CompSci 527 Midterm Exam — Sample Solution

1. ***(8 points)*** Let h be the sequence $(2, 0, 1)$. Fill the following table with the "valid" and "full" correlation and convolution of h with itself:

valid correlation	**(5)**
full correlation	**(2, 0, 5, 0, 2)**
valid convolution	**(4)**
full convolution	**(4, 0, 4, 0, 1)**

2. ***(4 points)*** Is the following convolution kernel separable? If so, separate it. If not, prove that it is not.

$$H = \begin{array}{|c|c|c|}\hline 2 & 0 & 1 \\ \hline 0 & 0 & 0 \\ \hline 1 & 0 & 2 \\ \hline \end{array}$$

Answer: *No, it is not. The first and last column are not proportional to each other, and H, as a matrix, has therefore rank 2. Thus, H cannot be written as the product of a column vector and a row vector, as would be required for separability.*

3. ***(9 points)*** Write simple expressions for the gradient $\mathbf{g}(x, y)$, magnitude of the gradient $m(x, y)$, and Hessian $H(x, y)$ of the following function:

$$f(x, y) = xy \; .$$

Answer:

$$\begin{aligned} \mathbf{g}(x, y) &= \begin{bmatrix} y \\ x \end{bmatrix} \\ m(x, y) &= \sqrt{x^2 + y^2} \\ H(x, y) &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{aligned}$$

4. ***(7 points)*** Prove that the function $f(x, y)$ in the previous question has a saddle point at $x_0 = y_0 = 0$ *without* computing eigenvalues of H.

Answer: *At the origin $x_0 = y_0 = 0$, the gradient is $\mathbf{g}(x_0, y_0) = 0$, so that the origin is a stationary point. The Hessian at that point, as anywhere else, is indefinite, because the product*

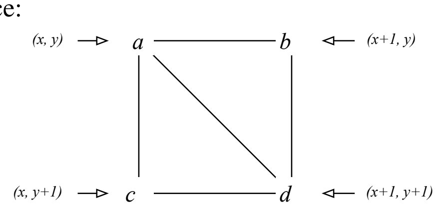
$$p(x, y) = \mathbf{x}^T H \mathbf{x} = 2xy$$

can be made positive, negative, or zero by changing the signs of x and y. For example, $p(1, 1) = 2$ and $p(1, -1) = -2$. A stationary point with an indefinite Hessian is a saddle point.

5. ***(9 points)*** Let image I have the following values at the given pixel coordinates:

$$I(x, y) = a \quad , \quad I(x+1, y) = b \quad , \quad I(x, y+1) = c \quad , \quad I(x+1, y+1) = d$$

A sketch is provided for your convenience:



Values of I inside the square formed by these four pixels are obtained by bilinear interpolation. Give the most general condition on the values of a, b, c, d that ensures that the graph of interpolated image values on the diagonal shown in the figure (connecting the pixel at (x, y) with the pixel at $(x + 1, y + 1)$) is a straight line. In other words, the interpolated value of a point p on the diagonal is an affine function of the position of p along the diagonal.

Answer: *Values at* $(x + \Delta x, y + \Delta x)$ *on the diagonal are given by*

$$J(\Delta x) = I(x + \Delta x, y + \Delta x) = a(1 - \Delta x)^2 + b\Delta x(1 - \Delta x) + c(1 - \Delta x)\Delta x + d(\Delta x)^2 \; .$$

The coefficient of $(\Delta x)^2$ *in this expression is* $a + d - b - c$*. For the graph of* $J(\Delta x)$ *to be a straight line we therefore require*

$$a + d = b + c \; .$$

6. ***(6 points)*** In a Gaussian pyramid, the `down` operator reduces the resolution of images, and the `up` operator increases it. We saw in class that $\texttt{up}(\texttt{down}(I))$ is not equal to I. Is $\texttt{down}(\texttt{up}(I))$ equal to I in general? Justify your answer briefly. Recall that the same scaling factor is used for the two operators.

   Answer: *No,* $\texttt{down}(\texttt{up}(I))$ *is generally not equal to* I*. The* `up` *operator up-samples by bilinear interpolation, and leaves the original samples unchanged. However, the* `down` *operator then smooths the resulting image with a Gaussian and down-samples by bilinear interpolation. Gaussian smoothing alters the image values even at integer grid points, because it is not an interpolator. Therefore, the original values in* I *are altered after the cascade of these two operators.*

7. ***(6 points)*** Give explicit, exact expressions for the matrices U, Σ, V of a full SVD of the following matrix.

$$A = \begin{bmatrix} 0 & 2 \\ -1 & 0 \\ 0 & 0 \end{bmatrix} .$$

   In a full SVD of A, the matrix U is 3×3, the matrix Σ is 3×2, and the matrix V is 2×2. Make sure you give V, not V^T.

   Answer: *The matrix* A *is almost its* Σ*, except for a column switch and a sign, so an SVD is as follows:*

$$\begin{aligned} U &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ \Sigma &= \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \\ V &= \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} . \end{aligned}$$

8. ***(4 points)*** The first two bins of a gradient-orientation histogram accumulate values in the intervals $[0, 10)$ and $[10, 20)$ respectively, in degrees. A gradient measurement with magnitude 1 and orientation θ degrees resulted into a bilinear vote of 0.7 units for the first bin, and a nonzero vote for the second bin. What is the vote for the second bin, and what is θ?

   Answer: *The vote for the second bin is 0.3 and* $\theta = 8$ *degrees.*

9. ***(9 points)*** Given n points in $\mathbb{R}^2$ arranged into the columns of an $2 \times n$ data matrix A, the one-dimensional PCA of A is given by a 2×1 vector $\boldsymbol{\mu}$, a 2×1 orthogonal matrix U, and a vector $\mathbf{s} = (s_1, s_2)$ of standard deviations for the two principal components. "Orthogonal" here means that $U^T U$ is the identity matrix, and "one-dimensional" means that we preserve only the first principal component. However, $\mathbf{s}$ contains standard deviations for both.

   Write exact formulas for $\boldsymbol{\mu}$, U, and s_2 for a one-dimensional PCA of the data matrix

$$A = \begin{bmatrix} x_1 & x_2 \\ y_1 & y_2 \end{bmatrix}$$

   where the two columns are distinct. No need to give a formula for s_1.

   [Hint: Think of the geometry, rather than calculating SVDs.]

   Answer:

$$\begin{aligned} \boldsymbol{\mu} &= \frac{1}{2} \begin{bmatrix} x_1 + x_2 \\ y_1 + y_2 \end{bmatrix} \\ U &= \frac{1}{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}} \begin{bmatrix} x_2 - x_1 \\ y_2 - y_1 \end{bmatrix} \\ s_2 &= 0 \end{aligned}$$

10. ***(8 points)*** A classification problem asks to find a classifier $h \in \mathcal{H}$ whose domain X and co-domain Y of h are defined as follows:

$$X = \{1, 3\} \quad \text{and} \quad Y = \{0, 1\}$$

and the hypothesis space $\mathcal{H}$ is the set of the following three functions

$$h_k(x) = \begin{cases} 0 & \text{if } x < k \\ 1 & \text{otherwise} \end{cases} \qquad \text{for} \quad k = 0, 2, 4 \ .$$

A training sample is a pair (x, y) with $x \in X$ and $y \in Y$.

The zero-one loss is used, and the following tiny training set is given:

$$T = \{(1, 0), (3, 0)\} \ .$$

Fill the table of the training risk $L_T(h_k)$ below. Express values in percent. Then use the table to determine the empirical risk minimizer $\hat{h}$ on T.

k	0	2	4
$L_T(h_k)$	**100**	**50**	**0**

Answer: *The empirical risk minimizer on T is $\hat{k} = h_4$.*

11. ***(8 points)*** Assume that the probability model $p(x, y)$ for the data in the previous question assigns equal probability to all possible samples (x, y). Thus, samples are not necessarily consistent. For instance, a training set can contain both samples $(1, 0)$ and $(1, 1)$. Under the same circumstances as in the previous question, and with the probability model above, fill the following table of the statistical risk $L_p(h_k)$. Express values in percent. What can you say about the statistical risk minimizer for this problem?

k	0	2	4
$L_p(h_k)$	**50**	**50**	**50**

Answer: *All classifiers have the same statistical risk, so they are all statistical-risk minimizers (or maximizers, for that matter).*

12. ***(8 points)*** Let $\mathbf{z} = [x, y]^T$. We use steepest descent to find a minimum of the function

$$f(\mathbf{z}) = \frac{x^2}{4} + \frac{y^2}{2}$$

starting at $\mathbf{z}_0 = [x_0, y_0]^T = [1, 1]^T$. If line search is used until full convergence (that is, until the bracketing interval is vanishingly small), what is the point $\mathbf{z}_1$ reached at the end of the first line search? Show your reasoning.

Answer: *The gradient of f at $\mathbf{z}_0$ is*

$$\mathbf{g}_0 = \begin{bmatrix} x_0/2 \\ y_0 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}$$

and the search line has parametric equation

$$\mathbf{z}(\alpha) = \mathbf{z}_0 - \alpha \mathbf{g}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \alpha \begin{bmatrix} 1/2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 - \alpha/2 \\ 1 - \alpha \end{bmatrix} \ .$$

The restriction of f to the search line is

$$h(\alpha) = f(\mathbf{z}(\alpha)) = \frac{(1 - \alpha/2)^2}{4} + \frac{(1 - \alpha)^2}{2} = \frac{1}{4}\left(\frac{9}{4}\alpha^2 - 5\alpha + 3\right)$$

and the derivative of h is

$$h'(\alpha) = \frac{1}{4}\left(\frac{9}{2}\alpha - 5\right) \ .$$

We have $h'(\alpha) = 0$ for

$$\alpha = \frac{10}{9}$$

so that

$$\mathbf{z}_1 = \mathbf{z}\left(\frac{10}{9}\right) = \begin{bmatrix} 1 - 5/9 \\ 1 - 10/9 \end{bmatrix} = \frac{1}{9}\begin{bmatrix} 4 \\ -1 \end{bmatrix} \ .$$

13. ***(6 points)*** Under the same circumstances as in the previous question, what is the point $\mathbf{z}_1$ reached from $\mathbf{z}_0$ after one Newton step? Explain briefly. [Hint: Think, don't calculate.]

    Answer: *The function f is quadratic, and a second-order Taylor expansion around $\mathbf{x}_0$ (or, for that matter, around any other point) is therefore exact. As a consequence, the first Newton step lands at the global minimum of f, that is,*

    $$\mathbf{z}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} .$$

14. ***(6 points)*** Fill in the blanks: The parameters of a machine learning system are optimized using a ___**training**___ set; its hyper-parameters are optimized using a ___**validation**___ set; its generalization performance is measured on a ___**test**___ set.

    Answer:

15. ***(2 points)*** In some cases, the training risk of a machine learning system is much smaller than its test risk. What is that phenomenon called?

    Answer: *Overfitting*