MLISP: Machine Learning in Signal Processing

Lectures 15, 16, and 17

Prof. V. I. Morgenshtern

Scribe: M. Elminshawi

Agenda:

1. The basic theorem in compressed sensing
2. Optimality conditions
3. Generalizing the gradient
4. The dual certificate
5. Construction of the dual certificate

1 The basic theorem in compressed sensing

Theorem 1 (Main theorem)**.** *Let $\mathbf{x}_0 \in \mathbb{R}^n$ be an s-sparse vector. Let $\mathbf{A}$ be an $m \times n$ random matrix with $A_{ij} \sim \mathcal{N}(0, \frac{1}{m})$, $m \geq 9s\log(n)$. Let $\mathbf{b} = \mathbf{A}\mathbf{x}_0$.*

Then $\mathbf{x}_0$ is the unique solution of

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{x}\|_1 \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \end{aligned}$$

with probability at least $1 - 3/n$.

The meaning of this theorem: If the number of measurements m is slightly larger that the information content of the signal, s, then we can recover $\mathbf{x}_0$ exactly from $\mathbf{b} = \mathbf{A}\mathbf{x}_0$ via linear programming with overwhelmingly high probability.

2 Optimality conditions

Consider the optimization problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \end{aligned} \tag{1}$$

Let $\mathbf{x}_*$ denote the unique optimal solution of this optimization problem, i.e., the solution to (1). What are the conditions that $\mathbf{x}_*$ should satisfy?

Note that

$$\{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}\} = \{\mathbf{x}_* + \mathbf{h} : \mathbf{h} \in \mathcal{N}(\mathbf{A})\}.$$

Therefore, $\mathbf{x}_*$ solves (1) iff

$$f(\mathbf{x}_* + \mathbf{h}) \geq f(\mathbf{x}_*) \text{ for all } \mathbf{h} \in \mathcal{N}(\mathbf{A})$$

where $\mathcal{N}(\cdot)$ denotes the null space.

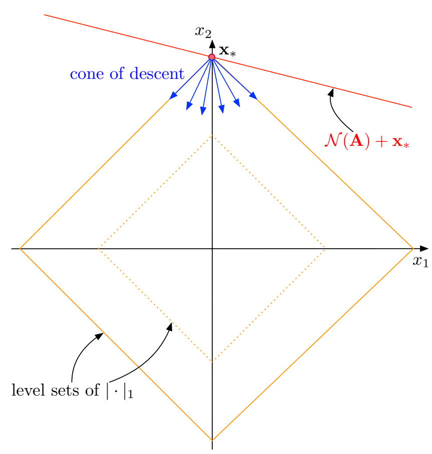
Define the cone of descent directions of $f(\cdot)$ at $\mathbf{x}_*$:

$$\mathcal{D} = \{\mathbf{d} : f(\mathbf{x}_* + \alpha\mathbf{d}) < f(\mathbf{x}_*) \text{ for some } \alpha > 0\}.$$

For

$$f(\mathbf{x}) = f(x_1, x_2) = \|\mathbf{x}\|_1 = |x_1| + |x_2|$$

this cone looks like this:



Since $\mathbf{x}_*$ is the optimum of (1), we must have:

$$\mathcal{D} \cap \mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}. \tag{2}$$

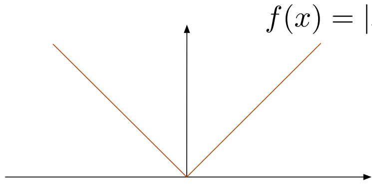
Indeed, if (2) is not satisfied, take $\mathbf{d} \neq \mathbf{0} \in \mathcal{D} \cap \mathcal{N}(\mathbf{A})$ and note that $f(\mathbf{x}_* + \alpha\mathbf{d}) < f(\mathbf{x}_*)$ and $\mathbf{A}(\mathbf{x}_* + \alpha\mathbf{d}) = \mathbf{A}\mathbf{x}_* = \mathbf{b}$. Therefore $\mathbf{x}_* + \alpha\mathbf{d}$ satisfies the constraints and makes the objective smaller. Contradiction with the assumption that $\mathbf{x}_*$ is optimum of (1).

How can we guarantee that (2) is satisfied? To express condition (2) in a more convenient form, we need to study subgradients.

3 Generalizing the gradient

Note that the $\ell 1$-norm, $f(\mathbf{x}) = \|\mathbf{x}\|_1$ is not differentiable everywhere: there is no gradient at the intersections with the axes.

For example, in 1D $f(x)$ is not differentiable for $x = 0$:

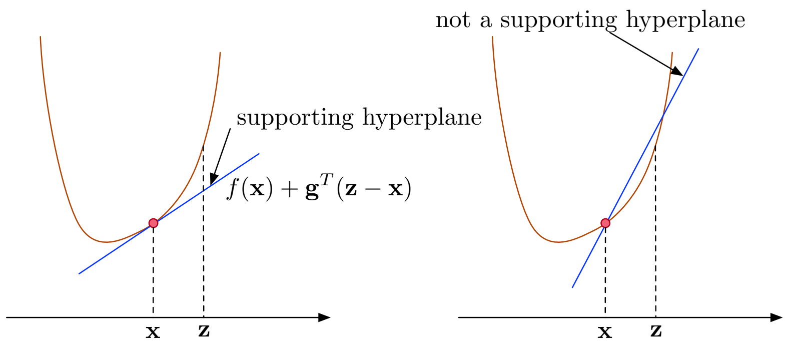


However $f(\mathbf{x}) = \|\mathbf{x}\|_1$ is convex, which allows us to define the generalized version of the gradient.

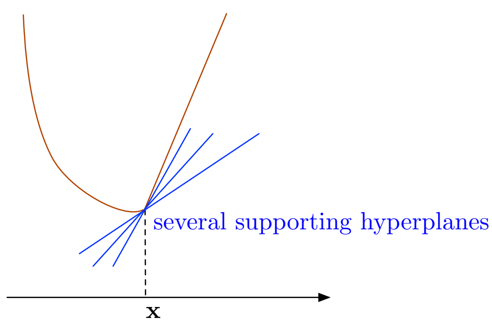
The subgradient: Vector $\mathbf{g}$ is a *subgradient* of $f(\cdot)$ at $\mathbf{x}$ if:

$$f(\mathbf{z}) \geq f(\mathbf{x}) + \mathbf{g}^\mathsf{T}(\mathbf{z} - \mathbf{x}).$$

For example if the function is smooth, the subgradient is unique and it is equal to the gradient:



If the function has a kink, there are infinitely many subradients:



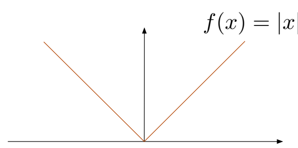
The subdifferential: The set of subgradients of $f(\cdot)$ at $\mathbf{x}$ is called the *subdifferential*:

$$\partial f(\mathbf{x}) = \{\mathbf{g} : \mathbf{g} \text{ is subgradients of } f \text{ at } \mathbf{x}\}.$$

If f is differentiable at $\mathbf{x}$, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

For example:

$$\partial|x| = \begin{cases} \{1\}, & x > 0 \\ \{-1\}, & x < 0 \\ [-1, 1], & x = 0 \end{cases}$$



Properties of the subdifferential:

1. $\partial\{af(\mathbf{x})\} = a\partial f(\mathbf{x})$

2. $\partial\left(\sum_{i=1}^{T} f_i(\mathbf{x})\right) = \sum_{i=1}^{T} \partial f_i(\mathbf{x})$ where $\sum$ is the Minkowski sum of sets.
   For example, if $\mathcal{S}_1 = \{a, b\}$ and $\mathcal{S}_2 = \{c, d, e\}$, then
   $$\mathcal{S}_1 + \mathcal{S}_2 = \{a + c, a + d, a + e, b + c, b + d, b + e\}.$$

3. If
   $$f(\mathbf{x}) = \max_{i \in I} \left(\mathbf{a}_i^\mathsf{T}\mathbf{x} + b_i\right)$$
   then
   $$\partial f(\mathbf{x}) = \text{conv}\{\mathbf{a}_i : f(\mathbf{x}) = \mathbf{a}_i^\mathsf{T}\mathbf{x} + b_i\}$$
   where conv$(\cdot)$ denotes the convex hull.
   For example if $\mathcal{A} = \{\mathbf{a}, \mathbf{b}\}$,
   $$\text{conv}(\mathcal{A}) = \{\theta\mathbf{a} + (1 - \theta)\mathbf{b} : 0 \leq \theta \leq 1\}.$$
   To illustrate this property, let $f(x) = |x| = \max(x, -x)$. Consider $x = 0$, then $f(x) = x$ and $f(x) = -x$. Therefore, $\partial f(x) = \text{conv}\{+1, -1\} = [-1, 1]$.

These 3 rules allow us to calculate most subdifferentials. Let's calculate the subdifferential of the $\ell 1$-norm.

Assume that

$$\begin{cases} x_i \neq 0, & i \in \mathcal{T} \\ x_i = 0, & i \in \mathcal{T}^\mathsf{c}. \end{cases}$$

where here and below $\mathcal{T}^\mathsf{c}$ denotes the complement of $\mathcal{T}$. Then $\mathbf{v} \in \partial \|\mathbf{x}\|_1$ iff

$$\begin{cases} v_i = \text{sign}(x_i), & i \in \mathcal{T} \\ v_i \in [-1, 1], & i \in \mathcal{T}^\mathsf{c}. \end{cases} \tag{3}$$

This can be seen as follows:

$$\|\mathbf{x}\|_1 = \sum_i \underbrace{|x_i|}_{f_i(\mathbf{x})}.$$

Therefore, when $x_i = 0$:

$$f_i(\mathbf{x}) = \max([0\ \dots\ 0\ 1\ 0\ \dots\ 0]\mathbf{x}, [0\ \dots\ 0\ -1\ 0\ \dots\ 0]\mathbf{x})$$

and

$$\partial f_i(\mathbf{x}) = \text{conv}\{[0\ \dots\ 0\ 1\ 0\ \dots\ 0], [0\ \dots\ 0\ -1\ 0\ \dots\ 0]\}.$$

From this (3) follows via property 2 of the subdifferential.

4 The dual certificate

Lemma 2. *Assume $f(\cdot)$ is convex. Then,* $\mathbf{x}$ *minimizes* $f(\cdot)$ *iff* $\mathbf{0} \in \partial f(\mathbf{x})$.

Proof. Let's prove the lemma in one direction.

Take any other $\mathbf{z}$, then by definition of subdifferential,

$$f(\mathbf{z}) \geq f(\mathbf{x}) + \mathbf{g}^\mathsf{T}(\mathbf{z} - \mathbf{x}) \text{ for every } \mathbf{g} \in \partial f(\mathbf{x}).$$

Since $\mathbf{0} \in \partial f(\mathbf{x})$ is a subgradient, the inequality is true for $\mathbf{g} = \mathbf{0}$. Hence, $f(\mathbf{z}) \geq f(\mathbf{x})$ for all $\mathbf{z}$. □

Lemma 3. *Assume $f(\cdot)$ is convex. Then,* $\mathbf{x}$ *minimizes* $f(\cdot)$ *over the affine set* $\{\mathbf{z} : \mathbf{A}\mathbf{z} = \mathbf{b}\}$ *iff there exists* $\boldsymbol{\lambda}$ *such that* $\mathbf{A}^\mathsf{T}\boldsymbol{\lambda} \in \partial f(\mathbf{x})$ *and* $\mathbf{A}\mathbf{x} = \mathbf{b}$. *The vector* $\boldsymbol{\lambda}$ *is called the* dual certificate.

Proof. Let's proof the lemma in one direction.

Every element from the affine set $\{\mathbf{z} : \mathbf{A}\mathbf{z} = \mathbf{b}\}$ can be written as $\mathbf{z} = \mathbf{x} + \mathbf{h}$ with $\mathbf{h} \in \mathcal{N}(\mathbf{A})$.

Since $\mathbf{A}^\mathsf{T}\boldsymbol{\lambda} \in \partial f(\mathbf{x})$, we have

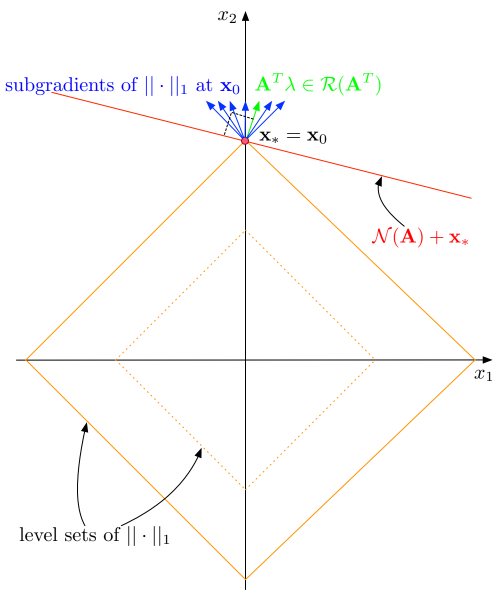
$$\begin{aligned} f(\mathbf{x} + \mathbf{h}) &\geq f(\mathbf{x}) + (\mathbf{A}^\mathsf{T}\boldsymbol{\lambda})^\mathsf{T}\mathbf{h} \\ &= f(\mathbf{x}) + \boldsymbol{\lambda}^\mathsf{T} \underbrace{\mathbf{A}\mathbf{h}}_{\mathbf{0}} \\ &= f(\mathbf{x}). \end{aligned}$$

Therefore, $\mathbf{x}$ minimizes $f(\cdot)$ over the affine set $\{\mathbf{z} : \mathbf{A}\mathbf{z} = \mathbf{b}\}$, as required. □

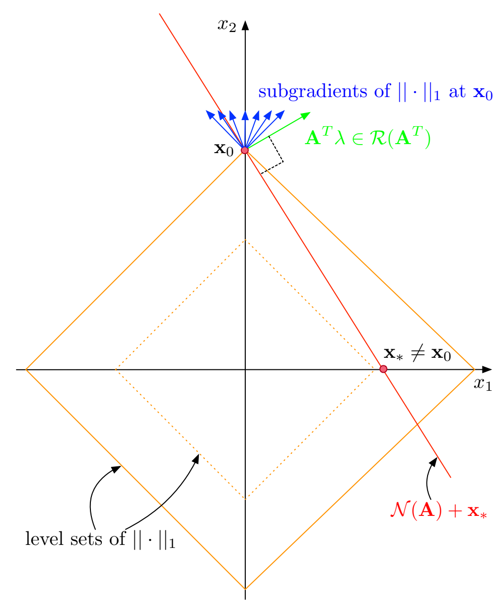
In the figure below we take $\mathbf{A} = [a_1\ a_2]$ so that we have one equation

$$[a_1\ a_2] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = b.$$

You can see the dual certificate in greed (the range $\mathcal{R}(\mathbf{A}^T)$ is orthogonal to $\mathcal{N}(\mathbf{A})$):



so that $\mathbf{x}$ is the optimal point. Here it is clear that dual certificated does not exist, because $\mathbf{A}^\mathsf{T}\boldsymbol{\lambda}$ does not belong to $\partial\|\mathbf{x}\|_1$ for any $\boldsymbol{\lambda}$:



so that $\mathbf{x}$ is not the optimal point.

Let's return to the proof of Theorem 1: the measurements are given by $\mathbf{b} = \mathbf{A}\mathbf{x}_0$, where $\mathbf{x}_0$ is the true signal we need to reconstruct. Our strategy in proving the main theorem is to explicitly construct the dual certificate $\boldsymbol{\lambda}$ such that $\mathbf{A}^\mathsf{T}\boldsymbol{\lambda} = \mathbf{v} \in \partial\|\mathbf{x}_0\|_1$, which, according to (3), is equivalent to

$$v_i = \begin{cases} \text{sign}([\mathbf{x}_0]_i), & [\mathbf{x}_0]_i \neq 0 \\ (-1, 1), & [\mathbf{x}_0]_i = 0. \end{cases}$$

Let's make the notation more compact. Let $\mathcal{T}$ denote the support of $\mathbf{x}_0$:

$$\mathcal{T} = \{i : [\mathbf{x}_0]_i \neq 0\}.$$

Let $\mathbf{e} \in \mathbb{R}^s$ be the vector of signs of $\mathbf{x}_0$ on its support $\mathcal{T}$: $\mathbf{e} = [\text{sign}(\mathbf{x}_0)]_{\mathcal{T}}$. With this notation our task is to find $\boldsymbol{\lambda}$ such that:

$$[\mathbf{A}^\mathsf{T}\boldsymbol{\lambda}]_{\mathcal{T}} = \mathbf{e} \tag{eq}$$

$$\left\| [\mathbf{A}^\mathsf{T}\boldsymbol{\lambda}]_{\mathcal{T}^c} \right\|_\infty < 1. \tag{bnd}$$

Above, the notation $[\cdots]_{\mathcal{T}}$ means the restriction of the components of the vector to the index set $\mathcal{T}$.

5 Construction of the dual certificate

We will take $\boldsymbol{\lambda}$ to be a specific solution of (eq). Then we will prove that this $\boldsymbol{\lambda}$ satisfies (bnd).

Without loss of generality we can permute the columns of $\mathbf{A}$ in such a way that the coordinates corresponding to $\mathcal{T}$ are the first coordinates of $\mathbf{A}$. Then we can partition $\mathbf{A}$ as follows:

$$\mathbf{A} = [\mathbf{A}_{\mathcal{T}} | \mathbf{A}_{\mathcal{T}^c}].$$

Note that the equation

$$[\mathbf{A}^\mathsf{T}\boldsymbol{\lambda}]_{\mathcal{T}} = \mathbf{e}$$

has infinitely many solution because $\mathbf{A}_{\mathcal{T}}^\mathsf{T}$ is an $s \times m$ matrix and $m > s$.

Among infinitely many solutions of (eq) we will choose the one with the smallest $\ell 2$-norm:

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & \|\boldsymbol{\lambda}\|_2 \\ \text{subject to} \quad & \mathbf{A}_{\mathcal{T}}^\mathsf{T}\boldsymbol{\lambda} = \mathbf{e} \end{aligned} \tag{4}$$

which gives us the closed form solution as follows: $\boldsymbol{\lambda} = \mathbf{A}_{\mathcal{T}}(\mathbf{A}_{\mathcal{T}}^\mathsf{T}\mathbf{A}_{\mathcal{T}})^{-1}\mathbf{e}$.

To see that this $\boldsymbol{\lambda}$ is indeed the solution of (4), note that every vector that satisfies the constraints in (4) can be written as $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda} + \mathbf{n}$ with $\mathbf{A}_{\mathcal{T}}^\mathsf{T}\mathbf{n} = \mathbf{0}$. Therefore, since

$$\langle \mathbf{n}, \boldsymbol{\lambda} \rangle = \left\langle \mathbf{n}, \mathbf{A}_{\mathcal{T}}(\mathbf{A}_{\mathcal{T}}^\mathsf{T}\mathbf{A}_{\mathcal{T}})^{-1}\mathbf{e} \right\rangle = \left\langle \mathbf{A}_{\mathcal{T}}^\mathsf{T}\mathbf{n}, (\mathbf{A}_{\mathcal{T}}^\mathsf{T}\mathbf{A}_{\mathcal{T}})^{-1}\mathbf{e} \right\rangle = 0,$$

we have, for $\mathbf{n} \neq \mathbf{0}$,

$$\begin{aligned} \left\|\tilde{\boldsymbol{\lambda}}\right\|_2^2 &= \|\boldsymbol{\lambda} + \mathbf{n}\|_2^2 \\ &= \|\boldsymbol{\lambda}\|_2^2 + \|\mathbf{n}\|_2^2 + 2\underbrace{|\langle \mathbf{n}, \boldsymbol{\lambda} \rangle|}_{0} \\ &= \|\boldsymbol{\lambda}\|_2^2 + \|\mathbf{n}\|_2^2 \\ &> \|\boldsymbol{\lambda}\|_2^2. \end{aligned}$$

We conclude that $\tilde{\boldsymbol{\lambda}}$ is not the optimal point unless $\mathbf{n} = \mathbf{0}$. Observe that it is not obvious that this particular solution of (eq) has a chance to satisfy (bnd). Intuitively, we minimize some norm of $\boldsymbol{\lambda}$, to make the vector shorter, so there is more chance for $\left\|[\mathbf{A}^\intercal \lambda]_{\mathcal{T}^c}\right\|_\infty < 1$ to be true. Minimizing the $\ell 2$-norm specifically is convenient because there is the closed form solution for it, as specified above. In general, the dual certificate is not unique, and other constructions also exist.

Define $\mathbf{z} = \mathbf{A}_{\mathcal{T}^c}^\intercal \lambda = \mathbf{A}_{\mathcal{T}^c}^\intercal \mathbf{A}_\mathcal{T}(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}\mathbf{e}$. To prove (bnd), it remains to show that: $|z_i| < 1$ for all i with high probability, where the probability is over the random choice of $\mathbf{A}$.

First note that $\mathbf{A}_{\mathcal{T}^c}$ is independent of $\mathbf{A}_\mathcal{T}(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}\mathbf{e}$. The vector $\mathbf{z}$ has a complicated distribution, but we can control it by controlling $\mathbf{A}_{\mathcal{T}^c}$ and $\mathbf{A}_\mathcal{T}(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}\mathbf{e}$ separately.

Let's calculate the $\ell 2$-norm of λ:

$$\begin{aligned}
\|\lambda\|_2^2 &= \left\|\mathbf{A}_\mathcal{T}(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}\mathbf{e}\right\|_2^2 \\
&= \mathbf{e}^\intercal(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-\intercal}\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T}(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}\mathbf{e} \\
&= \mathbf{e}^\intercal(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}\mathbf{e}.
\end{aligned}$$

In the following lemmas we will show that $\|\lambda\|_2^2$ is small with high probability.

Claim 4. $\|\lambda\|_2^2 = \mathbf{e}^\intercal(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}\mathbf{e}$ *has the same distribution as* $s[(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}]_{11}$.

Proof. The claim follows from the fact that the Gaussian distribution is symmetric w.r.t. change of basis as follows. Recall that $\mathbf{e} \in \mathbb{R}^s$ is a vector of $+1$'s and -1's. Therefore, $\|\mathbf{e}\|_2 = \sqrt{s}$. Let's change basis: $\mathbf{e} = \mathbf{U}\mathbf{e}_1$ where $\mathbf{e}_1 = [\sqrt{s}, 0, \dots, 0]^\intercal$ and $\mathbf{U}$ is a unitary matrix. Therefore,

$$\begin{aligned}
\mathbf{e}^\intercal(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}\mathbf{e} &= \mathbf{e}_1^\intercal \mathbf{U}^\intercal(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}\mathbf{U}\mathbf{e}_1 \\
&= \mathbf{e}_1^\intercal(\mathbf{U}\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T}\mathbf{U}^\intercal)^{-1}\mathbf{e}_1 \\
&\sim \mathbf{e}_1^\intercal(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}\mathbf{e}_1 \\
&= s[(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}]_{11}
\end{aligned}$$

where we have used that $\mathbf{A}_\mathcal{T}\mathbf{U}^\intercal$ and $\mathbf{A}_\mathcal{T}$ have the same distribution because $\mathbf{A}_\mathcal{T}$ is Gaussian and $\mathbf{U}$ is unitary; the notation $\sim$ means that the two random variables have the same distribution; and the last step follows from the definition of $\mathbf{e}_1$. □

Let us now recall the definition of χ^2 random variable.

Definition 5. *Let* $z_1, \dots, z_k$ *be i.i.d.* $\mathcal{N}(0, 1)$ *random variables. Then* $q = z_1^2 + \cdots + z_k^2$ *has* χ^2 *distribution with* k *degrees of freedom.*

Claim 6. $m/[(\mathbf{A}_\mathcal{T}^\intercal \mathbf{A}_\mathcal{T})^{-1}]_{11}$ *has* χ^2 *distribution with* $m - s + 1$ *degrees of freedom.*

Proof. To shorten notation let $\mathbf{B} = \mathbf{A}_\mathcal{T}$. Let $\mathbf{b}$ denote the first column of $\mathbf{B}$ and $\mathbf{C}$ be the matrix that contains all columns of $\mathbf{B}$, except for the first one: $\mathbf{B} = [\mathbf{b}\ \mathbf{C}]$. Then,

$$\mathbf{B}^\intercal \mathbf{B} = \begin{bmatrix} \mathbf{b}^\intercal \mathbf{b} & \mathbf{b}^\intercal \mathbf{C} \\ \mathbf{C}^\intercal \mathbf{b} & \mathbf{C}^\intercal \mathbf{C} \end{bmatrix}.$$

Using the Matrix Inversion Lemma, it follows:

$$[(\mathbf{B}^\intercal \mathbf{B})^{-1}]_{11} = 1/k$$

where

$$k = \mathbf{b}^\mathsf{T}\mathbf{b} - \mathbf{b}^\mathsf{T}\mathbf{C}(\mathbf{C}^\mathsf{T}\mathbf{C})^{-1}\mathbf{C}^\mathsf{T}\mathbf{b}.$$

Note that

$$\mathbf{p} = \mathbf{C}(\mathbf{C}^\mathsf{T}\mathbf{C})^{-1}\mathbf{C}^\mathsf{T}\mathbf{b}$$

is the projection of the vector $\mathbf{b} \in \mathbb{R}^m$ onto the column space of $\mathbf{C}$. Using that $\langle \mathbf{b} - \mathbf{p}, \mathbf{p} \rangle = 0$, we conclude that

$$k = \mathbf{b}^\mathsf{T}\mathbf{b} - \mathbf{b}^\mathsf{T}\mathbf{p} = \|\mathbf{b} - \mathbf{p}\|_2^2.$$

Therefore, k is the squared distance between a Gaussian vector with zero mean and $1/m$ variance and an $s-1$ dimensional subspace. Therefore, $mk = m/[(\mathbf{A}_\mathcal{T}^\mathsf{T}\mathbf{A}_\mathcal{T})^{-1}]_{11}$ has χ^2 distribution with $m - s + 1$ degrees of freedom. ☐

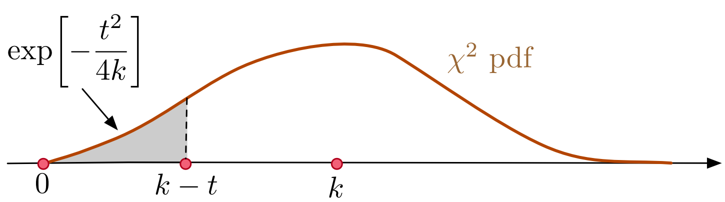
From Claims 4 and 6 we conclude that $ms/\|\boldsymbol{\lambda}\|_2^2$ has χ^2 distribution with $m - s + 1$ degrees of freedom.

Lemma 7. *Let q be a χ^2 distributed random variable with k degrees of freedom. Then, the following large deviation bound holds:*

$$\mathbb{P}[k - q > t] \le \exp\left[-\frac{t^2}{4k}\right] \tag{5}$$

for all $t > 0$.

The tail bound is illustrated in the plot below:



In your homework you will use the important Chernoff bound technique and derive a similar but somewhat weaker bound. If you are interested in the proof of the bound in Lemma 7, see [B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection", 2000, p. 1325].

Now we are ready to show the following tail bound for $\|\boldsymbol{\lambda}\|_2$:

Claim 8.

$$\mathbb{P}\left[\|\boldsymbol{\lambda}\|_2 > \sqrt{\frac{ms}{m - s + 1 - t}}\right] \le \exp\left[-\frac{t^2}{4(m - s + 1)}\right].$$

Proof. Since $ms/\|\boldsymbol{\lambda}\|_2^2$ is χ^2 distributed with $m - s + 1$ degrees of freedom, it follows from Lemma 7 that

$$\mathbb{P}\left[m - s + 1 - \frac{ms}{\|\boldsymbol{\lambda}\|_2^2} > t\right] \le \exp\left[-\frac{t^2}{4(m - s + 1)}\right].$$

Note that

$$\mathbb{P}\left[m - s + 1 - \frac{ms}{\|\boldsymbol{\lambda}\|_2^2} > t\right] = \mathbb{P}\left[\|\boldsymbol{\lambda}\|_2 > \sqrt{\frac{ms}{m-s+1-t}}\right]$$

which concludes the proof. □

Next, consider $\mathbf{A}_{\mathcal{T}^c}^{\mathsf{T}}\boldsymbol{\lambda}$ for a fixed $\boldsymbol{\lambda}$. This is a Gaussian random vector. Each component of this vector is $\mathcal{N}(0, \frac{1}{m}\|\boldsymbol{\lambda}\|_2^2)$. Therefore, using the Chernoff tail bound[1] for Gaussian $Q(\cdot)$ function that you will derive in your homework, we conclude:

$$\begin{aligned}
\mathbb{P}\left[|z_i| > 1 \middle| \|\boldsymbol{\lambda}\|_2 \leq \sqrt{\frac{ms}{m-s+1-t}}\right] &\leq \mathbb{P}\left[|w| > 1 \middle| w \sim \mathcal{N}\left(0, \frac{s}{m-s+1-t}\right)\right] \\
&= \mathbb{P}\left[|w|\sqrt{\frac{m-s+1-t}{s}} > \sqrt{\frac{m-s+1-t}{s}} \middle| w \sim \mathcal{N}\left(0, \frac{s}{m-s+1-t}\right)\right] \\
&= \mathbb{P}\left[|w| > \sqrt{\frac{m-s+1-t}{s}} \middle| w \sim \mathcal{N}(0,1)\right] \\
&\leq 2\exp\left[-\frac{m-s+1-t}{2s}\right].
\end{aligned}$$

Using the union bound and the fact that $\mathcal{T}^c$ contains $n-s$ elements we obtain:

$$\mathbb{P}\left[\left\|[\mathbf{A}^{\mathsf{T}}\boldsymbol{\lambda}]_{\mathcal{T}^c}\right\|_\infty > 1\right] \leq 2(n-s)\exp\left[-\frac{m-s+1-t}{2s}\right] + \exp\left[-\frac{t^2}{4(m-s+1)}\right].$$

This is an upper bound on the probability of failure. We would like to choose t so that this probability is less than $1/n$. First, let's find a condition on t so that the second term is less than $1/n$:

$$\begin{aligned}
&\exp\left[-\frac{t^2}{4(m-s+1)}\right] \leq \frac{1}{n} \\
\Leftrightarrow\; & -\frac{t^2}{4(m-s+1)} \leq -\log(n) \\
\Leftrightarrow\; & \frac{t^2}{4(m-s+1)} \geq \log(n) \\
\Leftrightarrow\; & t^2 \geq 4(m-s+1)\log(n)
\end{aligned}$$

so we can choose $t = 2\sqrt{(m-s+1)\log(n)}$ and plug this value into the first term. Now we want to find a value of m such that the first term is less than $2/n$:

$$2(n-s)\exp\left[-\frac{m-s+1-2\sqrt{(m-s+1)\log(n)}}{2s}\right] \leq \frac{2}{n}. \tag{6}$$

[1] $Q(v) \leq e^{-v^2/2}$

First show that

$$
\begin{aligned}
& m-s+1-2\sqrt{(m-s+1)\log(n)} \geq \frac{1}{2}(m-s+1) \\
& \Leftrightarrow m-s+1 \geq 4\sqrt{(m-s+1)\log(n)} \\
& \Leftrightarrow \sqrt{(m-s+1)} \geq 4\sqrt{\log(n)} \\
& \Leftrightarrow m-s+1 \geq 16\log(n)
\end{aligned}
$$

which is true because $s \geq 2$ and $m \geq 9s\log(n)$ by assumption:

$$
m-s+1 \geq 9s\log(n)-s \geq 8s\log(n) \geq 16\log(n).
$$

Therefore, to prove (6) it remains to show

$$
\begin{aligned}
& (n-s)\exp\left[-\frac{m-s+1}{4s}\right] \leq \frac{1}{n} \\
& \Leftarrow n\exp\left[-\frac{9s\log(n)-s}{4s}\right] \leq \frac{1}{n} \\
& \Leftarrow \exp\left[-\frac{8s\log(n)}{4s}\right] \leq \frac{1}{n^2} \\
& \Leftarrow \exp\left[-2\log(n)\right] = \frac{1}{n^2}.
\end{aligned}
$$

This completes the proof of the main theorem.