Solutions to Take-Home Quiz Four
ECE 271B - Winter 2022
Department of Electrical and Computer Engineering
University of California San Diego

Problem 1.

a) For this problem, the Lagrangian is

$$L(\mathbf{x}, \mu) = \frac{1}{2}||\mathbf{x}||^2 + \mu \left(\sum_i x_i + 3 \right)$$

and has zero gradient when

$$x_i + \mu = 0, \ \forall i.$$

We have two possibilities:

1. The constraint is inactive, $\sum_i x_i < -3$. In this case, $\mu = 0$ and $x_i = 0, \forall i$. This is a contradiction.
2. The constraint is active, $\sum_i x_i = -3$. In this case, $\mu > 0$ and $x_i = -\mu, \forall i$. It follows that $\mu = 3/n$ and $x_i = -3/n, \forall i$.

The Hessian of the Lagrangian is the identity and therefore always positive definite. Hence, the minimum is at

$$x_i^\star = -\frac{3}{n}, \forall i.$$

b) The problem of maximizing $\mathbf{y}^T\mathbf{x}$ subject to the constraint $\mathbf{x}^T\mathbf{Q}\mathbf{x} \leq 1$ has Lagrangian

$$L(\mathbf{x}, \mu) = \mathbf{y}^T\mathbf{x} + \mu \left(\mathbf{x}^T\mathbf{Q}\mathbf{x} - 1 \right),$$

which has zero gradient when

$$\mathbf{y} + 2\mu\mathbf{Q}\mathbf{x}^\star = 0.$$

Assuming that $\mathbf{y} \neq \mathbf{0}$, this rules out a solution with $\mu = 0$, from which the constraint must be active at the maximum, i.e. $(\mathbf{x}^\star)^T\mathbf{Q}\mathbf{x}^\star = 1$. Multiplying the equation above by $(\mathbf{x}^\star)^T$, we obtain

$$\mu = -\frac{1}{2}\mathbf{y}^T\mathbf{x}^\star.$$

Multiplying by $\mathbf{y}^T\mathbf{Q}^{-1}$, we obtain

$$\mathbf{y}^T\mathbf{Q}^{-1}\mathbf{y} + 2\mu\mathbf{y}^T\mathbf{x}^\star = 0$$

and, using the value of μ, it follows that the maximum is given by

$$\mathbf{y}^T\mathbf{x}^\star = \sqrt{\mathbf{y}^T\mathbf{Q}^{-1}\mathbf{y}}.$$

This implies that, when $\mathbf{x}^T\mathbf{Q}\mathbf{x} \leq 1$

$$(\mathbf{y}^T\mathbf{x})^2 \leq \mathbf{y}^T\mathbf{Q}^{-1}\mathbf{y}$$

and, exchanging the roles of $\mathbf{x}$ and $\mathbf{y}$, when $\mathbf{y}^T\mathbf{Q}^{-1}\mathbf{y} \leq 1$,

$$(\mathbf{y}^T\mathbf{x})^2 \leq \mathbf{x}^T\mathbf{Q}\mathbf{x}.$$

Problem 2.1. We consider the six problems.

1. Note that $f+g = 2x_1 - 1 \geq -1$ and $f-g = -x_2+1 \leq 1$. Hence, the set of feasible solutions is the set of (f,g) such that $-1-g \leq f \leq g+1$.
2. The feasible set is the set of (f,g) such that $f = \sqrt{g}$.
3. The feasible set is the set of (f,g) in $\{(-1/2,0),(1/2,-1)\}$.
4. The feasible set is the set of (f,g) on the line segment between $(-1/2,0)$ and $(1/2,-1)$.
5. Note that $f = 1/2[(g+1)^2+x_2^2]$, from which the feasible set is the set of (f,g) such that $f \geq \frac{(g+1)^2}{2}$.
6. Note that $f = |g| + x_2$, from which the feasible set is the set of (f,g) such that $f \geq |g|$.

Problem 2.2. In the lectures, we have seen that the Lagrange multiplier $\mu^\star$ defines the vector $w^\star = (\mu^\star, 1)$, which is normal to the hyper-plane that supports the set of feasible solutions

$$f(x) + \mu^\star g(x) - L^\star \geq 0, \ \ \forall x.$$

Furthermore, due to the constraint $\mu \geq 0$, $w^\star$ must be in the first quadrant of the (g,f) space. Finally, due to the constraint that $\mu^\star \neq 0$ only when the constraint is active, we have that one of the following two conditions must hold.

- Condition 1: Supporting plane is horizontal ($\mu^\star = 0$) and the point where it supports the set of feasible (f,g) has $g < 0$ (inactive constraint).
- Condition 2: $w^\star$ is anywhere in the first quadrant ($\mu^\star > 0$) and the support point has $g = 0$ (active constraint).

If you sketch the feasible regions derived above, you will see that the following holds for the six problems.

1. Condition 1 does not hold. Condition 2 holds when the supporting point is $(g,f) = (0,-1)$, which makes the supporting plane be the line $f = -g-1$, or $f+g+1 = 0$. Hence, $\mu^\star = 1$ and $f^\star = -1$.
2. The only plane supporting the set of feasible (f,g) is the vertical line through the origin. This is incompatible with the vector $w^\star$, which always has a component along the g-axis ($w^\star = (\mu^\star, 1)$). Hence, there is no Lagrange multiplier.
3. Condition 1 does not hold (no horizontal plane supports the two points and has support point with $g < 0$). Condition 2 does not hold because there is no point such that $g = 0$. Hence, there is no Lagrange multiplier.
4. Condition 1 still does not hold. Condition 2 now holds when the supporting plane is the segment itself and the supporting point is the intersection with the f-axis. $w^\star$ is the normal to the segment, and $\mu^\star$ can be computed from it.
5. Condition 2 does not hold, Condition 1 holds when the support plane is horizontal ($\mu^\star = 0$) and the support point $(g,f) = (-1,0)$. Note that $f^\star = 0$.
6. Condition 1 does not hold. Condition 2 holds for various values of $\mu^\star$, since there are many planes that support the feasible set at the point $(f,g) = (0,0)$. In fact, any line between the horizontal and $f = -g$ will be one such plane. Hence, the set of Lagrange multipliers is the interval $(0,1]$ and $f^\star = 0$.

Problem 2.3. Consider the six problems

1. The dual cost is

$$\begin{aligned} q(\mu) &= \min_{x_1 \geq 0, x_2 \geq 0} x_1 - x_2 + \mu(x_1 + x_2 - 1) \\ &= \min_{x_1 \geq 0, x_2 \geq 0} x_1(1+\mu) + x_2(\mu - 1) - \mu \\ &= \begin{cases} -\infty, & 0 \leq \mu < 1, \\ -\mu, & 1 \leq \mu \end{cases} \end{aligned}$$

   and the maximum is $q^\star = -1 = f^\star$. Hence, there is no duality gap.

2. The dual cost is

$$\begin{aligned} q(\mu) &= \min_x x + \mu x^2 \\ &= \begin{cases} -\infty, & \mu \leq 0, \\ -\frac{1}{4\mu}, & \mu > 0 \end{cases} \end{aligned}$$

   and there is no maximum ($q^\star \to 0$ as $\mu \to \infty$). Note that the solution of the primal problem is $f^* = 0$ and there is no duality gap. However, because there is no Lagrange multiplier, the dual problem has no solution.

3. The dual cost is

$$\begin{aligned} q(\mu) &= \min_{x \in \{0,1\}} -x + \mu(x - \frac{1}{2}) \\ &= \min_{\mu \geq 0} \left(-\frac{\mu}{2}, -1 + \frac{\mu}{2} \right) \end{aligned}$$

   and the maximum is $q^* = -1/2$. Note that the minimum of the primal problem is $f^* = 0$, and we have a duality gap. Hence, while the dual problem has a perfectly well-defined solution, this solution tells us nothing about the solution of the primal problem.

4. It can be checked that there is no duality gap in this case.

5. The dual cost is

$$\begin{aligned} q(\mu) &= \min_{x_1, x_2} \frac{1}{2}(x_1^2 + x_2^2) + \mu(x_1 - 1) \\ &= -\frac{1}{2}\mu^2 - \mu \end{aligned}$$

   and $q^\star = 0 = f^\star$. Hence, there is no duality gap.

6. The dual cost is

$$\begin{aligned} q(\mu) &= \min_{x_1, x_2 \geq 0} |x_1| + x_2 + \mu x_1 \\ &= \begin{cases} 0, & |\mu| \leq 1, \\ -\infty, & |\mu| > 1 \end{cases} \end{aligned}$$

   and $q^\star = 0 = f^\star$. Hence, there is no duality gap. Note that the set of dual optimal solutions is $0 < \mu^\star \leq 1$, confirming the previous observation that all these values of $\mu^\star$ are Lagrange multipliers.

Problem 3.

a) The Lagrangian is

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) = \frac{1}{2}||\mathbf{w}||^2 - \nu\rho + \frac{1}{n}\sum_i \xi_i - \sum_i \alpha_i[y_i(<\mathbf{x}_i, \mathbf{w}> +b) - \rho + \xi_i] - \sum_i \beta_i\xi_i - \gamma\rho.$$

Setting the derivatives w.r.t. to the primal variables $\mathbf{w}$, b, $\boldsymbol{\xi}, \rho$ to zero, we get

$$\begin{aligned}
\nabla_{\mathbf{w}} L = 0 \quad &\Leftrightarrow \quad \mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \\
\frac{\partial L}{\partial b} = 0 \quad &\Leftrightarrow \quad \sum_i \alpha_i y_i = 0 \\
\frac{\partial L}{\partial \xi_i} = 0 \quad &\Leftrightarrow \quad \alpha_i + \beta_i = \frac{1}{n} \\
\frac{\partial L}{\partial \rho} = 0 \quad &\Leftrightarrow \quad \sum_i \alpha_i - \gamma = \nu
\end{aligned}$$

and, plugging back in the Lagrangian,

$$\begin{aligned}
L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) &= -\frac{1}{2}||\mathbf{w}||^2 - \nu\rho + \frac{1}{n}\sum_i \xi_i - \sum_i \alpha_i(-\rho + \xi_i) - \sum_i \beta_i\xi_i - \gamma\rho \\
&= -\frac{1}{2}||\mathbf{w}||^2 - \nu\rho + \frac{1}{n}\sum_i \xi_i - \sum_i \xi_i(\alpha_i + \beta_i) + \rho(\sum_i \alpha_i - \gamma) \\
&= -\frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j > .
\end{aligned}$$

Noting that $\beta_i \geq 0$ implies that

$$\alpha_i \leq \frac{1}{n}$$

and $\gamma \geq 0$ that

$$\sum_i \alpha_i \geq \nu,$$

the dual problem is

$$\max_{\boldsymbol{\alpha}} \left(-\frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j > \right)$$

subject to

$$\begin{aligned}
0 \leq \alpha_i &\leq \frac{1}{n} \\
\sum_i \alpha_i &\geq \nu \\
\sum_i \alpha_i y_i &= 0.
\end{aligned}$$

The decision function is of the usual form

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i y_i < \mathbf{x}, \mathbf{x}_i > +b \right).$$

b) Consider the support vectors on the margin, i.e. with $0 < \alpha_i < 1/n$. Then, from the KKT conditions,

$$y_i(< \mathbf{x}_i, \mathbf{w} > +b) = \rho - \xi_i \quad \text{and} \quad \xi_i = 0,$$

from which

$$\rho = y_i(< \mathbf{x}_i, \mathbf{w} > +b).$$

Consider the set $\mathcal{A}$ of such SVs with $y_i = 1$ and the set $\mathcal{B}$ where $y_i = -1$. Then

$$\sum_{\mathbf{x}_i \in \mathcal{A}} (< \mathbf{x}_i, \mathbf{w} > +b) - \sum_{\mathbf{x}_i \in \mathcal{B}} (< \mathbf{x}_i, \mathbf{w} > +b) = (|\mathcal{A}| + |\mathcal{B}|)\rho$$

or

$$\sum_{\mathbf{x}_i \in \mathcal{A}} < \mathbf{x}_i, \mathbf{w} > - \sum_{\mathbf{x}_i \in \mathcal{B}} < \mathbf{x}_i, \mathbf{w} > +(|\mathcal{A}| - |\mathcal{B}|)b = (|\mathcal{A}| + |\mathcal{B}|)\rho.$$

Hence, if we pick two subsets $\mathcal{A}' \subset \mathcal{A}$ and $\mathcal{B}' \subset \mathcal{B}$ such that $|\mathcal{A}'| = |\mathcal{B}'|$, we can recover ρ from

$$\rho = \frac{1}{2|\mathcal{A}'|} \left[\sum_{\mathbf{x}_i \in \mathcal{A}'} < \mathbf{x}_i, \mathbf{w} > - \sum_{\mathbf{x}_i \in \mathcal{B}'} < \mathbf{x}_i, \mathbf{w} > \right].$$

Next, using the first equation again,

$$b = \rho - \frac{1}{|\mathcal{A}'|} \sum_{\mathbf{x}_i \in \mathcal{A}'} < \mathbf{x}_i, \mathbf{w} >= -\frac{1}{2|\mathcal{A}'|} \sum_{\mathbf{x}_i \in \mathcal{A}' \cup \mathcal{B}'} < \mathbf{x}_i, \mathbf{w} > .$$

c) From the KKT conditions, $\rho > 0$ implies that $\gamma = 0$, and

$$\nu = \sum_i \alpha_i.$$

Hence, if k is the number of i such that $\alpha_i = 1/n$, we must have

$$k\frac{1}{n} \leq \nu \;\Leftrightarrow\; \frac{k}{n} \leq \nu.$$

Furthermore, for all i such that $\xi_i > 0$, we must have $\alpha_i = 1/n$ since otherwise α_i could grow and decrease ξ_i. Hence,

$$\frac{1}{n}|\{i \,|\, \xi_i > 0\}| \leq \nu.$$

Bound 1 follows from the fact that the margin errors are the cases where $\xi_i > 0$ (otherwise, by the statement of the problem, $y_i g(\mathbf{x}_i) \geq \rho$). Bound 2 follows from the fact that

$$\sum_i \alpha_i \geq \nu$$

and each support vector can contribute at most $1/n$ to the summation. Hence, if the number of SVs is m,

$$\nu \leq \sum_i \alpha_i \leq \frac{m}{n}$$

and

$$\frac{m}{n} \geq \nu.$$

d) Consider the following procedure. We minimize the second problem. Then, we fix ρ to the optimal value and minimize over the remaining variables. Clearly, the solution will be the same. This means that values obtained for $\mathbf{w}$ and ξ minimize the cost of the first problem, with $C = 1$, under the constraints of the second. So, we only need to find a way to make

$$y_i(< \mathbf{x}_i, \mathbf{w} > +b) \geq \rho - \xi_i$$

equivalent to

$$y_i(< \mathbf{x}_i, \mathbf{w} > +b) \geq 1 - \xi_i,$$

which can be easily done by rescaling all variables so that

$$\begin{aligned} \mathbf{w}' &= \frac{1}{\rho}\mathbf{w} \\ \xi_i' &= \frac{1}{\rho}\xi_i \\ b' &= \frac{1}{\rho}b. \end{aligned}$$

Noting that

$$\min\left(\frac{1}{2}||\mathbf{w}||^2 + \frac{1}{n}\sum_i \xi_i\right) = \min\left(\frac{1}{2}||\mathbf{w}'||^2 + \frac{1}{n\rho}\sum_i \xi_i'\right)$$

leads to the desired equivalence.

Problem 4.a)

1) The test accuracy for each digit when $C = 2, 4, 8$ is below. Note that the value of C does not make a huge difference as long as one is close to the best. The performance of the individual classifiers varies quite a bit, reflecting the fact that some digits are much easier to identify than others. "1" appears to be the easiest digit and "8" and "9" the hardest ones.

digit 0	digit 1	digit 2	digit 3	digit 4	digit 5	digit 6	digit 7	digit 8	digit 9	overall
98.73%	99.32%	97.96%	97.47%	98.14%	97.06%	98.07%	98.28%	95.71%	96.31%	90.76%

Table 1: Accuracy for each digit when $C = 2$.

digit 0	digit 1	digit 2	digit 3	digit 4	digit 5	digit 6	digit 7	digit 8	digit 9	overall
98.68%	99.22%	97.92%	97.42%	98.02%	97.49%	98.04%	98.21%	95.59%	96.31%	90.43%

Table 2: Accuracy for each digit when $C = 4$.

digit 0	digit 1	digit 2	digit 3	digit 4	digit 5	digit 6	digit 7	digit 8	digit 9	overall
98.54%	99.13%	97.87%	97.48%	97.96%	97.36%	97.92%	98.08%	95.55%	96.36%	90.13%

Table 3: Accuracy for each digit when $C = 8$.

2) The number of support vectors for each digit when $C = 2, 4, 8$ is below. Note that the number of support vectors is proportional to how difficult the digits are to classify. In this case, "8" and "9" require a lot more support vectors than the remaining digits.

digit 0	digit 1	digit 2	digit 3	digit 4	digit 5	digit 6	digit 7	digit 8	digit 9
464	505	1211	1422	900	1326	686	779	2093	1848

Table 4: Number of support vectors for each digit when $C = 2$.

digit 0	digit 1	digit 2	digit 3	digit 4	digit 5	digit 6	digit 7	digit 8	digit 9
45 5	486	1201	1416	880	1292	671	770	2097	1836

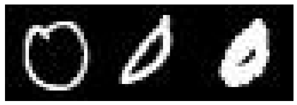
Table 5: Number of support vectors for each digit when $C = 4$.

digit 0	digit 1	digit 2	digit 3	digit 4	digit 5	digit 6	digit 7	digit 8	digit 9
436	455	1189	1400	869	1253	651	744	2077	1816

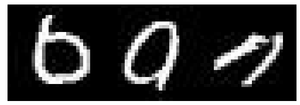
Table 6: Number of support vectors for each digit when $C = 8$.

3) The following plots show the 3 examples of largest Lagrange multiplier on both sides of the discriminant plane (y_i is the label for i^{th} example). These are the examples closest to the border and the hardest ones to get right.

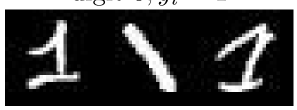
For $C = 2$,



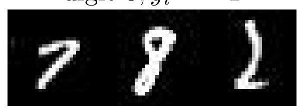
digit 0, $y_i = 1$



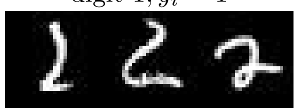
digit 0, $y_i = -1$



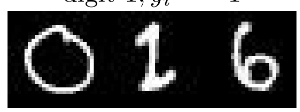
digit 1, $y_i = 1$



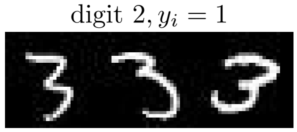
digit 1, $y_i = -1$



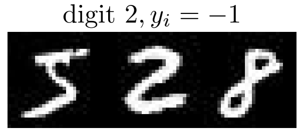
digit 2, $y_i = 1$



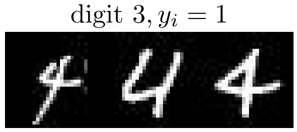
digit 2, $y_i = -1$



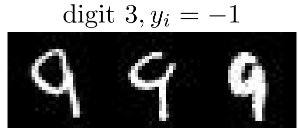
digit 3, $y_i = 1$



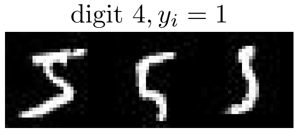
digit 3, $y_i = -1$



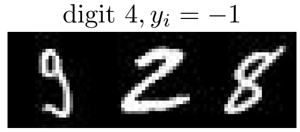
digit 4, $y_i = 1$



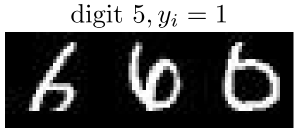
digit 4, $y_i = -1$



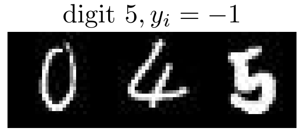
digit 5, $y_i = 1$



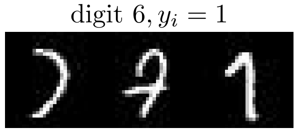
digit 5, $y_i = -1$



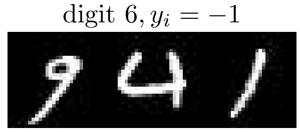
digit 6, $y_i = 1$



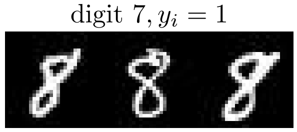
digit 6, $y_i = -1$



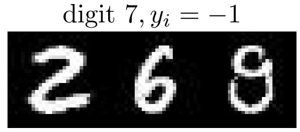
digit 7, $y_i = 1$



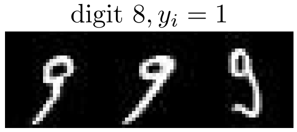
digit 7, $y_i = -1$



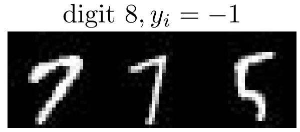
digit 8, $y_i = 1$



digit 8, $y_i = -1$

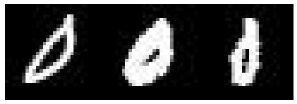


digit 9, $y_i = 1$

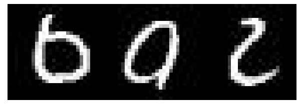


digit 9, $y_i = -1$

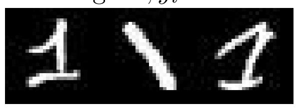
For $C = 4$,



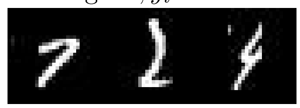
digit $0, y_i = 1$



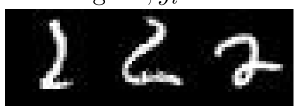
digit $0, y_i = -1$



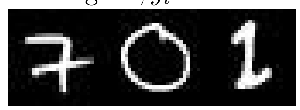
digit $1, y_i = 1$



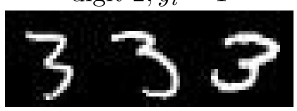
digit $1, y_i = -1$



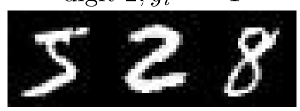
digit $2, y_i = 1$



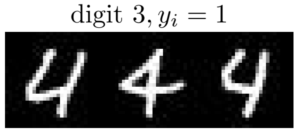
digit $2, y_i = -1$



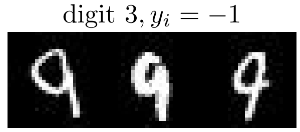
digit $3, y_i = 1$



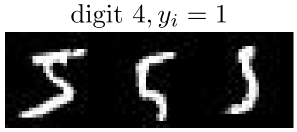
digit $3, y_i = -1$



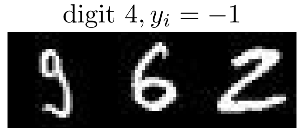
digit $4, y_i = 1$



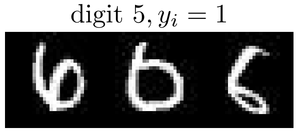
digit $4, y_i = -1$



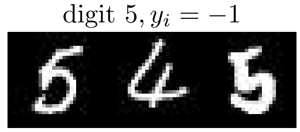
digit $5, y_i = 1$



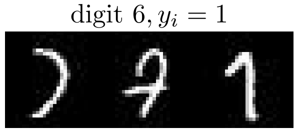
digit $5, y_i = -1$



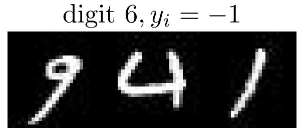
digit $6, y_i = 1$



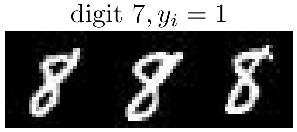
digit $6, y_i = -1$



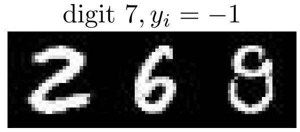
digit $7, y_i = 1$



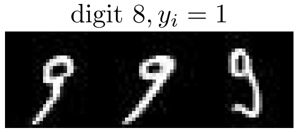
digit $7, y_i = -1$



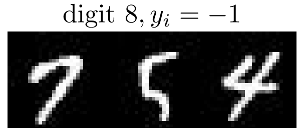
digit $8, y_i = 1$



digit $8, y_i = -1$

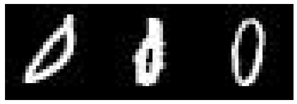


digit $9, y_i = 1$

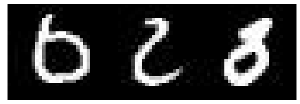


digit $9, y_i = -1$

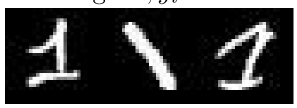
For $C = 8$,



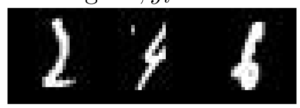
digit $0, y_i = 1$



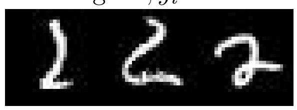
digit $0, y_i = -1$



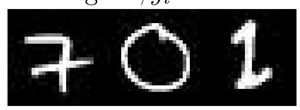
digit $1, y_i = 1$



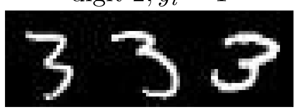
digit $1, y_i = -1$



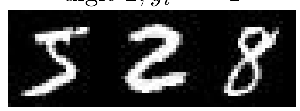
digit $2, y_i = 1$



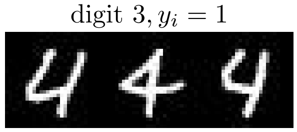
digit $2, y_i = -1$



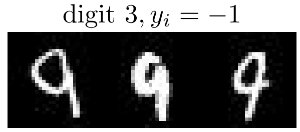
digit $3, y_i = 1$



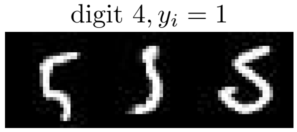
digit $3, y_i = -1$



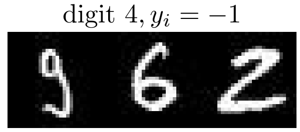
digit $4, y_i = 1$



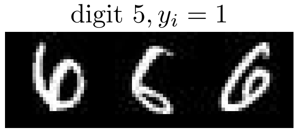
digit $4, y_i = -1$



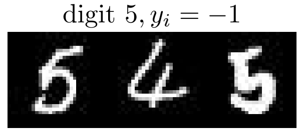
digit $5, y_i = 1$



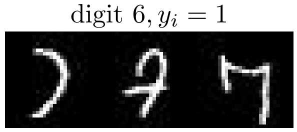
digit $5, y_i = -1$



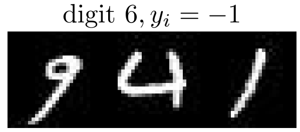
digit $6, y_i = 1$



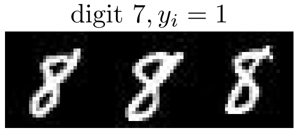
digit $6, y_i = -1$



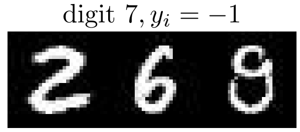
digit $7, y_i = 1$



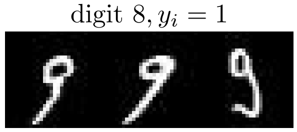
digit $7, y_i = -1$



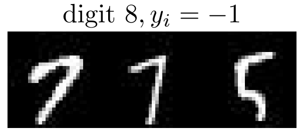
digit $8, y_i = 1$



digit $8, y_i = -1$

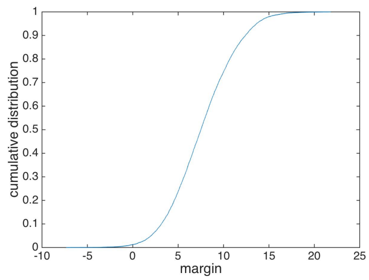


digit $9, y_i = 1$

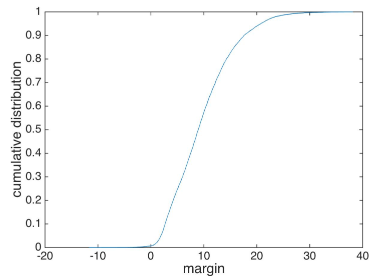


digit $9, y_i = -1$

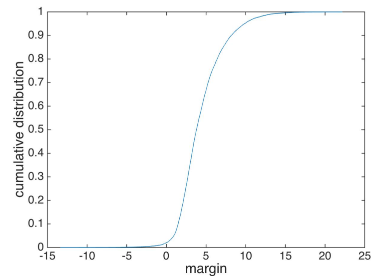
Problem 4.b) The plots of the cdf of the margin for every digit from 0 to 9 when $C = 2$ are as follows.



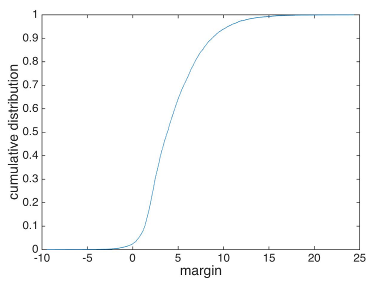
digit 0



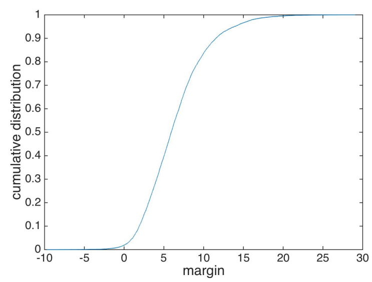
digit 1



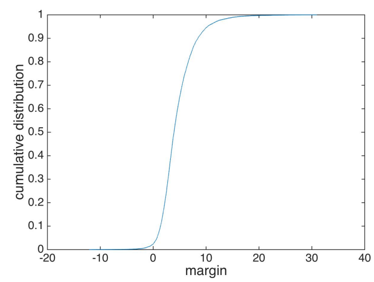
digit 2



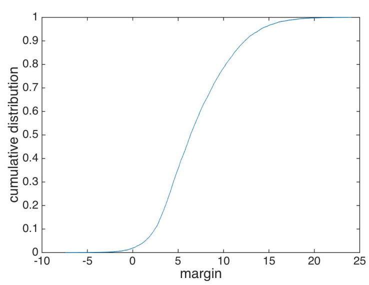
digit 3



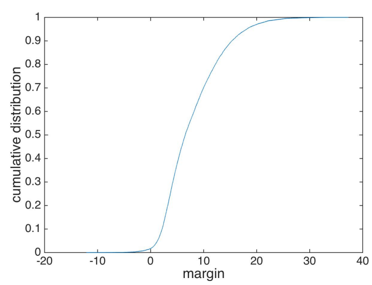
digit 4



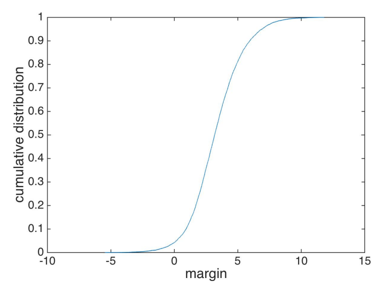
digit 5



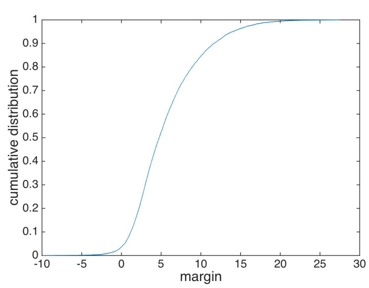
digit 6



digit 7

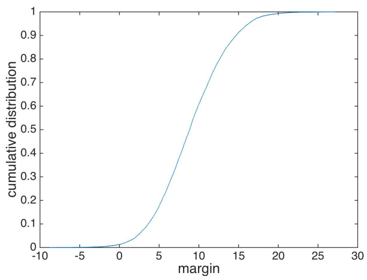


digit 8

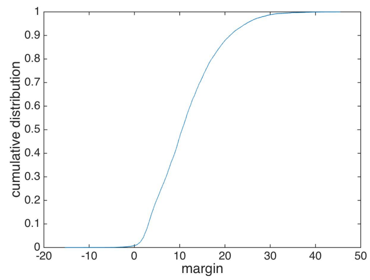


digit 9

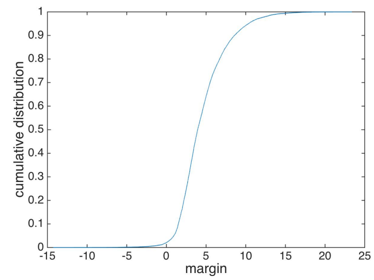
The plots of the cdf of the margin for every digit from 0 to 9 when $C = 4$ are as follows.



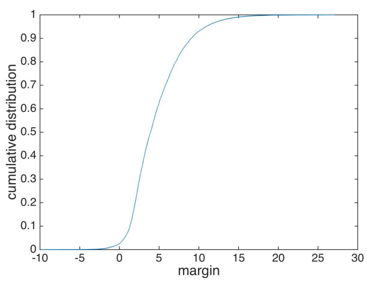
digit 0



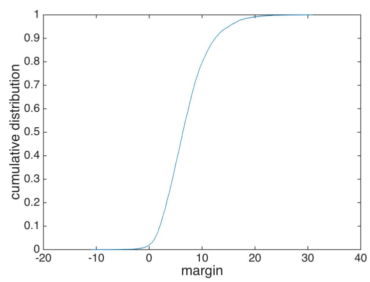
digit 1



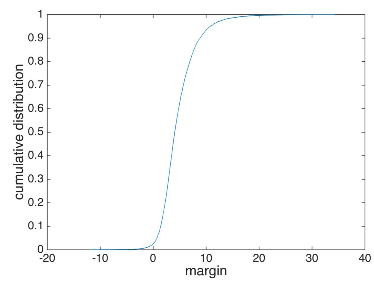
digit 2



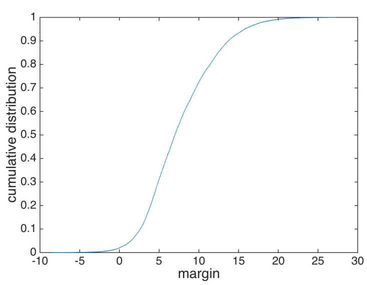
digit 3



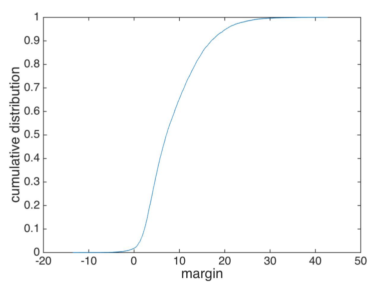
digit 4



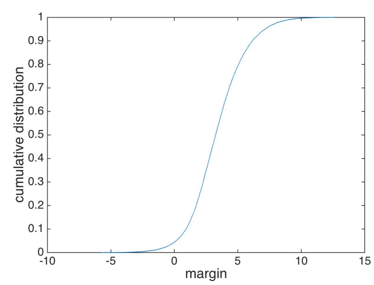
digit 5



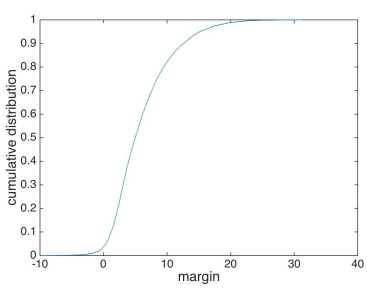
digit 6



digit 7

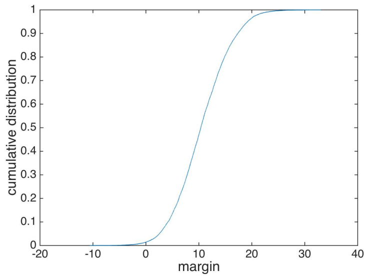


digit 8

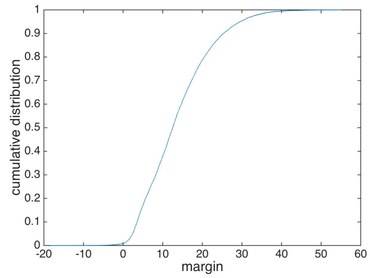


digit 9

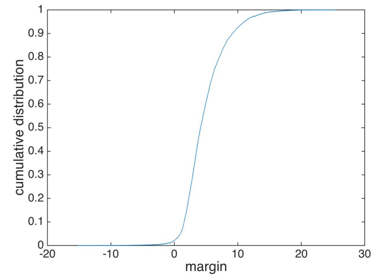
The plots of the cdf of the margin for every digit from 0 to 9 when $C = 8$ are as follows.



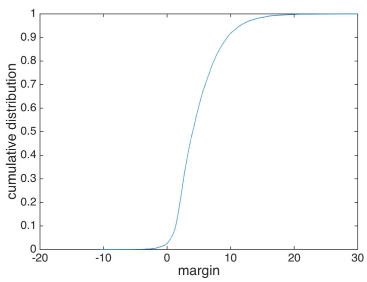
digit 0



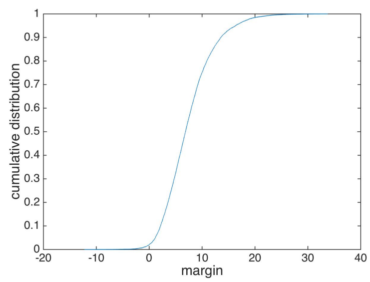
digit 1



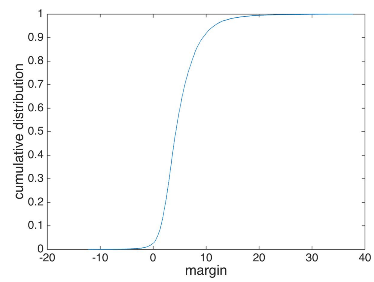
digit 2



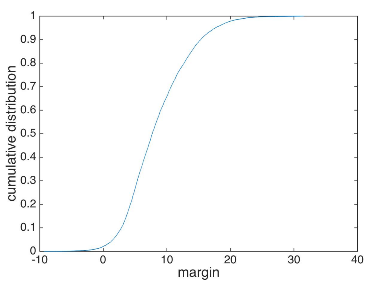
digit 3



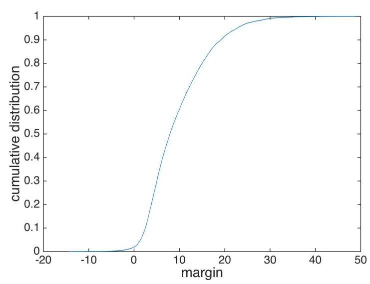
digit 4



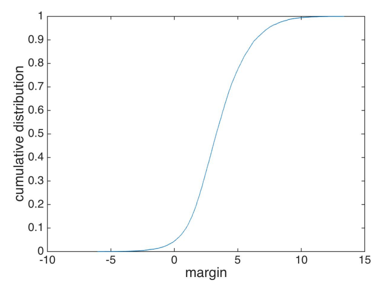
digit 5



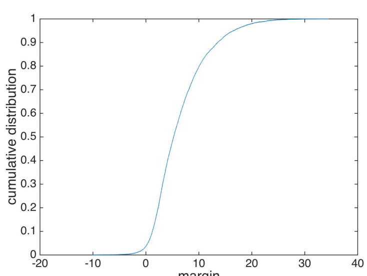
digit 6



digit 7



digit 8



digit 9

Problem 4.c)

1) After we run the script grid.py, we get the values of $C = 2$ and $\gamma = 0.0625$. The test accuracy for each digit is below. Note that the results are much better than for the linear SVM. This is a very common observation.

digit 0	digit 1	digit 2	digit 3	digit 4	digit 5	digit 6	digit 7	digit 8	digit 9	overall
99.53%	99.76%	98.88%	98.85%	99.15%	98.74%	99.30%	99.07%	98.37%	98.83%	97.42%

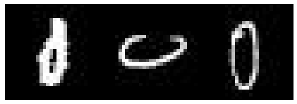
Table 7: Accuracy for each digit using radial basis function kernel, with $C = 2$ and $\gamma = 0.0625$.

2) The number of support vectors for each digit when $C = 2$ and $\gamma = 0.0625$ is below. This number is also much higher than for the linear SVM. This is the price you pay for the performance of the kernel SVM. It is not unheard of for a kernel-SVM to use 90% of the training set as support vectors, if the problem is really hard. This can be computationally very intensive.

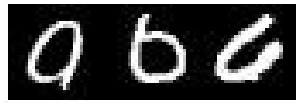
digit 0	digit 1	digit 2	digit 3	digit 4	digit 5	digit 6	digit 7	digit 8	digit 9
5861	2416	6892	7131	6302	6683	5648	5742	7672	6492

Table 8: Number of support vectors for each digit using radial basis function kernel, with $C = 2$ and $\gamma = 0.0625$.

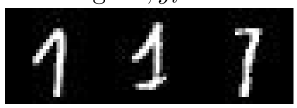
3) The support vectors with the 3 largest Lagrange multipliers for both sides of the discriminant plane are below (y_i is the label for i^{th} example), when $C = 2$.



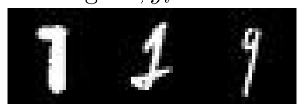
digit 0, $y_i = 1$



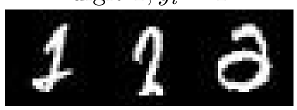
digit 0, $y_i = -1$



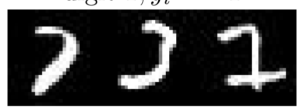
digit 1, $y_i = 1$



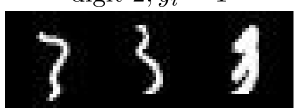
digit 1, $y_i = -1$



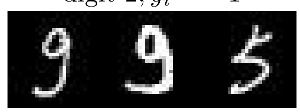
digit 2, $y_i = 1$



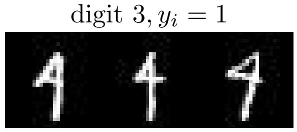
digit 2, $y_i = -1$



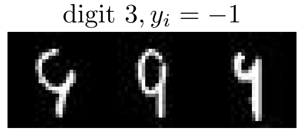
digit 3, $y_i = 1$



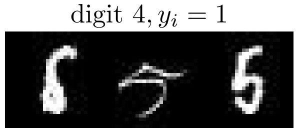
digit 3, $y_i = -1$



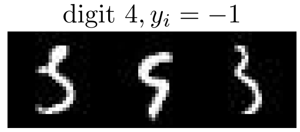
digit 4, $y_i = 1$



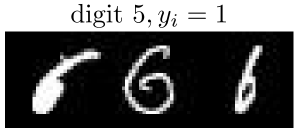
digit 4, $y_i = -1$



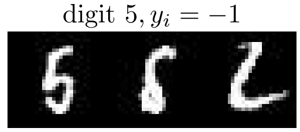
digit 5, $y_i = 1$



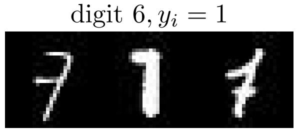
digit 5, $y_i = -1$



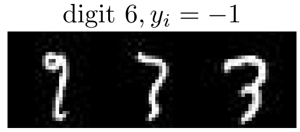
digit 6, $y_i = 1$



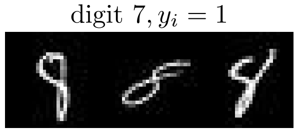
digit 6, $y_i = -1$



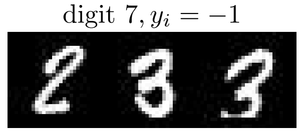
digit 7, $y_i = 1$



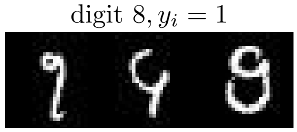
digit 7, $y_i = -1$



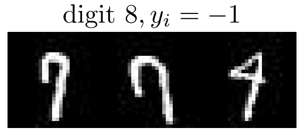
digit 8, $y_i = 1$



digit 8, $y_i = -1$

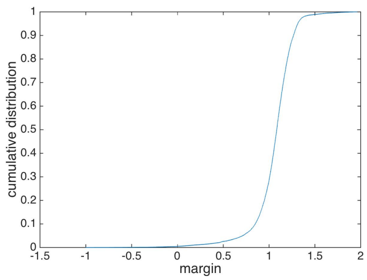


digit 9, $y_i = 1$

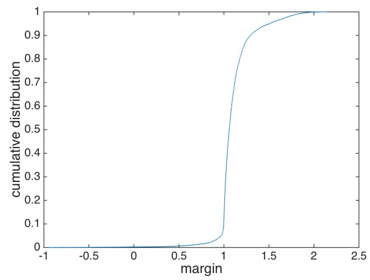


digit 9, $y_i = -1$

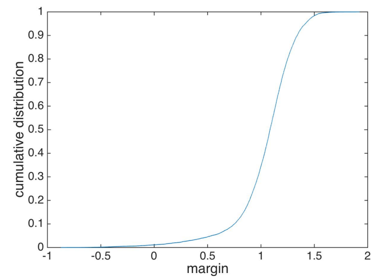
4) The plots of the cdf of the margin for every digit from 0 to 9 when $C = 2$ are as follows.



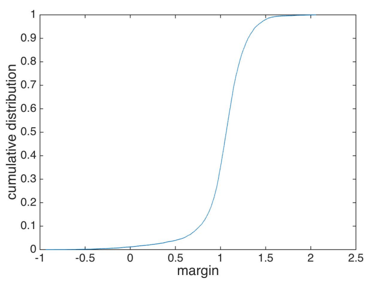
digit 0



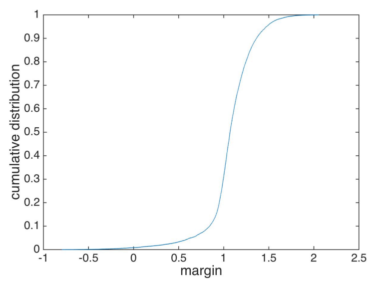
digit 1



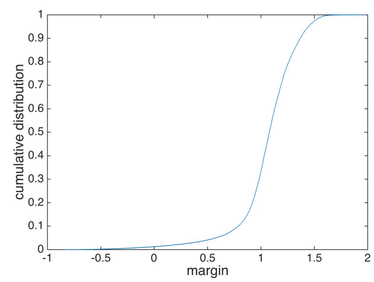
digit 2



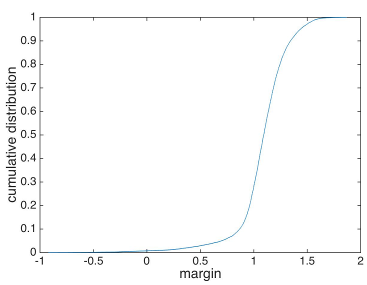
digit 3



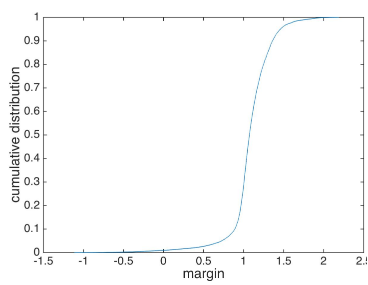
digit 4



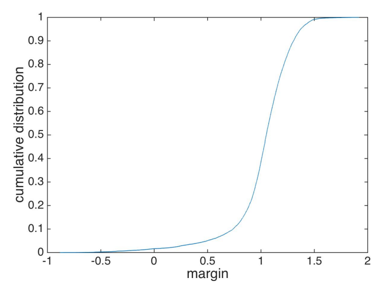
digit 5



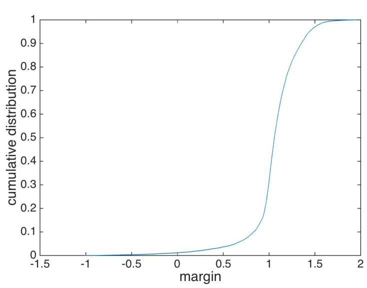
digit 6



digit 7



digit 8



digit 9