College of Engineering
School of Aerospace Engineering

ISYE7750: Mathematical Foundations of Machine Learning

Homework 3

Professor:
Ashwin Pananjady
Gtech ISYE Professor

Student:
Tomoki Koike
AE MS Student

October 7, 2022

Table of Contents

I Problem One **2**

II Problem Two **4**

III Problem Three **7**

I Problem One

The file `hw3p1_data.mat` contains two variables: `udata` and `ydata`. We will use this data to estimate a function $f : \mathbb{R}^2 \to \mathbb{R}$. The columns of `udata` contain sample locations, of which there are $M = 100$. The entries of $\mathbf{y}$ are the corresponding responses. We want to estimate f such that

$$f(\mathbf{u}_m) \approx y_m, \quad m = 1, \dots, M, \quad \text{where} \quad \mathbf{u}_m = \begin{bmatrix} s_m \\ t_m \end{bmatrix}.$$

We will restrict f to be a second-order polynomial on $[0,1] \times [0,1]$:

$$f(s,t) = \alpha_1 s^2 + \alpha_2 t^2 + \alpha_3 st + \alpha_4 s + \alpha_5 t + \alpha_6, \tag{I.1}$$

which means that f lies in a six dimensional subspace of $L_2([0,1]^2)$.

(a) Explain how to compute the 100×6 matrix $\mathbf{A}$ so that $\mathbf{y} \approx \mathbf{A}\boldsymbol{\alpha}$, where $\mathbf{y}$ contains the 100 response values in `ydata`. Write the code to compute $\mathbf{A}$ and turn it in.

(b) Solve

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^6} \|\mathbf{y} - \mathbf{A}\boldsymbol{\alpha}\|_2^2.$$

Turn in your code and the numerical value of your solution $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^6$.

(c) Make a contour plot of the corresponding

$$\hat{f}(s,t) = \hat{\alpha}_1 s^2 + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 st + \hat{\alpha}_4 s + \hat{\alpha}_5 t + \hat{\alpha}_6.$$

Include 50 contour lines, just so we have a very clear picture of what this function looks like.

Solution

Question (a)

To compute the $\mathbf{A}$ matrix we generate new column vectors corresponding to $\mathbf{s}^2$, $\mathbf{t}^2$, $\mathbf{st}$ and ones, then concatenate those to the original data matrix as shown in Figure 1. The code is submitted separately.

	ones	s	t	s^2	st	t^2	y
0	1.0	0.803180	0.24513	0.645098	0.196884	0.060089	1.8729
1	1.0	0.064122	0.26315	0.004112	0.016874	0.069248	1.6561
2	1.0	0.102720	0.48372	0.010551	0.049688	0.233985	1.6545
3	1.0	0.418880	0.38129	0.175460	0.159715	0.145382	2.1346
4	1.0	0.886770	0.42056	0.786361	0.372940	0.176871	1.8394
...	...	...	...	...	...	...	...
95	1.0	0.716310	0.92134	0.513100	0.659965	0.848867	3.1354
96	1.0	0.984020	0.98342	0.968295	0.967705	0.967115	3.3256
97	1.0	0.896310	0.86570	0.803372	0.775936	0.749436	3.1251
98	1.0	0.800960	0.55498	0.641537	0.444517	0.308003	2.3283
99	1.0	0.418870	0.12712	0.175452	0.053247	0.016159	1.8037

100 rows × 7 columns

Figure 1: Data with the $\mathbf{A}$ and output $\mathbf{y}$ columns concatenated.

Question (b)

From the code we found that the coefficients are

$$\alpha_1 = -0.6339, \quad \alpha_2 = 1.2282, \quad \alpha_3 = 0.1935$$
$$\alpha_4 = 0.9702, \quad \alpha_5 = 0.2255, \quad \alpha_6 = 1.2364$$

Question (c)

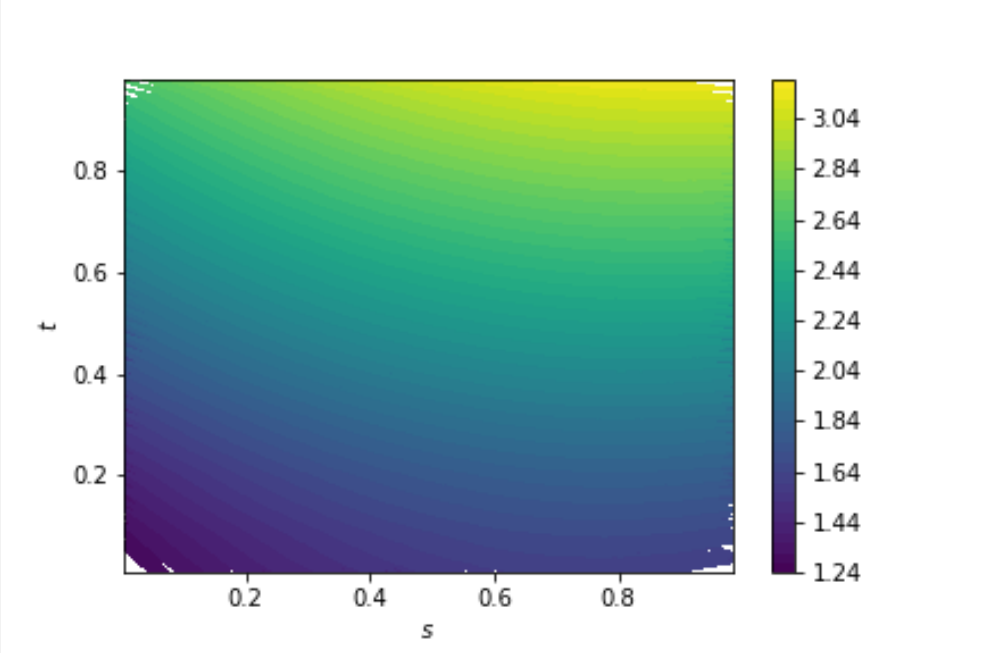

Figure 2: Contour plot of $f(s,t)$.

II Problem Two

Consider the space $\mathcal{P}_2$ of second-order polynomials on $[0,1]^2$ specified by $\alpha \in \mathbb{R}^6$ as in (I.1) above.

(a) At every point (s,t), the gradient $\nabla f(s,t)$ of a function $f \in \mathcal{P}_2$ is a vector in $\mathbb{R}^2$. As every $f \in \mathcal{P}_2$ is specified by a vector $\alpha \in \mathbb{R}^6$, we can think of the gradient at (s,t) as a mapping from $\mathbb{R}^6$ to $\mathbb{R}^2$. Show that this mapping is linear, which means, for a specified (s,t), there is a 2×6 matrix $\mathbf{G}_{s,t} \in \mathbb{R}^{2\times 6}$ such that

$$\nabla f(s,t) = \mathbf{G}_{s,t}\alpha$$

(b) Find the 6×6 matrix $\mathbf{H}_{s,t} \in \mathbb{R}^{6\times 6}$ such that (Hint: $\|\mathbf{G}_{s,t}\alpha\|_2^2 = \langle \mathbf{G}_{s,t}\alpha, \mathbf{G}_{s,t}\alpha\rangle = \cdots$.)

$$\|\nabla f(s,t)\|_2^2 = \alpha^\top \mathbf{H}_{s,t}\alpha.$$

What kinds of functions f are in the null space of $\mathbf{H}_{s,t}$ for all s and t? Why?

(c) Compute the matrix

$$\mathbf{Q} = \int_0^1 \int_0^1 \mathbf{H}_{s,t}\, dsdt.$$

(This is done simply by integrating each entry individually.)

(d) Describe how to set up and solve the optimization program

$$\min_{\mathbf{f}\in\mathcal{P}} \sum_{m=1}^{M} (y_m - f(s_m,t_m))^2 + \delta \int_0^1 \int_0^1 \|\nabla f(s,t)\|_2^2\, dsdt.$$

What is the regularizer above penalizing? What kinds of solutions do we expect for large δ?

(e) Apply your answer to part (d) to the data set from Problem 1. Play around with the value of δ, and produce estimates for three different δ that are interesting. Discuss why you think those values are indeed "interesting".

Solution

Question (a)

Given the function f as equation (I.1), if we take the gradient of this we have

$$\nabla f(s,t) = \begin{bmatrix} \frac{\partial f}{\partial s} \\ \frac{\partial f}{\partial t} \end{bmatrix} = \begin{bmatrix} 2s\alpha_1 + \alpha_3 t + \alpha_4 \\ 2t\alpha_2 + \alpha_3 s + \alpha_5 \end{bmatrix} = \begin{bmatrix} 2s & 0 & t & 1 & 0 & 0 \\ 0 & 2t & s & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{bmatrix} = \mathbf{G}_{s,t}\alpha. \tag{II.1}$$

Question (b)

Let $\mathbf{g}_i$ be the column vectors consisting $\mathbf{G}_{s,t}$. Then

$$\begin{aligned} \|\nabla f(s,t)\|_2^2 &= \|\mathbf{G}_{s,t}\alpha\|_2^2 = \langle \mathbf{G}_{s,t}\alpha, \mathbf{G}_{s,t}\alpha\rangle \\ &= \left\langle \sum_{i=1}^{6} \mathbf{g}_i\alpha_i, \sum_{j=1}^{6} \mathbf{g}_j\alpha_j \right\rangle = \sum_{i=1}^{6}\sum_{j=1}^{6} \alpha_i\alpha_j \langle \mathbf{g}_i, \mathbf{g}_j\rangle = \alpha^\top \mathbf{H}_{s,t}\alpha \end{aligned} \tag{II.2}$$

where $H_{ij} = \langle \mathbf{g}_i, \mathbf{g}_j \rangle$. Therefore, we have

$$\mathbf{H}_{s,t} = \begin{bmatrix} 4\,s^2 & 0 & 2\,s\,t & 2\,s & 0 & 0 \\ 0 & 4\,t^2 & 2\,s\,t & 0 & 2\,t & 0 \\ 2\,s\,t & 2\,s\,t & s^2+t^2 & t & s & 0 \\ 2\,s & 0 & t & 1 & 0 & 0 \\ 0 & 2\,t & s & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{II.3}$$

If we solve for $\text{null}(\mathbf{H}_{s,t})$ we obtain

$$\text{null}(\mathbf{H}_{s,t}) = \begin{bmatrix} -\frac{0.5000\,t}{s} & -\frac{0.5000}{s} & 0 & 0 \\ -\frac{0.5000\,s}{t} & 0 & -\frac{0.5000}{t} & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} -0.5000\,t^2 & -0.5000\,t & 0 & 0 \\ -0.5000\,s^2 & 0 & -0.5000\,s & 0 \\ s\,t & 0 & 0 & 0 \\ 0 & s\,t & 0 & 0 \\ 0 & 0 & s\,t & 0 \\ 0 & 0 & 0 & s\,t \end{bmatrix} \tag{II.4}$$

The functions of f that are in the null space of $\mathbf{H}_{s,t}$ are

$$\begin{aligned} \tilde{f}_1 &= -\frac{\alpha_1}{2}s^2 - \frac{\alpha_2}{2}t^2 + \alpha_3 st \\ \tilde{f}_2 &= \alpha_3 st - \frac{\alpha_5}{2}t \\ \tilde{f}_3 &= \alpha_3 st - \frac{\alpha_4}{2}s \\ \tilde{f}_4 &= \alpha_3 st \end{aligned} \tag{II.5}$$

These functions are in the null space since these are an orthogonal complement and do not suffice to solve the minization problem for the least square.

Question (c)

With simple integral operation we find that

$$\mathbf{Q} = \begin{bmatrix} 4/3 & 0 & 1/2 & 1 & 0 & 0 \\ 0 & 4/3 & 1/2 & 0 & 1 & 0 \\ 1/2 & 1/2 & 2/3 & 1/2 & 1/2 & 0 \\ 1 & 0 & 1/2 & 1 & 0 & 0 \\ 0 & 1 & 1/2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{II.6}$$

Question (d)

From the previous questions we can rewrite the setup as

$$\min_{f \in \mathcal{P}_2} \sum_{m=1}^{M} (y_m - f(s_m, t_m))^2 + \delta \alpha^\top \mathbf{Q} \alpha. \tag{II.7}$$

Since we have already computed the $\mathbf{A}$ matrix in the previous problem, we can solve this optimization problem by solely including the regularizer term. Hence, the solution of this becomes

$$\hat{\alpha} = (\mathbf{A}^\top \mathbf{A} + \delta \mathbf{Q})^{-1} \mathbf{A}^\top \mathbf{y}. \tag{II.8}$$

Here we know that the regularizer, i.e. the term with δ penalizes the size of the function, the function f in Hilbert space. Thus, we take the norm of the function to minimize it favoring a solution that is as small as possible with lower degree. If we consider a large δ value that is equivalent to relaxing the penalty, and thus, the approximation will become cruder.

Question (e)

For this question, we select the values $\delta = \{0.1, 5, 25\}$. The coefficients and costs are as follows.
$\delta = 0.1$: Selected because it shows how a small value of δ makes the results similar to Problem 1.

$$\alpha_1 = -0.5478, \quad \alpha_2 = 1.4129, \quad \alpha_3 = 0.1128$$
$$\alpha_4 = 0.9215, \quad \alpha_5 = 0.0673, \quad \alpha_6 = 1.2692$$

with cost of 0.0390.
$\delta = 5$: Selected to show somewhat of a deviations from the condition of no noise.

$$\alpha_1 = -0.5005, \quad \alpha_2 = 2.4727, \quad \alpha_3 = -0.4338$$
$$\alpha_4 = 1.0351, \quad \alpha_5 = -0.6494, \quad \alpha_6 = 1.2968$$

with cost of 0.0512.
$\delta = 25$: Selected because from here the difference is evident with cost over 0.1.

$$\alpha_1 = -0.5372, \quad \alpha_2 = 2.7897, \quad \alpha_3 = -0.4840$$
$$\alpha_4 = 0.8496, \quad \alpha_5 = -0.4538, \quad \alpha_6 = 1.0760$$

with cost of 0.1198.

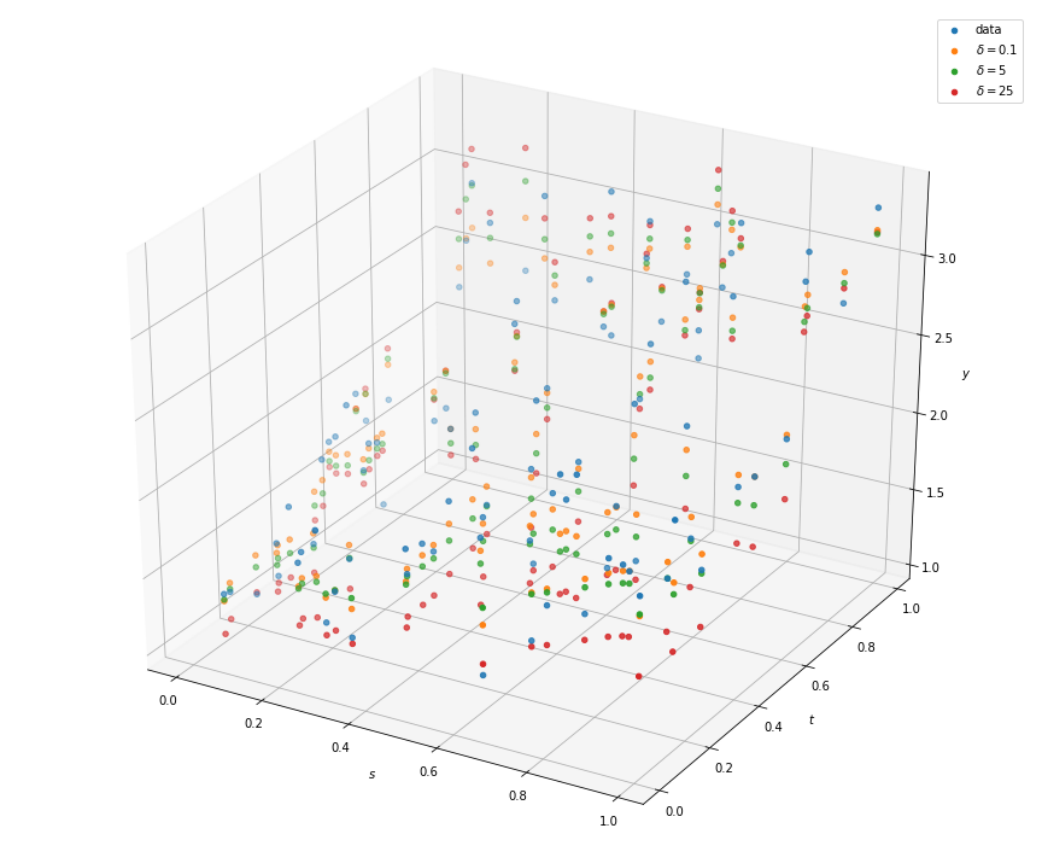

Figure 3: Caption

These δ values were interesting in that they show how increasing this values makes the discrepancy from the result in Problem 1 larger. That is making the value larger relaxes the optimization or makes it cruder by adding noise to the system. However, this allows to regularize the problem when we want to use a rather non-conservative and there is a large number of data points with a high degree of uncertainty.

III Problem Three

Let $\mathbf{A}$ be an $M \times N$ matrix with $\text{rank}(\mathbf{A}) < N$. We have seen in this case that the least-squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^N} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \tag{III.1}$$

has an infinite number of solutions. We have also seen, however, that the regularized least squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^N} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \delta\|\mathbf{x}\|_2^2 \tag{III.2}$$

has a unique solution for every $\delta > 0$. In this problem, we will show that as $\delta \to 0$, the regularized solution goes to the minimum norm solution of

$$\min_{\mathbf{x}\in\mathbb{R}^N} \|\mathbf{x}\|_2^2 \quad \text{subject to} \quad \mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{A}^\top\mathbf{y}. \tag{III.3}$$

(a) Start by showing that if $\mathbf{x}_1 \in \text{Row}(\mathbf{A})$ and $\mathbf{x}_2 \in \text{Row}(\mathbf{A})$ then $\mathbf{A}^\top\mathbf{A}\mathbf{x}_1 \neq \mathbf{A}^\top\mathbf{A}\mathbf{x}_2$ unless $\mathbf{x}_1 = \mathbf{x}_2$.

(b) Use part (a) to argue that the solution to (III.3) is always unique.

(c) In fact, something stronger than what we showed in part (a) is true.

There exists a constant $C > 0$ such that

$$\|\mathbf{A}^\top\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2)\|_2 \;\geq\; C\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \quad \text{for all} \quad \mathbf{x}_1, \mathbf{x}_2 \in \text{Row}(\mathbf{A}).$$

(This follows very easily from work we do later in the course, so we will defer its proof for now.) Use this fact to show that the solution of (III.2) goes to the solution of (III.3) as $\delta \to 0$. In particular, if $\mathbf{x}^\star$ is the (always unique) minimizer of (III.3), and $\hat{\mathbf{x}}_n$ is the (always unique) minimizer of (III.2) with (There is nothing special about taking $\delta = 1/n$ ... your argument should work for any sequence of δs that goes to zero.) $\delta = 1/n$, show that

$$\lim_{n\to\infty} \hat{\mathbf{x}}_n = \mathbf{x}^\star,$$

i.e. $\lim_{n\to\infty} \|\mathbf{x}^\star - \hat{\mathbf{x}}_n\|_2 = 0$.

Solution

Question (a)

Consider $\mathbf{A}^\top\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2) = 0$. This is feasible when $\mathbf{x}_1 - \mathbf{x}_2 \in \text{Null}(\mathbf{A}^\top\mathbf{A})$. If $x_1, x_2 \in \text{Row}(\mathbf{A})$ then from the linearity of vector spaces we know that $\{a\mathbf{x}_1 + b\mathbf{x}_2 \mid a, b \in \mathbb{R}\} \subseteq \text{Row}(\mathbf{A})$, and it follows that $\mathbf{x}_1 - \mathbf{x}_2 \in \text{Row}(\mathbf{A})$. Thus, this problem boils down to showing

$$\mathbf{x}_1 - \mathbf{x}_2 \in \text{Row}(\mathbf{A}) \bigcap \text{Null}(\mathbf{A}^\top\mathbf{A}) = 0. \tag{III.4}$$

If $\mathbf{A}^\top\mathbf{A}\mathbf{x} = 0$, then

$$\mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} = 0, \quad \text{or} \quad \|\mathbf{A}\mathbf{x}\| = 0,$$

i.e. $\mathbf{A}\mathbf{x} = 0$.

> **Proposition 1** *Let* $\mathbf{A} \in \mathbb{R}^{M\times N}$*, then* $\text{Null}(\mathbf{A}^\top\mathbf{A}) = \text{Null}(\mathbf{A})$.

From this proposition combined with the fact that $\text{Null}(\mathbf{A})$ is the orthogonal complement of $\text{Row}(\mathbf{A})$ we can conclude that (III.4) is true.

■

Question (b)

Let $\mathbf{A}^\top\mathbf{A}\mathbf{x}_1 = \mathbf{A}^\top\mathbf{A}\mathbf{x}_2 = \mathbf{A}^\top\mathbf{y}$. Now since Null($\mathbf{A}^\top\mathbf{A}$ = Null($\mathbf{A}$) and from the previous question, it follows that the equality only holds if $\mathbf{A}\mathbf{x}_1 = \mathbf{A}\mathbf{x}_2 = \mathbf{y}$. Thus, with $\mathbf{A}\mathbf{x}_1 \neq \mathbf{A}\mathbf{x}_2 \neq \mathbf{y}$ it shows that there exists a unique solution to (III.3).

Question (c)

We know that the explicit solution for the regularized least square problem (III.2) is

$$(\mathbf{A}^\top\mathbf{A} + \delta\mathbf{I}_N)\mathbf{x} = \mathbf{A}^\top\mathbf{y}. \tag{III.5}$$

Let $\hat{\mathbf{x}}_n,\ \mathbf{x}^\star \in \text{Row}(\mathbf{A})$, then

$$\begin{aligned}
\|\mathbf{x}^\star - \hat{\mathbf{x}}_n\| &\leq \frac{1}{C}\left\|\mathbf{A}^\top\mathbf{A}(\mathbf{x}^\star - \hat{\mathbf{x}}_n)\right\| \\
&= \frac{1}{C}\left\|\mathbf{A}^\top\mathbf{A}\mathbf{x}^\star + \left[(\mathbf{A}^\top\mathbf{A} + \delta\mathbf{I}_N)\hat{\mathbf{x}}_n - \mathbf{A}^\top\mathbf{y}\right] - \mathbf{A}^\top\mathbf{A}\hat{\mathbf{x}}_\mathbf{n}\right\| \\
&= \frac{1}{C}\left\|\mathbf{A}^\top\mathbf{A}\mathbf{x}^\star + \delta\hat{\mathbf{x}}_n - \mathbf{A}^\top\mathbf{y}\right\| \\
&= \frac{1}{C}\left\|\mathbf{A}^\top\mathbf{y} + \delta\hat{\mathbf{x}}_n - \mathbf{A}^\top\mathbf{y}\right\| \\
&= \frac{\delta}{C}\|\hat{\mathbf{x}}_n\|
\end{aligned}$$

Let $\delta = \frac{C}{n}$, then

$$\lim_{n\to\infty}\|\mathbf{x}^\star - \hat{\mathbf{x}}_n\| \leq \lim_{n\to\infty}\frac{1}{n}\|\hat{\mathbf{x}}_n\| = 0 \tag{III.6}$$