Homework 7
CS-526 Learning Theory

3. Variants of standard gradient descent; forward and backward schemes

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex Lipshitz continuous differentiable function with Lipshitz constant $\rho > 0$. Let S be a real symmetric strictly positive-definite $d \times d$ matrix with smallest eigenvalue $\lambda_{\min} > 0$. We consider a gradient descent iteration for $t \geq 1$ and step size $\eta > 0$:

$$x^{t+1} = x^t - \eta S^{-1} \nabla f(x^t) \tag{1}$$

with initial condition $x^1 = 0$. Further, define $x^* = \text{argmin}_{\|x\| \in B(0,R)} f(x)$, where $B(0, R)$ is the ball of radius R.

1. The update equation (**??**) is in the form of an Euler *forward* scheme. Write down the associated *backward* Euler scheme.

2. Consider the following iterations (assume the argmin exists and is unique)

$$x^{t+1} = \text{argmin}_x \left\{ f(x) + \frac{1}{2\eta}(x - x^t)^T S (x - x^t) \right\}$$

   Is the quantity in the bracket simply convex or strictly convex ? Show that this iteration is equivalent to one of the two Euler schemes.

3. Show that if we choose the step size $\eta = \frac{R\sqrt{\lambda_{\max}\lambda_{\min}}}{\rho\sqrt{T}}$ after T iterations we have

$$f\big(\frac{1}{T}\sum_{t=1}^{T} x^t\big) - f(x^*) \leq \frac{\rho R}{\sqrt{T}} \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

   Hint: recall that in class we proved this statement when $S = I$ the identity matrix. Here you can use an eigenvalue decomposition $S^{-1} = U^T \Lambda^{-1} U$. The following is also useful:

$$\left\langle \nabla f\left(x^t\right), x^t - x^* \right\rangle = \left\langle U \nabla f\left(x^t\right), U x^t - U x^* \right\rangle = \sum_{k=1}^{d} (U \nabla f)_k (x^t) \left(U x^t - U x^*\right)_k$$

   Justify why these steps can be used.