Optimization Algorithms: HW2

Lo Chun, Chou
R13922136

May 30, 2025

(-8)

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:

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

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

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

[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:

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

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 $(1/d, \dots, 1/d)$ (which is the notation used in the HW spec) in the following solution.

(-10)

~~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, consider the case that $0 < \alpha < 1$, then we have $x_\alpha[i] > 0$ for all $i = 1, \ldots, d$. For $\alpha = 0$, we have $x_\alpha[i] = x[i] \geq 0$ for all $i = 1, \ldots, d$, and since in order to use the previous inequality, we need $x \in \operatorname{int}\Delta_d$, thus each $x[i]$ is strictly positive, so we have $x_\alpha \in \operatorname{int}\Delta_d$ for $\alpha = 0$ (and also for $0 < \alpha < 1$).

Then, we have:

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

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:

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

■

(-10)

~~3~~

4

Solution. We need to show that:

$$
\begin{aligned}
x_{t+1} &= \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \oslash \left\{ \nabla f(x_t) + \begin{bmatrix} \frac{1}{x_t[1]} \\ \vdots \\ \frac{1}{x_t[d]} \end{bmatrix} + \begin{bmatrix} \lambda \\ \vdots \\ \lambda \end{bmatrix} \right\} \\
&= \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \oslash \begin{bmatrix} \frac{\nabla f(x_t) x_t[1] + 1 + \lambda x_t[1]}{x_t[1]} \\ \vdots \\ \frac{\nabla f(x_t) x_t[d] + 1 + \lambda x_t[d]}{x_t[d]} \end{bmatrix} \\
&= \begin{bmatrix} \frac{x_t[1]}{\nabla f(x_t) x_t[1] + 1 + \lambda x_t[1]} \\ \vdots \\ \frac{x_t[d]}{\nabla f(x_t) x_t[d] + 1 + \lambda x_t[d]} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\nabla f(x_t) + \frac{1}{x_t[1]} + \lambda} \\ \vdots \\ \frac{1}{\nabla f(x_t) + \frac{1}{x_t[d]} + \lambda} \end{bmatrix} \qquad (*)
\end{aligned}
$$

By the updating rule, we have:

$$
\begin{aligned}
& x_{t+1} \in \arg\min_{x \in \Delta_d} \{\langle \nabla f(x_t), x - x_t \rangle + D_h(x, x_t)\} \\
\to\ & x_{t+1} \in \arg\min_{x \in \Delta_d} \{\langle \nabla f(x_t), x - x_t \rangle + h(x) - [h(x_t) + \langle \nabla h(x_t), x - x_t \rangle]\} \\
\to\ & x_{t+1} \in \arg\min_{x \in \Delta_d} \{\langle \nabla f(x_t), x - x_t \rangle + h(x) - h(x_t) - \langle \nabla h(x_t), x - x_t \rangle\} \\
\to\ & x_{t+1} \in \arg\min_{x \in \Delta_d} \{\langle \nabla f(x_t), x - x_t \rangle + h(x) - h(x_t) - \langle \nabla h(x_t), x \rangle + \langle \nabla h(x_t), x_t \rangle\} \qquad (1)
\end{aligned}
$$

Recall we previously derived that:

$$
\nabla h(x_t) = \begin{bmatrix} -\frac{1}{x_t[1]} \\ -\frac{1}{x_t[2]} \\ \vdots \\ -\frac{1}{x_t[d]} \end{bmatrix}
$$

So:

$$
\langle \nabla h(x_t), x_t \rangle = \begin{bmatrix} -\frac{1}{x_t[1]} \cdots -\frac{1}{x_t[d]} \end{bmatrix} \begin{bmatrix} x_t[1] \\ \vdots \\ x_t[d] \end{bmatrix} = -\sum_{i=1}^{d} \frac{1}{x_t[i]} x_t[i] = -d
$$

Plugging this into (1), we have:

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

(Similarly, $\langle \nabla f(x_t), x_t \rangle$ would also be a constant.)

Since we need to find the x that gives the minimum, we can drop the terms that are constant or independent of x (the Green terms), so we have:

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

Combining the inner product terms, we have:

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

Expand this equation by what we previously derived:

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

we have:

$$x_{t+1} \in \arg\min_{x \in \Delta_d} \sum_{i=1}^{d} \left(\nabla f(x_t)[i] + \frac{1}{x_t[i]} \right) x[i] - \sum_{i=1}^{d} \log x[i]$$

In order to deal with the constraint $x \in \Delta_d$, we can use Lagrange multiplier λ and write the Lagrangian as:

$$L(x, \lambda) = \sum_{i=1}^{d} \left(\nabla f(x_t)[i] + \frac{1}{x_t[i]} \right) x[i] - \sum_{i=1}^{d} \log x[i] + \lambda \left(\sum_{i=1}^{d} x[i] - 1 \right)$$

Then taking the derivative w.r.t. $x[i]$ and set the result to 0 to find the optimal x, we have:

$$\frac{\partial L}{\partial x[i]} = \nabla f(x_t)[i] + \frac{1}{x_t[i]} - \frac{1}{x[i]} + \lambda = 0$$

Rearrange to solve for $x[i]$, we have:

$$\frac{1}{x[i]} = \nabla f(x_t)[i] + \frac{1}{x_t[i]} + \lambda$$
$$\to x[i] = \frac{1}{\nabla f(x_t)[i] + \frac{1}{x_t[i]} + \lambda}$$

And this matches with $(*)$, which is the given updating rule. ■

5

We need to show that the following function is self-concordant:

$$\varphi(u) = u - \sum_{i=1}^{d} \log(u + \nabla f(x_t)[i] + \frac{1}{x_t[i]})$$

Solution. In order to show that $\varphi(u)$ is self-concordant, since $\varphi(u)$ is univariate, we can directly use the following definition [2]:

Self-concordant for univariate functions

A function $f : \mathbb{R} \to \mathbb{R}$ is self-concordant on $\mathbb{R}$ if :

$$|f'''(x)| \leq 2f''(x)^{3/2}$$

Claim:

$$|\varphi'''(u)| \leq 2\varphi''(u)^{3/2}$$

Proof: Let us define:

$$y_i := u + \nabla f(x_t)[i] + \frac{1}{x_t[i]}, \qquad \forall i = 1, \ldots, d$$

Then, the original function $\varphi(u)$ can be rewritten as:

$$\varphi(u) = u - \sum_{i=1}^{d} \log y_i = u + \sum_{i=1}^{d} (-\log y_i)$$

Now we can compute the dirivatives of $\varphi(u)$:

[2] *Self-concordant function*, available at: `https://en.wikipedia.org/wiki/Self-concordant_function#Univariate_self-concordant_function`, accessed: May. 29, 2025.

$$\varphi'(u) = 1 - \sum_{i=1}^{d} \frac{1}{y_i}$$

and the second derivative:

$$\varphi''(u) = \sum_{i=1}^{d} \frac{1}{y_i^2}$$

and the third derivative:

$$\varphi'''(u) = -2 \sum_{i=1}^{d} \frac{1}{y_i^3}$$

Now we have:

$$\begin{aligned} |\varphi'''(u)| &= 2 \sum_{i=1}^{d} \frac{1}{y_i^3} \\ \varphi''(u) &= \sum_{i=1}^{d} \frac{1}{y_i^2} \end{aligned}$$

In order to let the original definition of $\varphi(u)$ be valid, $y_i \in (0, \infty)$ must hold, thus, if we further define $g(y_i) = -\log y_i$, then

$$g : \{y_i \in \mathbb{R} \mid y_i > 0\} \to \mathbb{R}$$

, and we have:

$$\begin{aligned} g'(y_i) &= \frac{d}{dy_i}(-\log y_i) = -\frac{1}{y_i} \\ g''(y_i) &= \frac{d}{dy_i}\left(-\frac{1}{y_i}\right) = \frac{1}{y_i^2} \\ g'''(y_i) &= \frac{d}{dy_i}\left(\frac{1}{y_i^2}\right) = -\frac{2}{y_i^3} \end{aligned}$$

And we have:

$$| g'''(y_i) | = | -\frac{2}{y_i^3} | = \frac{2}{y_i^3} \leq 2 \left(\frac{1}{y_i^2} \right)^{3/2} = 2 \left(\frac{1}{y_i^3} \right)$$

Which shows that $g(y_i)$ is self-concordant.

Then, using the following property [3]:

■ **Sum of self-concordant functions.** The set of self-concordant functions is closed under addition.

Theorem 2.2. Let $f_1 : \Omega_1 \to \mathbb{R}$ and $f_2 : \Omega_2 \to \mathbb{R}$ be self-concordant functions whose domains satisfy $\Omega_1 \cap \Omega_2 \neq \emptyset$. Then, the function $f + g : \Omega_1 \cap \Omega_2 \to \mathbb{R}$ is self-concordant.

Since $g(y_i)$ is self-concordant for all $i = 1, \ldots, d$, and they have the same domain, so $\bigcap_{i=1}^{d} \operatorname{dom} g(y_i) \neq \emptyset$, thus, their sum:

$$\sum_{i=1}^{d} g(y_i) = \sum_{i=1}^{d} (-\log y_i)$$

is also self-concordant.

Then, using another property:

■ **Addition of an affine function.** Addition of an affine function to a self-concordant functions does not affect the self-concordance property, since self-concordance depends only on the Hessian of the function, and the addition of affine functions does not affect the Hessian.

Theorem 2.3. Let $f : \Omega \to \mathbb{R}$ be self-concordant function. Then, the function $g(x) \coloneqq f(x) + \langle a, x \rangle + b$ is self-concordant on Ω.

If we let $h(u) = u$, then h is an affine function, then our self concordant function $\sum_{i=1}^{d} (-\log y_i)$ plussing the affine function h:

$$\varphi(u) = u + \sum_{i=1}^{d} (-\log y_i)$$

is also self-concordant.

■

[3] G. Farina, *Lecture 14A–B: Self-concordant functions*, MIT 6.7220/15.084 — Nonlinear Optimization, Apr. 16–18$^{\text{th}}$ 2024. Available at: `https://www.mit.edu/~gfarina/2024/67220s24_L14B_self_concordance/L14.pdf`, p. 4.

6

We're given that:

$$g_\mu(w) = \max_{v \in \mathcal{B}_\infty} \langle w, v \rangle - \frac{\mu}{2}\|v\|_2^2$$

where $\mathcal{B}_\infty$ is the unit l_∞ norm ball.

We need to show that g_μ is differentiable and:

$$\nabla g_\mu(w) = \begin{cases} 1 & \text{if } w[i] \geq \mu \\ \frac{w[i]}{\mu} & \text{if } -\mu \leq w[i] \leq \mu \\ -1 & \text{if } w[i] < -\mu \end{cases}$$

Solution. By the definition of l_∞ norm, we have:

$$\|v\|_\infty \leq 1 \iff \max_{i=1,\dots,d} |v[i]| \leq 1$$

Then the original $g_\mu(w)$ can be rewritten as:

$$g_\mu(w) = \max_{v \in \mathcal{B}_\infty} \sum_{i=1}^{d} \left(w[i]v[i] - \frac{\mu}{2} v[i]^2 \right), \qquad \text{where } \|v\|_\infty \leq 1$$

Since to find the v that maximizes the above expression, we can independently find each $v[i]$ that maximizes the component in the summation, so we can further define:

$$h_i(w[i]) = \max_{|v[i]| \leq 1} \left(w[i]v[i] - \frac{\mu}{2} v[i]^2 \right)$$

Then the original $g_\mu(w)$ can be rewritten as:

$$g_\mu(w) = \sum_{i=1}^{d} h_i(w[i])$$

Now we can prove the differentiability of $g_\mu(w)$ by proving the differentiability of each $h_i(w[i])$. Let:

$$f_{w[i]}(v[i]) = w[i]v[i] - \frac{\mu}{2} v[i]^2$$

Since $w[i]v[i]$ is linear in $v[i]$, and the quadratic term $-\frac{\mu}{2}v[i]^2 < 0$ (for $\mu > 0$), $f_{w[i]}(v[i])$ is concave in $v[i]$, which means that exists a unique $v^\star[i]$ that maximizes $f_{w[i]}(v[i])$, and we have:

$$\frac{d}{dv[i]} f_{w[i]}(v[i]) = w[i] - \mu v[i] = 0 \iff v[i]^* = \frac{w[i]}{\mu}$$

Thus, if we do not restrict the solution to be in the unit ball, the v that maximizes $\langle w, v \rangle - \frac{\mu}{2}\|v\|_2^2$ is:

$$v^* = \begin{bmatrix} \frac{w[1]}{\mu} \\ \vdots \\ \frac{w[d]}{\mu} \end{bmatrix} = \begin{bmatrix} v[1] \\ \vdots \\ v[d] \end{bmatrix}$$

To further impose the restriction that $\max_{i=1,\ldots,d} |v[i]| \leq 1$, the optimal v need to satisfy:

$$v[i] \in [-1, 1]$$

Thus, we need to project $v[i]$ to the interval $[-1, 1]$, by the following definition of Euclidean projection [4]:

- The Euclidean projection of x_0 on a rectangle $C = \{x \mid l \preceq x \preceq u\}$ (where $l \prec u$) is given by

$$P_C(x_0)_k = \begin{cases} l_k & x_{0k} \leq l_k \\ x_{0k} & l_k \leq x_{0k} \leq u_k \\ u_k & x_{0k} \geq u_k. \end{cases}$$

We have:

$$\text{proj}_{[-1,1]}(v[i]) = \begin{cases} -1 & \text{if } v[i] < -1 \\ v[i] & \text{if } -1 \leq v[i] \leq 1 \\ 1 & \text{if } v[i] > 1 \end{cases}$$

or equivalently:

$$\text{proj}_{[-1,1]}\left(\frac{w[i]}{\mu}\right) = v^\star(w[i]) = \begin{cases} -1 & \text{if } w[i] < -\mu \\ \frac{w[i]}{\mu} & \text{if } |w[i]| \leq \mu \\ 1 & \text{if } w[i] > \mu \end{cases} \tag{1}$$

[4] S. Boyd, *Convex Optimization*, 1st ed., Cambridge University Press, Cambridge, UK, 2004, p. 399.

And this matches the given $\nabla g_\mu(w)[i]$.

Then getting back to the part of proving differentiability, we have $h_i(w[i])$:

$$
\begin{aligned}
h_i(w[i]) &= \max_{|v[i]|\le 1}\left(w[i]v[i] - \frac{\mu}{2}v[i]^2\right)\\
&= \max_{|v[i]|\le 1}\left(f_{w[i]}(v[i])\right)\\
&= f_{w[i]}(v^\star(w[i]))\\
&= w[i]v^\star(w[i]) - \frac{\mu}{2}(v^\star(w[i]))^2
\end{aligned} \tag{2}
$$

Consider the three cases of $\mathrm{proj}_{[-1,1]}(v[i])$ in (1):

- Case 1: $w[i] < -\mu$

  Then $v^\star(w[i]) = -1$, and by plugging it into (2):

$$
\begin{aligned}
h_i(w[i]) &= w[i](-1) - \frac{\mu}{2}(-1)^2 = -w[i] - \frac{\mu}{2}\\
\rightarrow h_i'(w[i]) &= \frac{d}{dw[i]}\left(-w[i] - \frac{\mu}{2}\right) = -1
\end{aligned}
$$

- Case 2: $-\mu \le w[i] \le \mu$

  Then $v^\star(w[i]) = \frac{w[i]}{\mu}$, and by plugging it into (2):

$$
\begin{aligned}
h_i(w[i]) &= w[i]\frac{w[i]}{\mu} - \frac{\mu}{2}\left(\frac{w[i]}{\mu}\right)^2 = \frac{w[i]^2}{\mu} - \frac{\mu}{2}\frac{w[i]^2}{\mu^2} = \frac{w[i]^2}{2\mu}\\
\rightarrow h_i'(w[i]) &= \frac{d}{dw[i]}\left(\frac{w[i]^2}{2\mu}\right) = \frac{w[i]}{\mu}
\end{aligned}
$$

- Case 3: $w[i] > \mu$

  Then $v^\star(w[i]) = 1$, and by plugging it into (2):

$$
\begin{aligned}
h_i(w[i]) &= w[i](1) - \frac{\mu}{2}(1)^2 = w[i] - \frac{\mu}{2}\\
\rightarrow h_i'(w[i]) &= \frac{d}{dw[i]}\left(w[i] - \frac{\mu}{2}\right) = 1
\end{aligned}
$$

Thus, at the boundaries:

- $w[i] = \mu$

  Left derivative:

$$\lim_{w[i]\to\mu^-} h_i'(w[i]) = \lim_{w[i]\to\mu^-} \frac{w[i]}{\mu} = \frac{\mu}{\mu} = 1$$

  Right derivative:

$$\lim_{w[i]\to\mu^+} h_i'(w[i]) = 1$$

- $w[i] = -\mu$

  Left derivative:

$$\lim_{w[i]\to-\mu^-} h_i'(w[i]) = -1$$

  Right derivative:

$$\lim_{w[i]\to-\mu^+} h_i'(w[i]) = \lim_{w[i]\to-\mu^+} \frac{w[i]}{\mu} = \frac{-\mu}{\mu} = -1$$

And in the interior:

$$\begin{aligned} h_i'(w[i]) &= w[i]\frac{w[i]}{\mu} - \frac{\mu}{2}\left(\frac{w[i]}{\mu}\right)^2 \\ &= \frac{w[i]^2}{\mu} - \frac{w[i]^2}{2\mu} \\ &= \frac{w[i]^2}{2\mu} \end{aligned}$$

Which always exists and is unique.

Therefore, $h_i(w[i])$ is differentiable, and $g_\mu(w) = \sum_{i=1}^{d} h_i(w[i])$ is a sum of differentiable functions, so $g_\mu(w)$ is also differentiable.

■

7

We need to further prove that g_μ is $\frac{1}{\mu}$-smooth.

Solution. By the definition in the following image [5]:

Definition 2. Differentiable[2] $f : \mathbb{R}^n \to \mathbb{R}$ is *L-smooth* if and only for all $x, y \in \mathbb{R}^n$ we have that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \cdot \|x - y\|_2$$

Since we have already proved that $g_\mu(w)$ is differentiable, proving the following claim is equivalent to proving that $g_\mu(w)$ is $\frac{1}{\mu}$-smooth.

<u>Claim:</u>

$$\|\nabla g_\mu(w_1) - \nabla g_\mu(w_2)\|_2 \leq \frac{1}{\mu}\|w_1 - w_2\|_2, \qquad \forall w_1, w_2 \in \mathbb{R}^d$$

<u>Proof:</u> Following the notation in the previous subproblem, we have:

$$\nabla g_\mu(w) = \begin{bmatrix} \nabla g_\mu(w)[1] \\ \vdots \\ \nabla g_\mu(w)[d] \end{bmatrix}$$

Let:

$$w_1 = \begin{bmatrix} w_1[1] \\ \vdots \\ w_1[d] \end{bmatrix}, \quad w_2 = \begin{bmatrix} w_2[1] \\ \vdots \\ w_2[d] \end{bmatrix}$$

Then we have:

$$\nabla g_\mu(w_1) - \nabla g_\mu(w_2) = \begin{bmatrix} \nabla g_\mu(w_1)[1] - \nabla g_\mu(w_2)[1] \\ \vdots \\ \nabla g_\mu(w_1)[d] - \nabla g_\mu(w_2)[d] \end{bmatrix}$$

The maximum of $\|\nabla g_\mu(w_1) - \nabla g_\mu(w_2)\|_2$ happens when:

$$\nabla g_\mu(w_1) = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{and} \quad \nabla g_\mu(w_2) = \begin{bmatrix} -1 \\ \vdots \\ -1 \end{bmatrix}$$

[5] A. Sidford, *MS&E 213 / CS 269O: Chapter 2 — Smooth Functions*, Stanford University, Oct. 17, 2020. Available at: `https://web.stanford.edu/~sidford/courses/20fa_opt_theory/sidford_mse213_2020fa_chap_2_smoothness.pdf`, p. 2.

which implies that:

$$w_1[i] \geq \mu,\ w_2[i] \leq -\mu \qquad \forall i = 1, \ldots, d \tag{1}$$

and we'll have:

$$\|\nabla g_\mu(w_1) - \nabla g_\mu(w_2)\|_2 = \| \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} -1 \\ \vdots \\ -1 \end{bmatrix} \|_2 = 2\sqrt{d}$$

Under the condition in (1), if we want to find the minimium of $L\|w_1 - w_2\|_2$, we can set $w_1[i] = \mu$ and $w_2[i] = -\mu$ for all $i = 1, \ldots, d$, and we'll have:

$$L\|w_1 - w_2\|_2 = L\| \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} - \begin{bmatrix} -\mu \\ \vdots \\ -\mu \end{bmatrix} \|_2 = L\sqrt{4\mu^2 d} = L2\mu\sqrt{d}$$

Thus, we can set $L = \frac{1}{\mu}$, and we'll have:

$$\|\nabla g_\mu(w_1) - \nabla g_\mu(w_2)\|_2 = 2\sqrt{d} \leq 2\sqrt{d} = \frac{1}{\mu} 2\mu\sqrt{d} = \frac{1}{\mu}\|w_1 - w_2\|_2$$

■

8

We're asked to show that:

$$g_\mu(w) \leq g(w) \leq g_\mu(w) + \frac{\mu d}{2}$$

Solution. Since $g(w)$ is defined as $\|w\|_1$, it is equivalent to show that:

$$g_\mu(w) \leq \|w\|_1 = \sum_{i=1}^{d} |w[i]| \leq g_\mu(w) + \frac{\mu d}{2}$$

We prove this by showing the two inequalities separately.

- $\underline{g_\mu(w) \leq g(w)}$

  Since the dual norm of $\|\cdot\|_1$ is $\|\cdot\|_\infty$, we can rewrite the one norm $\|w\|_1$ as [6]:

[6] S. Boyd, *Convex Optimization*, 1st ed., Cambridge University Press, Cambridge, UK, 2004, p. 637.

$$g(w) = \|w\|_1 = \sup\{w^T v \mid \|v\|_\infty \le 1\}$$

Compared with the definition of $g_\mu(w)$, we have:

$$g_\mu(w) = \max_{v \in \mathcal{B}_\infty} \langle w, v \rangle - \frac{\mu}{2}\|v\|_2^2$$

Since μ is positive, and the square of a two norm $\|v\|_2^2$ is always non-negative, the term $\frac{\mu}{2}\|v\|_2^2$ is always non-negative, thus:

$$g_\mu(w) = \max_{v \in \mathcal{B}_\infty} \langle w, v \rangle - \frac{\mu}{2}\|v\|_2^2 \le \sup\{w^T v \mid \|v\|_\infty \le 1\} = g(w)$$

- $\underline{g(w) \le g_\mu(w) + \frac{\mu d}{2}}$

  Observe that:

$$g(w) = \|w\|_1 = \sum_{i=1}^{d} |w[i]| = [w[1] \cdots w[d]] \begin{bmatrix} \text{sign}(w[1]) \\ \vdots \\ \text{sign}(w[d]) \end{bmatrix} = \langle w, \text{sign}(w) \rangle$$

  where $\text{sign}(w)$ is the sign function, which is defined as:

$$\text{sign}(w[i]) = \begin{cases} 1 & \text{if } w[i] > 0 \\ -1 & \text{if } w[i] \le 0 \end{cases}$$

  Then let:

$$v^\star = \arg\max_{v \in \mathcal{B}_\infty} \langle w, v \rangle = \text{sign}(w)$$

  Since for all elements in $\text{sign}(w)$, its value is either 1 or -1, $\text{sign}(w) \in \mathcal{B}_\infty$. Thus, we can write:

$$\begin{aligned} g_\mu(w) &= \langle w, \text{sign}(w) \rangle - \frac{\mu}{2}\|\text{sign}(w)\|_2^2 \\ &= g(w) - \frac{\mu}{2} \cdot d \end{aligned}$$

And we can get the following inequality:

$$\begin{aligned} g_\mu(w) &= \max_{v \in \mathcal{B}_\infty} \langle w, v \rangle - \frac{\mu}{2}\|v\|_2^2 \\ &\geq \langle w, v^\star \rangle - \frac{\mu}{2}\|v^\star\|_2^2 \\ &= g(w) - \frac{\mu}{2}d \end{aligned}$$

Therefore:

$$\begin{aligned} g_\mu(w) &\geq g(w) - \frac{\mu}{2}d \\ \rightarrow g(w) &\leq g_\mu(w) + \frac{\mu d}{2} \end{aligned}$$

■

9

Solution. We're given:

$$F = f + \lambda g$$

<u>Claim:</u>

$$F(w_{T+1}) - F(w_\star) \leq \frac{\lambda\sqrt{d}}{2\sqrt{T}}\left(\|w_1 - w_\star\|_2^2 + 1\right) + \frac{L\|w_1 - w_\star\|_2^2}{2T}$$

<u>Proof:</u>

Let us define:

$$F_\mu(w) = f(w) + \lambda g_\mu(w)$$

which is a little bit different from the original F in the problem statement by replacing $g_\mu(w)$ with $g(w)$.

In the solution process, I aimed to use the following theorem [7]:

[7] Lecture 6 of *10-725: Optimization*, taught by Ryan Tibshirani at Carnegie Mellon University in Fall 2013. Scribed by Micol Marchetti-Bowick. URL: `https://www.stat.cmu.edu/~ryantibs/convexopt-F13/scribes/lec6.pdf`, p. 6-1

Theorem 6.1 *Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and that its gradient is Lipschitz continuous with constant $L > 0$, i.e. we have that $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$ for any x, y. Then if we run gradient descent for k iterations with a fixed step size $t \leq 1/L$, it will yield a solution $f^{(k)}$ which satisfies*

$$f(x^{(k)}) - f(x^*) \leq \frac{\|x^{(0)} - x^*\|_2^2}{2tk}, \tag{6.1}$$

Thus we need to prove the differentiability, convexity, and smoothness of $F_\mu(w)$.

$F_\mu(w)$ differentiable:

We're given that f is differentiable, and in the previous problem 6, we have proved that $g_\mu(w)$ is differentiable, thus $F_\mu(w)$ is differentiable.

$F_\mu(w)$ convex:

We're given that f is convex, so it remained to show that $g_\mu(w)$ is convex.

$$g_\mu(w) = \max_{v \in \mathcal{B}_\infty} \langle w, v \rangle - \frac{\mu}{2}\|v\|_2^2$$

Since $\langle w, v \rangle - \frac{\mu}{2}\|v\|_2^2$ is affine, it is convex, and taking the maximum of a convex function is convex, thus $g_\mu(w)$ is convex.

With both f and $g_\mu(w)$ being convex, and the operations of addition and multiplication by a constant preserve convexity, $F_\mu(w)$ is convex.

$F_\mu(w)$ smooth:

We're given that f is L-smooth, and in the previous problem 7, we have proved that $g_\mu(w)$ is $\frac{1}{\mu}$-smooth, thus by the definition of L-smoothness [8]:

Definition 2. Differentiable[2] $f : \mathbb{R}^n \to \mathbb{R}$ is *L-smooth* if and only for all $x, y \in \mathbb{R}^n$ we have that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \cdot \|x - y\|_2$$

we have:

$$\begin{aligned} \|\nabla f(w_1) - \nabla f(w_2)\|_2 &\leq L\|w_1 - w_2\|_2 \\ \|\nabla g_\mu(w_1) - \nabla g_\mu(w_2)\|_2 &\leq \frac{1}{\mu}\|w_1 - w_2\|_2 \end{aligned} \tag{1}$$

We then take gradient to the definition of $F_\mu(w)$:

[8] A. Sidford, *MS&E 213 / CS 269O: Chapter 2 — Smooth Functions*, Stanford University, Oct. 17, 2020. Available at: `https://web.stanford.edu/~sidford/courses/20fa_opt_theory/sidford_mse213_2020fa_chap_2_smoothness.pdf`, p. 2.

$$\nabla F_\mu(w) = \nabla f(w) + \lambda \nabla g_\mu(w)$$

with triangle inequality, and replace (1) in, we have:

$$\begin{aligned}
\|\nabla F_\mu(w_1) - \nabla F_\mu(w_2)\|_2 &= \|\nabla f(w_1) + \lambda \nabla g_\mu(w_1) - \nabla f(w_2) - \lambda \nabla g_\mu(w_2)\|_2 \\
&= \|\nabla f(w_1) - \nabla f(w_2) + \lambda \nabla g_\mu(w_1) - \lambda \nabla g_\mu(w_2)\|_2 \\
&\leq \|\nabla f(w_1) - \nabla f(w_2)\|_2 + \lambda \|\nabla g_\mu(w_1) - \nabla g_\mu(w_2)\|_2 \\
&\leq L\|w_1 - w_2\|_2 + \frac{\lambda}{\mu}\|w_1 - w_2\|_2 \\
&= \left(L + \frac{\lambda}{\mu}\right)\|w_1 - w_2\|_2
\end{aligned}$$

Thus we have derived that $F_\mu(w)$ is $L + \frac{\lambda}{\mu}$-smooth.

Thus we can use the theorem above and get:

$$F_\mu(w_{T+1}) - F_\mu(w_\star) \leq \frac{\left(L + \frac{\lambda}{\mu}\right)\|w_1 - w_\star\|_2^2}{2T}$$

From problem 8, we have proved that:

$$g_\mu(w) \leq g(w) \leq g_\mu(w) + \frac{\mu d}{2}$$

so we have:

$$\begin{aligned}
g(w_{T+1}) &\leq g_\mu(w_{T+1}) + \frac{\mu d}{2} \\
g_\mu(w_\star) &\leq g(w_\star) \qquad (\text{or } \; -g(w_\star) \leq -g_\mu(w_\star))
\end{aligned}$$

and we could finally derive:

$$
\begin{aligned}
F(w_{T+1}) - F(w_\star) &= f(w_{T+1}) + \lambda g(w_{T+1}) - f(w_\star) - \lambda g(w_\star) \\
&\leq f(w_{T+1}) + \lambda\left(g_\mu(w_{T+1}) + \frac{\mu d}{2}\right) - f(w_\star) - \lambda(g_\mu(w_\star)) \\
&= f(w_{T+1}) + \lambda\left(g_\mu(w_{T+1})\right) + \frac{\lambda \mu d}{2} - f(w_\star) - \lambda\left(g_\mu(w_\star)\right) \\
&= F_\mu(w_{T+1}) - F_\mu(w_\star) + \frac{\lambda \mu d}{2} \\
&\leq \frac{\left(L + \frac{\lambda}{\mu}\right)\|w_1 - w_\star\|_2^2}{2T} + \frac{\lambda \mu d}{2} \\
&= \frac{L\|w_1 - w_\star\|_2^2}{2T} + \frac{\frac{\lambda}{\mu}\|w_1 - w_\star\|_2^2}{2T} + \frac{\lambda \mu d}{2} \\
&= \frac{L\|w_1 - w_\star\|_2^2}{2T} + \frac{\lambda\left(\frac{1}{\mu}\|w_1 - w_\star\|_2^2 + T\mu d\right)}{2T} \qquad (*)
\end{aligned}
$$

By setting $\mu = \frac{1}{\sqrt{Td}}$, we have:

$$
\begin{aligned}
F(w_{T+1}) - F(w_\star) &\leq \frac{L\|w_1 - w_\star\|_2^2}{2T} + \frac{\lambda\left(\frac{1}{\mu}\|w_1 - w_\star\|_2^2 + T\mu d\right)}{2T} \\
&= \frac{L\|w_1 - w_\star\|_2^2}{2T} + \frac{\lambda\left(\sqrt{Td}\|w_1 - w_\star\|_2^2 + T\frac{1}{\sqrt{Td}}d\right)}{2T} \\
&= \frac{L\|w_1 - w_\star\|_2^2}{2T} + \frac{\lambda\left(\sqrt{Td}\|w_1 - w_\star\|_2^2 + \sqrt{Td}\right)}{2T} \\
&= \frac{L\|w_1 - w_\star\|_2^2}{2T} + \frac{\lambda\sqrt{Td}\left(\|w_1 - w_\star\|_2^2 + 1\right)}{2T} \\
&= \frac{L\|w_1 - w_\star\|_2^2}{2T} + \frac{\lambda\sqrt{d}\left(\|w_1 - w_\star\|_2^2 + 1\right)}{2\sqrt{T}}
\end{aligned}
$$

Which is the required bound. ■

10

Solution. We want to get a tighter bound using accelerated gradient descent and choose other value of μ. By the following theorem [9]:

[9] Sébastien Bubeck. *Convex Optimization: Algorithms and Complexity.* Foundations and Trends® in Machine Learning, Vol. 8, No. 3-4 (2015), pp. 231–357. Page 294. DOI: 10.1561/2200000050. Available at: `http://sbubeck.com/Bubeck15.pdf`

Theorem 3.19. Let f be a convex and β-smooth function, then Nesterov's accelerated gradient descent satisfies

$$f(y_t) - f(x^*) \leq \frac{2\beta\|x_1 - x^*\|^2}{t^2}.$$

Since we derived that our $F_\mu(w)$ is $L + \frac{\lambda}{\mu}$-smooth in the previous problem 9, we have the following inequality:

$$F_\mu(w_{T+1}) - F_\mu(w_\star) \leq \frac{2\left(L + \frac{\lambda}{\mu}\right)\|w_1 - w_\star\|_2^2}{T^2}$$

Using the same process as in problem 9 (please refer to the last part, especially the part with tag $(*)$ if needed) , we derive the bound on $F(w_{T+1}) - F(w_\star)$:

$$F(w_{T+1}) - F(w_\star) \leq \frac{2\left(L + \frac{\lambda}{\mu}\right)\|w_1 - w_\star\|_2^2}{T^2} + \frac{\lambda\mu d}{2}$$

To show the improved optimization error bound, we should take the derivative w.r.t. μ and set it to 0 to get the optimal minimum value of the righthand side:

$$\begin{aligned}
&\frac{d}{d\mu}\left(\frac{2\left(L + \frac{\lambda}{\mu}\right)\|w_1 - w_\star\|_2^2}{T^2} + \frac{\lambda\mu d}{2}\right) = 0 \\
&\rightarrow \frac{2\lambda\|w_1 - w_\star\|_2^2}{T^2}\frac{d}{d\mu}\left(\frac{1}{\mu}\right) = -\frac{\lambda d}{2} \\
&\rightarrow \frac{2\lambda\|w_1 - w_\star\|_2^2}{T^2\mu^2} = \frac{\lambda d}{2} \\
&\rightarrow \frac{4\|w_1 - w_\star\|_2^2}{dT^2} = \mu^2 \\
&\rightarrow \mu = \frac{2\|w_1 - w_\star\|_2}{\sqrt{d}T}
\end{aligned}$$

Substitute $\mu = \frac{2\|w_1 - w_\star\|_2}{\sqrt{d}T}$ into the bound on $F(w_{T+1}) - F(w_\star)$, we have:

$$
\begin{aligned}
F(w_{T+1}) - F(w_\star) &\leq \frac{2\left(L + \frac{\lambda\sqrt{d}T}{2\|w_1 - w_\star\|_2}\right)\|w_1 - w_\star\|_2^2}{T^2} + \frac{\lambda d 2\|w_1 - w_\star\|_2}{2\sqrt{d}T} \\
&= \frac{2L\|w_1 - w_\star\|_2^2}{T^2} + \frac{\frac{\lambda\sqrt{d}T}{\|w_1 - w_\star\|_2}\|w_1 - w_\star\|_2^2}{T^2} + \frac{\lambda\sqrt{d}\|w_1 - w_\star\|_2}{T} \\
&= \frac{2L\|w_1 - w_\star\|_2^2}{T^2} + \frac{\lambda\sqrt{d}\|w_1 - w_\star\|_2^2}{T} + \frac{\lambda\sqrt{d}\|w_1 - w_\star\|_2}{T} \\
&= \frac{2L\|w_1 - w_\star\|_2^2}{T^2} + \frac{\lambda\sqrt{d}\|w_1 - w_\star\|_2(\|w_1 - w_\star\|_2 + 1)}{T}
\end{aligned}
$$

■

11

Solution. The rate of convergence of FISTA is shown as in the following image [10]:

Again it is easy show that the rate of convergence of FISTA on $f + g$ is similar to the one of Nesterov's accelerated gradient descent on f, more precisely:

$$f(y_t) + g(y_t) - (f(x^*) + g(x^*)) \leq \frac{2\beta\|x_1 - x^*\|^2}{t^2}.$$

This means that the rate of convergence of FISTA is $O\left(\frac{1}{T^2}\right)$, however, in the previous problem 10, we have proved that the rate of convergence of NAG is $O\left(\frac{1}{T}\right)$.
Therefore, regarding the rate of convergence, FISTA is better than NAG.

In addition, from the resulting optimization error bound, we can see that there is $\sqrt{d}$ in the bound of NAG, which means that as the dimension d increases, the performance to approximate $w_\star$ by w_{T+1} will be worse. ■

[10] Sébastien Bubeck. *Convex Optimization: Algorithms and Complexity.* Page 311.