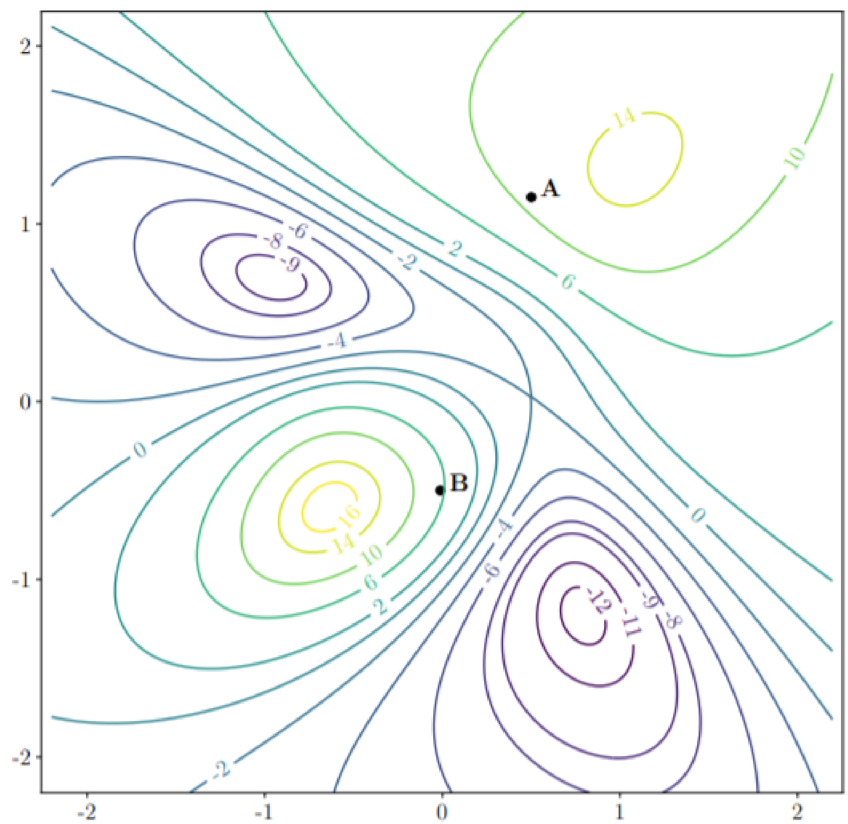
DS-GA 1014 - Homework 12

Eric Niblock

December 7th, 2020

1. (2 points). The following plot shows the contour lines of a function $f : \mathbb{R}^2 \to \mathbb{R}$.



(a) Give (approximately) the coordinates of the global/local minimizers/maximizers, saddle points of f.

Here are the following critical points of f:

- Global Maximizer: $\sim (-\frac{2}{3}, -\frac{2}{3})$
- Local Maximizers: $\sim (-\frac{2}{3}, -\frac{2}{3})$; $\sim (\frac{5}{4}, \frac{5}{4})$
- Global Minimizer: $\sim (\frac{4}{5}, -\frac{5}{4})$
- Local Minimizers: $\sim (\frac{4}{5}, -\frac{5}{4})$; $\sim (-\frac{7}{8}, \frac{2}{3})$
- Saddle Point: $\sim (\frac{1}{2}, 0)$

(b) **Assume that we run gradient descent to minimize f. Will gradient descent converge to the global minimizer of f when initialized at point A? At point B?**

When beginning at point A, gradient descent will not converge to the global minimizer, but instead to the local minimizer at $\sim (-\frac{7}{8}, \frac{2}{3})$.

When beginning at point B, gradient descent will converge to the global minimizer at $\sim (\frac{4}{5}, -\frac{5}{4})$.

2. (5 points). Let $M \in \mathbb{R}^{\mathrm{d}\times\mathrm{d}}$ be a positive semidefinite matrix, $\vec{\mathbf{b}} \in \mathbb{R}^{\mathrm{d}}$ and $c \in \mathbb{R}$. We aim at minimizing the quadratic function

$$f(\vec{\mathbf{x}}) = \tfrac{1}{2}\vec{\mathbf{x}}^T M \vec{\mathbf{x}} - \langle \vec{\mathbf{x}}, \vec{\mathbf{b}} \rangle + c$$

using gradient descent. We assume that M is positive definite (i.e. all its eigenvalues are positive). We let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d > 0$ be its eigenvalues and let $\vec{\mathbf{v}}_1, ..., \vec{\mathbf{v}}_d$ be an orthonormal basis of $\mathbb{R}^{\mathrm{d}}$ consisting of associated eigenvectors ($M\vec{\mathbf{v}}_i = \lambda_i \vec{\mathbf{v}}_i$ for all i). We write $L = \lambda_1$ and $= \lambda_d$.

We consider standard gradient descent with constant step-size β:

$$\vec{\mathbf{x}}_{t+1} = \vec{\mathbf{x}}_t - \beta \nabla f(\vec{\mathbf{x}}_t)$$

(a) **Show that f is L-smooth, μ-strongly convex and that $\vec{\mathbf{x}}^* = M^{-1}\vec{\mathbf{b}}$ is the unique minimizer of f.**

We have that,

$$\nabla f(\vec{\mathbf{x}}) = \frac{1}{2}\left(M + M^T\right)\vec{\mathbf{x}} - \vec{\mathbf{b}} \tag{1}$$

$$H_f(\vec{\mathbf{x}}) = \frac{1}{2}(M^T + M) \tag{2}$$

And since this Hessian is independent of $\vec{\mathbf{x}}$, there is some constant $\lambda_{max}(H_f(x))$ associated with f, which makes it obvious that we can find some L such that $\lambda_{max}(H_f(x)) \leq L$. This shows that f is L-smooth. By the same argument, it is obvious that we can find some μ such that $\lambda_{min}(H_f(x)) \geq \mu$. This shows that f is μ-strongly convex.

Now, we know that M is symmetric (because it is assumed to be positive definite) which yields,

$$\nabla f(\vec{\mathbf{x}}) = M\vec{\mathbf{x}} - \vec{\mathbf{b}} \tag{3}$$

And since we know that $f(\vec{\mathbf{x}})$ is at least convex, this implies that there exists a minimum. Setting the gradient equal to zero, we find,

$$\vec{\mathbf{x}}^* = M^{-1}\vec{\mathbf{b}} \tag{4}$$

Where we know that M is invertible, again, because it is assumed to be positive definite, and positive definite matrices are invertible.

(b) We now study the convergence of gradient descent to $\vec{\mathbf{x}}^*$. Show that for all $t \geq 0$,

$$\vec{\mathbf{x}}_{t+1} - \vec{\mathbf{x}}^* = (Id - \beta M)(\vec{\mathbf{x}}_t - \vec{\mathbf{x}}^*)$$

This follows readily by simplification,

$$\begin{aligned}
\vec{\mathbf{x}}_{t+1} - \vec{\mathbf{x}}^* &= \vec{\mathbf{x}}_t - \beta\nabla f(\vec{\mathbf{x}}_t) - \vec{\mathbf{x}}^* \\
&= \vec{\mathbf{x}}_t - \beta\left(M\vec{\mathbf{x}}_t - \vec{\mathbf{b}}\right) - \vec{\mathbf{x}}^* \\
&= \vec{\mathbf{x}}_t - \beta\left(M\vec{\mathbf{x}}_t - MM^{-1}\vec{\mathbf{b}}\right) - \vec{\mathbf{x}}^* \\
&= \vec{\mathbf{x}}_t - \beta\left(M\vec{\mathbf{x}}_t - M\vec{\mathbf{x}}^*\right) - \vec{\mathbf{x}}^* \\
&= \vec{\mathbf{x}}_t - \vec{\mathbf{x}}^* - \beta M\vec{\mathbf{x}}_t + \beta M\vec{\mathbf{x}}^* \\
&= (Id - \beta M)(\vec{\mathbf{x}}_t - \vec{\mathbf{x}}^*)
\end{aligned} \tag{5}$$

(c) From now, we set $\beta = \frac{1}{L}$. Deduce from the previous question that for all $t \geq 0$,

$$||\overrightarrow{\mathbf{x}}_t - \overrightarrow{\mathbf{x}}^*|| \leq \left(1 - \tfrac{\mu}{L}\right)^t ||\overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^*||$$

We have that,

$$\overrightarrow{\mathbf{x}}_{t+1} - \overrightarrow{\mathbf{x}}^* = (Id - \beta M)(\overrightarrow{\mathbf{x}}_t - \overrightarrow{\mathbf{x}}^*) \tag{6}$$

Then we can form the closed-solution,

$$\begin{aligned}
\overrightarrow{\mathbf{x}}_1 - \overrightarrow{\mathbf{x}}^* &= (Id - \beta M)(\overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^*) \\
\overrightarrow{\mathbf{x}}_2 - \overrightarrow{\mathbf{x}}^* &= (Id - \beta M)^2(\overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^*) \\
&\vdots \\
\overrightarrow{\mathbf{x}}_t - \overrightarrow{\mathbf{x}}^* &= (Id - \beta M)^t(\overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^*)
\end{aligned} \tag{7}$$

Then we can apply the norm to each side,

$$||\overrightarrow{\mathbf{x}}_t - \overrightarrow{\mathbf{x}}^*|| = ||(Id - \beta M)^t(\overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^*)|| \tag{8}$$

We have shown previously that $||A\overrightarrow{\mathbf{x}}|| \leq ||A||_{Sp}||\overrightarrow{\mathbf{x}}||$. In our case, this means,

$$||\overrightarrow{\mathbf{x}}_t - \overrightarrow{\mathbf{x}}^*|| = ||(Id - \beta M)^t(\overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^*)|| \leq ||(Id - \beta M)^t||_{Sp}||(\overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^*)|| \tag{9}$$

We also know that the eigenvalues of M are $L = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d = \mu$. Therefore the eigenvalues of $(Id - \beta M)^t$ are given by, $0 \leq ... \leq (1 - \frac{\mu}{L})^t$. Since all of the eigenvalues are non-zero, the spectral norm will just be the largest eigenvalue of $(Id - \beta M)^t$. Therefore, we have that,

$$||\overrightarrow{\mathbf{x}}_t - \overrightarrow{\mathbf{x}}^*|| = ||(Id - \beta M)^t(\overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^*)|| \leq (1 - \frac{\mu}{L})^t \, ||(\overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^*)|| \tag{10}$$

$$||\overrightarrow{\mathbf{x}}_t - \overrightarrow{\mathbf{x}}^*|| \leq (1 - \frac{\mu}{L})^t \, ||(\overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^*)|| \tag{11}$$

(d) We would like now to have something more precise than the error bound of the previous question. We define $\overrightarrow{\mathbf{w}}_t \overset{def}{=} \overrightarrow{\mathbf{x}}_t - \overrightarrow{\mathbf{x}}^*$. Let,

$$\alpha_1(t) = \langle \overrightarrow{\mathbf{v}}_1, \overrightarrow{\mathbf{w}}_t \rangle, ..., \alpha_d(t) = \langle \overrightarrow{\mathbf{v}}_d, \overrightarrow{\mathbf{w}}_t \rangle$$

be the coordinates of $\overrightarrow{\mathbf{w}}_t$ in the orthonormal basis $(\overrightarrow{\mathbf{v}}_1, ..., \overrightarrow{\mathbf{v}}_d)$. For $i \in \{1, ..., d\}$, express $\alpha_i(t)$ in terms of t, λ_i, L and $\alpha_i(0)$.

Since we have that,

$$\alpha_i(t) = \langle \overrightarrow{\mathbf{v}}_i, \overrightarrow{\mathbf{x}}_t - \overrightarrow{\mathbf{x}}^* \rangle \tag{12}$$

It is clear that $(\alpha_1(t), ..., \alpha_d(t))$ are the coordinates of $\overrightarrow{\mathbf{w}}_t$ in the orthonormal basis of $(\overrightarrow{\mathbf{v}}_1, ..., \overrightarrow{\mathbf{v}}_d)$. Furthermore, we have,

$$\overrightarrow{\mathbf{w}}_t = \begin{bmatrix} \alpha_1(t) \\ \vdots \\ \alpha_d(t) \end{bmatrix} = (Id - \beta M)^t \begin{bmatrix} \alpha_1(0) \\ \vdots \\ \alpha_d(0) \end{bmatrix} = (Id - \beta M)^t \overrightarrow{\mathbf{w}}_0 \tag{13}$$

Which implies that,

$$\alpha_1(t)\overrightarrow{\mathbf{v}}_1 + ... + \alpha_d(t)\overrightarrow{\mathbf{v}}_d = \alpha_1(0)(Id - \beta M)^t \overrightarrow{\mathbf{v}}_1 + ... + \alpha_d(0)(Id - \beta M)^t \overrightarrow{\mathbf{v}}_d \tag{14}$$

$$\alpha_1(t)\overrightarrow{\mathbf{v}}_1 + ... + \alpha_d(t)\overrightarrow{\mathbf{v}}_d = \alpha_1(0)(1 - \frac{\lambda_1}{L})^t \overrightarrow{\mathbf{v}}_1 + ... + \alpha_d(0)(1 - \frac{\lambda_d}{L})^t \overrightarrow{\mathbf{v}}_d \tag{15}$$

Therefore,

$$\vec{\mathbf{w}}_t = \begin{bmatrix} \alpha_1(t) \\ \vdots \\ \alpha_d(t) \end{bmatrix} = \begin{bmatrix} \alpha_1(0)(1-\frac{\lambda_1}{L})^t \\ \vdots \\ \alpha_d(0)(1-\frac{\lambda_d}{L})^t \end{bmatrix} \tag{16}$$

By the uniqueness of coordinates in a basis ($\vec{\mathbf{v}}_1, ..., \vec{\mathbf{v}}_d$). So, in general, we have that,

$$\alpha_i(t) = \alpha_i(0)(1-\frac{\lambda_i}{L})^t \tag{17}$$

(e) Using the previous question, justify the following sentence:

"Gradient descent converges towards the minimizer faster in directions given by the eigenvectors of the Hessian of f corresponding to large eigenvalues than in directions corresponding to eigenvectors with small eigenvalues"

We know that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$ and that,

$$\vec{\mathbf{w}}_t = \vec{\mathbf{x}}_t - \vec{\mathbf{x}}^* = \begin{bmatrix} \alpha_1(t) \\ \vdots \\ \alpha_d(t) \end{bmatrix} = \begin{bmatrix} \alpha_1(0)(1-\frac{\lambda_1}{L})^t \\ \vdots \\ \alpha_d(0)(1-\frac{\lambda_d}{L})^t \end{bmatrix} \tag{18}$$

Furthermore since,

$$0 \leq 1 - \frac{\lambda_1}{L} \leq ... \leq 1 - \frac{\lambda_d}{L} \leq 1 \tag{19}$$

$$0 \leq \left(1 - \frac{\lambda_1}{L}\right)^t \leq ... \leq \left(1 - \frac{\lambda_d}{L}\right)^t \leq 1 \tag{20}$$

Or, in other words,

$$\left|\frac{\partial}{\partial t}\left(1 - \frac{\lambda_1}{L}\right)^t\right| \geq ... \geq \left|\frac{\partial}{\partial t}\left(1 - \frac{\lambda_d}{L}\right)^t\right| \tag{21}$$

Which suggests that as t increases, $\alpha_1(t)$ updates more drastically that $\alpha_2(t)$, which updates more drastically than $\alpha_3(t)$, ... Therefore, the statement has been shown.

(f) Show that for all $t \geq 0$

$$||\overrightarrow{\mathbf{x}}_t - \overrightarrow{\mathbf{x}}^*|| = \sqrt{\sum_{i=1}^{d} \left(1 - \frac{\lambda_i}{L}\right)^{2t} \langle \overrightarrow{\mathbf{v}}_i, \overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^* \rangle^2} \tag{22}$$

We know that,

$$\begin{aligned} ||\overrightarrow{\mathbf{x}}_t - \overrightarrow{\mathbf{x}}^*|| = ||\overrightarrow{\mathbf{w}}_t|| &= \sqrt{\sum_{i=1}^{d} \alpha_i(0)^2 (1 - \frac{\lambda_i}{L})^{2t}} \\ &= \sqrt{\sum_{i=1}^{d} (1 - \frac{\lambda_i}{L})^{2t} \langle \overrightarrow{\mathbf{v}}_i, \overrightarrow{\mathbf{x}}_0 - \overrightarrow{\mathbf{x}}^* \rangle^2} \end{aligned} \tag{23}$$

Which is what we hoped to show.

3. (3 points). In this problem, you will implement and compare gradient descent with or without momentum to minimize the Ridge cost function.

The corresponding PDF is attached.

In [2]:
```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
plt.rc('font',family='serif')
```

In [3]:
```
d=1000 # d: dimension
n=2000 # n: number of points
A = np.random.normal(size=(n,d)) / np.sqrt(n) # matrix containing the data points
y = np.random.normal(size=n)
lambd = 1
I = np.identity(d)
```

We consider the Ridge cost function:

$$f(x) = \frac{1}{2}\|Ax - y\|^2 + \frac{\lambda}{2}\|x\|^2,$$

where $\lambda > 0$ is some regularization parameter that we take equal to 1. The matrix A and the vector y are defined in the cell above.

(a) Show that f is can be written in the format the function f of Problem 12.2, for some $M \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$ and $c \in \mathbb{R}$. Compute numerically the values of L and μ. Plot the eigenvalues of $H_f(x)$ using an histogram.

We can write $f(x)$ in the form of $f(x) = \frac{1}{2}x^T Mx - \langle x, b \rangle$ + c. Observe the following,

$$f(x) = \frac{1}{2}\|Ax - y\|^2 + \frac{\lambda}{2}\|x\|^2$$

$$f(x) = \frac{1}{2}(x^T A^T - y^T)(Ax - y) + \frac{\lambda}{2}x^T x$$

$$f(x) = \frac{1}{2}(x^T A^T Ax - y^T Ax - x^T A^T y + y^T y) + \frac{\lambda}{2}x^T x$$

$$f(x) = \frac{1}{2}(x^T A^T Ax - 2x^T A^T y + y^T y) + \frac{\lambda}{2}x^T x$$

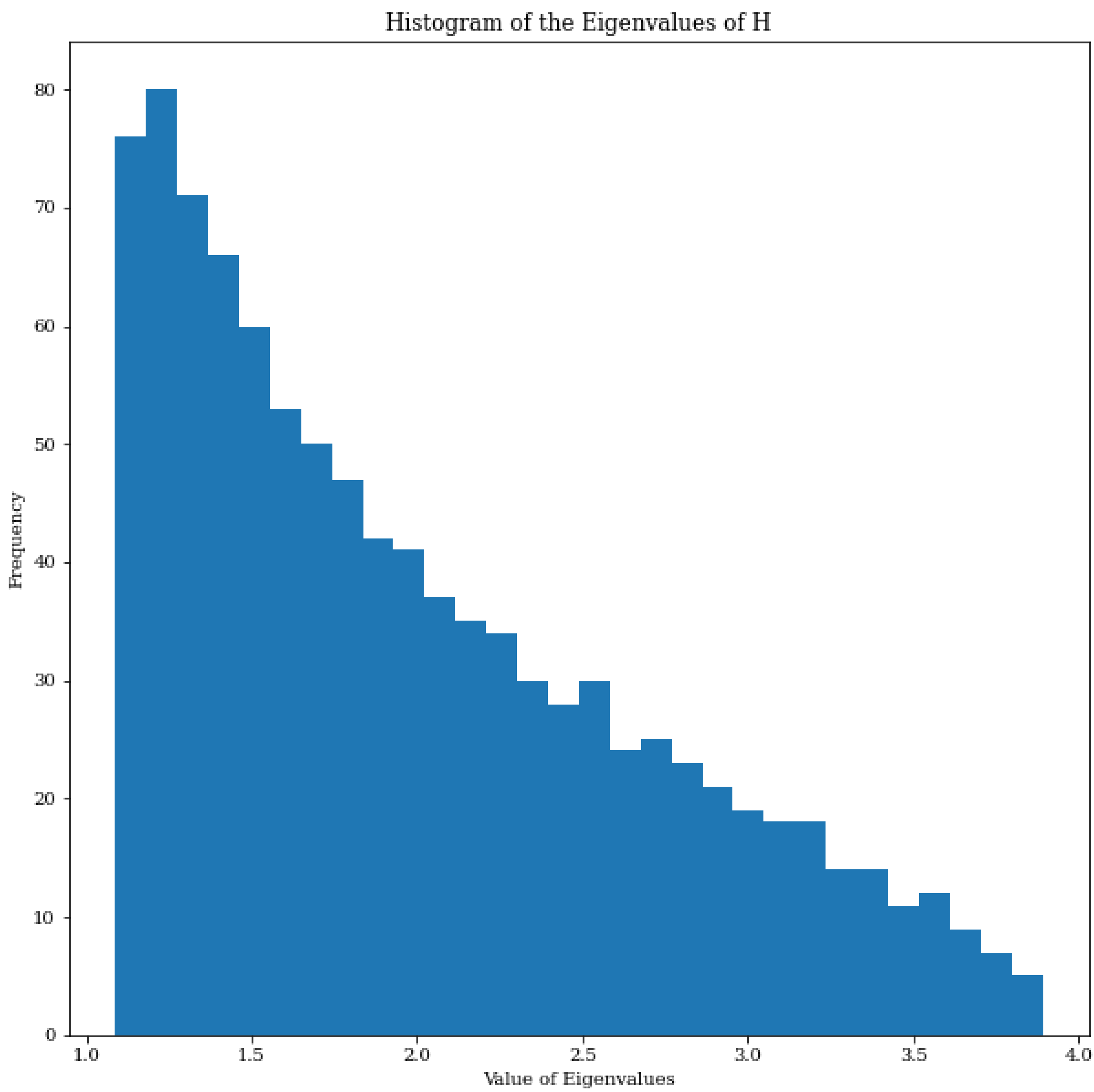
$$f(x) = \frac{1}{2}x^T(A^T A + \lambda Id)x - x^T A^T y + \frac{y^T y}{2}$$

$$f(x) = \frac{1}{2}x^T(A^T A + \lambda Id)x - \langle x, A^T y \rangle + \frac{y^T y}{2}$$

Then we see that $M = A^T A + \lambda Id$, $b = A^T y$, $c = \frac{y^T y}{2}$

In [27]:
```
H = A.T@A + lambd*I
vals,vect = np.linalg.eigh(H)
L = np.max(vals)
u = np.min(vals)
plt.figure(figsize=(10,10))
plt.hist(vals, bins=30)[2]
plt.title('Histogram of the Eigenvalues of H')
plt.xlabel('Value of Eigenvalues')
plt.ylabel('Frequency')
```

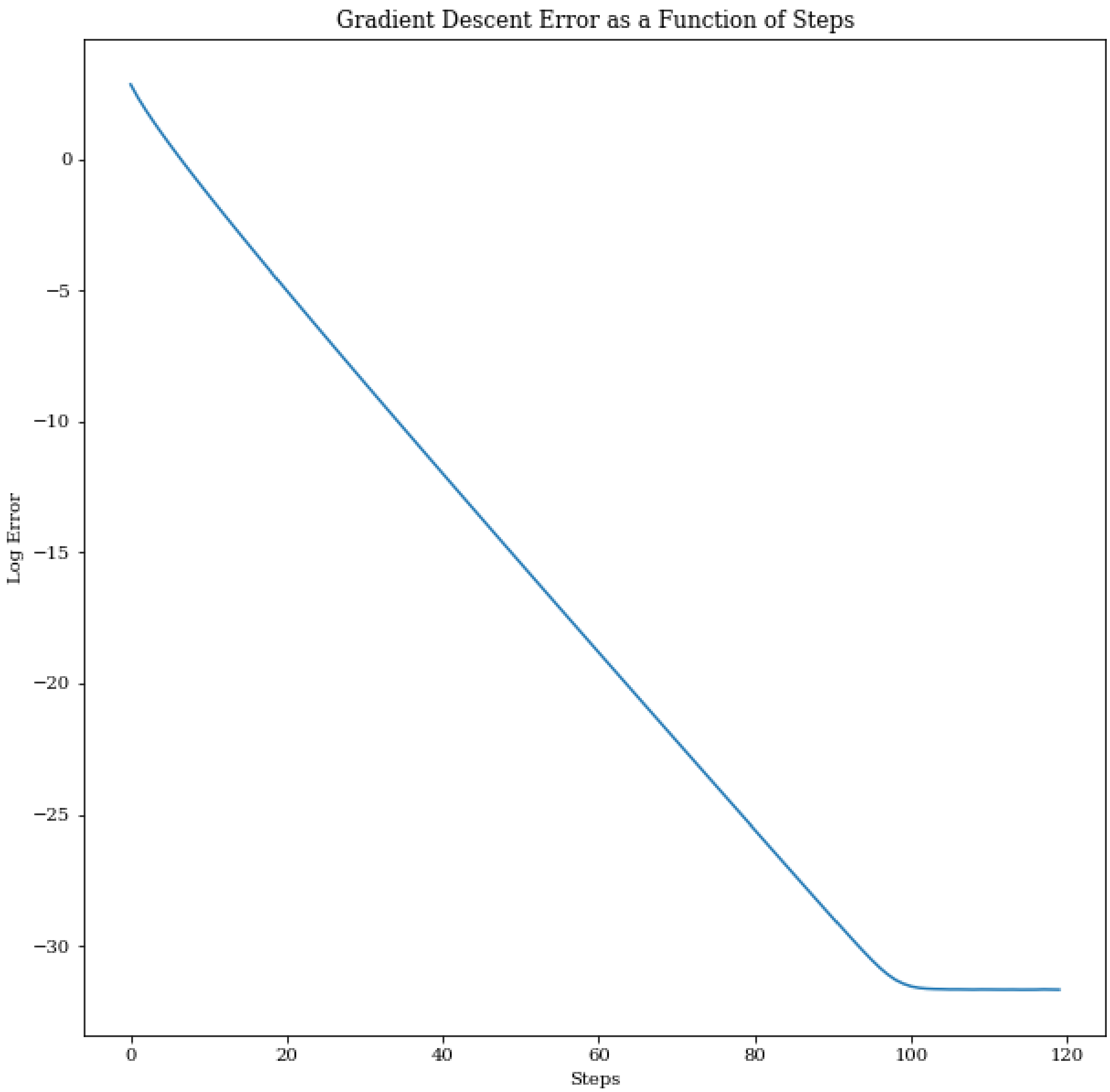
Out[27]:
```
Text(0, 0.5, 'Frequency')
```



(b) Implement gradient descent with constant step-size $\beta = 1/L$ (as in Problem 12.2), with random initial position x_0. Plot the log-error $\log(\|x_t - x_*\|)$ as a function of t.

In [24]:
```
steps = []
logerror = []
x = np.random.normal(size=d)
xmin = np.linalg.inv(H)@A.T@y
for i in range(120):
    x = x - ((1/L)*(H@x - A.T@y))
    steps.append(i)
    logerror.append(np.log(np.linalg.norm(x-xmin)))
plt.figure(figsize=(10,10))
plt.plot(steps,logerror)
plt.xlabel('Steps')
plt.ylabel('Log Error')
plt.title('Gradient Descent Error as a Function of Steps')
```

Out[24]:
```
Text(0.5, 1.0, 'Gradient Descent Error as a Function of Steps')
```

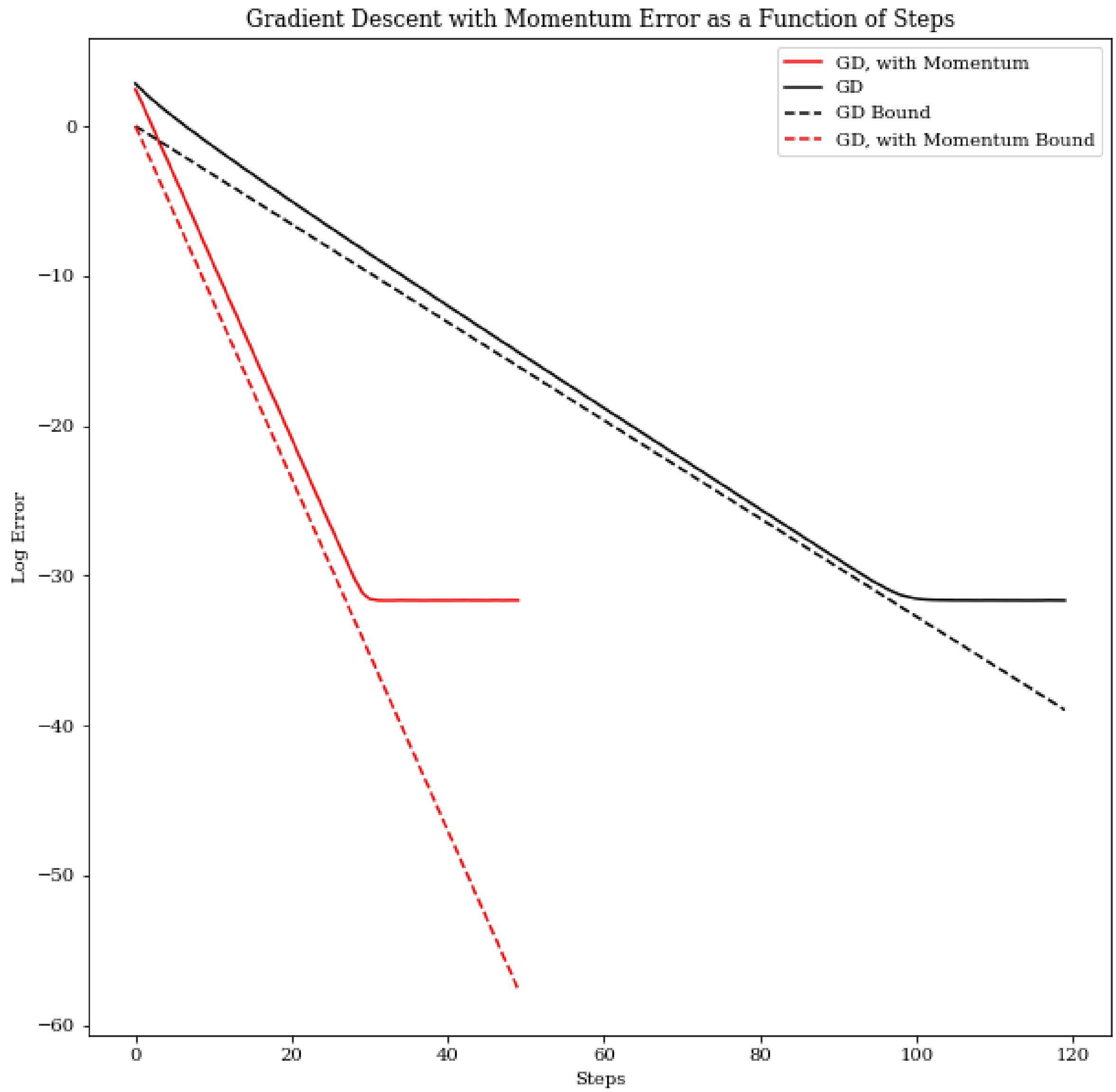


(c) Implement gradient descent with momentum, with the same parameters as in Problem 12.4. Plot the log-error $\log(\|x_t - x_*\|)$ as a function of t, on the same plot than the log-error of gradient descent without momentum. On the same plot, plot also the lines of equation

$$y = \log(1 - \mu/L) \times t \qquad \text{and} \qquad y = \log\left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right) \times t.$$

In [25]:
```
steps_log = []
logerror_log = []
x = np.random.normal(size=d)
xmin = np.linalg.inv(H)@A.T@y
xold = x
beta = 4/(((L**0.5)+(u**0.5))**2)
gamma = (((L**0.5)-(u**0.5))/((L**0.5)+(u**0.5)))**2
for i in range(120):
    temp = x - beta*(H@x - A.T@y) + gamma*(x - xold)
    xold=x
    x=temp
    steps_log.append(i)
    logerror_log.append(np.log(np.linalg.norm(x-xmin)))
plt.figure(figsize=(10,10))
plt.plot(steps_log[:50],logerror_log[:50], c='r', label='GD, with Momentum')
plt.plot(steps,logerror, c='k',label='GD')
plt.plot(steps_log, np.log(1-(u/L))*np.array(steps_log),'k--',label='GD Bound')
plt.plot(steps_log[:50], np.log(((L**0.5)-(u**0.5))/((L**0.5)+(u**0.5)))*np.array
         ,'r--',label='GD, with Momentum Bound')
plt.xlabel('Steps')
plt.ylabel('Log Error')
plt.title('Gradient Descent with Momentum Error as a Function of Steps')
plt.legend()
```

Out[25]: <matplotlib.legend.Legend at 0x12558e8f860>


Gradient Descent with Momentum Error as a Function of Steps
GD, with Momentum
GD
GD Bound
GD, with Momentum Bound
Log Error
0
−10
−20
−30
−40
−50
−60
0
20
40
60
80
100
120
Steps