Optimization Algorithms: HW2

Lo Chun, Chou
R13922136

May 21, 2025

1

We're given the following problem:

$$x_\star \in \arg\min_{x \in \Delta_d} f(x), \qquad f(x) = -\sum_{i=1}^{n} w_i \log\langle a_i, x\rangle$$

where:

1. $$x \in \Delta_d = \{x \in \mathbb{R}^d \mid x[i] \geq 0, \sum_{i=1}^{d} x[i] = 1\} \text{ (probabilit`y simplex)}$$

2. $$w_i \in \mathbb{R},\ w_i > 0,\ \sum_{i=1}^{n} w_i = 1$$

3. $$\begin{aligned} &a_i \in \mathbb{R}^d,\ a_i = \begin{bmatrix} a_i[1] \\ a_i[2] \\ \vdots \\ a_i[d] \end{bmatrix}, \\ &a_i[j] \geq 0\ \forall i = 1, \ldots, n,\ j = 1, \ldots, d \\ &a_i \neq 0\ \forall i = 1, \ldots, n \end{aligned}$$

We're asked to show that:
f is 1-smooth relative to the log-barrier, which is defined as:

$$h(x) = -\sum_{i=1}^{d} \log x[i]$$

Solution. By the following proposition [1]:

Proposition 1.1. *The following conditions are equivalent:*
(a-i) $f(\cdot)$ *is L-smooth relative to* $h(\cdot)$*;*
(a-ii) $Lh(\cdot) - f(\cdot)$ *is a convex function on* Q*;*
(a-iii) *under twice differentiability* $\nabla^2 f(x) \preceq L\nabla^2 h(x)$ *for any* $x \in \operatorname{int} Q$*;*
(a-iv) $\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L\langle \nabla h(x) - \nabla h(y), x - y\rangle$ *for all* $x, y \in \operatorname{int} Q$.

we could prove the required condition (which is (a-i), with $L = 1$) by proving its equivalent condition (a-iii, with $L = 1$).

First calculate $\nabla f(x)$:

$$\begin{aligned}
\nabla f(x) &= \frac{d}{dx}\left(-\sum_{i=1}^{n} w_i \log\langle a_i, x\rangle\right) \\
&= -\sum_{i=1}^{n} w_i \cdot \frac{d}{dx}\left(\log\langle a_i, x\rangle\right) \\
&= -\sum_{i=1}^{n} w_i \cdot \left(\frac{a_i}{\langle a_i, x\rangle}\right) \\
&= -\sum_{i=1}^{n} w_i \frac{a_i}{\langle a_i, x\rangle}
\end{aligned}$$

Then the Hessian of f is:

$$\begin{aligned}
\nabla^2 f(x) &= \frac{d}{dx}\left(-\sum_{i=1}^{n} w_i \frac{a_i}{\langle a_i, x\rangle}\right) \\
&= -\sum_{i=1}^{n} w_i \cdot \frac{d}{dx}\left(\frac{a_i}{\langle a_i, x\rangle}\right) \\
&= -\sum_{i=1}^{n} w_i \cdot \left(\frac{a_i}{\langle a_i, x\rangle^2} a_i^T\right)
\end{aligned}$$

Expanding the expression and writing in another form, we have:

$$\nabla^2 f(x) = -\sum_{i=1}^{n} \frac{w_i}{\langle a_i, x\rangle^2} a_i a_i^T \tag{1}$$

[1] *Relative Smooth Convex Optimization by First-Order Methods, and Applications*, MIT Lecture Notes, available at: `https://dspace.mit.edu/bitstream/handle/1721.1/120867/16m1099546.pdf`, accessed: May. 9, 2025, p. 336.

Then we shall do the same to $h(x)$

$$\begin{aligned}
\nabla h(x) &= \frac{d}{dx}\left(-\sum_{i=1}^{d}\log x[i]\right) \\
&= \begin{bmatrix} -\frac{1}{x[1]} \\ -\frac{1}{x[2]} \\ \vdots \\ -\frac{1}{x[d]} \end{bmatrix}
\end{aligned}$$

Then $\nabla^2 h(x)$ is:

$$\begin{aligned}
\nabla^2 h(x) &= \nabla \begin{bmatrix} -\frac{1}{x[1]} \\ -\frac{1}{x[2]} \\ \vdots \\ -\frac{1}{x[d]} \end{bmatrix} \\
&= \begin{bmatrix}
\frac{d}{dx[1]}\left(-\frac{1}{x[1]}\right) & \frac{d}{dx[2]}\left(-\frac{1}{x[1]}\right) & \cdots & \frac{d}{dx[d]}\left(-\frac{1}{x[1]}\right) \\
\frac{d}{dx[1]}\left(-\frac{1}{x[2]}\right) & \frac{d}{dx[2]}\left(-\frac{1}{x[2]}\right) & \cdots & \frac{d}{dx[d]}\left(-\frac{1}{x[2]}\right) \\
\vdots & & \ddots & \vdots \\
\frac{d}{dx[1]}\left(-\frac{1}{x[d]}\right) & \frac{d}{dx[2]}\left(-\frac{1}{x[d]}\right) & \cdots & \frac{d}{dx[d]}\left(-\frac{1}{x[d]}\right)
\end{bmatrix} \\
&= \begin{bmatrix}
\frac{1}{x[1]^2} & 0 & \cdots & 0 \\
0 & \frac{1}{x[2]^2} & \cdots & 0 \\
\vdots & & \ddots & \vdots \\
0 & 0 & \cdots & \frac{1}{x[d]^2}
\end{bmatrix} \qquad (2)
\end{aligned}$$

Observe $\nabla^2 f(x)$ in (1), since we're given $w_i > 0$, $x \in \Delta_d$, $a_i \neq 0$, and with proposition (a-iii) only requires dealing with $\operatorname{int}\Delta_d$, we can guarantee $x[i] > 0$, so the scalar $\frac{w_i}{\langle a_i, x\rangle^2} > 0$.

Also, we knew that for any $a_i \neq 0$, $a_i a_i^T$ is positive semidefinite, thus, each term in the summation is positive semidefinite, by summing up the n terms and adding a negative sign, we have $\nabla^2 f(x) \preceq 0$ as follows:

$$\nabla^2 f(x) = -\sum_{i=1}^{n} \frac{w_i}{\langle a_i, x\rangle^2} a_i a_i^T \preceq 0$$

Then, since $\nabla^2 h(x)$ is a diagonal matrix, and we're given that $x[i] \geq 0$, same as above, with proposition (a-iii) only requires dealing with $\operatorname{int}\Delta_d$, we can guarantee $x[i] > 0$ (so for each $\frac{1}{x[i]}$), and $\nabla^2 h(x)$ is positive definite.

Therefore, we have:

$$\nabla^2 f(x) \preceq 1 \cdot \nabla^2 h(x) \quad \text{for any } x \in \operatorname{int} \Delta_d$$

which means that (a-iii) is proved, and its equivalent condition (a-i) is also proved, and we have:

$$f \text{ is 1-smooth relative to the log-barrier } h$$

■

Denote the Bregman divergence associated with h as D_h, i.e.,

$$D_h(y, x) = h(y) - [h(x) + \langle \nabla h(x), (y - x) \rangle]$$

Consider solving the optimization problem (1) by the following algorithm:

- Let $x_1 = \begin{bmatrix} \frac{1}{d} \\ \vdots \\ \frac{1}{d} \end{bmatrix} \in \Delta_d$
- For every $t \in \mathbb{N}$, compute:

$$x_{t+1} \in \arg\min_{x \in \Delta_d} [\langle \nabla f(x_t), x - x_t \rangle + D_h(x, x_t)]$$

Note: I use $\begin{bmatrix} \frac{1}{d} \\ \vdots \\ \frac{1}{d} \end{bmatrix}$ to represent the vector $(1/d, \ldots, 1/d)$ (which is the notation used in the HW spec) in the following solution.

2

Show that for any $x \in \Delta_d$ and $0 \leq \alpha < 1$,

$$f(x_\alpha) \leq f(x) + \frac{\alpha}{1 - \alpha}, \quad \text{where } x_\alpha = (1 - \alpha)x + \alpha \begin{bmatrix} \frac{1}{d} \\ \vdots \\ \frac{1}{d} \end{bmatrix}$$

Solution. From the previous subproblem, we knew that f is 1-smooth relative to the log-barrier, so we have:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + D_h(y, x) \quad \forall x, y \in \operatorname{int} \Delta_d$$

To bound $f(x_\alpha)$, we first show that $x_\alpha \in \operatorname{int} \Delta_d$, and then let $y = x_\alpha,\ x = x$ so that we would have:

$$f(x_\alpha) \leq f(x) + \langle \nabla f(x), x_\alpha - x \rangle + D_h(x_\alpha, x)$$

By the definition of x_α, we knew that it is the convex combination of x and $\begin{bmatrix} \frac{1}{d} \\ \vdots \\ \frac{1}{d} \end{bmatrix}$, where $x \in \Delta_d$ and $\begin{bmatrix} \frac{1}{d} \\ \vdots \\ \frac{1}{d} \end{bmatrix} = x_1 \in \Delta_d$ as stated in the algorithm. Also, for each element in x_α, we have:

$$x_\alpha[i] = (1-\alpha)x[i] + \alpha \left(\frac{1}{d}\right) \qquad \forall i = 1, \ldots, d$$

Since $x[i] \geq 0$ and α is strictly smaller than 1, we have $x_\alpha[i] > 0$ for all $i = 1, \ldots, d$, so $x_\alpha \in \operatorname{int} \Delta_d$.

Then, we have:

$$f(x_\alpha) \leq f(x) + \langle \nabla f(x), x_\alpha - x \rangle + L D_h(x_\alpha, x)$$

To further simplify, we have:

$$x_\alpha - x = \left[(1-\alpha)x + \alpha \begin{bmatrix} \frac{1}{d} \\ \vdots \\ \frac{1}{d} \end{bmatrix} \right] - x = \alpha \left[\begin{bmatrix} \frac{1}{d} \\ \vdots \\ \frac{1}{d} \end{bmatrix} - x \right]$$

So we could expand the following expressions:

$$\begin{aligned}
\langle \nabla f(x), x_\alpha - x \rangle &= \langle -\sum_{i=1}^n w_i \frac{a_i}{\langle a_i, x \rangle}, \alpha \left(\begin{bmatrix} \frac{1}{d} \\ \vdots \\ \frac{1}{d} \end{bmatrix} - x \right) \rangle \\
&= -\frac{\alpha}{d} \sum_{i=1}^n \frac{w_i}{\langle a_i, x \rangle} \left[a_i[1] \cdots a_i[d] \right] \begin{bmatrix} 1 - x[1] \\ \vdots \\ 1 - x[d] \end{bmatrix} \\
&= -\frac{\alpha}{d} \sum_{i=1}^n \frac{w_i}{\langle a_i, x \rangle} \left(\sum_{j=1}^d a_i[j] - \sum_{j=1}^d a_i[j]x[j] \right) \\
&= -\frac{\alpha}{d} \sum_{i=1}^n \frac{w_i}{\sum_{k=1}^d a_i[k]x[k]} \left(\sum_{j=1}^d a_i[j] - \sum_{j=1}^d a_i[j]x[j] \right) \\
&= -\frac{\alpha}{d} \sum_{i=1}^n \frac{w_i \sum_{j=1}^d a_i[j]}{\sum_{k=1}^d a_i[k]x[k]} + \frac{\alpha}{d} \sum_{i=1}^n w_i \\
&= \frac{\alpha}{d} \left(1 - \sum_{i=1}^n w_i \sum_{j=1}^d \frac{a_i[j]}{a_i[j]x[j]} \right) \\
&= \frac{\alpha}{d} \left(1 - \sum_{i=1}^n w_i \sum_{j=1}^d \frac{1}{x[j]} \right) \qquad (1)
\end{aligned}$$

By the definition of D_h, we have:

$$
\begin{aligned}
D_h(x_\alpha, x) &= h(x_\alpha) - (h(x) + \langle \nabla h(x), (x_\alpha - x)\rangle) \\
&= h(x_\alpha) - \left(h(x) + \langle \begin{bmatrix} -\frac{1}{x[1]} \\ -\frac{1}{x[2]} \\ \vdots \\ -\frac{1}{x[d]} \end{bmatrix}, \alpha \left(\begin{bmatrix} \frac{1}{d} \\ \vdots \\ \frac{1}{d} \end{bmatrix} - x \right) \rangle \right) \\
&= -\sum_{i=1}^{d} \log x_\alpha[i] - \left(-\sum_{i=1}^{d} \log x[i] + \langle \begin{bmatrix} -\frac{1}{x[1]} \\ -\frac{1}{x[2]} \\ \vdots \\ -\frac{1}{x[d]} \end{bmatrix}, \alpha \left(\begin{bmatrix} \frac{1}{d} \\ \vdots \\ \frac{1}{d} \end{bmatrix} - x \right) \rangle \right) \\
&= -\sum_{i=1}^{d} \log x_\alpha[i] + \sum_{i=1}^{d} \log x[i] + \alpha[-\frac{1}{x[1]} \cdots - \frac{1}{x[d]}] \begin{bmatrix} \frac{1-dx[1]}{d} \\ \vdots \\ \frac{1-dx[d]}{d} \end{bmatrix} \\
&= \sum_{i=1}^{d} (\log x[i] - \log x_\alpha[i]) - \sum_{i=1}^{d} \frac{\alpha}{dx[i]} + \alpha d
\end{aligned}
\tag{2}
$$

Combining (1) and (2), we have:

$$
\begin{aligned}
&\langle \nabla f(x), x_\alpha - x \rangle + D_h(x_\alpha, x) \\
&= \frac{\alpha}{d} \left(1 - \sum_{i=1}^{n} w_i \sum_{j=1}^{d} \frac{1}{x[j]} \right) + \sum_{i=1}^{d} (\log x[i] - \log x_\alpha[i]) - \sum_{i=1}^{d} \frac{\alpha}{dx[i]} + \alpha d \\
&=
\end{aligned}
$$

■