Optimization Algorithms: HW1

Lo Chun, Chou
R13922136

April 27, 2025

1

(1)

First, consider a function $g(z) = \log(1 + e^{-z})$, taking its first and second derivatives:

$$g'(z) = \frac{d}{dz}\log(1 + e^{-z}) = \frac{-e^{-z}}{1 + e^{-z}}$$
$$g''(z) = \frac{d}{dz}\left(\frac{-e^{-z}}{1 + e^{-z}}\right) = \frac{e^{-z}}{(1 + e^{-z})^2}$$

We can see that it $g''(z)$ is nonnegative at all points, thus $g(z)$ is convex.
Now, consider $z = y_i\langle x_i, w\rangle$, and let $h(w) = \log(1 + e^{-y_i\langle x_i, w\rangle})$:

$$h : \mathbb{R}^d \to \mathbb{R}$$
$$h(w) = g(y_i\langle x_i, w\rangle)$$

Since $g(z)$ is convex, $h(w)$ is convex. [1] Also, since sum and scaling of convex functions are convex, the function $f(w)$ we're given is also convex. [2]

Next, we compute the gradient and Hessian of $f(w)$:

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*, 1st ed., Cambridge University Press, 2004, p. 79.
[2] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, NY, 2004, p. 82.

$$
\begin{aligned}
\nabla f(w) &= \nabla \left(\frac{1}{n} \sum_{i=1}^{n} \log \left(1 + \mathrm{e}^{-y_i \langle x_i, w \rangle}\right) \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \nabla \log \left(1 + \mathrm{e}^{-y_i \langle x_i, w \rangle}\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{-y_i x_i e^{-y_i \langle x_i, w \rangle}}{1 + e^{-y_i \langle x_i, w \rangle}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{-y_i x_i}{1 + e^{y_i \langle x_i, w \rangle}}
\end{aligned}
$$

$$
\begin{aligned}
\nabla^2 f(w) &= \nabla \left(\frac{1}{n} \sum_{i=1}^{n} \frac{-y_i x_i}{1 + e^{y_i \langle x_i, w \rangle}} \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \nabla \left(\frac{-y_i x_i}{1 + e^{y_i \langle x_i, w \rangle}} \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\left(\frac{\partial}{\partial w} \left(-y_i x_i\right)\right) \left(1 + e^{y_i \langle x_i, w \rangle}\right) - \left(-y_i x_i\right) \frac{\partial}{\partial w} \left(1 + e^{y_i \langle x_i, w \rangle}\right)}{\left(1 + e^{y_i \langle x_i, w \rangle}\right)^2} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\left(y_i x_i\right) \frac{\partial}{\partial w} \left(1 + e^{y_i \langle x_i, w \rangle}\right)}{\left(1 + e^{y_i \langle x_i, w \rangle}\right)^2} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\left(y_i x_i\right) \left(y_i e^{y_i \langle x_i, w \rangle} x_i\right)}{\left(1 + e^{y_i \langle x_i, w \rangle}\right)^2} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{e^{y_i \langle x_i, w \rangle} x_i x_i^T}{\left(1 + e^{y_i \langle x_i, w \rangle}\right)^2}
\end{aligned}
$$

Since every $x_i x_i^T$ is positive semidefinite, and is multiplied by a positive value $\frac{e^{y_i \langle x_i, w \rangle}}{(1+e^{y_i \langle x_i, w \rangle})^2}$, the average of them, which is the Hessian $\nabla^2 f(w)$, is positive semidefinite.

By the following theorem [3], we can derive the Lipschitz constant L of $\nabla f(w)$:

Theorem 2.1.6

Two times continuously differentiable function f belongs to $\mathsf{F}_L^{2,1}(\mathbb{R}^n)$

$$\Leftrightarrow 0 \preceq \nabla^2 f(x) \preceq L I_n \quad \forall x \in \mathbb{R}^n$$

[3] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, NY, 2004, p. 58.

This means that the eigenvalues of the Hessian should be in the range of $[0, L]$, and we could find L by finding the maximum eigenvalue of the Hessian.

Observe the structure of the Hessian, the scalar of $x_i x_i^T$ is:

$$\frac{e^{y_i \langle x_i, w \rangle}}{\left(1 + e^{y_i \langle x_i, w \rangle}\right)^2} \in (0, \frac{1}{4}]$$

Where $\frac{1}{4}$ happens when $y_i \langle x_i, w \rangle = 0$.

Thus, $L = \frac{1}{4n} \lambda_{max} (\sum_{i=1}^{n} x_i x_i^T)$

Then we could use the following scheme [4] to solve this optimization problem, since this scheme (2.2.6) is optimal for unconstrained minimization of the functions from $\mathsf{S}_{\mu,L}^{1,1}(\mathbb{R}^n),\ \mu \geq 0$ [5]

General scheme of optimal method

0. Choose $x_0 \in R^n$ and $\gamma_0 > 0$. Set $v_0 = x_0$.

1. kth iteration ($k \geq 0$).

a). Compute $\alpha_k \in (0, 1)$ from equation

$$L\alpha_k^2 = (1 - \alpha_k)\gamma_k + \alpha_k \mu.$$

Set $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k \mu$. (2.2.6)

b). Choose

$$y_k = \frac{\alpha_k \gamma_k v_k + \gamma_{k+1} x_k}{\gamma_k + \alpha_k \mu}$$

and compute $f(y_k)$ and $f'(y_k)$.

c). Find x_{k+1} such that

$$f(x_{k+1}) \leq f(y_k) - \frac{1}{2L} \parallel f'(y_k) \parallel^2$$

(see Section 1.2.3 for the step-size rules).

d). Set $v_{k+1} = \frac{(1-\alpha_k)\gamma_k v_k + \alpha_k \mu y_k - \alpha_k f'(y_k)}{\gamma_{k+1}}$.

In our case, we set $\mu = 0$ since we cannot guarantee strongly convexity, and L as the Lipschitz constant we just derived.

By Theorem 2.2.2 [6] , we have:

[4] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, NY, 2004, p. 76.

[5] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, NY, 2004, p. 77.

[6] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, NY, 2004, p. 77.

$$f(x_k) - f^* \leq L \min \left\{ \left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4}{(k+2)^2} \right\} \|x_0 - x^*\|^2$$

Since we have $\mu = 0$, we can write the above optimization error guarantee of our problem as:

$$f(w_k) - f^* \leq \frac{4L\|w_0 - w^*\|^2}{(k+2)^2}$$

2

(1)

Given a twice differentiable function $\varphi : \mathbb{R}^d \to [-\infty, \infty]$, assume that it is logarithmically homogeneous, then by the definition, the following holds:

$$\varphi(\gamma x) = \varphi(x) - \log \gamma, \quad \forall x \in \mathbb{R}^d, \gamma > 0 \tag{1}$$

Claim: $\langle \nabla\varphi(x), x\rangle = -1$

To derive the first equation, we first define the following:

$$F(\gamma) = \varphi(\gamma x)$$

Then the original equation (1) would become:

$$F(\gamma) = \varphi(x) - \log \gamma$$

Taking the derivative w.r.t. γ on both sides, we get:

$$\frac{dF}{d\gamma} = \frac{d}{d\gamma}\varphi(\gamma x) = \nabla\varphi(\gamma x) \cdot x = \langle \nabla\varphi(\gamma x), x\rangle \tag{2}$$

$$\frac{dF}{d\gamma} = \frac{d}{d\gamma}(\varphi(x) - \log \gamma) = -\frac{1}{\gamma} \tag{3}$$

Thus by (2) and (3), we have:

$$\langle \nabla\varphi(\gamma x), x\rangle = -\frac{1}{\gamma}$$

Then by plugging in $\gamma = 1$, we have:

$$\langle \nabla\varphi(x), x\rangle = -1 \qquad \square$$

Claim: $\nabla\varphi(x) = -\nabla^2\varphi(x)x$

From the previous part, we have:

$$\nabla\varphi(x)^T x = -1$$

Compute the gradient of both sides, for the left hand side, we have:

$$\begin{aligned}\nabla(\nabla\varphi(x)^T x) &= \nabla(\nabla\varphi(x))^T x + \nabla\varphi(x)^T \nabla x\\ &= \nabla^2\varphi(x)x + \nabla\varphi(x)^T\nabla x\end{aligned}$$

For the right hand side, we have:

$$\nabla(-1) = 0$$

Thus we have:

$$\begin{aligned}&\nabla^2\varphi(x)x + \nabla\varphi(x)^T\nabla x = 0\\ \Rightarrow&\nabla\varphi(x)^T\nabla x = -\nabla^2\varphi(x)x\\ \Rightarrow&\nabla\varphi(x)^T I_d = -\nabla^2\varphi(x)x\\ \Rightarrow&\nabla\varphi(x) = -\nabla^2\varphi(x)x \qquad \square\end{aligned}$$

Claim: $\langle x, \nabla^2\varphi(x)x\rangle = 1$

From the previous part, we have:

$$\nabla\varphi(x) = -\nabla^2\varphi(x)x$$

Multiply both sides by x^T, we have:

$$x^T\nabla\varphi(x) = -x^T\nabla^2\varphi(x)x$$

Which is equivalent to the following by using $\langle \nabla\varphi(x), x\rangle = -1$:

$$\langle x, \nabla^2\varphi(x)x\rangle = -\langle x, \nabla\varphi(x)\rangle = (-1)\times(-1) = 1 \qquad \square$$

(2)

Suppose that $\varphi : \mathbb{R}^d \to [-\infty, \infty]$ is a twice differentiable function, and is strictly convex and logarithmically homogeneous, then the following holds by the definition:

$$\nabla^2\varphi(x) > 0 \quad \forall x \in \mathbb{R}^d$$
$$\varphi(\gamma x) = \varphi(x) - \log\gamma, \quad \forall x \in \mathbb{R}^d, \gamma > 0.$$

Also, we have the following properties from the previous subsection:

$$\langle \nabla\varphi(x), x\rangle = -1 \tag{1}$$
$$\nabla\varphi(x) = -\nabla^2\varphi(x)x \tag{2}$$
$$\langle x, \nabla^2\varphi(x)x\rangle = 1 \tag{3}$$

Claim: $\nabla^2\varphi(x) \geq \nabla\varphi(x)\left(\nabla\varphi(x)\right)^T, \quad \forall x \in \text{dom } \varphi$

The claim is equivalent to proving that:

$$\nabla^2\varphi(x) - \nabla\varphi(x)\left(\nabla\varphi(x)\right)^T \succeq 0$$

where $\succeq 0$ denotes positive semidefinite.
Let z be any vector in $\mathbb{R}^d$, then we have:

$$\begin{aligned} z^T\left(\nabla^2\varphi(x) - \nabla\varphi(x)\left(\nabla\varphi(x)\right)^T\right)z &= z^T\nabla^2\varphi(x)z - z^T\nabla\varphi(x)\left(\nabla\varphi(x)\right)^T z \\ &= z^T\nabla^2\varphi(x)z - \left(\nabla\varphi(x)^T z\right)^2 \end{aligned} \tag{*}$$

Case 1: $x = z$

If $x = z$, then $(*)$ becomes the following using (1) and (2):

$$\begin{aligned} z^T\left(\nabla^2\varphi(x) - \nabla\varphi(x)\left(\nabla\varphi(x)\right)^T\right)z &= z^T\nabla^2\varphi(x)x - (\langle\nabla\varphi(x), x\rangle)^2 \\ &= x^T(-\nabla\varphi(x)) - (-1)^2 \\ &= x^T(-\nabla\varphi(x)) - 1 \\ &= -\langle\nabla\varphi(x), x\rangle - 1 \\ &= -(-1) - 1 \\ &= 0 \geq 0 \end{aligned}$$

Case 2: $x \neq z$

Using (2) to replace $\nabla\varphi(x)$ with $-\nabla^2\varphi(x)x$ in $(*)$, and using the fact that $(\nabla^2\varphi(x))^T = \nabla^2\varphi(x)$ (the Hessian is symmetric):

$$\begin{aligned}
z^T\left(\nabla^2\varphi(x) - \nabla\varphi(x)\left(\nabla\varphi(x)\right)^T\right)z &= z^T\nabla^2\varphi(x)z - \left(\nabla\varphi(x)^T z\right)^2 \\
&= z^T\nabla^2\varphi(x)z - \left((-\nabla^2\varphi(x)x)^T z\right)^2 \\
&= z^T\nabla^2\varphi(x)z - \left(-\underbrace{x^T}_{A^T}\underbrace{(\nabla^2\varphi(x))^T z}_{B}\right)^2 \\
&= z^T\nabla^2\varphi(x)z - \underbrace{[(\nabla^2\varphi(x))^T z]^T x x^T [(\nabla^2\varphi(x))^T z]}_{B^T A A^T B} \\
&= z^T\nabla^2\varphi(x)z - [x^T(\nabla^2\varphi(x))^T z]^T[x^T(\nabla^2\varphi(x))^T z] \\
&= z^T\nabla^2\varphi(x)z - ||x^T(\nabla^2\varphi(x))^T z||^2 \\
&= z^T\nabla^2\varphi(x)z - ||x^T(\nabla^2\varphi(x))z||^2
\end{aligned}$$

Let $H = \nabla^2\varphi(x)$, then the above expression is equivalent to:

$$z^T H z - (x^T H z)^2$$

Let's first check that we can define the function

$$h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad h(u,v) = u^T H v$$

as the inner product on $\mathbb{R}^d$. [7]

Using the fact that we assumed that $\nabla^2\varphi(x) > 0$, so H is positive definite, thus by theorem [8], there exists a one and only one positive definite matrix $H^{1/2}$ (also symmetric) such that $H = H^{1/2}H^{1/2}$.

- **Symmetry:** For any $u, v \in \mathbb{R}^d$, we have:

$$\begin{aligned}
h(u,v) &= u^T H v \\
&= u^T H^{1/2} H^{1/2} v \\
&= (H^{1/2}u)^T (H^{1/2}v) \\
&= (H^{1/2}v)^T (H^{1/2}u) \\
&= v^T H^{1/2} H^{1/2} u \\
&= v^T H u \\
&= h(v,u)
\end{aligned}$$

[7] H. Amann and J. Escher, *Analysis I*, 1st ed., Birkhäuser Basel, 2005, p. 153.

[8] "Square root of a matrix", Wikipedia, `https://en.wikipedia.org/wiki/Square_root_of_a_matrix`

- **Linearity:** For any $\lambda, \mu \in \mathbb{R}$ and $t, u, v \in \mathbb{R}^d$, we have:

$$\begin{aligned} h(t, \lambda u + \mu v) &= t^T H(\lambda u + \mu v) \\ &= t^T H(\lambda u + \mu v) \\ &= t^T H(\lambda u) + t^T H(\mu v) \\ &= \lambda t^T H u + \mu t^T H v \\ &= \lambda h(t, u) + \mu h(t, v) \end{aligned}$$

- **Positive definiteness:** For any $u \in \mathbb{R}^d$, we have:

$$h(u, u) = u^T H u > 0 \quad \text{since } H \text{ is positive definite}$$

[9]

Therefore, we have:

$$z^T H z - (x^T H z)^2 = \langle z, z \rangle_H - \langle x, z \rangle_H^2$$

Using the Cauchy-Schwarz inequality [10]:

Cauchy-Schwarz inequality

Let $(E, (\cdot \mid \cdot))$ be an inner product space. Then

$$|(x \mid y)|^2 \leq (x \mid x)(y \mid y), \qquad x, y \in E$$

We can derive the later equation using the fact that:

$$\langle x, x \rangle_H = x^T H x = x^T \nabla^2 \varphi(x) x = 1$$

(this is because property (3))
So we have:

$$\begin{aligned} \langle x, z \rangle_H^2 &\leq \langle x, x \rangle_H \langle z, z \rangle_H \\ &= 1 \times \langle z, z \rangle_H \\ &= \langle z, z \rangle_H \end{aligned}$$

Thus we have:

$$z^T H z - (x^T H z)^2 \geq 0 \qquad \square$$

[9] I later found that we have "A function $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is an inner product on $\mathbb{R}^n$ if and only if there exists a symmetric positive-definite matrix $\mathbf{M}$ such that $\langle x, y \rangle = x^\top \mathbf{M} y$ for all $x, y \in \mathbb{R}^n$." on "Inner product space", Wikipedia, `https://en.wikipedia.org/wiki/Inner_product_space`

[10] H. Amann and J. Escher, *Analysis I*, 1st ed., Birkhäuser Basel, 2005, p. 154.

(3)

We need to prove the following equivalence:

$$
\begin{aligned}
& (1) \quad e^{-\varphi(x)} \text{ is concave} \\
\iff & (2) \quad \varphi(y) \geq \varphi(x) - \log(1 - \langle \nabla\varphi(x), y - x \rangle), \quad \forall x, y \in \operatorname{dom}(\varphi) \\
\iff & (3) \quad \nabla^2\varphi(x) \succeq \nabla\varphi(x)\nabla\varphi(x)^\top, \quad \forall x \in \operatorname{dom}(\varphi)
\end{aligned}
$$

$\underline{\textbf{(1)} \implies \textbf{(2)}}$

Let $f : \mathbb{R}^d \to \mathbb{R}$ be defined as $f(x) = e^{-\varphi(x)}$.

Suppose that $f(x) = e^{-\varphi(x)}$ is concave, then by the definition of concavity [11]:

> **Convex**
>
> A continuously differentiable function $f(x)$ is called convex on $\mathbb{R}^n$ if for any $x, y \in \mathbb{R}^n$, we have:
>
> $$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$
>
> If $-f(x)$ is convex, then $f(x)$ is concave.

this means that our assumption is equivalent to saying that $-e^{-\varphi(x)}$ is convex. Let $g(x) = -f(x) = -e^{-\varphi(x)}$ a convex function, using the fact that:

$$\nabla g(x) = \frac{d}{dx}(-e^{-\varphi(x)}) = e^{-\varphi(x)}\nabla\varphi(x)$$

we have the following:

For any $x, y \in \mathbb{R}^d$:

$$
\begin{aligned}
& g(y) \geq g(x) + \langle \nabla g(x), y - x \rangle \\
\Rightarrow & -e^{-\varphi(y)} \geq -e^{-\varphi(x)} + \langle e^{-\varphi(x)}\nabla\varphi(x), y - x \rangle \\
\Rightarrow & e^{-\varphi(y)} \leq e^{-\varphi(x)} - e^{-\varphi(x)}\langle \nabla\varphi(x), y - x \rangle \\
\Rightarrow & e^{-\varphi(y)} \leq e^{-\varphi(x)}(1 - \langle \nabla\varphi(x), y - x \rangle) \\
\Rightarrow & -\varphi(y) \leq -\varphi(x) + \log(1 - \langle \nabla\varphi(x), y - x \rangle) \\
\Rightarrow & \varphi(y) \geq \varphi(x) - \log(1 - \langle \nabla\varphi(x), y - x \rangle)
\end{aligned}
$$

$\underline{\textbf{(2)} \implies \textbf{(3)}}$

[11] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, NY, 2004, p. 52.

Suppose (2) holds, so we have:

$$\varphi(y) \geq \varphi(x) - \log(1 - \langle \nabla\varphi(x), y - x\rangle), \quad \forall x, y \in \text{dom}(\varphi)$$

By plugging in $y = x + h$ ($h = y - x$), with $||h|| \to 0$, we have:

$$\varphi(x+h) \geq \varphi(x) - \log(1 - \langle \nabla\varphi(x), h\rangle) \tag{1}$$

Then by using the second-order approximation [12]:

Second-order approximation

Let f be twice differentiable at $\bar{x}$. Then

$$f(y) = f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x}\rangle + \frac{1}{2}\langle \nabla^2 f(\bar{x})(y - \bar{x}), y - \bar{x}\rangle + o(||y - \bar{x}||^2)$$

Since φ is twice differentiable on its domain, we have:

$$\varphi(x+h) = \varphi(x) + \langle \nabla\varphi(x), h\rangle + \frac{1}{2}\langle \nabla^2\varphi(x)h, h\rangle + o(||h||^2) \tag{2}$$

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

$$\begin{aligned}
&\varphi(x) + \langle \nabla\varphi(x), h\rangle + \frac{1}{2}\langle \nabla^2\varphi(x)h, h\rangle + o(||h||^2) \geq \varphi(x) - \log(1 - \langle \nabla\varphi(x), h\rangle) \\
&\Rightarrow \langle \nabla\varphi(x), h\rangle + \frac{1}{2}\langle \nabla^2\varphi(x)h, h\rangle + o(||h||^2) \geq -\log(1 - \langle \nabla\varphi(x), h\rangle) \\
&\Rightarrow \langle \nabla\varphi(x), h\rangle + \frac{1}{2}\langle \nabla^2\varphi(x)h, h\rangle + o(||h||^2) \geq -(-\sum_{n=1}^{\infty} \frac{\langle \nabla\varphi(x), h\rangle^n}{n}) \\
&\Rightarrow \langle \nabla\varphi(x), h\rangle + \frac{1}{2}\langle \nabla^2\varphi(x)h, h\rangle + o(||h||^2) \geq \langle \nabla\varphi(x), h\rangle + \sum_{n=2}^{\infty} \frac{\langle \nabla\varphi(x), h\rangle^n}{n} \\
&\Rightarrow \frac{1}{2}\langle \nabla^2\varphi(x)h, h\rangle + o(||h||^2) \geq \sum_{n=2}^{\infty} \frac{\langle \nabla\varphi(x), h\rangle^n}{n} \\
&\Rightarrow \frac{1}{2}\langle \nabla^2\varphi(x)h, h\rangle + o(||h||^2) \geq \frac{\langle \nabla\varphi(x), h\rangle^2}{2} + \frac{\langle \nabla\varphi(x), h\rangle^3}{3} + \cdots \qquad (*)
\end{aligned}$$

Examine the terms on the right hand side by Cauchy-Schwarz inequality:

[12] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, NY, 2004, p. 19.

$$\frac{(\langle \nabla\varphi(x), h\rangle)^3}{3} \leq \frac{(||\nabla\varphi(x)|| \cdot ||h||)^3}{3}$$

Since $||h|| \to 0$ by our assumption, we can write:

$$\frac{\langle \nabla\varphi(x), h\rangle^2}{2} + \frac{\langle \nabla\varphi(x), h\rangle^3}{3} + \cdots = o(||h||^2)$$

Substituting this bound back into $(*)$, we have:

$$\begin{aligned}
&\frac{1}{2}\langle \nabla^2\varphi(x)h, h\rangle + o(||h||^2) \geq \frac{\langle \nabla\varphi(x), h\rangle^2}{2} + o(||h||^2) \\
\Rightarrow\ & \frac{1}{2}\langle \nabla^2\varphi(x)h, h\rangle \geq \frac{\langle \nabla\varphi(x), h\rangle^2}{2} \\
\Rightarrow\ & \langle \nabla^2\varphi(x)h, h\rangle \geq \langle \nabla\varphi(x), h\rangle^2 \\
\Rightarrow\ & (\nabla^2\varphi(x)h)^T h \geq (\nabla\varphi(x)^T h)^T(\nabla\varphi(x)^T h) \\
\Rightarrow\ & h^T(\nabla^2\varphi(x))^T h \geq h^T \nabla\varphi(x)(\nabla\varphi(x))^T h \\
\Rightarrow\ & h^T((\nabla^2\varphi(x))^T - \nabla\varphi(x)(\nabla\varphi(x))^T)h \geq 0 \\
\Rightarrow\ & \nabla^2\varphi(x) - \nabla\varphi(x)(\nabla\varphi(x))^T \succeq 0 \qquad \text{(since the Hessian is symmetric)}
\end{aligned}$$

Thus, we have proved that:

$$\nabla^2\varphi(x) \geq \nabla\varphi(x)(\nabla\varphi(x))^T, \quad \forall x, y \in \operatorname{dom}\varphi$$

(3) $\implies$ (1)

Suppose (3) holds, so we have:

$$\nabla^2\varphi(x) \geq \nabla\varphi(x)(\nabla\varphi(x))^T, \quad \forall x, y \in \operatorname{dom}\varphi$$

Since we need to show that $e^{-\varphi(x)}$ is concave, similar to the previous proof, we can define $g(x) = -f(x) = -e^{-\varphi(x)}$ (where $f(x) = e^{-\varphi(x)}$), and show that $g(x)$ is convex.

By theorem [13], we have:

[13] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, NY, 2004, p. 55.

Theorem 2.1.4

Two times continuously differentiable function $f \in \mathcal{F}^2(\mathbb{R}^n)$ iff for any $x \in \mathbb{R}^n$, we have:

$$f''(x) \succeq 0$$

Therefore, we need to show that $\nabla^2 g(x) \succeq 0$. We derive the following using the Scalar-by-vector identity [14]:

If $u = u(x)$ and $v = v(x)$ are vector functions of x, then:

$$\nabla(u \cdot v) = (\nabla u)v^T + u^T(\nabla v)$$

Hence, we have:

$$\begin{aligned}
\nabla^2 g(x) &= \nabla(e^{-\varphi(x)}\nabla\varphi(x)) \\
&= \left[\frac{d}{dx}(e^{-\varphi(x)})\right](\nabla\varphi(x))^T + e^{-\varphi(x)}\nabla^2\varphi(x) \\
&= -e^{-\varphi(x)}(\nabla\varphi(x))(\nabla\varphi(x))^T + e^{-\varphi(x)}\nabla^2\varphi(x) \\
&= e^{-\varphi(x)}\left[\nabla^2\varphi(x) - (\nabla\varphi(x))(\nabla\varphi(x))^T\right]
\end{aligned}$$

By our assumption, we knew that $\nabla^2\varphi(x) - \nabla\varphi(x)(\nabla\varphi(x))^T \succeq 0$, and multiplying by $e^{-\varphi(x)} > 0$ would not change the sign, therefore we have:

$$\nabla^2 g(x) \succeq 0$$

And the equivalence of the three statements is proved. $\square$

(3)

We're given:
The ratio of the d stocks on the t-th day:

$$x_t \in \Delta = \left\{x = (x[1], \ldots, x[d]) \in \mathbb{R}^d_+ : \sum_{i=1}^{d} x[i] = 1\right\},$$

The price relative on the t-th day:

$$a_t = (a_t[1], \ldots, a_t[d]) = \left(\frac{p_t^c[1]}{p_t^o[1]}, \ldots, \frac{p_t^c[d]}{p_t^o[d]}\right) \in \mathbb{R}^d_+$$

[14]"Matrix calculus", Wikipedia, `https://en.wikipedia.org/wiki/Matrix_calculus`

where:

$$p_t^c[i] : \text{the closing price of the } i\text{-th stock on the } t\text{-th day}$$
$$p_t^o[i] : \text{the opening price of the } i\text{-th stock on the } t\text{-th day}$$

Suppose $a_1, \ldots, a_T$ are i.i.d. random vectors, following known common probability distribution P.

Strategy:

$$x_t \in \operatorname*{argmin}_{x \in \Delta} f(x); \quad f(x) := \mathsf{E}\left[-\log\langle a_t, x\rangle\right], \quad \forall t \in \mathbb{N}$$

Assume f strictly convex.

(1)

Since Alice has one unit of wealth before the first day, let $W_0 = 1$. And let W_{t-1} be the wealth of Alice before the t-th day.

So after the end of the t-th day, Alice would have her wealth W_t:

$$\begin{aligned} W_t &= W_{t-1} \cdot x_t[1] \cdot a_t[1] + W_{t-1} \cdot x_t[2] \cdot a_t[2] + \cdots + W_{t-1} \cdot x_t[d] \cdot a_t[d] \\ &= W_{t-1} \cdot \langle a_t, x_t \rangle \end{aligned}$$

For example, if $a_t[1] = 2$, then the price of the first stock on day t is twice as high as the price on day $t-1$, we can then calculate how much Alice invests in the first stock on day t, which is $W_{t-1} \cdot x_t[1]$, and multiply this price relative to get the wealth on day t.
Using this formula, we knew that:

$$\begin{aligned} W_1 &= W_0 \cdot \langle a_1, x_1 \rangle \\ W_2 &= W_1 \cdot \langle a_2, x_2 \rangle = W_0 \cdot \langle a_1, x_1 \rangle \cdot \langle a_2, x_2 \rangle \\ W_3 &= W_2 \cdot \langle a_3, x_3 \rangle = W_0 \cdot \langle a_1, x_1 \rangle \cdot \langle a_2, x_2 \rangle \cdot \langle a_3, x_3 \rangle \\ &\vdots \\ W_T &= W_0 \cdot \langle a_1, x_1 \rangle \cdot \langle a_2, x_2 \rangle \cdot \cdots \cdot \langle a_T, x_T \rangle \end{aligned}$$

Which is the same as required since $W_0 = 1$, and we have:

$$W_T = \langle a_1, x_1 \rangle \cdot \langle a_2, x_2 \rangle \cdot \cdots \cdot \langle a_T, x_T \rangle \qquad \square$$

(2)

Our aim is to show that:

$$\lim_{T\to\infty} \mathsf{E}\left[\frac{W_T(x)}{W_T^\star}\right] \le 1, \quad \forall x \in \Delta \tag{*}$$

where $W_T(x)$ (given by the problem statement) and $W_T^\star$ (by the previous sub-problem) are defined as:

$$\begin{aligned} W_T(x) &:= \langle a_1, x\rangle \cdots \langle a_T, x\rangle = \prod_{t=1}^{T} \langle a_t, x\rangle \\ W_T^\star &:= \langle a_1, x_1\rangle \cdots \langle a_T, x_T\rangle = \prod_{t=1}^{T} \langle a_t, x_t\rangle \end{aligned}$$

From Alice's strategy, we have:

$$x_t \in \underset{x\in\Delta}{\operatorname{argmin}} f(x); \quad f(x) := \mathsf{E}\left[-\log\langle a_t, x\rangle\right], \quad \forall t \in \mathbb{N}$$

This means that Alice decides the ratio of the t-th day by finding the x that minimizes the function f, which is the expected loss of using x as the ratio.

By the fact that f is strictly convex and $x_t \in \underset{x\in\Delta}{\operatorname{argmin}} f(x)$, we have:

$$f(x_t) < f(x) \quad \forall x \in \Delta \setminus \{x_t\}$$

Replace by the definition of f, and using the fact that expectation is linear:

$$\begin{aligned} & \mathsf{E}\left[-\log\langle a_t, x_t\rangle\right] < \mathsf{E}\left[-\log\langle a_t, x\rangle\right] \quad \forall x \in \Delta \setminus \{x_t\} \\ \Rightarrow\ & -\mathsf{E}\left[\log\langle a_t, x_t\rangle\right] < -\mathsf{E}\left[\log\langle a_t, x\rangle\right] \quad \forall x \in \Delta \setminus \{x_t\} \\ \Rightarrow\ & \mathsf{E}\left[\log\langle a_t, x\rangle\right] - \mathsf{E}\left[\log\langle a_t, x_t\rangle\right] < 0 \quad \forall x \in \Delta \setminus \{x_t\} \\ \Rightarrow\ & \mathsf{E}\left[\log\langle a_t, x\rangle - \log\langle a_t, x_t\rangle\right] < 0 \quad \forall x \in \Delta \setminus \{x_t\} \\ \Rightarrow\ & \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathsf{E}\left[\log\langle a_t, x\rangle - \log\langle a_t, x_t\rangle\right] < 0 \quad \forall x \in \Delta \setminus \{x_t\} \\ \Rightarrow\ & \lim_{T\to\infty} \frac{1}{T}\mathsf{E}\left[\sum_{t=1}^{T} \left(\log\langle a_t, x\rangle - \log\langle a_t, x_t\rangle\right)\right] < 0, \quad \forall x \in \Delta \setminus \{x_t\} \end{aligned}$$

The above inequality would be equality only when $x_t = x$, so if we modify the set to not exclude x_t, we would have the following:

$$\lim_{T\to\infty} \frac{1}{T} \mathsf{E}\left[\sum_{t=1}^{T} \left(\log\langle a_t, x\rangle - \log\langle a_t, x_t\rangle\right)\right] \le 0, \quad \forall x \in \Delta$$

And this implies:

$$\begin{aligned}
&\lim_{T\to\infty} \frac{1}{T} \mathsf{E}\left[\sum_{t=1}^{T} \left(\log\langle a_t, x\rangle - \log\langle a_t, x_t\rangle\right)\right] \le 0, \quad \forall x \in \Delta \\
\Rightarrow\ &\lim_{T\to\infty} \frac{1}{T} \mathsf{E}\left[\sum_{t=1}^{T} \log\langle a_t, x\rangle - \sum_{t=1}^{T} \log\langle a_t, x_t\rangle\right] \le 0, \quad \forall x \in \Delta \\
\Rightarrow\ &\lim_{T\to\infty} \frac{1}{T} \mathsf{E}\left[\log \prod_{t=1}^{T} \langle a_t, x\rangle - \log \prod_{t=1}^{T} \langle a_t, x_t\rangle\right] \le 0, \quad \forall x \in \Delta \\
\Rightarrow\ &\lim_{T\to\infty} \frac{1}{T} \mathsf{E}\left[\log\left(\frac{\prod_{t=1}^{T} \langle a_t, x\rangle}{\prod_{t=1}^{T} \langle a_t, x_t\rangle}\right)\right] \le 0, \quad \forall x \in \Delta \\
\Rightarrow\ &\lim_{T\to\infty} \frac{1}{T} \mathsf{E}\left[\log\left(\frac{W_T(x)}{W_T^\star}\right)\right] \le 0, \quad \forall x \in \Delta \\
\Rightarrow\ &\mathsf{E}\left[\log\left(\frac{W_T(x)}{W_T^\star}\right)\right] \le 0, \quad \forall x \in \Delta
\end{aligned}$$

By Jensen's inequality [15]:

Jensen's inequality

If x is a random variable such that $x \in \operatorname{dom} f$ with probability one, and f is convex, then we have:

$$f\left(\mathsf{E}[x]\right) \le \mathsf{E}\left[f(x)\right]$$

(3)

(-14)

(4)

We're given:

[15] Boyd, S. P., and L. Vandenberghe, *Convex Optimization*, 1st ed., Cambridge University Press, Cambridge, UK, 2004, p. 77-78.

$$
\begin{aligned}
&f : \mathbb{R}^d \to \mathbb{R} \qquad \text{differentiable, may be non-convex} \\
&\nabla f : \ L\text{-Lipschitz}, \ L > 0 \quad i.e. \\
&\qquad \|\nabla f(y) - \nabla f(x)\|_* \leq L\|y - x\|, \quad \forall x, y \in \mathbb{R}^d \\
&\qquad \text{where } \|u\|_* := \max_{x \in \mathbb{R}^d, \|x\| \leq 1} \langle u, x \rangle
\end{aligned}
$$

And the definition of a point x being ϵ-stationary for some $\epsilon > 0$ if:

$$
\|\nabla f(x)\|_* \leq \varepsilon
$$

(1)

Need to show:

$$
f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d
$$

The thought is to use the proof process of Lemma 1.2.3 [16]:

Lemma 1.2.3

Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Then for any $x, y \in \mathbb{R}^n$, we have:

$$
|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2
$$

Let $g(\tau) = x + \tau(y - x), \quad$ where $\tau \in [0, 1]$, which means that $g(0) = x$ and $g(1) = y$. Then we have:

$$
\begin{aligned}
\frac{d}{d\tau} g(\tau) &= y - x \\
\nabla f(g(\tau)) &= \nabla f(x + \tau(y - x))
\end{aligned}
$$

Then, for all $x, y \in \mathbb{R}^d$, we have:

$$
\begin{aligned}
f(y) - f(x) &= \int_x^y \nabla f(g(\tau)) \cdot dg(\tau) \\
&= \int_0^1 \nabla f(x + \tau(y - x)) \cdot (y - x) d\tau \\
&= \int_0^1 \langle \nabla f(x + \tau(y - x)), y - x \rangle d\tau
\end{aligned}
$$

[16] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, NY, 2004, p. 22-23.

Which is the same as the following, using the fact that the integral is linear, and $f(x), y-x$ are not functions of τ:

$$\begin{aligned}
&f(y) = f(x) + \int_0^1 \langle \nabla f(x+\tau(y-x)), y-x\rangle d\tau \\
\Rightarrow\ &f(y) = f(x) + \int_0^1 \langle \nabla f(x+\tau(y-x)) + \nabla f(x) - \nabla f(x), y-x\rangle d\tau \\
\Rightarrow\ &f(y) = f(x) + \int_0^1 \langle \nabla f(x), y-x\rangle d\tau + \int_0^1 \langle \nabla f(x+\tau(y-x)) - \nabla f(x), y-x\rangle d\tau \\
\Rightarrow\ &f(y) = f(x) + \langle \nabla f(x), y-x\rangle + \int_0^1 \langle \nabla f(x+\tau(y-x)) - \nabla f(x), y-x\rangle d\tau
\end{aligned} \tag{*}$$

And we're given for any $x, y \in \mathbb{R}^d$:

$$\begin{aligned}
&\| \underbrace{\nabla f(y) - \nabla f(x)}_{\text{u}} \|_* \le L\|y-x\| \\
\Rightarrow\ &\max_{z\in\mathbb{R}^d, \|z\|\le 1} \langle \nabla f(y) - \nabla f(x), z\rangle \le L\|y-x\|
\end{aligned}$$

Let $y = x + \tau(y-x)$, then we have:

$$\begin{aligned}
&\|\nabla f(y) - \nabla f(x)\|_* = \|\nabla f(x+\tau(y-x)) - \nabla f(x)\|_* \le L\|x+\tau(y-x) - x\| = L\tau\|y-x\| \\
\Rightarrow\ &\|\nabla f(x+\tau(y-x)) - \nabla f(x)\|_* \le L\tau\|y-x\| \qquad (1)
\end{aligned}$$

Going back to the definition $\|u\|_* := \max_{x\in\mathbb{R}^d, \|x\|\le 1} \langle u, x\rangle$, this means that for any $z \in \mathbb{R}^d, \|z\| \le 1$:

$$\langle u, z\rangle \le \|u\|_*$$

If we want to expand the definition to arbitrary $v \in \mathbb{R}^d$ (not necessarily $\|v\| \le 1$), we can let $z = \frac{v}{\|v\|}$, then we have:

$$\begin{aligned}
&\langle u, z\rangle = \langle u, \frac{v}{\|v\|}\rangle = \frac{\langle u, v\rangle}{\|v\|} \le \|u\|_* \\
\Rightarrow\ &\langle u, v\rangle \le \|u\|_*\|v\| \qquad \forall u, v \in \mathbb{R}^d
\end{aligned}$$

Let $u = \nabla f(x+\tau(y-x)) - \nabla f(x)$, $v = y - x$, then we have:

$$\langle \nabla f(x+\tau(y-x)) - \nabla f(x), y-x \rangle \le \|\nabla f(x+\tau(y-x)) - \nabla f(x)\|_* \|y-x\| \tag{2}$$

Multiply the result of (1) by $\|y-x\|$, we have:

$$\begin{aligned} &\|\nabla f(x+\tau(y-x)) - \nabla f(x)\|_* \le L\tau\|y-x\| \\ \Rightarrow\ &\|\nabla f(x+\tau(y-x)) - \nabla f(x)\|_* \|y-x\| \le L\tau\|y-x\|^2 \end{aligned} \tag{3}$$

Combining (2) and (3), we have:

$$\langle \nabla f(x+\tau(y-x)) - \nabla f(x), y-x \rangle \le L\tau\|y-x\|^2$$

Substituting this back into $(*)$, we have:

$$\begin{aligned} &f(y) \le f(x) + \langle \nabla f(x), y-x \rangle + \int_0^1 L\tau\|y-x\|^2 d\tau \\ \Rightarrow\ &f(y) \le f(x) + \langle \nabla f(x), y-x \rangle + L\|y-x\|^2 \int_0^1 \tau d\tau \\ \Rightarrow\ &f(y) \le f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}\|y-x\|^2 \qquad \square \end{aligned}$$

(2)

We're given the algorithm (generalization of gradient descent):

$$\begin{aligned} &x_1 \in \mathbb{R}^d \\ &\text{for every } t \in \mathbb{N} \\ &\qquad x_{t+1} \in \operatorname*{argmin}_{x\in\mathbb{R}^d} \langle \nabla f(x_t), x - x_t \rangle + \frac{L}{2}\|x-x_t\|^2 \end{aligned}$$

Need to show:

$$f(x_{t+1}) - f(x_t) \le -\frac{1}{2L}\|\nabla f(x_t)\|_*^2, \quad \forall t \in \mathbb{N}$$

Let the function to minimize in the update rule be g:

$$\begin{aligned} &g : \mathbb{R}^d \to \mathbb{R} \\ &g(x) = \langle \nabla f(x_t), x - x_t \rangle + \frac{L}{2}\|x-x_t\|^2 \end{aligned}$$

Let $z = x - x_t$, then we have:

$$g(x) = \langle \nabla f(x_t), z \rangle + \frac{L}{2}\|z\|^2$$

So:

$$\arg\min_x g(x) = x_t + \arg\min_z \{\langle \nabla f(x_t), z \rangle + \frac{L}{2}\|z\|^2\}$$

Let $h(z) = \langle \nabla f(x_t), z \rangle + \frac{L}{2}\|z\|^2$, we can rearrange the equation as:

$$h(z) - \frac{L}{2}\|z\|^2 = \langle \nabla f(x_t), z \rangle$$

Using the following proposition [17]:

> **Proposition: Equivalent conditions of strong convexity**
>
> A differentiable function f is strongly convex with constant $\mu > 0$
>
> $$\Leftrightarrow g(x) = f(x) - \frac{\mu}{2}\|x\|^2 \text{ is convex}, \forall x$$

Since $\langle \nabla f(x_t), z \rangle$ is affine, $h(z) - \frac{L}{2}\|z\|^2$ is convex, so $h(z)$ is strongly convex with convexity parameter L.

Then, taking the gradient of $h(z)$ with respect to z, we have:

$$\begin{aligned} \partial h(z) &= \frac{d}{dz}\left(\langle \nabla f(x_t), z \rangle + \frac{L}{2}\|z\|^2\right) \\ &= \nabla f(x_t) + \partial\left(\frac{L}{2}\|z\|^2\right) \end{aligned}$$

Here we can be sure that the derivative of $\langle \nabla f(x_t), z \rangle$ is $\nabla f(x_t)$, since this term is linear in z, and a lienar map is differentiable, however, we need to take the subdifferential for $\frac{L}{2}\|z\|^2$, since the norm is uncertain. [18]

Using the following theorem [19]:

[17] *Strong Convexity*, available at: `https://xingyuzhou.org/blog/notes/strong-convexity`, accessed: Apr. 21, 2025.

[18] The definition of subdifferential is from Nesterov, Y. N., *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, NY, 2004, p. 126.

[19] Nesterov, Y. N., *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, NY, 2004, p. 129.

Theorem 3.1.15

We have $f(x^*) = \min_{x \in \operatorname{dom} f} f(x)$ iff.

$$0 \in \partial f(x^*)$$

Since we knew that $h(z)$ is strongly convex, this means that the above equation is equivalent to saying there exists an unique minimizer z^* for $h(z)$ such that:

$$0 \in \partial h(z^*)$$

so there exists $u \in \partial(\frac{1}{2}\|z\|^2)$ such that:

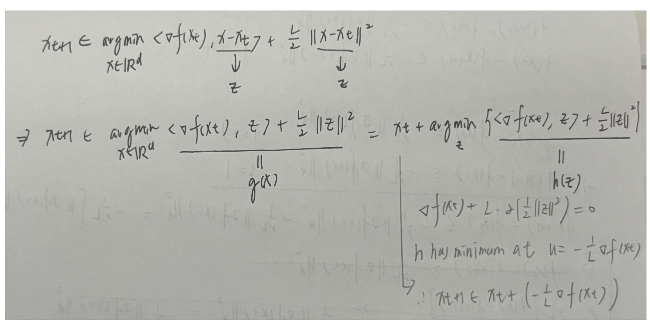
$$\nabla f(x_t) + Lu = 0$$
$$\Rightarrow u = -\frac{1}{L}\nabla f(x_t)$$

(-6) You need z not u in \partial (llzll).

Thus, the optimal update rule is:

$$x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t)$$

A clearer presentation of the above process is as the image below:



Define:

$$\phi : \mathbb{R}^d \to \mathbb{R}$$
$$\phi(z) = \frac{L}{2}\|z\|^2$$

Then the conjugate [20] of ϕ is defined as:

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004, p. 91. Available online at `https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`.

$$
\begin{aligned}
\phi^*(v) &= \sup_{z \in \mathbb{R}^d} \left(\langle v, z \rangle - \phi(z) \right) \\
&= \sup_{z \in \mathbb{R}^d} \left(\langle v, z \rangle - \frac{L}{2} \|z\|^2 \right)
\end{aligned}
$$

有沒有覺得很像 update rule / 4.1 的右半，
但是差一個負號

Let $z = \alpha z'$, where $\|\|z'\| = 1$, then:

$$
\begin{aligned}
\phi^*(v) &= \sup_{z \in \mathbb{R}^d} \left(\alpha \langle v, z' \rangle - \frac{L}{2} \langle \alpha z', \alpha z' \rangle \right) \\
&= \sup_{\alpha \in \mathbb{R}} \left(\alpha \langle v, z' \rangle - \frac{L\alpha^2}{2} \right)
\end{aligned}
$$

And by the definition of dual norm, we can derive the inequality (generalization of Cauchy-Schwarz inequality) [21] :

$$
\alpha \langle v, z' \rangle \leq \alpha \|v\|_* \|z'\| = \alpha \|v\|_*
$$

And the original conjugate can be rewritten as:

$$
\phi^*(v) = \sup_{\alpha \in \mathbb{R}} \left(\alpha \|v\|_* - \frac{L}{2} \alpha^2 \right)
$$

Taking the derivative:

$$
\frac{d}{d\alpha} \left(\alpha \|v\|_* - \frac{L}{2} \alpha^2 \right) = \|v\|_* - L\alpha \implies \alpha = \frac{\|v\|_*}{L}
$$

Plugging back in:

$$
\begin{aligned}
\phi^*(v) &= \frac{\|v\|_*}{L} \cdot \|v\|_* - \frac{L}{2} \cdot \frac{\|v\|_*^2}{L^2} \\
&= \frac{\|v\|_*^2}{L} - \frac{\|v\|_*^2}{2L} \\
&= \frac{\|v\|_*^2}{2L}
\end{aligned}
$$

(+2) 接近了…

By Fenchel's inequality [22] , which is stated as follows:

[21] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004, p. 637.
[22] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004, p. 94.

Fenchel's inequality

For all x, y:

$$f(x) + f^*(y) \geq \langle x, y \rangle$$

Therefore, we have for all $z, v \in \mathbb{R}^d$:

$$\begin{aligned} \phi(z) + \phi^*(v) &\geq \langle z, v \rangle \\ \Rightarrow \frac{L}{2}\|z\|^2 + \frac{\|v\|_*^2}{2L} &\geq \langle z, v \rangle \end{aligned}$$

Let $v = \nabla f(x_t)$, and since $z = x - x_t$, which means that choosing the optiaml z is equivalent to choosing the optimal x, which is x_{t+1}, so $z = x_{t+1} - x_t$, and we have:

$$\frac{L}{2}\|x_{t+1} - x_t\|^2 + \frac{1}{2L}\|\nabla f(x_t)\|_*^2 \geq \langle x_{t+1} - x_t, \nabla f(x_t) \rangle$$

By the result of subproblem (1), and plugging in $y = x_{t+1}$ and $x = x_t$, we have:

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ \Rightarrow f(x_{t+1}) - f(x_t) &\leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ \Rightarrow f(x_{t+1}) - f(x_t) &\leq \frac{1}{2L}\|\nabla f(x_t)\|_*^2 + \frac{L}{2}\|x_{t+1} - x_t\|^2 + \frac{L}{2}\|x_{t+1} - x_t\|^2 \end{aligned}$$

---

(3)

Claim:

$$\min_{1 \leq \tau \leq t} \|\nabla f(x_\tau)\|_*^2 \leq \frac{2L\left[f(x_1) - f(x_{t+1})\right]}{t}, \quad \forall t \in \mathbb{N}$$

By the result of the second subproblem, we have:

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L}\|\nabla f(x_t)\|_*^2 \qquad \forall t \in \mathbb{N}$$

And since this holds for all $t \in \mathbb{N}$, let $t = 1, \dots, t$:

$$
\begin{aligned}
f(x_{t+1}) - \cancel{f(x_t)} &\le -\frac{1}{2L}\|\nabla f(x_t)\|_*^2 \qquad (t = t) \\
\cancel{f(x_t)} - \cancel{f(x_{t-1})} &\le -\frac{1}{2L}\|\nabla f(x_{t-1})\|_*^2 \qquad (t = t-1) \\
&\vdots \\
\cancel{f(x_3)} - \cancel{f(x_2)} &\le -\frac{1}{2L}\|\nabla f(x_2)\|_*^2 \qquad (t = 2) \\
\cancel{f(x_2)} - f(x_1) &\le -\frac{1}{2L}\|\nabla f(x_1)\|_*^2 \qquad (t = 1)
\end{aligned}
$$

Summing up these inequalities, the terms $f(x_t)$ to $f(x_2)$ on the left hand side will cancel out, and we have:

$$
\begin{aligned}
& f(x_{t+1}) - f(x_1) \le -\frac{1}{2L}\sum_{i=1}^{t}\|\nabla f(x_i)\|_*^2 \\
\Rightarrow\ & f(x_1) - f(x_{t+1}) \ge \frac{1}{2L}\sum_{i=1}^{t}\|\nabla f(x_i)\|_*^2 \\
\Rightarrow\ & \frac{2L\left[f(x_1) - f(x_{t+1})\right]}{t} \ge \frac{1}{t}\sum_{i=1}^{t}\|\nabla f(x_i)\|_*^2 \quad \forall t \in \mathbb{N} \qquad (1)
\end{aligned}
$$

Since:

$$
\begin{aligned}
\min_{1\le\tau\le t}\|\nabla f(x_\tau)\|_*^2 &\le \|\nabla f(x_1)\|_*^2 \\
\min_{1\le\tau\le t}\|\nabla f(x_\tau)\|_*^2 &\le \|\nabla f(x_2)\|_*^2 \\
&\vdots \\
\min_{1\le\tau\le t}\|\nabla f(x_\tau)\|_*^2 &\le \|\nabla f(x_t)\|_*^2
\end{aligned}
$$

Summing up these inequalities, we have:

$$
\begin{aligned}
& t\min_{1\le\tau\le t}\|\nabla f(x_\tau)\|_*^2 \le \sum_{i=1}^{t}\|\nabla f(x_i)\|_*^2 \\
\Rightarrow\ & \min_{1\le\tau\le t}\|\nabla f(x_\tau)\|_*^2 \le \frac{1}{t}\sum_{i=1}^{t}\|\nabla f(x_i)\|_*^2
\end{aligned}
$$

Plugging this back into (1), we have:

$$\min_{1\le\tau\le t} \|\nabla f(x_\tau)\|_*^2 \le \frac{2L\left[f(x_1) - f(x_{t+1})\right]}{t}, \quad \forall t \in \mathbb{N} \qquad \square$$

(4) (-10) Show that this update rule is the algorithm in 4.2 with ℓ_∞-norm

We're given the algorithm:

$$x_1 \in \mathbb{R}^d$$

$$\text{For every } t \in \mathbb{N}, \qquad x_{t+1} = x_t - \frac{\|\nabla f(x_t)\|_1}{L}\text{sign}(\nabla f(x_t))$$

where:

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x \ge 0 \\ -1 & \text{if } x < 0 \end{cases} \text{ for any } x \in \mathbb{R}$$

$$\text{sign}(v) = \begin{bmatrix} \text{sign}(v[1]) \\ \vdots \\ \text{sign}(v[d]) \end{bmatrix} \text{ for any } v = \begin{bmatrix} v[1] \\ \vdots \\ v[d] \end{bmatrix} \in \mathbb{R}^d$$

Denote:

$$\nabla f(x_t) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x_t) \\ \vdots \\ \frac{\partial f}{\partial x_d}(x_t) \end{bmatrix} \in \mathbb{R}^d$$

Then:

$$\text{sign}(\nabla f(x_t)) = \begin{bmatrix} \text{sign}(\frac{\partial f}{\partial x_1}(x_t)) \\ \vdots \\ \text{sign}(\frac{\partial f}{\partial x_d}(x_t)) \end{bmatrix} \quad , \text{ and } \|\nabla f(x_t)\|_1 = \sum_{i=1}^{d} \left| \frac{\partial f}{\partial x_i}(x_t) \right|$$

Note that l_1-norm is nondifferentiable, so to find the subgradient, we consider l_1-norm in the following form [23]

$$\|x\|_1 = \{\max s^T x \mid s_i \in \{-1, 1\}\}$$

[23]S. Boyd and L. Vandenberghe, *Subgradients*, Notes for EE364b, Stanford University, Winter 2006–07, Apr. 13, 2008. Available at: `https://see.stanford.edu/materials/lsocoee364b/01-subgradients_notes.pdf`, p. 5.

And we could find the unique s by choosing $s_i = +1$ if $x_i \geq 0$, and $s_i = -1$ if $x_i < 0$, this is equivalent to saying that for the case $x = \nabla f(x_t)$, if $\frac{\partial f}{\partial x_i}(x_t) \geq 0$, then $s_i = 1$, otherwise $s_i = -1$, and we could see that s is actually $\text{sign}(\nabla f(x_t))$.

Thus, the update rule of this algorithm can be rewritten as:

$$x_{t+1} = x_t - \frac{s^T \nabla f(x_t)}{L} s$$

In the second subproblem, we've shown that the optimal update rule is:

$$x_{t+1} = x_t - \frac{1}{L} \nabla f(x_t)$$