Labs
Optimization for Machine Learning
Spring 2024

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Nicolas Flammarion
github.com/epfml/OptML_course

Problem Set 4 — Solutions (Subgradient Descent)

Subgradient Descent

Solve Exercises 28, 29 , 30 , 32 and 27 from the lecture notes.

Exercise 28. *Prove Lemma 4.2, meaning that a function that is differentiable at* $\mathbf{x}$ *has at most one subgradient there, namely* $\nabla f(\mathbf{x})$.

Solution: Let $\mathbf{g}$ be a subgradient at $\mathbf{x}$. Together with differentiabilty at $\mathbf{x}$ (Definition 1.5), we derive the inequality

$$(\mathbf{g} - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) \leq r_{\mathbf{x}}(\mathbf{y} - \mathbf{x})$$

for all $\mathbf{y}$ in some neighborhood of $\mathbf{x}$, where $r_{\mathbf{x}}$ is a sublinear error function ($r_{\mathbf{x}}(\mathbf{v})/\|\mathbf{v}\| \to 0$ as $\mathbf{v} \to 0$). Then it should also hold for all $\mathbf{y}_\varepsilon = \varepsilon \mathbf{e}_i + \mathbf{x}$ for small enough ε, where $\mathbf{e}_i$ is the i-th coordinate vector. Substituting $\mathbf{y}_\varepsilon$ and dividing both sides with $\|\mathbf{y} - \mathbf{x}\|$ we get

$$\frac{(\mathbf{g} - \nabla f(\mathbf{x}))^\top (\varepsilon \mathbf{e}_i)}{\varepsilon \|\mathbf{e}_i\|} \leq \frac{r_{\mathbf{x}}(\varepsilon \mathbf{e}_i)}{\|\varepsilon \mathbf{e}_i\|}$$

We see that on the left hand side ε cancels and the term does not depend on it, while the right part goes to zero as $\varepsilon \to 0$ since r_x is sublinear function. This means that the left part has to be zero, i.e. $(\mathbf{g} - \nabla f(\mathbf{x}))^\top \mathbf{e}_i = 0$ and this should hold for any i. This is possible only when $\mathbf{g} = \nabla f(\mathbf{x})$.

Exercise 29. *Prove the easy direction of Lemma 4.3, meaning that the existence of subgradients everywhere implies convexity!*

Solution: Let's assume that we have subgradients everywhere. With $\mathbf{g} \in \partial f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y})$, (4.1) yields

$$\begin{aligned} f(\mathbf{x}) &\geq f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) + \mathbf{g}^\top ((1-\lambda)(\mathbf{x} - \mathbf{y})), \\ f(\mathbf{y}) &\geq f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) + \mathbf{g}^\top (\lambda(\mathbf{y} - \mathbf{x})). \end{aligned}$$

Adding up these two inequalities with multiples λ and $1 - \lambda$ cancels the subgradient terms and yields

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}),$$

which is convexity.

Exercise 30. *Prove Lemma 4.4 (Lipschitz continuity and bounded subgradients).*

Solution: We assume that $\mathbf{dom}(f) = \mathbb{R}^d$ and hint at the general case. $ii \implies i$: Given any $\mathbf{x} \in \mathbb{R}^d$ ("harder" alternative: $\mathbf{x}$ in a convex domain $D = \mathbf{dom}(f)$), consider $\mathbf{g}$ an element of $\partial f(\mathbf{x})$. Let $\mathbf{z} = \mathbf{x} + \mathbf{g}$ (alternative: let $\eta > 0$ such that $\mathbf{z} = \mathbf{x} + \eta \mathbf{g}$ is still in D).

Since f is B-Lipschitz, we have

$$f(\mathbf{z}) - f(\mathbf{x}) \leq B \cdot \|\mathbf{z} - \mathbf{x}\| = B \cdot \|\mathbf{g}\|.$$

(Alternative $\cdots \leq \eta \cdot \|\mathbf{g}\|$.)

Using the definition of subgradient, we have:

$$f(\mathbf{z}) - f(\mathbf{x}) \geq \mathbf{g}^\top (\mathbf{z} - \mathbf{x}) = \|\mathbf{g}\|^2.$$

(Alternative: $\cdots \geq \eta \cdot \|\mathbf{g}\|^2$.)

Combining the inequalities, we have $\|\mathbf{g}\| \leq B$ (the η is simplified on both sides in the alternative situation when $\mathbf{x}$ is drawn from a domain D and not from all $\mathbb{R}^d$ and we get the same result.)

$i \implies ii$:

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and let $\mathbf{g}$ be any element in $\partial f(\mathbf{x})$, by definition of subgradient we have: $f(\mathbf{y}) - f(\mathbf{x}) \geq \mathbf{g}^\top(\mathbf{y}-\mathbf{x})$, therefore, by inversing the signs in the inequality, then using Cauchy-Schwartz and finally the bound on the norm of the subgradient, we have:

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{y}) &\leq \mathbf{g}^\top(\mathbf{x}-\mathbf{y}) \\ &\leq \|\mathbf{g}\| \cdot \|\mathbf{x}-\mathbf{y}\| \\ &\leq B \cdot \|\mathbf{x}-\mathbf{y}\| \end{aligned}$$

which is the desired inequality to conclude that ii holds.

Note: in the case where f is defined on a convex domain D, the latter is assumed to be open in the alternative situation described above. If not, the reasoning applies for any $\mathbf{x}$ in the interior of D.

Exercise 32. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is convex and satisfies*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2$$

for all $\mathbf{x}$ such that $\nabla f(\mathbf{x})$ exists, and for all $\mathbf{y}$. Prove that this implies

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}_\mathbf{x}^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2$$

for all $\mathbf{x}$, all $\mathbf{g}_\mathbf{x} \in \partial f(\mathbf{x})$ and all $\mathbf{y}$.

Solution: We first show that the conclusion holds for all limit subgradients $\mathbf{g}$ of the form $\mathbf{g} = \lim_{n\to\infty} \nabla f(\mathbf{x}_n)$ where $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$. We have

$$f(\mathbf{y}) \geq f(\mathbf{x}_n) + \nabla f(\mathbf{x}_n)^\top(\mathbf{y}-\mathbf{x}_n) + \frac{\mu}{2}\|\mathbf{x}_n-\mathbf{y}\|^2, \quad n \in \mathbb{N},$$

so this inequality also holds in the limit. Continuity of f and $\|\cdot\|^2$, convergence of gradients, and the fact that limits and products commute, implies that

$$\begin{aligned} \lim_{n\to\infty} f(\mathbf{x}_n) &= f(\mathbf{x}), \\ \lim_{n\to\infty} \frac{\mu}{2}\|\mathbf{x}_n-\mathbf{y}\|^2 &= \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2, \\ \lim_{n\to\infty} \nabla f(\mathbf{x}_n)^\top(\mathbf{y}-\mathbf{x}_n) &= \mathbf{g}^\top(\mathbf{y}-\mathbf{x}). \end{aligned}$$

This yields the statement for any limit subgradient $\mathbf{g}$ at $\mathbf{x}$, i.e., it holds that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2.$$

By Theorem 4.6, every subgradient at $\mathbf{x}$ is a convex combination of limit subgradients, $\mathbf{g}_\mathbf{x} = \sum_i \lambda_i \mathbf{g}_i$, $\sum_i \lambda_i = 1$, $\lambda_i \geq 0$ for all i. Hence, using the above statement for limit subgradients, we get

$$\begin{aligned} f(\mathbf{y}) = \sum_i \lambda_i f(\mathbf{y}) &\geq \sum_i \lambda_i f(\mathbf{x}) + \sum_i \lambda_i \mathbf{g}_i^\top(\mathbf{y}-\mathbf{x}) + \sum_i \lambda_i \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2 \\ &= f(\mathbf{x}) + g_\mathbf{x}^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2. \end{aligned}$$

Exercise 27. *Prove Theorem 3.14!*

Solution: From (3.17), the proximal step could be written as

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^d}\{g(\mathbf{x}_t) + \nabla g(\mathbf{x}_t)^\top(\mathbf{y}-\mathbf{x}_t) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}_t\|^2 + h(\mathbf{y})\} = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^d}\{\psi(\mathbf{y})\},$$

where the function $\psi(\mathbf{y}) = g(\mathbf{x}_t) + \nabla g(\mathbf{x}_t)^\top(\mathbf{y}-\mathbf{x}_t) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}_t\|^2 + h(\mathbf{y})$ is strongly convex with the parameter L. This means that $\psi(\mathbf{y}) \geq \psi(\mathbf{x}_{t+1}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}_{t+1}\|^2$. This is equivalent to

$$\nabla g(\mathbf{x}_t)^\top(\mathbf{y}-\mathbf{x}_t) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}_t\|^2 + h(\mathbf{y}) \geq \nabla g(\mathbf{x}_t)^\top(\mathbf{x}_{t+1}-\mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + h(\mathbf{x}_{t+1}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}_{t+1}\|^2,$$

Rearranging terms and subtracting $h(\mathbf{x}_t)$ from both sides,

$$\nabla g(\mathbf{x}_t)^\top(\mathbf{y}-\mathbf{x}_t)+\frac{L}{2}\|\mathbf{y}-\mathbf{x}_t\|^2-\frac{L}{2}\|\mathbf{y}-\mathbf{x}_{t+1}\|^2+h(\mathbf{y})-h(\mathbf{x}_t) \geq \nabla g(\mathbf{x}_t)^\top(\mathbf{x}_{t+1}-\mathbf{x}_t)+\frac{L}{2}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2+h(\mathbf{x}_{t+1})-h(\mathbf{x}_t)$$

As the function g is L-smooth, we can estimate the right side as $g(\mathbf{x}_t)^\top(\mathbf{x}_{t+1}-\mathbf{x}_t)+\frac{L}{2}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 \geq g(\mathbf{x}_{t+1})-g(\mathbf{x}_t)$, and because g is convex, on the left side we estimate $\nabla g(\mathbf{x}_t)^\top(\mathbf{y}-\mathbf{x}_t) \leq g(\mathbf{y})-g(\mathbf{x}_t)$. Putting this together

$$f(\mathbf{y})-f(\mathbf{x}_t)+\frac{L}{2}\|\mathbf{y}-\mathbf{x}_t\|^2-\frac{L}{2}\|\mathbf{y}-\mathbf{x}_{t+1}\|^2 \geq f(\mathbf{x}_{t+1})-f(\mathbf{x}_t)$$

This holds for any $\mathbf{y} \in \mathbb{R}^d$. Lets take $\mathbf{y}=\mathbf{x}^*$ and sum up the inequation above from $t=0$ to $t=T-1$

$$\sum_{t=0}^{T-1}(f(\mathbf{x}^*)-f(\mathbf{x}_t))+\frac{L}{2}\|\mathbf{x}^*-\mathbf{x}_0\|^2-\frac{L}{2}\|\mathbf{x}^*-\mathbf{x}_T\|^2 \geq f(\mathbf{x}_T)-f(\mathbf{x}_0)$$

or equivalently,

$$\sum_{t=1}^{T}(f(\mathbf{x}_t)-f(\mathbf{x}^*)) \leq \frac{L}{2}\|\mathbf{x}^*-\mathbf{x}_0\|^2-\frac{L}{2}\|\mathbf{x}^*-\mathbf{x}_T\|^2 \leq \frac{L}{2}\|\mathbf{x}^*-\mathbf{x}_0\|^2$$

Note that $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$, as $\psi(\mathbf{x}_{t+1}) \leq \psi(\mathbf{x}_t)$ for each $0 \leq t \leq T$

$$f(\mathbf{x}_t)-f(\mathbf{x}^*) \leq \frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{x}_t)-f(\mathbf{x}^*)) \leq \frac{L}{2T}\|\mathbf{x}^*-\mathbf{x}_0\|^2.$$

Random Walks

Gradient descent turns up in a surprising number of situations which apriori have nothing to do with optimization. In this exercise, we will see how performing a random walk on a graph can be seen as a special case of gradient descent.

We are given an *undirected* graph $G(V,E)$ with vertices $V=[n]$ labelled 1 through n, and edges $E \subseteq [n]^2$ such that if $(i,j) \in E$, then $(j,i) \in E$. Further, we assume that the graph is *regular* in the sense that every edge has the same degree. Let d be the degree of each node such that if we denote $\mathcal{N}(i)=\{j:(i,j)\in E\}$ to be the neighbors of i, then $|\mathcal{N}(i)|=d$. We assume that every node is connected to itself and so $(i,i)\in\mathcal{N}(i)$.

Now we start our random walk from node 1, jumping randomly from a node to its neighbor. More precisely, suppose at time step t we are at node i_t. Then i_{t+1} is picked uniformly at random from $\mathcal{N}(i)$. If we run this random walk for a large enough T steps, we expect that $\Pr(i_T=j)=1/n$ for any $j\in[n]$. This is called the stationary distribution.

Problem A. Let us represent the position at time step t in the graph with $\mathbf{e}_{i_t} \in \mathbb{R}^n$ where the i_tth coordinate is 1 and all others are 0. Then, the vector $\mathbf{x}_t=\mathbb{E}[\mathbf{e}_{i_t}]$ denotes the probability distribtion over the n nodes of the graph. Further, let us denote $\mathbf{G}\in\mathbb{R}^{n\times n}$ be the transition probability matrix such that

$$\mathbf{G}_{i,j}=\begin{cases}\frac{1}{d} & \text{if } (i,j)\in E\\ 0 & \text{otherwise}\,.\end{cases}$$

Show that

$$\mathbf{x}_{t+1}=\mathbf{G}\mathbf{x}_t \tag{1}$$

Solution: Let look at one coordinate j of random vector $\mathbf{x}_{t+1}=\mathbb{E}[\mathbf{e}_{i_{t+1}}]$. Then by the low of total probability, the expectation of this coordinate would be

$$[\mathbf{x}_{t+1}]_j=\mathbb{E}[\mathbf{e}_{i_{t+1}}]_j=\Pr\left([\mathbf{e}_{i_{t+1}}]_j=1\right)=\sum_k \Pr(i_{t+1}=j|i_t=k)\Pr(i_t=k)=\sum_k \Pr(i_{t+1}=j|i_t=k)\Pr\left([\mathbf{e}_{i_t}]_k=1\right)$$
$$=\sum_k \Pr(i_{t+1}=j|i_t=k)\mathbb{E}[\mathbf{e}_{i_t}]_k=\sum_k \Pr(i_{t+1}=j|i_t=k)[\mathbf{x}_t]_j$$

Note, that for $k: \ (j,k) \notin E$, $\Pr(i_{t+1} = j | i_t = k) = 0 = \mathbf{G}_{j,k}$ and for $k: \ (j,k) \in E$, $\Pr(i_{t+1} = j | i_t = k) = \frac{1}{d} = \mathbf{G}_{j,k}$. This means that

$$[\mathbf{x}_{t+1}]_j = \sum_k \mathbf{G}_{jk} [\mathbf{x}_t]_k ,$$

or equivalently

$$\mathbf{x}_{t+1} = \mathbf{G}\mathbf{x}_t \tag{2}$$

Problem B. Simulate the random walk above over a torus and confirm that we indeed converge to a uniform distribution over the nodes. What is the *rate* at which this convergence occurs?

Follow the Python notebook provided here:

github.com/epfml/OptML_course/tree/master/labs/ex04/

Problem C. Define $\mu = \frac{1}{n}\mathbf{1}_n$ be a vector of all $1/n$, and a objective function $f : \mathcal{S} \to \mathbb{R}$ as

$$f(\mathbf{x}) = (\mathbf{x} - \mu)^\top (\mathbf{I} - \mathbf{G})(\mathbf{x} - \mu) ,$$

defined over the probability simplex $\mathcal{S} \subseteq \mathbb{R}^n$ where $\mathcal{S} = \{\mathbf{v} : \mathbf{1}_n^\top \mathbf{v} = 1, v_i \geq 0\}$.

1. Show that f defined above is convex and compute its smoothness constant.
2. Show that running gradient descent on f with the correct step-size is equivalent to the random walk step (1).
3. Prove that $\mathbf{x}_t$ converges to the distribution μ at a linear rate i.e. for the random walk on a torus with n nodes,

$$\|\mathbf{x}_t - \mu\|_2^2 \leq \left(1 - \frac{1}{n}\right)^t \|\mathbf{x}_0 - \mu\|_2^2 \leq \left(1 - \frac{1}{n}\right)^t .$$

Hint: Use that the two largest eigenvalues of $\mathbf{G}$ are 1 and $1 - \frac{1}{n}$. Also $\mathbf{G}\mu = \mu$ and so μ is the eigenvector corresponding to eigenvalue 1.

Solution:

1. By the second order characterization of convexity (Lemma 1.18) the function is convex if its hessian is positive semidefinite. Lets show that

$$\nabla^2 f(\mathbf{x}) = 2(\mathbf{I} - \mathbf{G}) \succeq 0$$

For any vector $\mathbf{z}$

$$\begin{aligned}
\mathbf{z}^\top (\mathbf{I} - \mathbf{G})\mathbf{z} &= \sum_{i=1}^n z_i^2 - \sum_{i=1}^n \sum_{j=1}^n \mathbf{G}_{ij} z_i z_j = d \sum_{i=1}^n \frac{1}{d} z_i^2 - \sum_{i=1}^n \sum_{j:(i,j)\in E} \frac{1}{d} z_i z_j = \\
&= (d-1) \sum_{i=1}^n \frac{1}{d} z_i^2 - \sum_{i=1}^n \sum_{j<i:(i,j)\in E} \frac{2}{d} z_i z_j = \sum_{i=1}^n \frac{1}{d} \sum_{j<i:(i,j)\in E} z_i^2 + z_j^2 - 2 z_i z_j \\
&= \sum_{i=1}^n \frac{1}{d} \sum_{j<i:(i,j)\in E} (z_i - z_j)^2 \geq 0.
\end{aligned}$$

where we used that the $\mathbf{G}$ is symmetric because the graph is undirected and that every row of $\mathbf{G}$ had exactly d non-zero elements.

Let us prove now that the function f is L-smooth with smoothness constant $L = 2$. From Exercise 14 we know that $L = 2\|I - G\|$, and we claim that $\|I - G\|$ is less than 1. As we already showed above,

$$\mathbf{z}^\top (\mathbf{I} - \mathbf{G})\mathbf{z} = \sum_{i=1}^n \frac{1}{d} \sum_{j<i:(i,j)\in E} (z_i - z_j)^2.$$

Using that $z_i > 0 \; \forall i$,

$$\mathbf{z}^\top(\mathbf{I}-\mathbf{G})\mathbf{z} \le \frac{1}{d}\sum_{i=1}^{n}\sum_{j<i:(i,j)\in E} z_i^2 + z_j^2 = \frac{d-1}{d}\sum_{i=1}^{n} z_i^2 < \|\mathbf{z}_i\|^2$$

This means that $\|\mathbf{I}-\mathbf{G}\| < 1$.

2. The gradient of f is

$$\nabla f(\mathbf{x}) = 2(\mathbf{I}-\mathbf{G})(\mathbf{x}_t-\mu) = 2(\mathbf{I}-\mathbf{G})\mathbf{x}_t - 2(\mu-\mathbf{G}\mu) = 2(\mathbf{I}-\mathbf{G})\mathbf{x}_t\,.$$

The last equality followed since $\mathbf{G}\mu = \mu$. With the stepsize $\gamma = \frac{1}{L} = \frac{1}{2}$ gradient descent will take form

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{2}\nabla f(\mathbf{x}_t) = \mathbf{x}_t - \frac{1}{2}2(\mathbf{I}-\mathbf{G})\mathbf{x}_t = \mathbf{G}\mathbf{x}_t\,.$$

Since our problem is constrained to the set $\mathcal{S}$, we have to make sure $\mathbf{x}_{t+1}$ also lies in $\mathcal{S}$. This is easy to verify.

3. To show the linear convergence rate, we first will prove that function f restricted to the set $\mathcal{S}$ is strongly convex with parameter $\frac{2}{n}$. Then, the convergence rate would follow from the Theorem 2.11.

To find strong convexity coefficient we need to show a lower bound on $(\mathbf{y}-\mathbf{x})^\top\nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}) = (\mathbf{y}-\mathbf{x})^\top 2(\mathbf{I}-\mathbf{G})(\mathbf{y}-\mathbf{x})$ for $\mathbf{x},\mathbf{y}\in\mathcal{S}$. For that we will find the minimum

$$\min_{\mathbf{y},\mathbf{x}\in\mathcal{S}} \frac{(\mathbf{y}-\mathbf{x})^\top(\mathbf{I}-\mathbf{G})(\mathbf{y}-\mathbf{x})}{\|\mathbf{y}-\mathbf{x}\|^2}$$

First, let's show that $\mathbf{y}-\mathbf{x}\perp\mu \; \forall \mathbf{x},\mathbf{y}\in\mathcal{S}$. Indeed,

$$(\mathbf{y}-\mathbf{x})^\top\mu = \mathbf{y}^\top\mu - \mathbf{x}^\top\mu = \frac{1}{n}-\frac{1}{n} = 0\,.$$

Here we used that $\sum_i y_i = 1$ and $\sum_i x_i = 1$.

Then

$$\min_{\mathbf{y},\mathbf{x}\in\mathcal{S}} \frac{(\mathbf{y}-\mathbf{x})^\top(\mathbf{I}-\mathbf{G})(\mathbf{y}-\mathbf{x})}{\|\mathbf{y}-\mathbf{x}\|^2} \ge \min_{\mathbf{z}\perp\mu} \frac{\mathbf{z}^\top(\mathbf{I}-\mathbf{G})\mathbf{z}}{\|\mathbf{z}\|^2}\,.$$

Recall that μ is the principal eigenvector. Then, the right hand side of the above equation characterizes the second largest eigenvalue. In the basis of orthonormal eigenvectors $\{\mathbf{v}_i\}_{i=1}^n$ of $\mathbf{I}-\mathbf{G}$ vector $\mathbf{z}$ is represented as $\mathbf{z} = \sum_{i=2}^n \alpha_i\mathbf{v}_i$, because it is orthogonal to $\mathbf{v}_1 = \mu$. Then

$$\min_{\mathbf{z}\perp\mu} \frac{\mathbf{z}^\top(\mathbf{I}-\mathbf{G})\mathbf{z}}{\|\mathbf{z}\|^2} = \min_{\alpha_2,\dots,\alpha_n} \frac{\sum_{i=2}^n \alpha_i^2\lambda_i}{\sum_{i=2}^n \alpha_i^2} = \lambda_2 = \frac{1}{n}.$$

This shows that f is $\frac{2}{n}$ strongly convex when restricted to $\mathcal{S}$.