COS-511: Learning Theory **Spring 2017**

Lecture 20: Online Newton Step Analysis

Lecturer: Roi Livni

20.1 ONS

Recall the ONS algorithm

Algorithm 1 Online Newton Step ONS

Inititalization $\mathbf{x}_1 \in \mathcal{K}$, parameters $\gamma, \epsilon > 0$, $A_0 = \epsilon \mathbf{Id}$.
for $t = 1, 2 \ldots T$ **do**
Play $\mathbf{x}_t$ and observe cost $f_t(\mathbf{x}_t)$
Rank 1 update $A_t = A_{t-1} + \nabla_t \nabla_t^\top$.
Newton step and projection:

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \frac{1}{\gamma} A_t^{-1} \nabla_t$$
$$\mathbf{x}_{t+1} = \Pi_{\mathcal{K}}^{A_t}(\mathbf{y}_{t+1})$$

end for
return

Theorem 20.1. *Alg. 1 with parameters* $\gamma = \min\{\frac{1}{4GD}, \alpha\}$ *and* $\epsilon = \frac{1}{\gamma^2 D^2}$*, guarantees (for* $T > 4$*):*

$$\text{Regret}_T \leq 5(\frac{1}{\alpha} + GD) n \log T$$

To prove Thm. 20.1 we begin by proving the following::

Lemma 20.2. *Let* f *be an* $\alpha-\exp$*-concave function, and* D*,* G *denote bounds on the diameter of* $\mathcal{K}$ *and on the (sub)gradient of* f *respectively. The following holds for all* $\gamma \leq \frac{1}{2}\min\{\frac{1}{4DG}, \alpha\}$*, and all* $\mathbf{x}, \mathbf{y} \in \mathcal{K}$*:*

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y}) \cdot (\mathbf{x} - \mathbf{y}) + \frac{\gamma}{2} (\mathbf{x} - \mathbf{y})^\top \nabla f(\mathbf{y}) \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$$

Proof. Since $e^{-\alpha f}$ is $\alpha-\exp$-concave , it follows by Lem. **??** that for $2\gamma \leq \alpha$, the function $h = \exp^{-2\gamma f}$ is also concave. By concavity of h we have that:

$$h(\mathbf{x}) \leq h(\mathbf{y}) + \nabla h(\mathbf{y}) \cdot (\mathbf{x} - \mathbf{y})$$

Plugging $\nabla h(\mathbf{y}) = -2\gamma \exp(-2\gamma f(\mathbf{y})) \nabla f(\mathbf{y})$ and taking log:

$$f(\mathbf{x}) \geq f(\mathbf{y}) - \frac{1}{2\gamma} \log(1 - 2\gamma \nabla f(\mathbf{y}) \cdot (\mathbf{x} - \mathbf{y}))$$

Next, note that $|2\gamma\nabla f(\mathbf{y})\cdot(\mathbf{x}-\mathbf{y})| \leq 2\gamma GD \leq \frac{1}{4}$, and that for $|z|\leq\frac{1}{4}$. , $-\log(1-z)\geq z+\frac{1}{4}z^2$, Applying the inequality over $z=2\gamma\nabla f(y)\cdot(\mathbf{x}-\mathbf{y})$ gives the lemma.

□

The proof now relies on the following result

Lemma 20.3. *The regret of ONS (with appropriate choice of parameters) is bounded by*

$$\mathrm{Regret}_T \leq 4(\frac{1}{\alpha}+GD)(\sum_{t=1}^{T}\nabla_t A_t^{-1}\nabla_t+1)$$

Proof. Let $\mathbf{x}^*\in\mathcal{K}$ be the best decision in hindesight. By Lem. 20.2 we have for our choice of γ:

$$f_t(\mathbf{x}_t)-f_t(\mathbf{x}^*)\leq\nabla_t\cdot(\mathbf{x}_t-\mathbf{x}^*)-\frac{\gamma}{2}(\mathbf{x}^*-\mathbf{x})\nabla_t\nabla_t^\top(\mathbf{x}^*-\mathbf{x}_t):=R_t$$

By definition of $\mathbf{y}_{t+1}$, we can write

$$A_t(\mathbf{y}_{t+1}-\mathbf{x}^*)=A_t(\mathbf{x}_t-\mathbf{x}^*)-\frac{1}{\gamma}\nabla_t$$
$$(\mathbf{y}_{t+1}-\mathbf{x}^*)=(\mathbf{x}_t-\mathbf{x}^*)-\frac{1}{\gamma}A_t^{-1}\nabla_t$$

Multiplying the transpose of the two equalitites we obtain:

$$(\mathbf{y}_{t+1}-\mathbf{x}^*)^\top A_t(\mathbf{y}_{t+1}-\mathbf{x}^*)=(\mathbf{x}_t-\mathbf{x}^*)^\top A_t(\mathbf{x}_t-\mathbf{x}^*)-\frac{2}{\gamma}\nabla_t^\top(\mathbf{x}_t-\mathbf{x}^*)+\frac{1}{\gamma^2}\nabla_t^\top A_t^{-1}\nabla_t \tag{20.1}$$

Since $\mathbf{x}_t$ is the projection of $\mathbf{y}_t$ induced by the norm of A_t:

$$(\mathbf{x}_t-\mathbf{x}^*)A_t(\mathbf{x}_t-\mathbf{x}^*)\leq(\mathbf{y}_t-\mathbf{x}^*)A_t(\mathbf{y}_t-\mathbf{x}^*)$$

Pluggin the inequality to Eq. 20.1 we obtain:

$$\nabla_t^\top(\mathbf{x}_t-\mathbf{x}^*)\leq\frac{1}{2\gamma}\nabla_t^\top A_t^{-1}\nabla_t+\frac{\gamma}{2}(\mathbf{x}_t-\mathbf{x}^*)^\top A_t(\mathbf{x}_t-\mathbf{x}^*)-\frac{\gamma}{2}(\mathbf{x}_{t+1}-\mathbf{x}^*)^\top A_t(\mathbf{x}_{t+1}-\mathbf{x}^*)$$

Summing up we obtain:

$$\begin{aligned}
\sum_{t=1}^{T}\nabla_t^\top(\mathbf{x}_t-\mathbf{x}^*) &\leq \sum_{t=1}^{T}\frac{1}{2\gamma}\nabla_t^\top A_t^{-1}\nabla_t+\frac{\gamma}{2}(\mathbf{x}_t-\mathbf{x}^*)^\top A_t(\mathbf{x}_t-\mathbf{x}^*)-\frac{\gamma}{2}(\mathbf{x}_{t+1}-\mathbf{x}^*)^\top A_t(\mathbf{x}_{t+1}-\mathbf{x}^*)\\
&\leq \sum_{t=1}^{T}\frac{1}{2\gamma}\nabla_t^\top A_t^{-1}\nabla_t+\frac{\gamma}{2}(\mathbf{x}_1-\mathbf{x}^*)^\top A_1(\mathbf{x}_1-\mathbf{x}^*)\\
&+\frac{\gamma}{2}\sum_{t=2}^{T}(\mathbf{x}_t-\mathbf{x}^*)(A_t-A_{t-1})(\mathbf{x}_t-\mathbf{x}^*)\\
&-\frac{\gamma}{2}(\mathbf{x}_{T+1}-\mathbf{x}^*)^\top A_T(\mathbf{x}_{T+1}-\mathbf{x}^*)\\
&\leq\sum_{t=1}^{T}\frac{1}{2\gamma}\nabla_t^\top A_t^{-1}\nabla_t+\frac{\gamma}{2}(\mathbf{x}_1-\mathbf{x}^*)^\top(A_1-\nabla_1\nabla_1^\top)(\mathbf{x}_1-\mathbf{x}^*)+\frac{\gamma}{2}(\mathbf{x}_1-\mathbf{x}^*)^\top\nabla_t\nabla_t^\top(\mathbf{x}_1-\mathbf{x}^*)
\end{aligned}$$

Where we used the fact that $A_t - A_{t-1} = \nabla_t \nabla_t^\top$, and that A_T is p.s.d hence the last term is negative. Overall we have that

$$\sum_{t=1}^T R_t \leq \sum_{t=1}^T \frac{1}{2\gamma} \nabla_t^\top A_t^{-1} \nabla_t + \frac{\gamma}{2}(\mathbf{x}_1 - \mathbf{x}^*)^\top (A_1 - \nabla_1 \nabla_1^\top)(\mathbf{x}_1 - \mathbf{x}^*)$$

We have that $A_t - \nabla_1 \nabla_1^\top = \epsilon \mathbf{Id}$, $\epsilon = \frac{1}{\gamma^2 D^2}$ hence

$$\begin{aligned} \text{Regret}_T \leq \sum R_t &\leq \frac{1}{2\gamma} \sum_{t=1}^T \nabla_t A_t^{-1} \nabla_t + \frac{\gamma}{2} D^2 \epsilon \\ &\leq \frac{1}{2\gamma} \sum_{t=1}^T \nabla_t A_t^{-1} \nabla_t + \frac{1}{2\gamma} \end{aligned}$$

Since $\gamma \leq \frac{1}{8}(\frac{1}{\alpha + GD})$, the result follows. □

proof of Thm. 20.1. We will use the following fact about p.s.d matrices:

$$\text{Tr}(A^{-1}(A - B)) \leq \log \frac{|A|}{|B|} \quad \forall A, B \succ 0 \tag{20.2}$$

Where $|A|$ stands for the determinant of A. Using this fact we have

$$\begin{aligned} \sum_{t=1}^T \nabla_t A_t^{-1} \nabla_t &= \sum_{t=1}^T \text{Tr}\left(A^{-1} \nabla_t \nabla_t^\top\right) \\ &= \sum_{t=1}^T \text{Tr}\left(A_t^{-1}(A_t - A_{t-1})\right) \\ &\leq \sum_{t=1}^T \log \frac{|A_t|}{|A_{t-1}|} = \log \frac{|A_T|}{|A_0|} \end{aligned}$$

Since $A_T = \sum \nabla_t \nabla_t^\top + \epsilon \mathbf{Id}$ and $\|\nabla_t\| \leq G$, the largest eigenvalue of A_T is at most $TG^2 + \epsilon$. Hence the determinant of A_T can be bounded by $|A_T| \leq (TG^2 + \epsilon)^n$. Hence by choice of ϵ and γ for $T \geq 4$ we have that

$$\log \frac{|A_T|}{|A_0|} \leq \log \frac{(TG + \epsilon)^n}{\epsilon} \leq n \log(TG^2 \gamma^2 D^2 + 1) \leq n \log T.$$

Pluggin this to Lem. **??** we obtain the desried result. □