Solution 7
CS-526 Learning Theory

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

1. The backward Euler scheme is

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

2. The first term f is convex. The second term is strictly convex because S is positive definite (with $\lambda_{min} > 0$). Thus the sum is strictly convex.

   Since f is differentiable we can differentiate the gradient of the quantity in the bracket in order to find the argmin:

$$\nabla f(x) + \eta^{-1}S(x - x^t) = 0$$

   which implies the backward Euler scheme:

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

3. Let $S^{-1} = U^T\Lambda^{-1}U$ with U an orthogonal matrix, and $\Lambda = \mathrm{Diag}\,(\lambda_1 \cdots \lambda_d)$. With $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x^t$, we have

$$\begin{aligned}
f(\bar{x}) - f\left(x^*\right) &\leqslant \frac{1}{T}\sum_{t=1}^{T}\left(f\left(x^t\right) - f\left(x^*\right)\right) \qquad \text{convexity} \\
&\leqslant \frac{1}{T}\sum_{t=1}^{T}\left\langle \nabla f\left(x^t\right), x^t - x^*\right\rangle \qquad \text{convexity} \\
&= \frac{1}{T}\sum_{t=1}^{T}\left\langle U\nabla f\left(x^t\right), Ux^t - Ux^*\right\rangle \\
&= \sum_{k=1}^{d}\frac{1}{T}\sum_{t=1}^{T}(U\nabla f)_k(x^t)\left(U\left(x^t - x^*\right)\right)_k \\
&= \sum_{k=1}^{d}\frac{\lambda_k}{\eta T}\sum_{t=1}^{T}\left(\frac{\eta}{\lambda_k}\right)(U\nabla f)_k(x^t)\left(U\left(x^t - x^*\right)\right)_k
\end{aligned}$$

$$= \sum_{k=1}^{d}\frac{\lambda_k}{2\eta T}\sum_{t=1}^{T}\left\{-\left(\left(U\left(x^t - x^*\right)\right)_k - \frac{\eta}{\lambda_k}(U\nabla f)_k(x^t)\right)^2 + \left(U\left(x^t - x^*\right)\right)_k^2 + \frac{\eta^2}{\lambda_k^2}\left(U\nabla_f\right)_k\left(x^t\right)^2\right\}$$

Now, from the backward equation we have:

$$\begin{aligned}
&x^{t+1} = x^t - \eta U^T \Lambda^{-1} U \nabla (x^t) \\
&\Rightarrow U x^{t+1} = U x^t - \eta \Lambda^{-1} U \nabla f(x^t) \\
&\left(U x^{t+1}\right)_k = \left(U x^t\right)_k - \frac{\eta}{\lambda_k} (U \nabla f)_k (x^t)
\end{aligned}$$

From which we get

$$\begin{aligned}
f(\bar{x}) - f\left(x^*\right) &\leq \sum_{k=1}^{d} \frac{\lambda_k}{2\eta T} \sum_{t=1}^{T} \left\{ -\left(U\left(x^{t+1} - x^*\right)\right)_k^2 + \left(U\left(x^t - x^*\right)\right)_k^2 + \frac{\eta^2}{\lambda_k^2} \left(U \nabla_f\right)_k \left(x^t\right)^2 \right\} \\
&= \sum_{k=1}^{d} \frac{\lambda_k}{2\eta T} \left[\left(U\left(x^1 - x^*\right)\right)_k^2 - \left(U\left(x^{T+1} - x^*\right)\right)_k^2 \right] + \sum_{k=1}^{d} \frac{\lambda_k}{2\eta T} \sum_{t=1}^{T} \frac{\eta^2}{\lambda_k^2} (U \nabla f)_k (x^t)^2 \\
&\leq \frac{\lambda_{\max}}{2\eta T} \sum_{k=1}^{d} \left(U\left(x^1 - x^*\right)\right)_k^2 + \frac{\eta}{2T\lambda_{\min}} \sum_{t=1}^{T} \|U \nabla f\|^2 \\
&= \frac{\lambda_{\max}}{2\eta T} \left\|U\left(x^1 - x^*\right)\right\|^2 + \frac{\eta}{2\lambda_{\min}} \|\nabla f\|^2 \\
&\leq \frac{\lambda_{\max}}{2\eta T} R^2 + \frac{\eta}{2\lambda_{\min}} \rho^2
\end{aligned}$$

where we used that $x^1 = 0$ and $\|x^*\|^2 \leq R^2$ (by assumption) in the last inequality.
Set

$$\eta^2 = \frac{\lambda_{\max} \lambda_{\min} R^2}{\rho^2 T}$$

Then, we find:

$$\begin{aligned}
f(\bar{x}) - f\left(x^*\right) &\leq \frac{\lambda_{\max} R^2 \rho \sqrt{T}}{2\sqrt{\lambda_{\max}\lambda_{\min}} R T} + \frac{\sqrt{\lambda_{\max}\lambda_{\min}} R}{\rho\sqrt{T}} \frac{\rho^2}{2\lambda_{\min}} \\
&= \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\rho R}{2\sqrt{T}} + \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\rho R}{2\sqrt{T}} \\
&= \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\rho R}{\sqrt{T}}
\end{aligned}$$