MLT Homework 11

Ana Borovac
Bas Haver

November 26, 2018

Question 1

Mirror Descent and Continuous Exponential Weights
In this exercise we look at Online Gradient Descent on $U = \mathbb{R}^d$, i.e. without any projections. Then Online Gradient Descent plays iterates $w_1 = 0$ and

$$w_{t+1} = w_t - \eta \nabla f_t(w_t) \tag{1}$$

Subquestion 1.1

Show that the OGD iterate w_{t+1} is the minimiser of the problem

$$\min_{w \in \mathbb{R}^d} \langle w, \nabla f_t(w) \rangle + \frac{1}{2\eta} ||w - w_t||^2.$$

Solution

Let's define a function $g : \mathbb{R}^d \to \mathbb{R}$:

$$\begin{aligned} g(w) &= \langle w, \nabla f_t(w_t) \rangle + \frac{1}{2\eta} ||w - w_t||^2 \\ &= w_1 \frac{\partial f_t}{\partial w_1}(w_t) + \dots + w_d \frac{\partial f_t}{\partial w_d}(w_t) + \frac{1}{2\eta} \left((w_1 - w_{t1})^2 + \cdots (w_d - w_{td})^2 \right) \end{aligned}$$

We would like to find an extreme point, so $\frac{\partial g}{\partial w_i}(w^*) = 0;\ \forall i \in \{1, \dots, d\}$.

$$\frac{\partial g}{\partial w_i} = \frac{\partial f_t}{\partial w_i}(w_t) + \frac{1}{\eta}(w_i - w_{ti})$$

$$\begin{aligned} \Rightarrow \qquad & w_i^* = w_{ti} - \eta \frac{\partial f_t}{\partial w_i}(w_t) \\ \Rightarrow \qquad & w^* = w_t - \eta \nabla f_t(w_t) \end{aligned}$$

To show that calculated extreme is a minimum, we are going to show that g is a convex function.

$$\frac{\partial^2 g}{\partial w_i \partial w_i} = \frac{1}{\eta}$$
$$\frac{\partial^2 g}{\partial w_i \partial w_j} = 0; \quad i \neq j$$

The second derivative of g is non-negative for every $w \in \mathbb{R}^d$, therefore it is convex.

Subquestion 1.2

Next we look at Exponential Weights (with learning rate η) on the continuous space $\mathbb{R}^d$. We start with the spherical Gaussian prior density

$$p_1(u) = (2\pi)^{-d/2} e^{-\frac{||u||^2}{2}}$$

and we update the density using the exponential weights update

$$p_{t+1}(u) = \frac{p_t(u) e^{-\eta \langle u, \nabla f_t(w_t) \rangle}}{\textit{normalisation}}$$

where we change each point $u \in \mathbb{R}$ the linearized loss $\langle u, \nabla f_t(w_t) \rangle$ (and not the actual loss $f_t(u)$). Let $\mu_t = \int_{\mathbb{R}^d} u p_t(u) du$ be the mean of p_t. Let w_t be the iterates of Online Gradient Descent (1)*. Show that $\mu_t = w_t$ for all t.*

Solution

In the following calculations we used two known integrals:

$$\int_{-\infty}^{\infty} x e^{-ax^2 + bx} dx = \frac{\sqrt{\pi} b}{2a^{3/2}} e^{\frac{b^2}{4a}}; \quad (\mathrm{Re}(a) > 0)$$
$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \frac{1}{2}\sqrt{\frac{\pi}{a}}; \quad (a > 0)$$

$$\begin{aligned}
\int_{\mathbb{R}^d} u p_1(u) du &= \int_{\mathbb{R}^d} u \; (2\pi)^{-d/2} e^{-\frac{||u||^2}{2}} du \\
&= (2\pi)^{-d/2} \int_{\mathbb{R}^d} u \; e^{-\frac{u_1^2 + \cdots u_d^2}{2}} du \\
&= (2\pi)^{-d/2} \Bigg(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u_1 \; e^{-\frac{u_1^2 + \cdots u_d^2}{2}} du_1 \dots du_d, \\
&\qquad \dots, \\
&\qquad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u_d \; e^{-\frac{u_1^2 + \cdots u_d^2}{2}} du_1 \dots du_d \Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= (2\pi)^{-d/2} \Bigg(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{u_2^2 + \cdots u_d^2}{2}} \left(\int_{-\infty}^{\infty} u_1 \, e^{-\frac{u_1^2}{2}} du_1 \right) du_2 \ldots du_d, \\
&\qquad \ldots, \\
&\qquad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u_d \, e^{-\frac{u_2^2 + \cdots u_{d-1}^2}{2}} \left(\int_{-\infty}^{\infty} e^{-\frac{u_1^2}{2}} du_1 \right) du_2 \ldots, du_d \Bigg) \\
&= (2\pi)^{-d/2} \Bigg(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{u_2^2 + \cdots u_d^2}{2}} \left(\frac{\sqrt{\pi} \cdot 0}{2(\frac{1}{2})^{3/2}} e^{\frac{0^2}{-4\frac{1}{2}}} \right) du_2 \ldots du_d, \\
&\qquad \ldots, \\
&\qquad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u_d \, e^{-\frac{u_2^2 + \cdots u_{d-1}^2}{2}} \left(\frac{1}{2} \sqrt{\frac{\pi}{1/2}} \right) du_2 \ldots, du_d \Bigg) \\
&= (2\pi)^{-d/2} \Bigg(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{u_2^2 + \cdots u_d^2}{2}} \, (0) \, du_2 \ldots du_d, \\
&\qquad \ldots, \\
&\qquad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u_d \, e^{-\frac{u_2^2 + \cdots u_{d-1}^2}{2}} \left(\sqrt{\frac{\pi}{2}} \right) du_2 \ldots, du_d \Bigg) \\
&= \cdots \\
&= (0, \ldots, 0) \\
&= w_1
\end{aligned}
$$

$$
\begin{aligned}
p_{t+1}(u) &= \frac{p_t(u) e^{-\eta \langle u, \nabla f_t(w_t) \rangle}}{N_t} \\
&= \frac{p_{t-1}(u) e^{-\eta \langle u, \nabla f_{t-1}(w_{t-1}) \rangle} e^{-\eta \langle u, \nabla f_t(w_t) \rangle}}{N_{t-1} N_t} \\
&= p_1(u) \frac{e^{-\eta \langle u, \nabla f_1(w_1) \rangle} \cdots e^{-\eta \langle u, \nabla f_t(w_t) \rangle}}{N_1 \cdots N_t} \\
&= p_1(u) \frac{e^{-\eta \sum_{i=1}^{t} \langle u, \nabla f_i(w_i) \rangle}}{N_1 \cdots N_t} \\
&= (2\pi)^{-d/2} \frac{e^{-1/2 \sum_{j=1}^{d} u_j^2} \; e^{-\eta \sum_{i=1}^{t} \langle u, \nabla f_i(w_i) \rangle}}{N_1 \cdots N_t}
\end{aligned}
$$

$$
\begin{aligned}
\mu_{t+1} &= \int_{\mathbb{R}^d} u \; (2\pi)^{-d/2} \frac{e^{-1/2 \sum_{j=1}^{d} u_j^2} \; e^{-\eta \sum_{i=1}^{t} \langle u, \nabla f_i(w_i) \rangle}}{N_1 \cdots N_t} du \\
\mu_{t+1,k} &= \int_{\mathbb{R}^d} u_k \; (2\pi)^{-d/2} \frac{e^{-1/2 \sum_{j=1}^{d} u_j^2} \; e^{-\eta \sum_{i=1}^{t} \langle u, \nabla f_i(w_i) \rangle}}{N_1 \cdots N_t} du \\
&= \frac{(2\pi)^{-d/2}}{N_1 \cdots N_t} \int_{\mathbb{R}^d} u_k \; e^{-1/2 \sum_{j=1}^{d} u_j^2} \; e^{-\eta \sum_{i=1}^{t} \langle u, \nabla f_i(w_i) \rangle} du \\
&= \frac{(2\pi)^{-d/2}}{N_1 \cdots N_t} \int_{\mathbb{R}^{d-1}} \left(\int_{-\infty}^{\infty} u_k \; e^{-1/2 \sum_{j=1}^{d} u_j^2} \; e^{-\eta \sum_{i=1}^{t} \langle u, \nabla f_i(w_i) \rangle} du_1 \right) du_2 \ldots du_d
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{(2\pi)^{-d/2}}{N_1 \cdots N_t} \int_{\mathbb{R}^{d-1}} u_k \; e^{-1/2 \sum_{j=2}^{d} u_j^2} \; e^{-\eta \sum_{i=1}^{t} \sum_{j=2}^{d} u_j \cdot \frac{\partial f_i}{\partial u_j}(w_i)} \left(\int_{-\infty}^{\infty} e^{-1/2 u_1^2} \; e^{-\eta \sum_{i=1}^{t} u_1 \cdot \frac{\partial f_i}{\partial u_1}(w_i)} du_1 \right) du_2 \ldots \\
&= \frac{(2\pi)^{-d/2}}{N_1 \cdots N_t} \int_{\mathbb{R}^{d-1}} u_k \; e^{-1/2 \sum_{j=2}^{d} u_j^2} \; e^{-\eta \sum_{i=1}^{t} \sum_{j=2}^{d} u_j \cdot \frac{\partial f_i}{\partial u_j}(w_i)} \left(\sqrt{2\pi} e^{\eta^2 (\sum_{i=1}^{t} \frac{\partial f_i}{\partial u_j}(w_i))^2} \right) du_2 \ldots du_d \\
&= \frac{(2\pi)^{-d/2}}{N_1 \cdots N_t} \left((\sqrt{2\pi})^{d-1} e^{\eta^2 \sum_{j=1, j \neq k}^{d} (\sum_{i=1}^{t} \frac{\partial f_i}{\partial u_j}(w_i))^2} \right) \int_{-\infty}^{\infty} u_k \; e^{-1/2 u_k^2} \; e^{-\eta \sum_{i=1}^{t} u_k \cdot \frac{\partial f_i}{\partial u_k}(w_i)} du_k \\
&= \frac{(2\pi)^{-1/2}}{N_1 \cdots N_t} \left(e^{\eta^2 \sum_{j=1, j \neq k}^{d} (\sum_{i=1}^{t} \frac{\partial f_i}{\partial u_j}(w_i))^2} \right) \left(\frac{\sqrt{\pi}(-\eta) \sum_{i=1}^{t} \frac{\partial f_i}{\partial u_k}(w_i)}{2 \frac{1}{2\sqrt{2}}} e^{\frac{\eta^2 (\sum_{i=1}^{t} \frac{\partial f_i}{\partial u_k}(w_i))^2}{2}} \right) \\
&= \frac{1}{N_1 \cdots N_t} \left(e^{\eta^2 \sum_{j=1}^{d} (\sum_{i=1}^{t} \frac{\partial f_i}{\partial u_j}(w_i))^2} \right) \left((-\eta) \sum_{i=1}^{t} \frac{\partial f_i}{\partial u_k}(w_i) \right)
\end{aligned}
$$

Question 2

Strongly Convex Online To Batch Conversion

Subquestion 2.1

Consider loss functions of the form $f_t(u) = \frac{1}{2}(u - y_t)^2$ for $u, y_t \in \mathbb{R}$. Show that f_t is strongly convex for degree $\alpha = 1$.

Solution

For strongly convex function f it holds:

$$
f(y) \geq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{\alpha}{2} ||x - y||^2
$$

Since, in our case $\alpha = 1$ and $f_t : \mathbb{R} \to \mathbb{R}$, we need to prove for any $u_2, u_1 \in \mathbb{R}$ the following:

$$
f_t(u_2) \geq f_t(u_1) + (u_2 - u_1) \cdot f_t'(u_1) + \frac{1}{2}(u_1 - u_2)^2; \quad f_t'(u) = u - y_t
$$

So:

$$
\begin{aligned}
\frac{1}{2}(u_2 - y_t)^2 &\geq \frac{1}{2}(u_1 - y_t)^2 + (u_2 - u_1)(u_1 - y_t) + \frac{1}{2}(u_1 - u_2)^2 \\
(u_2 - y_t)^2 &\geq (u_1 - y_t)^2 + 2(u_2 - u_1)(u_1 - y_t) + (u_1 - u_2)^2 \\
u_2^2 - 2u_2 y_t + y_t^2 &\geq u_1^2 - 2u_1 y_t + y_t^2 + 2u_1 u_2 - 2u_2 y_t - 2u_1^2 + 2u_1 y_t + u_1^2 - 2u_1 u_2 + u_2^2 \\
0 &\geq 0
\end{aligned}
$$

Subquestion 2.2

Construct an estimator $\hat{w}_T(y_1, \dots, y_T)$ *(by online to batch conversion) and show that its excess risk is at most*

$$\mathbb{E}_{y_1,\dots,y_T,y}\left[\frac{1}{2}\left(\hat{w}_T(y_1,\dots,y_T) - y\right)^2 - \frac{1}{2}(u^* - y)^2\right] \leq \frac{1 + \ln T}{2T}$$

Solution

Theorem 4 of the lecture notes gives us that for a learning rate $\eta = \frac{1}{t}$ we have $R_T \leq \frac{G^2}{2}(1 + \ln T$, so lets first calculate G:

$$G = \max_{u, y_t \in [-1,1]} ||\nabla f_t(u)|| = \max_{u, y_t \in [-1,1]} ||u - y|| = 2$$

So if we obtain a learning rate $\eta = \frac{1}{t}$ then we have $R_T \leq 2(1 + \ln T)$. When we now pick $\hat{\omega}$ to be the average iterate estimator, then we have our desired learning rate. Theorem 3 from the lecture notes now gives us

$$\mathbb{E}_{y_1,\dots,y_T,y}\left[\frac{1}{2}\left(\hat{w}_T(y_1,\dots,y_T) - y\right)^2 - \frac{1}{2}(u^* - y)^2\right] \leq \frac{1 + \ln T}{2T}$$

Subquestion 2.3

S how that Online Gradient Descent for 1-strongly convex losses results in iterates

$$w_{t+1} = \frac{\sum_{s=1}^{t} y_s}{t}.$$

Solution

We have

$$\begin{aligned}\omega_{t+1} &= \Pi_{\mathcal{U}}(\omega_t - \eta_t \nabla f_t(\omega_t)) \\ &= \Pi_{\mathcal{U}}(\omega_t - \frac{1}{t}(\omega_t - y_t))\end{aligned}$$

Now for $\omega_1 = 0$ we have $\omega_2 = \Pi_{\mathcal{U}}(\frac{1}{t} y_t) = \Pi_{\mathcal{U}}(y_1) = y_1$, so it holds for at least one case. Now suppose that

$$\omega_t = \frac{\sum_{s=1}^{t-1}}{t-1}$$

now we would like to show that also

$$\omega_{t+1} = \frac{\sum_{s=1}^{t}}{t}.$$

We obtain

$$\begin{aligned}
\omega_{t+1} &= \Pi_{\mathcal{U}}(\omega_t - \frac{1}{t}(\omega_t - y_t)) \\
&= \Pi_{\mathcal{U}}(\frac{\sum_{s=1}^{t-1} y_s}{t-1} - \frac{1}{t}(\frac{\sum_{s=1}^{t-1} y_s}{t-1} - y_t)) \\
&= \Pi_{\mathcal{U}}((1 - \frac{1}{t})\frac{\sum_{s=1}^{t-1} y_s}{t-1} + \frac{1}{t} y_t) \\
&= \Pi_{\mathcal{U}}(\frac{t-1}{t}\frac{\sum_{s=1}^{t=1} y_s}{t-1} + \frac{1}{t} y_t) \\
&= \Pi_{\mathcal{U}}(\frac{\sum_{s=1}^{t-1} y_s}{t} + \frac{1}{t} y_t) \\
&= \Pi_{\mathcal{U}}(\frac{\sum_{s=1}^{t} y_s}{t})
\end{aligned}$$

So we find $\omega_{t+1} = \frac{\sum_{s=1}^{t}}{t}$, which is what we wanted to prove.

Subquestion 2.4

S how that, in this case, the *final iterate* estimator $\hat{\omega}_T(y_1, \ldots, y_T) = \omega_{T+1}$, results in excess risk at most

$$\mathbb{E}_{y_1,\ldots,y_T,y}[\frac{1}{2}(\hat{\omega}_T(y_1, \ldots, y_T) - y)^2 - \frac{1}{2}(u^* - y)^2] \leq \frac{Var(y)}{2T}.$$

Solution

$\mathbb{E}_{y_1,\ldots,y_T,y}[\frac{1}{2}(\hat{\omega}_T(y_1, \ldots, y_T) - y)^2 - \frac{1}{2}(u^* - y)^2] = \mathbb{E}_{y_1,\ldots,y_T,y}[\frac{1}{2}(\frac{\sum_{s=1}^{T} y_s}{T} - y)^2 - \frac{1}{2}(u^* - y)^2]$