DAI Assignment 3

Ansh Garg - 24B1066
Kashyap Khandelwal - 24B0932
Pranav Patil - 24B1074

September 2025

Question 1

Let us first underline what each mathematical symbol implies:

- N: Total number of images ($N = 10000$),
- (x, y): The coordinates of a pixel,
- $I(x, y)$: True (unknown) intensity at that pixel,
- $J_k(x, y)$: Measured intensity at pixel (x, y) in the k-th image, where $k = 1, 2, \ldots, N$,
- $W_k(x, y)$: Noise or disturbance at pixel (x, y) in the k-th image, where $k = 1, 2, \ldots, N$.

It is given that the measured intensity is modeled as

$$J_k(x, y) = I(x, y) + W_k(x, y).$$

Since $I(x, y)$ is unknown, we estimate it and denote the estimate by $\widehat{I}(x, y)$. From the model,

$$I(x, y) = J_k(x, y) - W_k(x, y).$$

We define the estimator

$$\widehat{I}(x, y) = \frac{1}{N} \sum_{k=1}^{N} J_k(x, y).$$

This method works because $\mathbb{E}[W_k(x, y)] = 0$. According to the law of large numbers, as N becomes very large, the average of random samples converges to their mean, i.e.

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} W_k(x, y) = \mathbb{E}[W_k(x, y)] = 0.$$

Hence, in our case with $N = 10000$, the above approximation holds.
Now, since we know $\widehat{I}(x, y)$, we can calculate the noise samples:

$$\widehat{W}_k(x, y) = J_k(x, y) - \widehat{I}(x, y).$$

If the image dimensions are $B \times L$, then we will have $10000 \times B \times L$ estimated noise samples.
To determine the noise distribution, we plot a histogram with the horizontal axis representing the noise values and the vertical axis representing the frequency. We may then use kernel density estimation (KDE) to obtain a smooth estimate of the distribution.
Using the kernel function we obtain the probability density as'

$$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{W_i - x}{h}\right)$$

where n is $10000 \times B \times L$ and h is the smoothing function.

Question 2

a) Consider the random variable v. Its distribution function $F_v(y)$ is given by:

$$\begin{aligned} F_v(y) &= P(v \leq y) \\ &= P(F^{-1}(u) \leq y) \\ &= P(u \leq F(y)) \\ &= F(y) \end{aligned}$$

Using the fact that $P(\mathcal{U} \leq x) = x$ for $x \in [0, 1]$ for given uniform distribution, $u_i \sim \mathcal{U}[0, 1]$. Note that since the distribution function, F, is monotonically increasing we can take F on both sides as done in the derivation Thus, $\{v_i\}_{i=1}^n$ follows the distribution F.

b)

$$\begin{aligned} P(D \geq d) &= P\left\{\max_x \left|\frac{1}{n}\sum_i \mathbf{1}(Y_i \leq x) - F(x)\right| \geq d\right\} \\ &= P\left\{\max_x \left|\frac{1}{n}\sum_i \mathbf{1}(F(Y_i) \leq F(x)) - F(x)\right| \geq d\right\} \\ &\quad \text{(since F is a monotonically increasing function)} \\ &= P\left\{\max_x \left|\frac{1}{n}\sum_i \mathbf{1}(U_i \leq F(x)) - F(x)\right| \geq d\right\} \\ &\quad \text{(let } y = F(x)\text{)} \\ &= P\left\{\max_{0 \leq y \leq 1} \left|\frac{1}{n}\sum_i \mathbf{1}(U_i \leq y) - y\right| \geq d\right\} \\ &= P(E \geq d). \end{aligned}$$

Hence, proven.

c) We have shown that

$$P(D \geq d) = P(E \geq d),$$

which means that the distribution of the statistic

$$D = \max_x |F_e(x) - F(x)|$$

is independent of the underlying distribution F.
This result has important practical significance as it allows us to check whether the empirical distribution of given data matches a known distribution F. If indeed the data come from F, then the observed deviation D should not exceed (with high probability) the corresponding deviation between the empirical CDF of Unif$(0, 1)$ random variables and the true

Unif$(0, 1)$ distribution.
Essentially, we can say that that the distribution of the maximum absolute deviation between an empirical CDF and the true CDF does not depend on which continuous CDF F we are testing, because you can transform any continuous F-sample to uniform via the probability integral transform.

Question 3

a) According to the question, we have accurate information about the X and Y co-ordinates. The Z co-ordinates are corrupted (independently) by Gaussian noise. Modeling this, we can write:

$$z_i = a \cdot x_i + b \cdot y_i + c + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Thus,

$$z_i \sim \mathcal{N}(a \cdot x_i + b \cdot y_i + c, \sigma^2)$$

The joint likelihood function comes out to be:

$$P(z_i\,; x_i, y_i, a, b, c) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(z_i - (a \cdot x_i + b \cdot y_i + c))^2}{2\sigma^2} \right)$$

The log likelihood function is:

$$\mathcal{L} = -\sum_{i=1}^{n} \frac{(z_i - (a \cdot x_i + b \cdot y_i + c))^2}{2\sigma^2} + \text{constants}$$

To maximize this, we set derivatives with respect to a, b, c zero:

$$\frac{\partial \mathcal{L}}{\partial a} = \sum_{i=1}^{n} \frac{z_i - (a \cdot x_i + b \cdot y_i + c)}{\sigma^2} \cdot x_i = 0$$

$$\sum_{i=1}^{n} z_i \cdot x_i = \hat{a} \sum_{i=1}^{n} x_i^2 + \hat{b} \sum_{i=1}^{n} x_i \cdot y_i + \hat{c} \sum_{i=1}^{n} x_i \tag{1}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n} \frac{z_i - (a \cdot x_i + b \cdot y_i + c)}{\sigma^2} \cdot y_i = 0$$

$$\sum_{i=1}^{n} z_i \cdot y_i = \hat{a} \sum_{i=1}^{n} x_i \cdot y_i + \hat{b} \sum_{i=1}^{n} y_i^2 + \hat{c} \sum_{i=1}^{n} y_i \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial c} = \sum_{i=1}^{n} \frac{z_i - (a \cdot x_i + b \cdot y_i + c)}{\sigma^2} = 0$$

$$\sum_{i=1}^{n} z_i = \hat{a} \sum_{i=1}^{n} x_i + \hat{b} \sum_{i=1}^{n} y_i + n\hat{c} \tag{3}$$

Writing these out in matrix form, we get:

$$\begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} = \begin{pmatrix} x_i & y_i & 1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

Compact form of this equation can be written as

$$Z = X \cdot \theta$$

where Z, X, θ are the respective matrices.
Our aim is to minimize $\|Z - X \cdot \theta\|^2$. Taking derivative and setting it to zero we get,

$$-2X^T \cdot (Z - X \cdot \hat{\theta}) = 0$$

$$X^T \cdot Z = (X^T \cdot X) \cdot \hat{\theta}$$

The three rows in the above matrix representation give our 3 equations, as required.

$$\begin{pmatrix} \sum_{i=1}^n z_i \cdot x_i \\ \sum_{i=1}^n z_i \cdot y_i \\ \sum_{i=1}^n z_i \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \cdot y_i & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i \cdot y_i & \sum_{i=1}^n y_i^2 & \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n y_i & 1 \end{pmatrix} \cdot \begin{pmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{pmatrix}$$

b) According to the question, we have accurate information about the X and Y coordinates. The Z coordinates are corrupted (independently) by Gaussian noise. Modeling this, we can write:

$$z_i = a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6 + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Thus,

$$z_i \sim \mathcal{N}(a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6, \sigma^2)$$

The joint likelihood function is:

$$P(z_i\,; x_i, y_i, a_1, \ldots, a_6) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(z_i - (a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6))^2}{2\sigma^2}\right)$$

The log-likelihood function is:

$$\mathcal{L} = -\sum_{i=1}^n \frac{(z_i - (a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6))^2}{2\sigma^2} + \text{constants}$$

To maximize this, we set derivatives with respect to $a_1, \ldots, a_6$ to zero:

$$\frac{\partial \mathcal{L}}{\partial a_1} = \sum_{i=1}^n \frac{z_i - \hat{z}_i}{\sigma^2} \cdot x_i^2 = 0$$

$$\sum_{i=1}^n z_i x_i^2 = \hat{a}_1 \sum_{i=1}^n x_i^4 + \hat{a}_2 \sum_{i=1}^n x_i^2 y_i^2 + \hat{a}_3 \sum_{i=1}^n x_i^3 y_i + \hat{a}_4 \sum_{i=1}^n x_i^3 + \hat{a}_5 \sum_{i=1}^n x_i^2 y_i + \hat{a}_6 \sum_{i=1}^n x_i^2$$

$$\frac{\partial \mathcal{L}}{\partial a_2} = \sum_{i=1}^n \frac{z_i - \hat{z}_i}{\sigma^2} \cdot y_i^2 = 0$$

$$\sum_{i=1}^n z_i y_i^2 = \hat{a}_1 \sum_{i=1}^n x_i^2 y_i^2 + \hat{a}_2 \sum_{i=1}^n y_i^4 + \hat{a}_3 \sum_{i=1}^n x_i y_i^3 + \hat{a}_4 \sum_{i=1}^n x_i y_i^2 + \hat{a}_5 \sum_{i=1}^n y_i^3 + \hat{a}_6 \sum_{i=1}^n y_i^2$$

$$\frac{\partial \mathcal{L}}{\partial a_3} = \sum_{i=1}^n \frac{z_i - \hat{z}_i}{\sigma^2} \cdot x_i y_i = 0$$

$$\sum_{i=1}^n z_i x_i y_i = \hat{a}_1 \sum_{i=1}^n x_i^3 y_i + \hat{a}_2 \sum_{i=1}^n x_i y_i^3 + \hat{a}_3 \sum_{i=1}^n x_i^2 y_i^2 + \hat{a}_4 \sum_{i=1}^n x_i^2 y_i + \hat{a}_5 \sum_{i=1}^n x_i y_i^2 + \hat{a}_6 \sum_{i=1}^n x_i y_i$$

$$\frac{\partial \mathcal{L}}{\partial a_4} = \sum_{i=1}^n \frac{z_i - \hat{z}_i}{\sigma^2} \cdot x_i = 0$$

$$\sum_{i=1}^n z_i x_i = \hat{a}_1 \sum_{i=1}^n x_i^3 + \hat{a}_2 \sum_{i=1}^n x_i y_i^2 + \hat{a}_3 \sum_{i=1}^n x_i^2 y_i + \hat{a}_4 \sum_{i=1}^n x_i^2 + \hat{a}_5 \sum_{i=1}^n x_i y_i + \hat{a}_6 \sum_{i=1}^n x_i$$

$$\frac{\partial \mathcal{L}}{\partial a_5} = \sum_{i=1}^n \frac{z_i - \hat{z}_i}{\sigma^2} \cdot y_i = 0$$

$$\sum_{i=1}^n z_i y_i = \hat{a}_1 \sum_{i=1}^n x_i^2 y_i + \hat{a}_2 \sum_{i=1}^n y_i^3 + \hat{a}_3 \sum_{i=1}^n x_i y_i^2 + \hat{a}_4 \sum_{i=1}^n x_i y_i + \hat{a}_5 \sum_{i=1}^n y_i^2 + \hat{a}_6 \sum_{i=1}^n y_i$$

$$\frac{\partial \mathcal{L}}{\partial a_6} = \sum_{i=1}^n \frac{z_i - \hat{z}_i}{\sigma^2} = 0$$

$$\sum_{i=1}^n z_i = \hat{a}_1 \sum_{i=1}^n x_i^2 + \hat{a}_2 \sum_{i=1}^n y_i^2 + \hat{a}_3 \sum_{i=1}^n x_i y_i + \hat{a}_4 \sum_{i=1}^n x_i + \hat{a}_5 \sum_{i=1}^n y_i + n\hat{a}_6$$

Writing these out in matrix form, we get:

$$\begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} = \begin{pmatrix} x_1^2 & y_1^2 & x_1 y_1 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^2 & y_n^2 & x_n y_n & x_n & y_n & 1 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{pmatrix}$$

Compact form:

$$Z = X \cdot \theta$$

where Z, X, θ are the respective matrices. Our aim is to minimize $\|Z - X \cdot \theta\|^2$. Taking derivative and setting it to zero:

$$-2X^T(Z - X\hat{\theta}) = 0$$

$$X^T Z = X^T X \cdot \hat{\theta}$$

The six rows of this matrix equation give the six linear equations corresponding to $\partial\mathcal{L}/\partial a_j = 0$ for $j = 1, \ldots, 6$.

c) No, we do not need knowledge of noise variance in estimating the equation of the plane. The term of noise variance, σ^2 does not appear anywhere in the equations derived above.
To estimate the noise variance, differentiate the log likelihood function with respect to variance and set it to zero:

$$\frac{\partial\mathcal{L}}{\partial\sigma^2} = \sum_{i=1}^{n} \frac{(z_i - \hat{z}_i)^2}{2(\hat{\sigma^2})^2} - \frac{N}{2\hat{\sigma^2}} = 0$$

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}{N}$$

Estimate the other parameters using the above equations and substitute in above equation to get estimate, $\hat{\sigma^2}$ of the noise variance.

d) On running the code uploaded alongside, values of a, b and c obtained are 10.0022, 19.998 and 29.9516. Thus, the estimated equation of the plane is

$$z = 10.0022x + 19.998y + 29.9516$$

The estimated noise variance is: 23.057

e) We first estimate the plane parameters a, b and c by solving the least squares equations on the full dataset and compute the residuals, $\epsilon_i = z_i - (ax_i + by_i + c)$.
Using these residuals, we estimate the noise variance $\hat{\sigma^2}$.
To detect outliers caused by data corruption (swapped coordinates), we set a threshold, say $2\hat{\sigma^2}$ and flag all points with noise greater than this, i.e. $|\epsilon_i| \geq 2\hat{\sigma^2}$, since such large deviations are unlikely under the Gaussian noise model.
These flagged points are then removed from the dataset. Finally, we re-estimate the plane parameters on the cleaned dataset using the same parametric estimation procedure.
This two-step approach should work because the majority of points lie close to the true plane, and eliminating outliers prevents them from biasing the parameter estimates.

Question 4

a) In MATLAB

b) Using the samples in set T, our estimate of the pdf, $\hat{p}_n(x;\sigma)$, is given by:

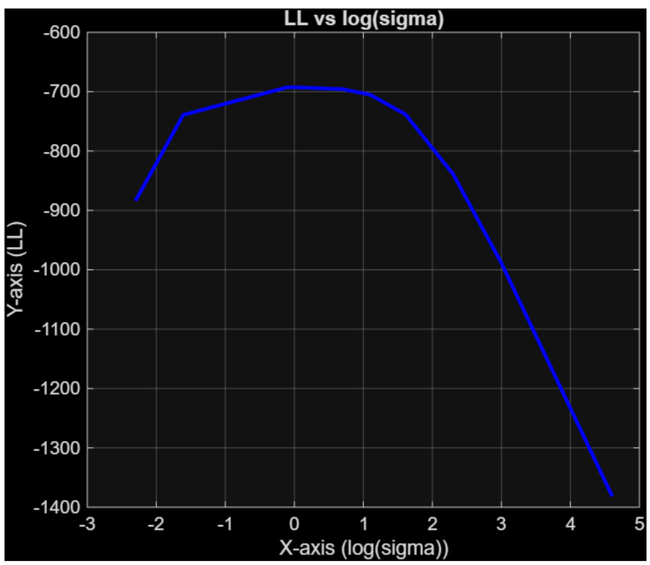
$$\hat{p}_n(x;\sigma) = \frac{1}{750\sigma\sqrt{2\pi}} \sum_{i=1}^{750} \exp\left(-\frac{(x-t_i)^2}{2\sigma^2}\right).$$

The Joint Likelihood LL of the samples in V, denoted by $\hat{p}_n(v_1, v_2, \ldots, v_{250};\sigma)$:

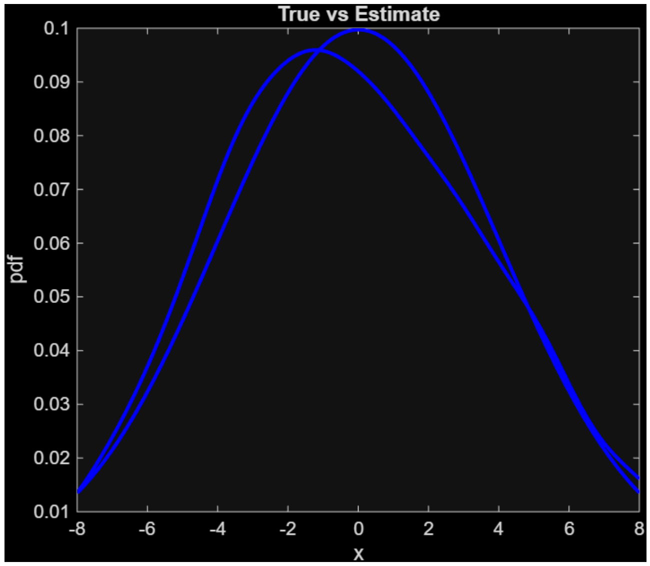
$$\hat{p}_n(v_1, v_2, \ldots, v_n;\sigma) = \prod_{j=1}^{n} \hat{p}_n(v_j;\sigma)$$

$$\hat{p}_n(v_1, v_2, \ldots, v_{250};\sigma) = \prod_{j=1}^{250} (\frac{1}{750\sigma\sqrt{2\pi}} \sum_{i=1}^{750} \exp\left(-\frac{(v_j-t_i)^2}{2\sigma^2}\right))$$

c) The plot of the log likelihood versus $\log\sigma$ is:

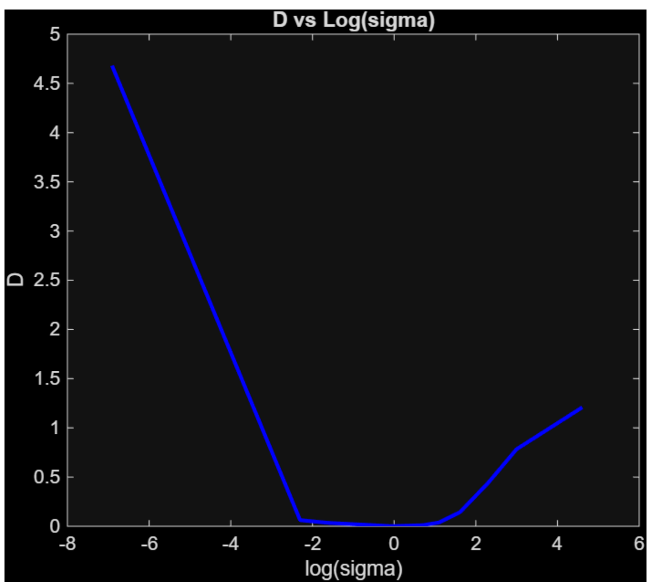


Best value of LL (maximum) is -692.60219 and is achieved at 1.000
Using this value of σ, we plot a graph of $\hat{p}_n(x;\sigma)$:

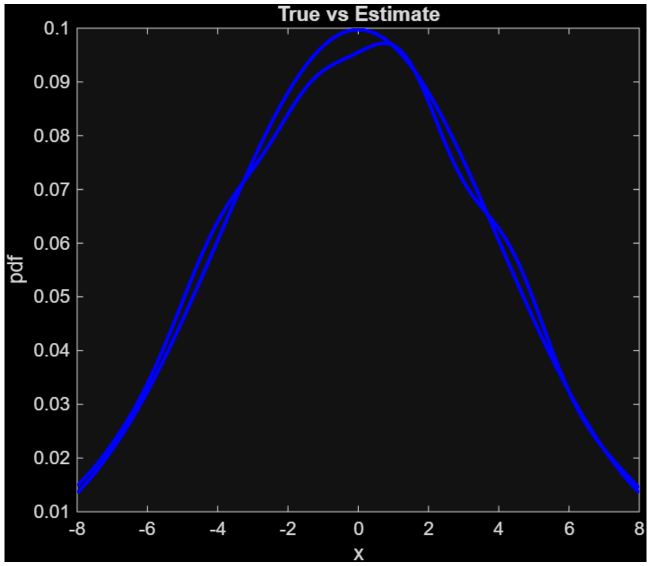


Overlay of graphs of true density and our estimated PDF

d) The plot of D versus $\log \sigma$ is:



Best value of D (minimum) and is 0.00226 is achieved at 1.000
Using this value of σ, we plot a graph of $\hat{p}_n(x;\sigma)$:



Overlay of graphs of true density and our estimated PDF

e) In the case when $V = T$, the Joint Likelihood, for a general size of T suppose n, is given by:

$$\hat{p}_n(v_1 = t_1, v_2 = t_2, \ldots, v_n = t_n; \sigma) = \prod_{j=1}^{n} \hat{p}_n(v_i = t_i; \sigma)$$

$$\hat{p}_n(v_1 = t_1, v_2 = t_2, \ldots, v_n = t_n; \sigma) = \prod_{j=1}^{n} \hat{p}_n(t_i; \sigma)$$

Notice that, for $\hat{p}_n(t_i; \sigma)$:

$$\begin{aligned}
\hat{p}_n(t_i; \sigma) &= \frac{1}{n\sigma\sqrt{2\pi}} \sum_{j=1}^{n} \exp\left(-\frac{(t_i - t_j)^2}{2\sigma^2}\right) \\
&\geq \frac{1}{n\sigma\sqrt{2\pi}} \exp\left(-\frac{(t_i - t_i)^2}{2\sigma^2}\right) \\
&= \frac{1}{n\sigma\sqrt{2\pi}}
\end{aligned}$$

Thus, all $\hat{p}_n(t_i;\sigma)$ is lower bounded by $\frac{1}{n\sigma\sqrt{2\pi}}$:

$$\prod_{j=1}^{n} \hat{p}_n(t_j;\sigma) \geq \prod_{j=1}^{n} \frac{1}{n\sigma\sqrt{2\pi}}$$

$$\prod_{j=1}^{n} \hat{p}_n(t_j;\sigma) \geq (\frac{1}{n\sigma\sqrt{2\pi}})^n$$

$$\log \prod_{j=1}^{n} \hat{p}_n(t_j;\sigma) \geq -n \log n\sigma\sqrt{2\pi}$$

Now, notice that as $\sigma \to 0$, $\log \sigma \to -\infty$ and $-n \log n\sigma\sqrt{2\pi} \to \infty$, thus the Log Likelihood (LL) is lower bounded by ∞, thus $LL \to \infty$ as $\sigma \to 0$. This implies, that practically, LL gets larger and larger as we take σ to 0, until it goes to infinity.

Going by the cross-validation method, When $T = V$, it would yield σ as 0 or the closest value to 0 in the set that we're experimenting over. In such a case, $\hat{p}_n(x;\sigma)$ is zero for any value other than t_i, and infinite otherwise.

This can directly be observed just by setting $V = T$ in the code for (c) part.