1. (e) A polynomial in x, y of order r has $r+1$ terms with power r. For example, $r = 2$ yields 3 quadratic terms: x^2, xy, and y^2. Therefore, a polynomial of order r has (excluding the constant term)
$$(1+1) + (2+1) + \cdots + (r-1+1) + (r+1) \qquad \text{terms.}$$
This is
$$r + \frac{r(r+1)}{2} \qquad \text{terms.}$$
In the case $r = 10$, this is
$$10 + 5 \cdot 11 = 65 \qquad \text{terms}$$

2. (d) In Logistic regression, we take a linear signal $s = \mathbf{w} \cdot \mathbf{x}$ and place it in a sigmoid and take a sign.
$$y = \operatorname{sgn}\left(\frac{1}{1+e^{-s}}\right).$$
This is clearly not linear, and so taking expectation values (which is linear) would not guarantee that our resulting signal is logistic.

3. (d) In overfitting, we tend to fit the noise instead of estimating results that are out-of-sample. This tends to give a small E_{in} and a large E_{out}. The goal is to get a small E_{out}, so we do not really care if we have a larger E_{in}. In fact, we will always choose a hypothesis that yields a larger E_{in} at the expense of a smaller E_{out}. As a result, measuring $E_{\text{out}} - E_{\text{in}}$ does not say if we are overfitting. Yes, it is true that as we overfit, $E_{\text{out}} - E_{\text{in}}$ increases. However, in the opposite spectrum, if we are underfitting, $E_{\text{out}} - E_{\text{in}}$ can be pretty small, however increasing the complexity of our hypothesis can increase $E_{\text{out}} - E_{\text{in}}$ while decreasing E_{out} simultaneously. Therefore, we cannot determine overfitting by only comparing $E_{\text{out}} - E_{\text{in}}$.

4. (d) Stochastic noise is noise due to random fluctuations and not due to the nature of the target function.

5. (a) $\mathbf{w}_{\text{lin}}$ minimizes the mean-squared error without any constraints. If $\mathbf{w}_{\text{lin}}$ minimizes the mean-squared error while also satisfying the constraint, then it must be the minimum in the constraint-region.

6. (b) Suppose we want to minimize
$$E_{\text{in}} = \frac{1}{N}(\mathbf{Zw} - \mathbf{y})^T(\mathbf{Zw} - \mathbf{y})$$
with respect to the constraint
$$\mathbf{w}^T \Gamma^T \Gamma \mathbf{w} \leq C.$$
From KKT Lagrange multipliers, we have
$$\nabla E_{\text{in}} \propto 2\Gamma^T \Gamma \mathbf{w}$$
implying that we can minimize without constraint
$$E_{\text{aug}} = \frac{1}{N}(\mathbf{Zw} - \mathbf{y})^T(\mathbf{Zw} - \mathbf{y}) + \frac{\lambda}{N}\mathbf{w}^T \Gamma^T \Gamma \mathbf{w}.$$

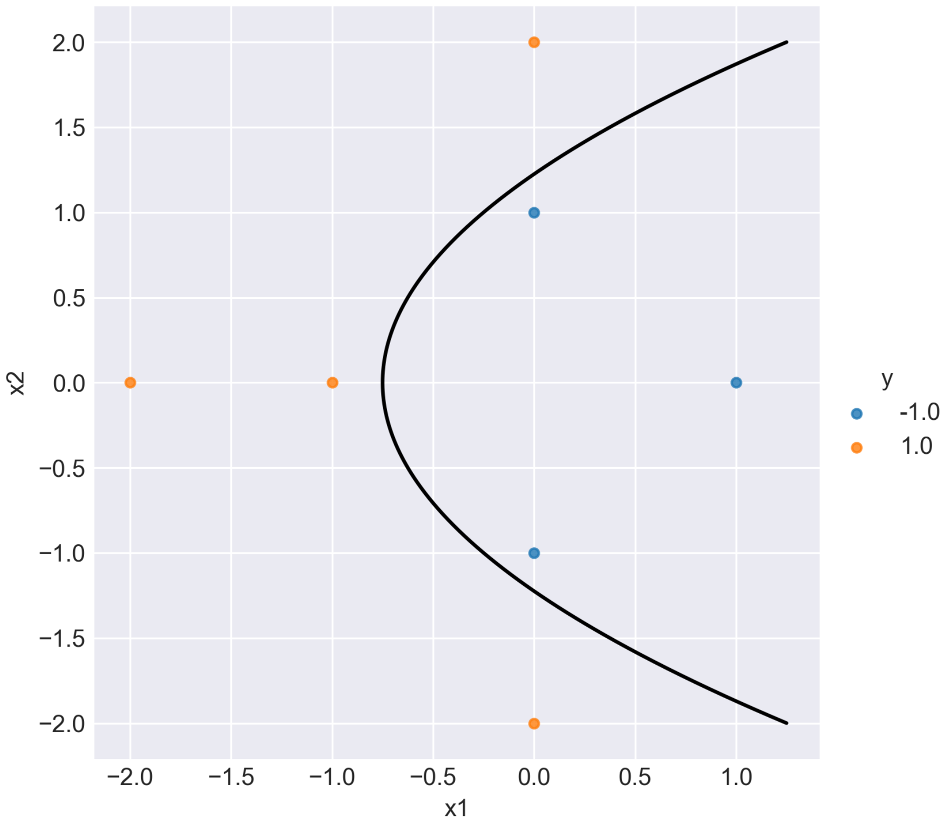

Figure 1: Problem 11

7. (d) 8 vs ALL had the lowest E_{in} with $\lambda = 1$ at 0.074.

8. (b) 1 vs ALL had the lowest E_{out} with $\lambda = 1$ at 0.02.

9. (e) $E_{\text{out}} = 0.07972$ without the transform and $E_{\text{out}} = 0.07922$ with the transform.

10. When comparing the 1 vs 5 classifier between $\lambda = 0.01$ and $\lambda = 1$, we see that the $\lambda = 0.01$ has a lower E_{in} but a higher E_{out}, implying that we have overfitting.

11. (c) We see from Figure 1 that our desired solution should look something like a parabola given that the transformation is quadratic. Given the symmetry of our datapoints, we should expect our separating curve to be symmetric about the $x_2 = 0$ axis. Solving $\mathbf{w}^T\mathbf{z} + b = 0$, we get

$$w_1(x_2^2 - 2x_1 - 1) + w_2(x_1^2 - 2x_2 + 1) + b = 0$$

By imposing our symmetry condition, the vertex of our parabola must be along the $x_2 = 0$ axis, so we force the linear term of x_2 to be zero so that $w_2 = 0$. Therefore,

$$2x_1 = x_2^2 + \frac{b}{w_1} - 1$$

We want to obtain the ratio b/w_1. The minimum value of this ratio would be when the vertex of the parabola touches $(x_1, x_2) = (-1, 0)$. This yields

$$\left(\frac{b}{w_1}\right)_{\text{min}} = -1$$

The maximum value of this ratio would be when the parabola hits $(0, \pm 1)$. This gives

$$\left(\frac{b}{w_1}\right)_{\text{max}} = 0$$

The separating curve given by SVMs that maximizes the margin would be the midpoint of these two ratios, so that

$$\left(\frac{b}{w_1}\right)_{\text{SVM}} = -0.5$$

so $w_1 = 1$ and $b = -0.5$.

12. The Lagrangian we would like to maximize is given by

$$\mathcal{L}(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m K_{nm}(\mathbf{x}_n^T \mathbf{x}_m)$$

where $K_{nm}(\mathbf{x}_n, \mathbf{x}_m) = (1 + \mathbf{x}_n^T \mathbf{x}_m)^2$ subject to the constraints

$$\alpha_n \geq 0 \qquad \text{and} \qquad \sum_{n=1}^{N} \alpha_n y_n = 0.$$

When running SVM on this data, we have 5 support vectors with values

$$\alpha_1 = 0, \alpha_2 = 0.4857, \alpha_3 = 0.92, \alpha_4 = 0.8887, \alpha_5 = 0.15, \alpha_6 = 0.36, \alpha_7 = 0.$$

13. (a) When training on only 100 samples, $E_{\text{in}} = 0$ with probably almost 1.

14. (e) The kernel form beats the regular form 91 percent of the time.

15. (d) The kernel form beats the regular form 83 percent of the time.

16. (d) On average, E_{in} and E_{out} both go down.

17. (c) On average, E_{in} and E_{out} both go up.

18. (a) E_{in} is almost never 0.

19. (b) The value of f is our parameter we would like to estimate. It follows some probability distribution $g(f)\,df$, and we have from Bayes' rule

$$g(f \mid x = 1)\,df = \frac{g(f) p(x = 1 \mid f)\,df}{p(x = 1)} = \frac{g(f) p(x = 1 \mid f)\,df}{\int_0^1 df\, p(x = 1 \mid f) g(f)}$$

Now $g(f) = 1$ since we assumed it's a continuous prior, and $p(x = 1 \,|\, f) = f$. Therefore,

$$g(f \,|\, x = 1)\, df = 2f\, df$$

so that

$$g(f \,|\, x = 1) = 2f$$

20. (c) The MSE of $g_1(\mathbf{x}), g_2(\mathbf{x})$ are

$$\text{MSE}(g_i(\mathbf{x})) = \mathbb{E}[(g_i - f)^2] = \mathbb{E}[g_i^2] + f^2 - 2\hat{g}_i f$$

So the average of these MSE's is

$$\frac{1}{2}\mathbb{E}[g_1^2 + g_2^2] + f^2 - (\hat{g}_1 + \hat{g}_2)f$$

If we take the MSE of $(g_1 + g_2)/2$, then

$$\begin{aligned} \text{MSE}\left[\frac{1}{2}(g_1 + g_2)\right] &= \mathbb{E}\left[\left(\frac{1}{2}(g_1 + g_2) - f\right)^2\right] \\ &= \mathbb{E}\left[\frac{1}{4}g_1^2 + \frac{1}{4}g_2^2 + \frac{1}{2}g_1 g_2 - (g_1 + g_2)f + f^2\right] \end{aligned}$$

If we subtract this from the average of the MSEs, then we get

$$\frac{1}{2}(\text{MSE}(g_1) + \text{MSE}(g_2)) - \text{MSE}\left[\frac{1}{2}(g_1 + g_2)\right] = \frac{1}{2}\mathbb{E}\left[g_1^2 + g_2^2 - g_1 g_2\right] \geq 0$$

since $x^2 + y^2 - xy \geq 0$ for all x, y.