MIT 6.7220/15.084 — Nonlinear Optimization Spring 2025

Problem Set 6

Due: May 9th 2025 11:59pm ET

Collaboration policy. We encourage working together whenever possible: in the recitations, problem sets, and general discussion of the material and assignments. Keep in mind, however, that for the problem sets the solutions you hand in should reflect your own understanding of the class material, and should be written solely by you. It is not acceptable to copy (in whole or in part) a solution that somebody else has written.

1. A simple example of AdaGrad [20pts]

Let function $f(x, y) = ax^2 + y^2$, with $a > 0$. Suppose we use Adagrad to optimize f with learning rate $\eta = 1$ and initialization $(x_0, y_0) = (c, c)$ with $c \neq 0$.

(1.a) [20pts] Derive the ratio (as a function of a and t) of the pre-conditioner:

$$\gamma_t := \sqrt{\frac{\sum_{s=0}^{t} \left[\nabla_x f(x_s, y_s)\right]^2}{\sum_{s=0}^{t} \left[\nabla_y f(x_s, y_s)\right]^2}}.$$

Solution. The Adagrad has the update:

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \begin{pmatrix} x_t \\ y_t \end{pmatrix} + M_t^{-1} \nabla f(x_t, y_t),$$

where

$$M_t = \begin{pmatrix} \sqrt{\sum_{s=0}^{t} \left[\nabla_x f(x_s, y_s)\right]^2} & 0 \\ 0 & \sqrt{\sum_{s=0}^{t} \left[\nabla_y f(x_s, y_s)\right]^2} \end{pmatrix}$$

$$= \begin{pmatrix} 2a\sqrt{\sum_{s=0}^{t} x^2} & 0 \\ 0 & 2\sqrt{\sum_{s=0}^{t} y^2} \end{pmatrix}$$

So we have

$$x_{t+1} = x_t - \frac{2ax_t}{2a\sqrt{\sum_{s=0}^{t} x^2}} = x_t - \frac{x_t}{\sqrt{\sum_{s=0}^{t} x^2}} =$$

$$y_{t+1} = y_t - \frac{2y_t}{2\sqrt{\sum_{s=0}^{t} y^2}} = y_t - \frac{y_t}{\sqrt{\sum_{s=0}^{t} y^2}}$$

Note that x_t and y_t have the same update rule and initialization, so we have $x_t = y_t$ for all $t \geq 0$ and hence

$$\gamma_t = \sqrt{\frac{\sum_{s=0}^{t} \left[\nabla_x f(x_s, y_s)\right]^2}{\sum_{s=0}^{t} \left[\nabla_y f(x_s, y_s)\right]^2}} = \sqrt{\frac{\sum_{s=0}^{t} \left[2ax_s\right]^2}{\sum_{s=0}^{t} \left[2y_s\right]^2}} = a$$

for all $t \geq 0$. ◂

2. Computing analytic centers [30pts]

For an open polyhedral set given by

$$\Omega := \{x \in \mathbb{R}^n : a_i^\top x > b_i \quad i = 1, 2, ..., m\}$$

(assuming there are no redundant constraints), the analytic center of Ω is the solution of

$$\min_{x \in \Omega} f(x) := -\sum_{i=1}^{m} \log(a_i^\top x - b).$$

Consider using the damped Newton method:

$$x_{t+1} = x_t - \eta_t (\nabla^2 f(x_t))^{-1} \nabla f(x_t)$$

with $\eta_t \in (0, 1)$. One standard choice of η_t is given by:

$$\eta_t = \frac{1}{1 + \lambda_f(x_t)}, \quad \text{with} \quad \lambda_f(x) := \sqrt{\nabla f(x_t)^\top [\nabla^2 f(x_t)]^{-1} \nabla f(x_t)}$$

(2.a) [15pts] Consider Ω being the triangle in $\mathbb{R}^2$ with three vertices $(0, 0)$, $(1, 0)$ and $(0, 1)$. Compute the analytic center of Ω using the damped Newton's method described above.

> *Solution.* See the jupyter notebook. ◂

(2.b) [15pts] Consider Ω being the polytope in $\mathbb{R}^2$ with four vertices $(0, 0)$, $(1, 2)$ $(2, 1)$ and $(2, 0)$. Compute the analytic center of Ω using the damped Newton's method described above.

> *Solution.* See the jupyter notebook. ◂

Check your solution using CVXPY to ensure your implementation converges to the correct point.

3. Self-concordance theory [50pts]

In this problem, you will tie up some loose ends regarding our analysis of self-concordant functions.

(3.a) [10pts] Prove that the function $f(x) = -\log(1 - \|x\|_2^2)$ is self-concordant on the open ball $\Omega := \{x : \|x\|_2 < 1\}$.

▸ *Hint:* To keep the calculations tidy, compute the second and third derivative of the restriction $t \mapsto f(x_0 + td)$ for t close to 0 by expanding the resulting expression using Taylor series. This is significantly more pleasant than brute forcing the derivative.

Solution. The Hessian of the given function can be written as

$$\nabla^2 f(x) = \frac{2}{1 - \|x\|^2} \cdot I_n + \frac{4}{(1 - \|x\|^2)^2} \cdot xx^\top$$

Theorem L19.1 implies that we only need to show that for any x_0 and direction d,

$$\varphi'''(\lambda)^2 \leq 4\varphi''(\lambda)^3$$

Where $\varphi(\lambda) = f(x_0 + \lambda d)$, note that it satisfies to show this property for $\lambda = 0$. Computing derivatives and simplifying leads to

$$\varphi'(0) = \frac{2\langle x_0, d\rangle}{1 - \|x_0\|^2}$$

$$\varphi''(0) = \frac{(2\langle x_0, d\rangle)^2 + 2(1 - \|x_0\|^2)\|d\|^2}{\left(1 - \|x_0\|^2\right)^2}$$

$$\varphi'''(0) = \frac{6 \cdot (2\langle x_0, d\rangle) \cdot \left(1 - \|x_0\|^2\right) \cdot \|d\|^2 - 2(2\langle x_0, d\rangle)^3}{\left(1 - \|x_0\|^2\right)^3}$$

Define

$$\alpha = \frac{\left(1 - \|x_0\|^2\right) \cdot \|d\|^2}{(2\langle x_0, d\rangle)^2}$$

and rewrite the inequality based on it. One way to prove the resulting inequality is to show the derivative is positive and the inequality holds at $\alpha = 0$. ◂

(3.b) [15pts] Prove that if the self-concordant function f is bounded below, then f has a minimizer.

▸ *Hint:* Choose x that satisfies $f(x) - \frac{1}{1000} < \inf_x f(x)$ and consider $y := x + \frac{1}{10\|n(x)\|_x} n(x)$. Upper bound $\|n(x)\|_x$ and use a theorem to guarantee that f must have a minimum in the proximity of x.

Solution.

Since f is bounded below we can choose x that satisfies $f(x) - \frac{1}{1000} < \inf_x f(x)$. Now consider $y := x + \frac{1}{10\|n(x)\|_x} n(x)$. Note that $\|y - x\|_x = \frac{1}{10} < 1$, so Theorem L19.5 implies

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\langle y - x, \nabla^2 f(x)(y - x) \rangle + \frac{\|y - x\|_x^3}{3(1 - \|y - x\|_x)}$$
$$= f(x) - \frac{1}{10}\|n(x)\|_x + \frac{1}{200} + \frac{1}{2700}$$

Since $f(x) - \frac{1}{1000} < \inf_x f(x) \le f(y)$ we have

$$0.1\|n(x)\|_x < \frac{1}{100} + \frac{1}{200} + \frac{1}{2700} < \frac{1}{9}$$

Hence, Theorem 19.6 would prove the proximity to a minimum

$$0.1\|z - x\|_x < 3\|n(x)\|_x$$

For some minimizer z.

◂

(3.c) [10pts] Let f be self-concordant, attaining a minimum at $x^\star$. Establish that if $\|x_t - x^\star\|_{x_t} < 1/2$, then a single step of Newton's method (with $\eta = 1$) guarantees

$$\|x_{t+1} - x^\star\|_{x_t} \le \frac{\|x_t - x^\star\|_{x_t}^2}{1 - \|x_t - x^\star\|_{x_t}}.$$

Explain why this fact alone does not imply quadratic convergence to the optimum.

▸ *Hint:* Use the Newton's method lemma we saw in Lecture 17, together with the fact that self-concordance implies slow-changing Hessian to bound the integral.

Solution. From L17.1 we have

$$x_{t+1} - x^* = (I - H_t)(x_t - x^*),$$

taking the intrinsic norm on both side we get

$$\|x_{t+1} - x^*\|_{x_t} \le \|I - H_t\|_{x_t} \, \|x_t - x^*\|_{x_t},$$

where the intrinsic operator norm $\|I - H_t\|_{x_t}$ is defined as

$$\sup_{v\neq 0}\frac{\|(I-H_t)v\|_{x_t}}{\|v\|_{x_t}} \leq 1+\sup_{v\neq 0}\frac{\|H_t v\|_{x_t}}{\|v\|_{x_t}},$$

where we used the triangle inequality. We now expand the intrinsic norm and the definition of H_t to get

$$\begin{aligned}
\|H_t\|_{x_t} &= \sup_{v\neq 0}\frac{\sqrt{\left\langle v, \nabla^2 f(x_t)[\nabla^2 f(x_t)]^{-1}\int_0^1 \nabla^2 f(x^*+\lambda(x_t-x^*))\,\mathrm{d}\lambda v\right\rangle}}{\|v\|_{x_t}} \\
&= \sup_{v\neq 0}\frac{\sqrt{\int_0^1 \langle v, \nabla^2 f(x^*+\lambda(x_t-x^*))\,\mathrm{d}\lambda v\rangle}}{\|v\|_{x_t}} \\
&= \sup_{v\neq 0}\frac{\sqrt{\int_0^1 \|v\|^2_{x^*+\lambda(x_t-x^*)}\,\mathrm{d}\lambda}}{\|v\|_{x_t}} \\
&\leq \sup_{v\neq 0}\frac{\sup_{\lambda\in[0,1]}\|v\|_{x^*+\lambda(x_t-x^*)}}{\|v\|_{x_t}} \\
&= \sup_{v\neq 0,\lambda\in[0,1]}\frac{\|v\|_{x^*+\lambda(x_t-x^*)}}{\|v\|_{x_t}}.
\end{aligned}$$

By assuming $\|x_t-x^*\|_{x_t}<\frac{1}{2}$, we have $x^*+\lambda(x_t-x^*)\in W(x_t)$, thus the third point in L19.1 implies that

$$1-\|x^*+\lambda(x_t-x^*)-x_t\|_{x_t} \leq \frac{\|v\|_{x^*+\lambda(x_t-x^*)}}{\|v\|_{x_t}} \leq \frac{1}{1-\|x^*+\lambda(x_t-x^*)-x_t\|_{x_t}}.$$

Thus we have

$$\|H_t\|_{x_t} \leq \frac{1}{1-\|x^*-x_t\|_{x_t}},$$

which yields the desired result.

To get the quadratic rate we need to bound $\|x_{t+1}-x^*\|_{x_{t+1}}$.

◂

(3.d) [15pts] (Theorem L19.8) Let $f:\Omega\to\mathbb{R}$ be self-concordant. Show that if a point $x_t\in\Omega$ is such that $\|n(x_t)\|_{x_t}<1$, then a single step of Newton's method (with $\eta=1$) guarantees

$$\|n(x_{t+1})\|_{x_{t+1}} \leq \left(\frac{\|n(x_t)\|_{x_t}}{1-\|n(x_t)\|_{x_t}}\right)^2.$$

Solution. With $\eta = 1$, a Newton's step update becomes $x_{t+1} = x_t + n(x_t)$. Note that,

$$\nabla f(x_{t+1}) - \nabla f(x_t) = \int_0^1 \nabla^2 f(x_t + \lambda(x_{t+1} - x_t))(x_{t+1} - x_t) d\lambda.$$

Plugging the value of $n(x_t)$ into this equation and reorganizing it, we have,

$$\nabla f(x_{t+1}) - \nabla f(x_t) = \int_0^1 \nabla^2 f(x_t + \lambda n(x_t)) n(x_t) d\lambda$$

$$\nabla f(x_{t+1}) = \int_0^1 [\nabla^2 f(x_t + \lambda n(x_t)) - \nabla f^2(x_t)] n(x_t) d\lambda.$$

From the definition of intrinsic norm, we write

$$\|n(x_{t+1})\|_{x_{t+1}}^2 = \nabla f(x_{t+1})^T [\nabla^2 f(x_{t+1})]^{-1} \nabla f(x_{t+1}).$$

Since f is self-concordant, $[\nabla f^2(x_{t+1})]^{-1}$ is PSD and hence, can be factored as $[\nabla f^2(x_{t+1})]^{-1} = [\nabla f^2(x_{t+1})]^{-\frac{1}{2}} [\nabla f^2(x_{t+1})]^{-\frac{1}{2}}$.

$$\begin{aligned}
\|n(x_{t+1})\|_{x_{t+1}} &= \left\| [\nabla^2 f(x_{t+1})]^{-\frac{1}{2}} \nabla f(x_{t+1}) \right\|_2 \\
&= \left\| [\nabla^2 f(x_{t+1})]^{-\frac{1}{2}} \int_0^1 [\nabla^2 f(x_t + \lambda n(x_t)) - \nabla^2 f(x_t)] n(x_t) d\lambda \right\|_2 \\
&\leq \int_0^1 \left\| [\nabla^2 f(x_{t+1})]^{-\frac{1}{2}} [\nabla^2 f(x_t + \lambda n(x_t)) - \nabla^2 f(x_t)] n(x_t) \right\|_2 d\lambda \\
&\leq \int_0^1 \left\| [\nabla^2 f(x_{t+1})]^{-\frac{1}{2}} \left[\frac{1}{1 - \lambda \|n(x_t)\|_{x_t}} - 1 \right] \nabla^2 f(x_t)] n(x_t) \right\|_2 d\lambda
\end{aligned}$$

where the final inequality follows from being almost quadratic, *i.e.*,

$$\frac{\nabla^2 f(x_t)}{1 - \lambda \|n(x_t)\|_{x_t}} \succeq \nabla^2 f(x_t + \lambda n(x_t)).$$

Plugging in the definition of $n(x_t)$ and rearranging the right hand side, we get,

$$
\begin{aligned}
\left\|n(x_{t+1})\right\|_{x_{t+1}} &\leq \left\|\left[\nabla^2 f(x_{t+1})\right]^{-\frac{1}{2}} \nabla f(x_t)\right\|_2 \int_0^1 \frac{1}{\left(1-\lambda\|n(x_t)\|\right)^2} - 1 \quad d\lambda \\
&= \left\|\left[\nabla^2 f(x_{t+1})\right]^{-\frac{1}{2}} \nabla f(x_t)\right\|_2 \frac{\|n(x_t)\|_{x_t}}{\left(1-\|n(x_t)\|\right)_{x_t}} \\
&= \|n(x_t)\|_{x_t} \frac{\|n(x_t)\|_{x_t}}{\left(1-\|n(x_t)\|\right)_{x_t}} \\
&\leq \left(\frac{\|n(x_t)\|_{x_t}}{\left(1-\|n(x_t)\|\right)_{x_t}}\right)^2,
\end{aligned}
$$

where the second-to-last inequality follows from the same procedure we used to rewrite $\left\|n(x_{t+1})\right\|_{x_{t+1}}$ and the last inequality follows from $\|n(x_t)\|_{x_t} < 1$. ◂

6.7220 Nonlinear Optimization

Spring 2025