Exercise sheet 3

You can return your solutions to questions 1 and 2 to get them marked. If so, please return them before Thursday 28/11 at 3pm in my pigeonhole. Otherwise you can return them by email to hf323@cam.ac.uk before Saturday 30/11 midnight.

1. (*Mirror maps*) Let $C \subset \mathbb{R}^n$ be a closed convex set and let $\phi : C \to \mathbb{R}$ be strongly convex. To minimize a convex function $f(x)$ on C we consider the following algorithm:

$$x_{k+1} = \nabla\phi^*(\nabla\phi(x_k) - t_k g_k) \tag{1}$$

where $g_k \in \partial f(x_k)$ and where ϕ^* denotes the conjugate function of ϕ. Show that (1) is equivalent to the mirror descent algorithm.

Remark. *Mirror descent was proposed originally by Nemirovski and Yudin in the 1980s in the form (1). The function ϕ (or rather $\nabla\phi$) was called a mirror map because $\nabla\phi$ maps vector in $\mathbb{R}^n$ to the dual space $(\mathbb{R}^n)^*$, and $\nabla\phi^*$ is the inverse map (check that $\nabla\phi^*(\nabla\phi(x)) = x$).*

2. (*Exponentiated gradient descent*) We want to minimize a function f on the simplex

$$\Delta_n = \left\{ x \in \mathbb{R}^n : x_i \geq 0 \ \forall i = 1, \ldots, n \text{ and } \sum_{i=1}^n x_i = 1 \right\},$$

and we know that $\|g\|_\infty \leq G$ for all $g \in \partial f(x)$ for all $x \in \Delta_n$. We will show that, in this situation, mirror descent can be better adapted than the projected subgradient descent.

   (a) Let $\phi(x) = \sum_{i=1}^n x_i \log x_i$. Show that the iterates of mirror descent, assuming we start with $x_0 \in \Delta_n$ satisfying $(x_0)_i > 0$ for all $i = 1, \ldots, n$, take the form

$$x_{k+1} = \frac{x_k \odot e^{-t_k g_k}}{\mathbf{1}^T (x_k \odot e^{-t_k g_k})} \qquad (g_k \in \partial f(x_k)), \tag{2}$$

   where $x \odot y = (x_i y_i)_{1 \leq i \leq n}$ is the componentwise product and $e^z = (e^{z_i})_{1 \leq i \leq n}$ is the componentwise exponential function. Verify that the iterates belong to Δ_n.

   (b) Show that ϕ is 1-strongly convex with respect to the ℓ_1 norm.

   (c) Show that for any $x \in \Delta_n$ we have $D_\phi(x \| \frac{1}{n}\mathbf{1}) \leq \log n$ where $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$.

   (d) Deduce that, with the right choice of step size, and with $x_0 = \frac{1}{n}\mathbf{1}$, the iterates (2) satisfy $f_{\text{best},k} - f^* \leq \frac{G\sqrt{\log n}}{\sqrt{k}}$.

   (e) Work out an upper bound on $f_{\text{best},k} - f^*$ if we use the projected subgradient method ($x_{k+1} = P_{\Delta_n}(x_k - t_k g_k)$). Your upper bound should only depend on G, n and k. How does it compare with the answer in part (*d*) when n is large?

3. (*Newton's method*) Let $f : \mathbb{R}^n \to \mathbb{R}$ be m-strongly convex and L-smooth (i.e., $mI \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^n$). Consider Newton's method with constant step size $t_k = m/L$

$$x^+ = x - \frac{m}{L}\nabla^2 f(x)^{-1}\nabla f(x).$$

Show that $f(x^+) - f(x) \leq -c\|\nabla f(x)\|_2^2$ for some constant $c > 0$ that depends only on m and L that you should specify.

4. Show that if f is self-concordant, then αf is self-concordant for $\alpha \geq 1$.

5. Show that if f_1, f_2 are self-concordant, then $f_1 + f_2$ is self-concordant with $\mathrm{dom}(f_1 + f_2) = \mathrm{dom}(f_1) \cap \mathrm{dom}(f_2)$.

6. Let f be a self-concordant function with $\mathrm{dom}(f) \subset \mathbb{R}^n$, and let $A \in \mathbb{R}^{n \times m}$ an injective linear map.

   (a) Show that $g(x) = f(Ax)$ is self-concordant

   (b) Show that $\max_{x \in \mathrm{dom}(g)} \lambda_g(x)^2 \leq \max_{z \in \mathrm{dom}(f)} \lambda_f(z)^2$. [*Hint: For $H \succ 0$, $H^{1/2}A(A^THA)^{-1}A^TH^{1/2}$ is a projector matrix.*]

7. Show that $f(x) = -\sum_{i=1}^n \log(x_i)$ with $\mathrm{dom}(f) = \mathbb{R}^n_{++}$ is self-concordant.

8. (*Quadratic convergence of Newton's method for self-concordant functions*) Let f be a self-concordant function, $x \in \mathrm{dom}(f)$ such that $\lambda(x) < 1$. Let $x^+ = x - \nabla^2 f(x)^{-1} \nabla f(x)$. We want to prove that

$$\lambda(x^+) \leq \frac{\lambda(x)^2}{(1 - \lambda(x))^2}. \tag{3}$$

   For any $x \in \mathrm{dom}(f)$ we let, for convenience, $H(x) = \nabla^2 f(x)$. Also we let $h = x^+ - x$ so that $\|h\|_x = \lambda(x)$.

   (a) Prove that $\lambda(x^+) \leq \frac{1}{1 - \|h\|_x} \|H(x)^{-1} \nabla f(x^+)\|_x$. In the following parts we will focus on bounding $\|H(x)^{-1} \nabla f(x^+)\|_x$.

   (b) Show that

$$\begin{aligned} \|H(x)^{-1} \nabla f(x^+)\|_x &\leq \int_0^1 \|H(x)^{-1}(H(x + th) - H(x))h\|_x dt \\ &= \int_0^1 \|E(t) H(x)^{1/2} h\|_2 dt \end{aligned}$$

   where $E(t) = H(x)^{-1/2}(H(x + th) - H(x))H(x)^{-1/2}$.

   (c) Using self-concordance of f show that

$$((1 - t\|h\|_x)^2 - 1) I \preceq E(t) \preceq \left(\frac{1}{(1 - t\|h\|_x)^2} - 1 \right) I.$$

   Deduce that

$$E(t)^2 \preceq \left(\frac{1}{(1 - t\|h\|_x)^2} - 1 \right) I.$$

   (d) Deduce that $\|H(x)^{-1} \nabla f(x^+)\|_x \leq \frac{\|h\|_x^2}{1 - \|h\|_x}$. Conclude.