Exercise sheet 2, Theme "Convex Optimization"

Aufgabe 1 (Gradient descent with varying step size)
Consider the result of the lecture analyzing the convergence of (averaged) gradient descent for T steps in the general case (the assumption on the objective function f besides convexity is just that it has subgradients bounded in norm by a constant L) using the step size $\eta = \frac{R}{L\sqrt{T}}$ (which is constant but depends on the total number of steps T).
Analyze the convergence (under the same hypotheses) of the gradient with varying and decreasing step size $\eta_t = \frac{R}{L\sqrt{t}}$. (Look at the proof in the lecture and try to adapt it to this case.)
Analyze the same method when stochastic gradient is applied instead of deterministic gradient.

Aufgabe 2 (Gradient descent under α-strong convexity)
In the lecture we have seen that additional hypotheses on f (compared to the "basic" case where it is just assumed to be Lipschitz) allows to obtain faster convergence for (appropriate variants of) the gradient descent method. In particular we have analyzed separately the cases where the function f to optimize is β-smooth, or when f is both β-smooth and α-strongly convex. In this exercise we consider the case where f is only α-strongly convex.
More precisely, we assume that:

- f is convex and its domain of definition $\mathcal{X}$ is a compact convex subset of $\mathbb{R}^d$;
- f is L-Lipschitz (i.e. $\forall g_x \in \partial f(x), \|g_x\| \leq L$);
- f is α-strongly convex.

We consider the (projected) gradient descent iterations with varying step-sizes $\eta_t > 0$. Denote $\delta_t := \eta_t^{-1}$ and assume for simplicity of notation below that $\delta_0 > 0$ is defined.

a) Looking at the proof in the lecture of the basic L-Lipschitz case, prove that under the additional α-strongly convex assumption one has for any $t \geq 1$:

$$f(x_t) - f(x^*) \leq \frac{1}{2}\left(\eta_t L^2 + (\delta_t - \alpha)\|x_t - x^*\|^2 - \delta_t \|x_{t+1} - x^*\|^2\right).$$

b) Assume that the step-sizes are decreasing and let $S_T := \sum_{t=1}^T \delta_{t-1}$. Deduce from the above that

$$f\left(\frac{1}{S_T}\sum_{t=1}^T \delta_{t-1} x_t\right) - f(x^*) \leq \frac{1}{2S_T}\left(TL^2 + \delta_0(\delta_1 - \alpha)\|x_1 - x^*\|^2\right),$$

provided $\delta_{t-1} \geq \delta_{t+1} - \alpha$, for $t \geq 1$.

c) Deduce that the gradient descent with step size $\eta_t = 2/(\alpha(t+1))$ satisfies under the considered assumptions:

$$f\left(\frac{1}{T(T+1)}\sum_{t=1}^T t x_t\right) \leq \frac{2L^2}{\alpha(T+1)}.$$

d) It can seem surprising that the bound does not depend at all on an a priori bound on $|f|$, nor on the diameter of $\mathcal{X}$ (as did the other bounds obtained in the lecture). However, prove that in fact the assumptions made for the above result imply that the diameter of $\mathcal{X}$ must be less than $4L/\alpha$.

Aufgabe 3 (Properties of Bregman divergences)
Recall that if $\Phi : \mathcal{D} \to \mathbb{R}$ is a differentiable strictly convex function on an open convex set $\mathcal{D} \subset \mathbb{R}^d$, the Bregman divergence associated to Φ is defined as

$$D(y, x) = \Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x \rangle$$

which is always a nonnegative quantity by convexity.

a) Let $\mathcal{X} \subset \mathcal{D}$ be a compact convex subset, and

$$\Pi_{\mathcal{X}}(x) = \operatorname*{Arg\,Min}_{y \in \mathcal{D}} D(y, x)$$

denote the projection on $\mathcal{X}$ in the Φ-Bregman divergence sense.

Prove that

$$\langle \nabla\Phi(\Pi_{\mathcal{X}}(x)) - \nabla\Phi(x), \Pi_{\mathcal{X}}(x) - y \rangle \leq 0,$$

for any $x \in \mathcal{D}, y \in \mathcal{X}$; and as a consequence

$$D(y, \Pi_{\mathcal{X}}(x)) + D(\Pi_{\mathcal{X}}(x), x) \leq D(y, x)\,.$$

b) We consider the following particular case: $\mathcal{D} = \mathbb{R}^d_{>0}$, $\mathcal{X} = \left\{ x \in \mathbb{R}_{>0} : \sum_{i=1}^d x_i = 1 \right\}$ the open simplex, and $\Phi(x) = \sum_{i=1}^d x_i \log x_i$ the Shannon entropy. (It can be checked that although $\mathcal{X}$ is not compact, the results of the previous question apply.)

(a) What is the Bregman divergence associated to Φ?

(b) Prove that $\Pi_{\mathcal{X}}(x) = \frac{x}{\|x\|_1}$ for $x \in \mathcal{D}$.

(c) Prove that Φ is 1-strongly convex on $\mathcal{X}$ with respect to $\|\cdot\|_1$, i.e

$$\sum_{i=1}^d x_i \log \frac{x_i}{y_i} \geq \frac{1}{2} \|x - y\|_1^2$$

for $x, y \in \mathcal{X}$ (Pinsker's inequality).

- Prove the inequality by elementary means in the case $d = 2$.
- Reduce the inequality in the case $d > 2$ to the case $d = 2$ by using the "log-sum inequality":

$$\sum_i x_i \log \frac{x_i}{y_i} \geq (\sum_i x_i) \log \frac{\sum_i x_i}{\sum_i y_i}.$$

(Prove this inequality, and apply it separately to the set of indices $\{i : x_i \geq y_i\}$, and its complementary).