1. Consider the optimization problem:

$$\min_x f(x), \text{ where } f(x) = x^4 - 8x^2 + 3x + 16.$$

a) Plot $f(x)$ between $x = -3$ and $x = 3$. Is this a convex problem?

b) Perform gradient descent for 2000 steps with a step size of $\eta = 0.001$, starting at $x_0 = 3$. That is,

$$x_{n+1} = x_n - \eta f'(x_n).$$

Plot the curve of $(x_n, f(x_n))$ on top of the plot you already generated. Does this algorithm approach the global minimum?

c) Now add random noise to the update. Use the following update rule:

$$x_{n+1} = x_n - \eta f'(x_n) + \xi_n,$$

where ξ_n is a Gaussian random variable with zero mean and variance σ_n^2. Let $\sigma_0 = 0.2$ and $\sigma_{n+1} = 0.999\sigma_n$ so that the variance of the noise slowly decays. Run this new algorithm several times – does it approach the global minimum?

2. Suppose a linear, hard-margin support vector machine must classify the points $\mathbf{r}^1 = (-1, -2)$, $\mathbf{r}^2 = (1, 3)$, and $\mathbf{r}^3 = (-1, 0)$ into one class $\ell = 1$, and the points $\mathbf{r}^4 = (2, -3)$, $\mathbf{r}^5 = (4, 2)$ and $\mathbf{r}^6 = (1, -2)$ into another class $\ell = -1$.

a) Plot the points belonging to the two classes (using two different colors) in a 2-d plane.

b) If $\mathbf{w}$ and θ define a linear classifier $\ell = \text{sign}(\mathbf{w} \cdot \mathbf{r} - \theta)$, write the set of constraints that these parameters must satisfy if the classifier is to correctly classify the points.

c) Write the expression for the parameters of the SVM solution in the standard form for a quadratic optimization problem:

$$\mathbf{x}^* = \text{argmin}_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T Q \mathbf{x} + \mathbf{c}^T \mathbf{x}, \text{ s.t } A\mathbf{x} \leq \mathbf{b}.$$

That is, determine the values of the matrices Q and A, and vectors $\mathbf{c}$ and $\mathbf{b}$, that make this optimization problem equivalent to the optimal solution for the SVM. Hint: $\mathbf{x}$ should be a length three vector corresponding to the parameters that define the classifier.

c) **Optional:** Use a numerical solver such as MATLAB's `quadprog` to find the solution to this problem, and plot the classification boundary in the 2-d plane from part (a).

d) **Optional:** We can define the "Lagrangian dual" of an optimization problem that corresponds to maximizing over Lagrange multipliers rather than minimizing over $\mathbf{x}$. Let's do this for a linear, hard-margin SVM.

If we define Lagrange multipliers for the inequality constraints of an SVM, we get a Lagrangian:

$$\mathcal{L}(\mathbf{w}, \theta, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_\mu \lambda_\mu[(\mathbf{w} \cdot \mathbf{r}^\mu - \theta)\ell^\mu - 1].$$

The Lagrangian dual function is defined as:

$$G(\boldsymbol{\lambda}) = \inf_{\mathbf{w},\theta} \mathcal{L}(\mathbf{w}, \theta, \boldsymbol{\lambda}).$$

If $\lambda_\mu \geq 0$ for each μ, then $G(\boldsymbol{\lambda}) \leq f^*$, where f^* is the minimum of $\mathbf{w}^T\mathbf{w}$. This is because each term in the sum over constraints in $\mathcal{L}$ must be positive if $\lambda_\mu \geq 0$ and the constraint is satisfied. Thus, the maximum of G gives us a lower bound for f^*.

By taking partial derivatives of $\mathcal{L}$ with respect to w_i and θ and setting them equal to zero, show that at the infimum:

$$\mathbf{w} - \sum_\mu \lambda_\mu \mathbf{r}^\mu \ell^\mu = 0, \tag{1}$$

$$\sum_\mu \lambda_\mu \ell^\mu = 0. \tag{2}$$

Note that the first expression tells us that $\mathbf{w}$ is a sum over the $\mathbf{r}^\mu$ for which $\lambda_\mu \neq 0$: the "support vectors."

e) **Optional:** Use the expressions you found in (d) and substitute them into $\mathcal{L}$ to show that:

$$G(\boldsymbol{\lambda}) = -\frac{1}{2}\sum_{\mu,\nu} \lambda_\mu \lambda_\nu \ell^\mu \ell^\nu (\mathbf{r}^\mu)^T \mathbf{r}^\nu + \sum_\mu \lambda^\mu. \tag{3}$$

This implies that we can obtain a lower bound for f^* with the following optimization problem:

$$\max_{\boldsymbol{\lambda}} -\frac{1}{2}\sum_{\mu,\nu} \lambda_\mu \lambda_\nu \ell^\mu \ell^\nu (\mathbf{r}^\mu)^T \mathbf{r}^\nu + \sum_\mu \lambda^\mu, \text{ s.t. } \sum_\mu \lambda_\mu \ell^\mu = 0, \ \lambda_\mu \geq 0. \tag{4}$$

In fact, because the original problem is convex, this lower bound is tight and gives us the optimal SVM solution. This is an optimization over P variables, where P is the number of data points (constraints), and is therefore an improvement when there are fewer data points than dimensions. It also depends only on the dot products $(\mathbf{r}^\mu)^T\mathbf{r}^\nu$ rather than the $\mathbf{r}$s explicitly, which leads naturally to kernel methods, to be discussed later.