}}$ are diagonal matrices with elements $\tilde{\boldsymbol{o}}$ and $\tilde{\boldsymbol{z}}$. Note that O is typically of low rank $D_{\tilde{\boldsymbol{o}}}$ and $D_{\boldsymbol{z}}$ can be partitioned into block matrices

$$D_{\tilde{\boldsymbol{o}}} = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \text{ and } D_{\tilde{\boldsymbol{z}}} = \begin{bmatrix} \Sigma & 0 \\ 0 & \Delta \end{bmatrix}. \tag{59}$$

Here Σ contains the k non-zero singular values of O. Due to Lemma A.2 the $D_{\tilde{\boldsymbol{z}}}$ also contains this block. The matrix Δ contains the singular values of $\tilde{\boldsymbol{z}}$ that correspond to zeros in $\tilde{\boldsymbol{o}}$. We can make a corresponding partition of the U and V matrices into

$$U = \begin{bmatrix} \bar{U} & \bar{U}_\perp \end{bmatrix} \text{ and } V = \begin{bmatrix} \bar{V} & \bar{V}_\perp \end{bmatrix}, \tag{60}$$

where $\bar{U}$ and $\bar{V}$ are the first k columns of U and V respectively. Note that only $\bar{U}$ and $\bar{V}$ are uniquely determined by O. The matrices $\bar{U}_\perp$ and $\bar{V}_\perp$ can be selected arbitrarily as long as they are orthogonal to $\bar{U}$ and $\bar{V}$ respectively. Any choice of $\bar{U}$, $\bar{V}$ and Δ, where the elements of Δ are less than $\min\{\tilde{o}_k, \sqrt{g_{k+1}}\}$ gives us a Z that is in the sub differential. Consequently we have

$$Z - O = \begin{bmatrix} \bar{U} & \bar{U}_\perp \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \Delta \end{bmatrix} \begin{bmatrix} \bar{V}^T \\ \bar{V}_\perp^T \end{bmatrix} = \bar{U}_\perp \Delta \bar{V}_\perp^T. \tag{61}$$

□

We now consider term $\mathcal{A}^*(\mathcal{A}O - \boldsymbol{b})$. We have

$$\begin{aligned} \|\mathcal{A}(O + tH) - \boldsymbol{b}\|^2 =& t^2\|\mathcal{A}H\|^2 + 2t\langle H, \mathcal{A}^*(\mathcal{A}O - \boldsymbol{b})\rangle \\ &+ \|\mathcal{A}O - \boldsymbol{b}\|^2. \end{aligned} \tag{62}$$

Recall that O minimizes the left hand side over all matrices with rank at most k. Since the linear term dominates the quadratic one for small t we must have

$$\langle H, \mathcal{A}^*(\mathcal{A}O - \boldsymbol{b})\rangle \geq 0 \tag{63}$$

for all H such that $\text{rank}(O + tH) \leq k$. Since Σ has full rank it is clear that any matrix of the form

$$H = \begin{bmatrix} \bar{U} & \bar{U}_\perp \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & 0 \end{bmatrix} \begin{bmatrix} \bar{V}^T \\ \bar{V}_\perp^T \end{bmatrix} \tag{64}$$

fulfills this requirement. It is now easy to see that

$$-\mathcal{A}^*(\mathcal{A}O-\boldsymbol{b})=U_\perp M V_\perp^T, \tag{65}$$

where M is some matrix. Furthermore since $U_\perp$ and $V_\perp$ can be selected freely (as long as they are perpendicular to U and V respectively) we can assume that M is diagonal. What remains is therefore to estimate its singular values, which similarly to the vector case is done by

$$\|\mathcal{A}^*(\mathcal{A}O-\boldsymbol{b})\|_2 \leq \|\mathcal{A}\|\|\boldsymbol{\epsilon}\| \leq \min\{\tilde{o}_k, \sqrt{g_{k+1}}\}. \tag{66}$$

C. Proof of Theorem 3.2

In this section we prove our main theorem. The proof requires a growth estimate of the subgradients of f^{**} which we give in the following lemmas.

Lemma C.1. *If $\boldsymbol{z}\in\partial f^{**}(\boldsymbol{x})$ and $\boldsymbol{z}'\in\partial f^{**}(\boldsymbol{x}')$ and $d\leq 1$ then*

$$\langle \boldsymbol{z}'-\boldsymbol{z}, \boldsymbol{x}'-\boldsymbol{x}\rangle > d\|\boldsymbol{x}'-\boldsymbol{x}\|^2, \tag{67}$$

if

$$\langle \pi\tilde{\boldsymbol{z}}'-\tilde{\boldsymbol{z}}, \pi\tilde{\boldsymbol{x}}'-\tilde{\boldsymbol{x}}\rangle > d\|\pi\tilde{\boldsymbol{x}}'-\tilde{\boldsymbol{x}}\|^2, \tag{68}$$

for all permutation matrices π.

Proof. We have that (67) can be written

$$C-\langle \boldsymbol{z}'-d\boldsymbol{x}', \boldsymbol{x}\rangle - \langle \boldsymbol{z}-d\boldsymbol{x}, \boldsymbol{x}'\rangle > 0, \tag{69}$$

where

$$C=\langle \boldsymbol{z}'-d\boldsymbol{x}', \boldsymbol{x}'\rangle + \langle \boldsymbol{z}-d\boldsymbol{x}, \boldsymbol{x}\rangle. \tag{70}$$

Note that since the elements of $\boldsymbol{z}'$ and $\boldsymbol{x}'$ have the same signs and $\tilde{z}_i'\geq\tilde{x}_i'$ for all i the term C is independent of signs. For fixed magnitudes and permutations the term $\langle \boldsymbol{z}'-d\boldsymbol{x}', \boldsymbol{x}\rangle + \langle \boldsymbol{z}-d\boldsymbol{x}, \boldsymbol{x}'\rangle$ is clearly maximized when $\boldsymbol{z}'$ and $\boldsymbol{z}$ have the same signs. In which case we have

$$\langle \boldsymbol{z}'-\boldsymbol{z}, \boldsymbol{x}'-\boldsymbol{x}\rangle = \langle \pi\tilde{\boldsymbol{z}}'-\tilde{\boldsymbol{z}}, \pi\tilde{\boldsymbol{x}}'-\tilde{\boldsymbol{x}}\rangle \tag{71}$$

and $\|\boldsymbol{x}'-\boldsymbol{x}\|^2 = \|\pi\tilde{\boldsymbol{x}}'-\tilde{\boldsymbol{x}}\|^2$ for some permutation π. □

In the matrix case we have $Z\in\partial f^{**}(X)$ and $Z'\in\partial f^{**}(X')$. Recall that here $\tilde{\boldsymbol{x}}$, $\tilde{\boldsymbol{z}}$, $\tilde{\boldsymbol{x}}'$, $\tilde{\boldsymbol{z}}'$, are the singular values of the matrices X,Z,X',Z' respectively. The corresponding statement is then that

$$\langle Z'-Z, X'-X\rangle > d\|X'-X\|^2 \tag{72}$$

holds whenever (69) holds. The proof is however more complicated than the vector case. We therefore refer the reader to Proposition 4.5 [11] from which it is clear that the above statement holds.

We are now ready to establish the growth estimates on the directional derivatives needed to prove Theorem 3.2. We will first consider directional derivatives between points where the relaxation is tight, that is $f_g(\boldsymbol{x}) = f_g^{**}(\boldsymbol{x})$. In the subsequent result we then relax this assumption to only be valid for one of the points (namely the stationary point we want to prove is unique).

Lemma C.2. *Suppose that $2\boldsymbol{z}\in\partial f^{**}(\boldsymbol{x})$ and $2\boldsymbol{z}'\in\partial f^{**}(\boldsymbol{x}')$, and that neither $\tilde{\boldsymbol{x}}$ nor $\tilde{\boldsymbol{x}}'$ have values in $(0,\sqrt{g_i})$. If the elements of $\tilde{\boldsymbol{z}}$ fulfill*

$$\tilde{z}_i \notin \left[(1-d)\sqrt{g_k}, \frac{\sqrt{g_k}}{(1-d)}\right] \text{ and } \tilde{z}_{k+1} < (1-2d)\tilde{z}_k, \tag{73}$$

where k is defined so that $\tilde{x}_i\geq\sqrt{g_i}$ for $i\leq k$ and $\tilde{x}_i=0$ if $i>k$, then

$$\langle \boldsymbol{z}'-\boldsymbol{z}, \boldsymbol{x}'-\boldsymbol{x}\rangle > d\|\boldsymbol{x}'-\boldsymbol{x}\|^2. \tag{74}$$

Proof. We need show that

$$\langle \pi\tilde{\boldsymbol{z}}'-\tilde{\boldsymbol{z}}, \pi\tilde{\boldsymbol{x}}'-\tilde{\boldsymbol{x}}\rangle > d\|\pi\tilde{\boldsymbol{x}}'-\tilde{\boldsymbol{x}}\|^2, \tag{75}$$

where π is a permutation. For ease of notation let $\boldsymbol{z}'=\pi\tilde{\boldsymbol{z}}'$ and $\boldsymbol{x}'=\pi\tilde{\boldsymbol{x}}'$. We let the $\tilde{I}=\{i;\ \tilde{x}_i\neq 0\}=\{i; i\leq k\}$ and $I'=\{i;\ x_i'\neq 0\}$. Then

$$\begin{aligned}\langle \boldsymbol{z}'-\tilde{\boldsymbol{z}}, \boldsymbol{x}'-\tilde{\boldsymbol{x}}\rangle = &\sum_{i\in\tilde{I}, i\in I'}(x_i'-\tilde{x}_i)^2 + \sum_{i\in\tilde{I}, i\notin I'}\tilde{x}_i(\tilde{x}_i-z_i')\\ &+\sum_{i\notin\tilde{I}, i\in I'} x_i'(x_i'-\tilde{z}_i).\end{aligned} \tag{76}$$

Note that

$$d\|\boldsymbol{x}'-\tilde{\boldsymbol{x}}\|^2 = \sum_{i\in\tilde{I}, i\in I'} d(x_i'-\tilde{x}_i)^2 + \sum_{i\in\tilde{I}, i\notin I'} d\tilde{x}_i^2 + \sum_{i\notin\tilde{I}, i\in I'} d{x_i'}^2. \tag{77}$$

We first consider pairs of terms from the second and third sums of (76). If $i\in\tilde{I}, i\notin I'$ and $j\notin\tilde{I}, j\in I'$ we have

$$\tilde{x}_i(\tilde{x}_i-z_i') + x_j'(x_j'-\tilde{z}_j) = \tilde{x}_i^2 + {x_j'}^2 - \tilde{x}_i z_i' - x_j'\tilde{z}_j. \tag{78}$$

Since $j\notin\tilde{I}$ and $i\in\tilde{I}$ we have $\tilde{z}_j<(1-2d)\tilde{z}_k\leq(1-2d)\tilde{z}_i=(1-2d)\tilde{x}_i$. Similarly, since $j\in I'$ and $i\notin I'$ we have $z_i'\leq z_j'=x_j'$. Therefore

$$\tilde{x}_i z_i' \leq \tilde{x}_i x_j' \leq \frac{\tilde{x}_i^2+{x_i'}^2}{2} \tag{79}$$

$$\tilde{x}_j z_j < (1-2d)\tilde{x}_j x_i \leq (1-2d)\frac{x_i^2+\tilde{x}_i^2}{2} \tag{80}$$

which gives

$$\tilde{x}_i(\tilde{x}_i - z'_i) + x'_j(x'_j - \tilde{z}_j) < d(\tilde{x}_i^2 + {x'_i}^2). \tag{81}$$

If the number of elements in $\tilde{I}$ and I' are the same the two last sums of (76) have the same number of terms. Then (81) shows that (76) larger is than (77) since clearly $(x'_i - \tilde{x}_i)^2 > d(x'_i - \tilde{x}_i)^2$. It therefore remains to consider the two cases when I has more elements than I' and vice versa.

Let k' be the number of elements in I'. Suppose first that $\tilde{I}$ has more elements than I', that is, $k > k'$. Then the middle sums of (76) and (77) have more terms than the third ones. Therefore we need to show that

$$\tilde{x}_i(\tilde{x}_i - z'_i) > d\tilde{x}_i^2, \tag{82}$$

for $i \in \tilde{I}$ and $i \notin I'$. Suppose that $\pi_{ij} = 1$, that is element j of $\tilde{\boldsymbol{z}}'$ is moved to element i of $\boldsymbol{z}'$ by the permutation π. By Corollary A.4 we have that $z'_i = \tilde{z}'_j \leq \tilde{x}'_{k'} \leq \sqrt{g_{k'}} \leq \sqrt{g_k}$. Since $i < k$ we have by assumption (73) that $\tilde{x}_i = \tilde{z}_i > \frac{\sqrt{g_k}}{(1-d)}$. Therefore

$$\tilde{x}_i - z'_i = (1-d)\tilde{x}_i + d\tilde{x}_i - z'_i > \sqrt{g_k} + d\tilde{x}_i - z'_i \geq d\tilde{x}_i, \tag{83}$$

which gives (82).

Now suppose instead that $\tilde{I}$ has fewer elements than I', that is, $k' > k$. Then we need to show that

$$x'_i(x'_i - \tilde{z}_i) \geq d{x'_i}^2, \tag{84}$$

for $i \notin \tilde{I}$ and $i \in I'$. Suppose again that $\pi_{ij} = 1$, that is element j of $\tilde{\boldsymbol{z}}'$ is moved to element i of $\boldsymbol{z}'$ by the permutation π. By Corollary A.4 we have $x'_i = \tilde{z}'_j > \tilde{z}'_{k'} = \tilde{x}'_{k'} \geq \sqrt{g_{k'}}$. Furthermore by assumption (73) we have $\tilde{z}_i < (1-d)\sqrt{g_k} \leq (1-d)\sqrt{g_{k'}}$ since $k < k'$. Therefore

$$x'_i - \tilde{z}_i = (1-d)x'_i + dx'_i - \tilde{z}_i \geq (1-d)\sqrt{g_{k'}} + dx'_i - \tilde{z}_i > dx'_i. \tag{85}$$

which gives (84). □

Lemma C.3. *Suppose that $\boldsymbol{x}$ fulfills the assumptions of Lemma C.2. If $2\boldsymbol{z}' \in \partial f^{**}(\boldsymbol{x}')$ (without any additional assumptions on the values of $\boldsymbol{x}'$ or $\boldsymbol{z}'$) then (74) holds.*

Proof. By Lemma A.1 we have $f^{**}(\boldsymbol{x}') = \sum_j \lambda_j f^{**}(\boldsymbol{x}^j)$, where $\boldsymbol{x}^j$ are points where $f^{**}(\boldsymbol{x}^j) = f(\boldsymbol{x}^j)$, that is $\boldsymbol{x}^j$ has no elements in $(0, \sqrt{g_i})$. Then by Lemma C.2, for any $\boldsymbol{z}' \in \partial f^{**}(\boldsymbol{x}') \subset \bigcap_j \partial f^{**}(\boldsymbol{x}^j)$ we have

$$\langle \boldsymbol{z}' - \boldsymbol{z}, \boldsymbol{x}^j - \boldsymbol{x} \rangle > d\|\boldsymbol{x}^j - \boldsymbol{x}\|^2. \tag{86}$$

By convexity of $\|(\cdot) - X\|^2$ we now get

$$\begin{aligned} \langle \boldsymbol{z}' - \boldsymbol{z}, \boldsymbol{x}' - \boldsymbol{x} \rangle &> d \sum_j \lambda_j \|\boldsymbol{x}^i - \boldsymbol{x}\|^2 \\ &\geq d\| \sum_j \lambda_j \boldsymbol{x}^j - \boldsymbol{x}\|^2 = d\|\boldsymbol{x}' - \boldsymbol{x}\|^2. \end{aligned} \tag{87}$$

□

Proof of Theorem 3.2. We will show that $\nabla h(\boldsymbol{x}') = 2(I - A^T A)\boldsymbol{x}' + 2A^T b \notin \partial f^{**}(\boldsymbol{x}')$. Suppose that $2\boldsymbol{z}' \in \partial f^{**}(\boldsymbol{x}')$. Since $\boldsymbol{x}$ is stationary we have $2\boldsymbol{z} + \nabla h(\boldsymbol{x}) = 0$

$$\langle \boldsymbol{z}' + \nabla h(\boldsymbol{x}'), \boldsymbol{x}' - \boldsymbol{x} \rangle = \langle \boldsymbol{z}' - \boldsymbol{z}, \boldsymbol{x}' - \boldsymbol{x} \rangle + \langle \nabla h(\boldsymbol{x}') - \nabla h(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x} \rangle. \tag{88}$$

For the second term we have

$$\begin{aligned} \langle \nabla h(\boldsymbol{x}') - \nabla h(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x} \rangle &= \|\boldsymbol{x}' - \boldsymbol{x}\|_F^2 - \|A(\boldsymbol{x}' - \boldsymbol{x})\|^2 \\ &\leq \delta_r \|\boldsymbol{x} - \boldsymbol{x}'\|^2, \end{aligned} \tag{89}$$

if $\operatorname{rank}(\boldsymbol{x} - \boldsymbol{x}') \leq r$, which clearly holds if $\operatorname{card}(\boldsymbol{x}') \leq r - k$. On the other hand we also have by Lemma C.2 that

$$\langle \boldsymbol{z}' - \boldsymbol{z}, \boldsymbol{x}' - \boldsymbol{x} \rangle > \delta_r \|\boldsymbol{x}' - \boldsymbol{x}\|^2, \tag{90}$$

and therefore (88) is positive and $\boldsymbol{x}'$ cannot be a stationary point.

Suppose now that $\boldsymbol{x}$ is a point that has $\operatorname{card}(\boldsymbol{x}) < \frac{r}{2}$. We will consider the directional derivatives along the line $\boldsymbol{x} + t\boldsymbol{v}$, where $\boldsymbol{v} = \frac{\boldsymbol{x}' - \boldsymbol{x}}{\|\boldsymbol{x}' - \boldsymbol{x}\|_F}$. Since f^{**} is convex (and finite) the directional derivative of the objective function exists and is given by

$$\sup_{\boldsymbol{z}' \in \partial f^{**}(\boldsymbol{x} + t\boldsymbol{v})} \langle \boldsymbol{z}' + \nabla h(\boldsymbol{x} + t\boldsymbol{v}), \boldsymbol{v} \rangle. \tag{91}$$

Since the $\operatorname{card}(\boldsymbol{v}) \leq r$ it is clear by the arguments above that this is positive. □

D. Proof of Theorem 3.3

Proof. We will let $\boldsymbol{x}$ be a global solution to $\min_{\operatorname{card}(\boldsymbol{x}) \leq k} \|A\boldsymbol{x} - \boldsymbol{b}\|$ and show that this point will be stationary under the conditions above. To do this we need to show that $2\boldsymbol{z} \in \partial f(\boldsymbol{x})$ for $\boldsymbol{z} = (I - A^T A)\boldsymbol{x} + A^T \boldsymbol{b}$. We first note that since $\|A\| < 1$ the vector $\boldsymbol{x}$ will be the global minimizer of (7) for the fixed-cardinality relaxation, that is, the special case $g_i = 0$ if $i \leq k$ and $g_i = \infty$ if $i > k$. This shows that $\boldsymbol{x}$ is stationary in (7) for this particular choice of g. In particular $\boldsymbol{x} = D_s \pi \tilde{x}$ and $\boldsymbol{z} = D_s \pi \tilde{z}$ with the same s and π. (In the matrix case the corresponding statement is that the SVD's of X and Z have the same U and V matrices.) Furthermore, since $\operatorname{card}(\boldsymbol{x}) \leq k$ and $g_i = 0$ when $i \leq k$ it is clear from Lemma A.2 that the $\tilde{z}_i = \tilde{x}_i$ for $i \leq k$.

To show that $\boldsymbol{x}$ is stationary for a general choice of g fulfilling (18) it is enough to show that $\sqrt{g_i} \leq \tilde{x}_i$ for $i \leq k$ and $\sqrt{g_i} \geq \tilde{x}_k$ for $i > k$ by Lemma A.2. This is however implied by the stricter constraints (14) and we therefore proceed by proving these directly.

First we show that $\tilde{x}_i$ is close to $\tilde{y}_i$. Since $\|A\boldsymbol{x}-\boldsymbol{b}\| \leq \|A\boldsymbol{y}-\boldsymbol{b}\| = \|\boldsymbol{\epsilon}\|$ we have

$$\begin{aligned}\sqrt{1-\delta_{2k}}\|\boldsymbol{x}-\boldsymbol{y}\| &\leq \|A(\boldsymbol{x}-\boldsymbol{y})\| \\ &\leq \|A\boldsymbol{x}-\boldsymbol{b}\| + \|A\boldsymbol{y}-\boldsymbol{b}\| \\ &\leq 2\|\boldsymbol{\epsilon}\|.\end{aligned} \tag{92}$$

Therefore

$$|x_i - y_i| \leq \frac{2}{\sqrt{1-\delta_{2k}}}\|\boldsymbol{\epsilon}\|. \tag{93}$$

Furthermore

$$\begin{aligned}\|\boldsymbol{z}-\boldsymbol{x}\| &= \|A^T A(\boldsymbol{x}-\boldsymbol{y}) - A^T\boldsymbol{\epsilon}\| \\ &\leq \|A^T\|\|A\|\|\boldsymbol{x}-\boldsymbol{y}\| + \|A^T\|\|\boldsymbol{\epsilon}\| \\ &\leq \|\boldsymbol{x}-\boldsymbol{y}\| + \|\boldsymbol{\epsilon}\| \\ &\leq \frac{3}{\sqrt{1-\delta_{2k}}}\|\boldsymbol{\epsilon}\|.\end{aligned} \tag{94}$$

And since $\tilde{x}_{k+1} = 0$ this means that

$$\tilde{z}_{k+1} \leq \frac{3}{\sqrt{1-\delta_{2k}}}\|\boldsymbol{\epsilon}\|. \tag{95}$$

Now inserting the above estimates in $\boldsymbol{z}_{k+1} < (1-2\delta_k)\boldsymbol{z}_k$ shows (after some simplification) that this constraint holds if

$$\tilde{y}_k > \frac{5-4\delta_k}{\sqrt{1-\delta_{2k}}(1-2\delta_{2k})}\|\boldsymbol{\epsilon}\|, \tag{96}$$

which is implied by (17) since $\delta_k \geq \delta_{2k} > 0$. Furthermore since $\sqrt{g}_i$ is non-decreasing (18) and (93) implies that $\tilde{z}_i = \tilde{x}_i > \frac{\sqrt{g_k}}{1-\delta_k}$ for $i \leq k$, while (18) and (95) implies that $\boldsymbol{z}_i < (1-\delta_k)\sqrt{g_i}$ for $i > k$. □