MLT Homework 11

Ana Borovac
Bas Haver

December 3, 2018

Question 1

Online Newton Step for large dimension
We saw that GD has a $GD\sqrt{T}$ regret bound for convex loss functions, while ONS has a regret bound of order d ln T for exp-concave loss functions. So can it be that GD is preferable for exp-concave losses in large dimensions $d > \sqrt{T}$? In this exercise you will show that this is not possible with appropriate tuning. Starting from the intermediate ONS regret bound

$$R_T \leq \frac{\gamma}{2}\epsilon D^2 + \frac{d}{2\gamma}\ln\left(1 + \frac{TG^2}{\epsilon d}\right)$$

prove that carefully tuned ONS also guarantees

$$R_T \leq GD\sqrt{T}$$

Solution

In the proof we used:

- $\ln(1 + x) \leq x$
- $d > \sqrt{T} \quad \Rightarrow \quad \frac{\sqrt{T}}{d} < 1 \quad \Rightarrow \quad \frac{T}{d} < \sqrt{T}$
- $\epsilon = \frac{G\sqrt{T}}{\gamma D}$

It follows:

$$
\begin{aligned}
R_T &\leq \frac{\gamma}{2}\epsilon D^2 + \frac{d}{2\gamma}\ln\left(1+\frac{TG^2}{\epsilon d}\right) \\
&\leq \frac{\gamma}{2}\epsilon D^2 + \frac{d}{2\gamma}\cdot\frac{TG^2}{\epsilon d} \\
&\leq \frac{\gamma}{2}\cdot\frac{G\sqrt{T}}{\gamma D}D^2 + \frac{d}{2\gamma}\cdot\frac{TG^2}{\frac{G\sqrt{T}}{\gamma D}d} \\
&= \frac{1}{2}\cdot\frac{G\sqrt{T}}{1}D + \frac{\sqrt{T}GD}{2} \\
&= GD\sqrt{T}
\end{aligned}
$$

Question 2

ONS as Exponential Weights
In this exercise we recover Online Newton Step as an instance of exponential weights. For simplicity, here we consider ONS without projections (the result is also true with projections). Recall that exp-concave functions f_t satisfy

$$f_t(u) \geq f_t(x_t) + \langle u - x_t, \nabla_t\rangle + \frac{\gamma}{2}\langle u - x_t, \nabla_t\rangle^2.$$

Now consider multivariate Gaussian prior $p_1(u) = N(0, \epsilon^{-1}I)$ and exponential weights update with learning rate $\frac{1}{\gamma}$ applied to the above quadratic lower bound:

$$p_{t+1}(u) \propto p_t(u)e^{\frac{1}{\gamma}(f_t(x_t)+\langle u-x_t,\nabla_t\rangle+\frac{\gamma}{2}\langle u-x_t,\nabla_t\rangle^2)}.$$

Show that the weights remain multivariate Gaussian, i.e. $p_t = N(\mu_t, \sum_t)$. Show that $\mu_t = x_t$ and $\sum_t = A_{t-1}^{-1}$.

Question 3

Determine whether the following functions are exp-concave, and to what degree.

Subquestion 3.1

$x \mapsto (x-y)^2$ *on $x, y \in [-Y, Y]$ First consider a fixed y and then determine the worst-case over all $y \in [-Y, Y]$.*

Solution

Let's recall the definition of α-exp-concave function: A function f is α-exp-concave if $x \mapsto e^{-\alpha f(x)}$ is concave.

In our case we get:

$$\begin{aligned} x \mapsto e^{-\alpha(x-y)^2} \quad &\Rightarrow \quad g(x) = e^{-\alpha(x-y)^2} \\ &\Rightarrow \quad g'(x) = -2\alpha(x-y)e^{-\alpha(x-y)^2} \\ &\Rightarrow \quad g''(x) = -2\alpha e^{-\alpha(x-y)^2} + 4\alpha^2(x-y)^2 e^{-\alpha(x-y)^2} \end{aligned}$$

If g is concave then $g''(x) \leq 0$ for $\forall x$.

$$\begin{aligned} -2\alpha e^{-\alpha(x-y)^2} + 4\alpha^2(x-y)^2 e^{-\alpha(x-y)^2} &\leq 0 \\ \underbrace{e^{-\alpha(x-y)^2}}_{\geq 0}(-2\alpha + 4\alpha^2(x-y)^2) &\leq 0 \\ 2\alpha(-1 + 2\alpha(x-y)^2) &\leq 0 \\ -1 + 2\alpha(x-y)^2 &\leq 0 \\ \alpha &\leq \frac{1}{2(x-y)^2} \end{aligned}$$

If we want to inequality to hold for $\forall x$ we need to observe when the fraction on the right is the smallest. So, when does $(x-y)^2$ reach maximum? It depands on y, but it is obvious that it is reached when x is Y or $-Y$, so:

$$\alpha = \frac{1}{2\max\{(Y-y)^2, (-Y-y)^2\}}$$

In the worst case the difference between x and y is $2Y$. From that it follows:

$$\alpha = \frac{1}{8Y^2}$$

Subquestion 3.2

$x \mapsto |x|^p$ on $x \in [-1, 1]$ for $p > 1$.

Solution

$$g(x) = e^{-\alpha|x|^p}$$

Since, g is not differentiable for $x = 0$, we are going analyse concavity on two intervals: $[-1, 0)$ and $(0, 1]$.

- $x \in [-1, 0)$

$$\begin{aligned}
g(x) &= e^{-\alpha(-x)^p} \\
g'(x) &= \alpha p(-x)^{p-1} e^{-\alpha(-x)^p} \\
g''(x) &= -\alpha p(p-1)(-x)^{p-2} e^{-\alpha(-x)^p} + \alpha p(-x)^{p-1} \alpha p(-x)^{p-1} e^{-\alpha(-x)^p}
\end{aligned}$$

$\Rightarrow$

$$\begin{aligned}
g''(x) &\leq 0 \\
-\alpha p(p-1)(-x)^{p-2} e^{-\alpha(-x)^p} + \alpha p(-x)^{p-1} \alpha p(-x)^{p-1} e^{-\alpha(-x)^p} &\leq 0 \\
-(p-1)(-x)^{-1} + (-x)^{p-1} \alpha p &\leq 0 \\
(-x)^{p-1} \alpha p &\leq \frac{p-1}{-x} \\
\alpha &\leq \frac{p-1}{p(-x)(-x)^{p-1}} \\
\alpha &\leq \frac{p-1}{p(-x)^p}
\end{aligned}$$

$\Rightarrow$

$$\begin{aligned}
\alpha &= \min_{x \in [-1,0)} \frac{p-1}{p(-x)^p} \\
\alpha &= \frac{p-1}{p}
\end{aligned}$$

- $x \in (0, 1]$

$$\begin{aligned}
g(x) &= e^{-\alpha x^p} \\
g'(x) &= -\alpha p x^{p-1} e^{-\alpha x^p} \\
g''(x) &= -\alpha p(p-1) x^{p-2} e^{-\alpha x^p} + \alpha p x^{p-1} \alpha p x^{p-1} e^{-\alpha x^p}
\end{aligned}$$

$\Rightarrow$

$$\begin{aligned}
g''(x) &\leq 0 \\
-\alpha p(p-1)x^{p-2}e^{-\alpha x^p} + \alpha p x^{p-1}\alpha p x^{p-1}e^{-\alpha x^p} &\leq 0 \\
-(p-1)x^{-1} + x^{p-1}\alpha p &\leq 0 \\
x^{p-1}\alpha p &\leq \frac{p-1}{x} \\
\alpha &\leq \frac{p-1}{pxx^{p-1}} \\
\alpha &\leq \frac{p-1}{px^p}
\end{aligned}$$

$\Rightarrow$

$$\begin{aligned}
\alpha &= \min_{x\in(0,1]} \frac{p-1}{px^p} \\
\alpha &= \frac{p-1}{p}
\end{aligned}$$

We can conclude that g is $\frac{p-1}{p}$-exp-concave on intervals $[-1,0)$ and $(0,1]$. Since g is continuous function on $[-1,1]$, it is $\frac{p-1}{p}$-exp-concave on whole $[-1,1]$.

Question 4

Subquestion 4.1

Let $f(w) = g(\langle w, x\rangle)$ for some fixed x. Show that f is α-exp-concave when g is α-exp-concave.

Solution

We have

$$\begin{aligned}
e^{-\alpha g(\langle t\mathbf{w}+(1-t)\mathbf{u},\mathbf{x}\rangle)} &= e^{-\alpha g(t\langle \mathbf{w},\mathbf{x}\rangle+(1-t)\langle \mathbf{u},\mathbf{x}\rangle)} \\
&\leq te^{-\alpha g(\langle \mathbf{w},\mathbf{x}\rangle)} + (1-t)e^{-\alpha g(\langle \mathbf{u},\mathbf{x}\rangle)}
\end{aligned}$$

Where the inequality holds because g is α-exp-concave.

Subquestion 4.2

Let $f(w,y) = g(\langle w, x\rangle, y)$ for some fixed x. Show that f is η-mixable when g is η-mixable.

Solution

Because g is η-mixable, we have that $\forall P \in \Delta_{\mathcal{A}} \exists a_P \in \mathcal{A}$ such that $\forall y \in \mathcal{Y}$:

$$g(a_P, y) \leq -\eta \ln \mathbb{E}_{a \sim P}[e^{-\eta g(a,y)}]$$

Now we would like to take b_P such that $\langle b_P, x \rangle = a_P$ in order to get $\forall P \in \Delta_{\mathcal{A}} \exists b_P \in \mathcal{A}$ such that $\forall y \in \mathcal{Y}$

$$g(\langle b_P, x \rangle, y) \leq -\eta \ln \mathbb{E}_{\langle b,x \rangle \sim P}[e^{-\eta g(b,y)}]$$