Idan Drori
Ohad Sapir

Question #1

(i) Let $x, y \in S,\ g_1, g_2 \in \partial f(x),\ \lambda \in [0,1]$.Define $g = \lambda g_1 + (1-\lambda)g_2$.
From the subgradient inequality:

$$\lambda f(y) \geq \lambda f(x) + \lambda g_1^\top (y-x)$$

$$(1-\lambda) f(y) \geq (1-\lambda) f(x) + (1-\lambda) g_2^\top (y-x)$$

By combining the two equations:

$$\begin{aligned} f(y) = \lambda f(y) + (1-\lambda) f(y) &\geq \lambda f(x) + \lambda g_1^\top (y-x) + (1-\lambda) f(x) + (1-\lambda) g_2^\top (y-x) \\ &= f(x) + \lambda g_1^\top (y-x) + (1-\lambda) g_2^\top (y-x) \\ &= f(x) + g^\top (y-x) \end{aligned}$$

So $g \in \partial f(x)$, meaning $\partial f(x)$ is convex.

(ii) Let $x, y \in S,\ \lambda \in [0,1],\ z = \lambda x + (1-\lambda) y,\ g \in \partial f(z)$. From the subgradient inequality:

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

(iii) Let $x, y \in S,\ g_x \in \partial f(x), g_y \in \partial f(y)$. From the subgradient inequality:

$$f(y) + f(x) \geq f(x) + g_x^\top (y-x) + f(y) + g_y^\top (x-y)$$

$$0 \geq g_x^\top (y-x) + g_y^\top (x-y)$$

$$0 \leq (+g_y - +g_x)^\top (y-x)$$

(iv) Let $b = f(0), h(x) = -b, g(x) = f(x) + h(x)$.
g is a sum of two convex and concave, and therefore g is also convex and concave. we'll prove that g is linear.
Homogeneity: Let $x \in \mathbb{R}^d$.
Let $\lambda \in [0,1]$.

$$g(0) = f(0) + h(0) = b - b = 0$$

Since g is both convex and concave,

$$g(\lambda x + (1-\lambda)0) = \lambda g(x) + (1-\lambda) g(0)$$

$$g(\lambda x) = \lambda g(x)$$

Let $\lambda > 1$. $\lambda^{-1} \in [0, 1]$

$$\lambda g(x) = \lambda g(\lambda^{-1} \lambda x) = \lambda \lambda^{-1} g(\lambda x) = g(\lambda x)$$

Let $\lambda < 0$. Let $y \in \mathbb{R}^d$.

$$0.5g(y) + 0.5g(-y) = g(0.5y - 0.5y) = g(0) = 0$$

$$-g(y) = g(-y)$$

$$\lambda g(x) = -(-\lambda)g(x) = -g(-(\lambda x)) = -(-g(\lambda x)) = g(\lambda x)$$

Additivity: Let $x, y \in \mathbb{R}^d$.

$$g(x + y) = g(0.5(2x) + 0.5(2y) = 0.5g(2x) + 0.5g(2y) = g(x) + g(y)$$

So g is linear, meaning there exists $a \in \mathbb{R}^d$, such that $g(x) = a^\top x$. And now:

$$f(x) = g(x) - h(x)$$

$$f(x) = a^\top x + b$$

Question #2

(i) We'll define $f_x(z) = f(z) - \nabla f(x)^T \cdot z$ and $f_y(z) = f(z) - \nabla f(y)^T \cdot z$. Calculating their gradients we get:

$$\nabla f_x(z) = \nabla f(z) - \nabla f(x)$$
$$\nabla f_y(z) = \nabla f(z) - \nabla f(y)$$

From the fact that f is β-smooth, we get that ∇f is β-Lipschitz, and so both $\nabla f_x(z)$ and $\nabla f_y(z)$ are β-Lipschitz. We can also see that $z = x, z = y$ are minimizers of $f_x(z)$ and $f_y(z)$ respectively ($\nabla f_x(x) = 0$, $\nabla f_y(y) = 0$).
We've seen that given f is β-smooth and x^* is its minimizer:

$$\forall x \in \mathbb{R}^d, \ \frac{1}{2\beta}\|\nabla f(x)\|^2 \leq f(x) - f(x^*) \leq \frac{\beta}{2}\|x - x^*\|^2$$

From the left hand side of that inequality we get:

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^T \cdot (y - x) = f_x(y) - f_x(x) &\geq \frac{1}{2\beta}\|\nabla f_x(y)\|^2 \\ &= \frac{1}{2\beta}\|\nabla f(y) - \nabla f(x)\|^2 \end{aligned}$$

And similarly because $z = y$ is a minimizer for $f_y(z)$ we get:

$$f(x) - f(y) - \nabla f(y)^T \cdot (x - y) \geq \frac{1}{2\beta}\|\nabla f(y) - \nabla f(x)\|^2$$

Combining the two inequalities we get:

$$\frac{1}{2\beta}\|\nabla f(y) - \nabla f(x)\|^2 + \frac{1}{2\beta}\|\nabla f(y) - \nabla f(x)\|^2 \leq$$
$$f(y) - f(x) - \nabla f(x)^T \cdot (y - x) + f(x) - f(y) - \nabla f(y)^T \cdot (x - y)$$
$$\Rightarrow \frac{1}{\beta}|\nabla f(y) - \nabla f(x)\|^2 \leq (\nabla f(y) - \nabla f(x))^T \cdot (y - x)$$

As required.

(ii)

$$\begin{aligned}\|x_1^+ - x_2^+\|^2 &= \|(x_1 - x_2) - \eta(\nabla f(x_1) - \nabla f(x_2))\|^2 \\ &= \|x_1 - x_2\|^2 - \eta(\nabla f(x_1) - \nabla f(x_2)) \cdot (x_1 - x_2) + \eta^2\|\nabla f(x_1) - \nabla f(x_2)\|^2 \\ &\underset{(*)}{\leq} \|x_1 - x_2\|^2 - \eta \cdot \frac{1}{\beta} \cdot \|\nabla f(x_1) - \nabla f(x_2)\|^2 + \frac{\eta}{\beta}\|\nabla f(x_1) - \nabla f(x_2)\|^2 \\ &= \|x_1 - x_2\|^2\end{aligned}$$

$(*)$ Inequality from co-coercity and the fact that $\eta \leq \frac{1}{\beta}$

(iii) x^* is a minimizer, so $\nabla f(x^*) = 0$, and so $(x^*)^+ = x^* - \eta\nabla f(x^*) = x^*$. Using (ii) we get that:

$$\|x^+ - x^*\| = \|x^+ - (x^*)^+\| \leq \|x - x^*\|$$

Question #3

(i) We'll prove it. The function $g(x) = \frac{\beta}{2}\|x\|^2$ is β-strongly convex. So we have for all $x, y \in \mathbb{R}^d$:

$$g(y) - g(x) - \nabla g(x) \cdot (y - x) \geq \frac{\beta}{2}\|y - x\|^2$$
$$|f(y) - f(x) - \nabla f(x) \cdot (y - x)| \leq \frac{\beta}{2}\|y - x\|^2$$

And so we get:

$$\begin{aligned}h(y) - h(x) - \nabla h(x) \cdot (y - x) &= f(y) - f(x) - \nabla f(x) \cdot (y - x) + g(y) - g(x) - \nabla g(x) \cdot (y - x) \\ &\geq -\frac{\beta}{2}\|y - x\|^2 + \frac{\beta}{2}\|y - x\|^2 = 0 \\ &\Rightarrow h(y) \geq h(x) + \nabla h(x) \cdot (y - x)\end{aligned}$$

Meaning, h is convex, as required.

(ii) We'll disprove it: We've seen that a $(\beta - \alpha)$-smooth function h satisfies for every $x, y \in R^d$:

$$-\frac{\beta}{2}\|y - x\|^2 \leq f(y) - f(x) - \nabla f(x) \cdot (y - x) \leq \frac{\beta}{2}\|y - x\|^2$$

Let $f(x) = \frac{1}{2}x^T Ax$ where $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Because $A \preceq 1$, we conclude that f is 1-smooth and convex. So we get that $h(x) = \frac{1}{2}x^T Ax - \frac{\alpha}{2}\|x\|^2$, and calculating the gradient we get:

$$\nabla h(x) = Ax - \alpha x$$

We'll set $\alpha = \frac{3}{4} < \beta, x = 0, y = e_2$, then we get:

$$\begin{aligned}
-\frac{1}{8} = -\frac{1}{8}\|y\|^2 = -\frac{(\beta-\alpha)}{2}\|y-x\|^2 &\overset{?}{\leq} h(y) - h(x) - \nabla h(x) \cdot (y-x) \\
&= \frac{1}{2}y^T Ay - \frac{\alpha}{2}\|y\|^2 - \frac{1}{2}x^T Ax + \frac{\alpha}{2}\|x\|^2 - Ax + \alpha x \\
&= 0 - \frac{3}{2} \cdot 1 - 0 + \frac{3}{2} \cdot 0 - 0 + 0 = -\frac{3}{2}
\end{aligned}$$

Meaning we got that $-\frac{1}{8} \leq -\frac{3}{2}$, in contradiction to the lower bound of the alternative characterization of smooth functions.

Question #4

(i)

$$f(0) = \frac{1}{n}\sum_{i=1}^{n} \max\{1 - y_i(0^\top x_i), 0\} + \lambda\|0\|^2 = \frac{1}{n}\sum_{i=1}^{n} \max\{1, 0\} + 0 = 1$$

$$f(w^\star) = \frac{1}{n}\sum_{i=1}^{n} \max\{1 - y_i(w^{\star\top} x_i), 0\} + \lambda\|w^\star\|^2 \geq \lambda\|w^\star\|^2$$

$$1 = f(0) \geq f(w^\star) \geq \lambda\|w^\star\|^2$$

$$\|w^\star\|^2 \leq \frac{1}{\lambda}$$

(ii) Since $\|w^\star\|^2 \leq \frac{1}{\lambda}$, we can take $S = \{w : \|w\|^2 \leq \frac{1}{\lambda}\}$.
Define $h_i(w) = \max\{1 - y_i(w^\top x_i), 0\}$. Let $w \in S, g_i \in \partial h_i(w)$.
From what we've seen in class about subgradients of finite maximum:
if $1 - y_i(w^\top x_i) > 0$, $\partial h_i(w) = \{-y_i x_i\}$, and $\|g_i\| = \|x_i\|$.
if $1 - y_i(w^\top x_i) < 0$, $\partial h_i(w) = \{0\}$, and $\|g_i\| = 0$.
if $1 - y_i(w^\top x_i) = 0$, $\partial h_i(w) = \{\lambda(-y_i x_i) : \lambda \in [0,1]\}$, and $\|g_i\| = \lambda\|x_i\| \leq \|x_i\|$.
So in any case $\|g_i\| \leq \|x_i\|$.

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} h_i(w) + \lambda\|w\|^2$$

$$\partial f(w) = \frac{1}{n}\sum_{i=1}^{n} \partial h_i(w) + \{2\lambda w\}$$

Let $g \in \partial f(w)$. there exist $g_1, \ldots, g_n$ such that $g = \frac{1}{n}\sum_{i=1}^{n} g_i + 2\lambda w$.

$$\|g\| = \frac{1}{n}\sum_{i=1}^{n} \|g_i\| + 2\lambda\|w\| \le \frac{1}{n}\sum_{i=1}^{n} \|x_i\| + 2\sqrt{\lambda}$$

So f is G-Lipschitz for $G = \frac{1}{n}\sum_{i=1}^{n} \|x_i\| + 2\sqrt{\lambda}$.

(iii) The algorithm would be PGD:

$$w_{t+1} = \Pi_S[w_t - \eta_t \nabla g_t]$$

initialization - $w_0 = 0$
Subgradient oracle - $g_t = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y_i(w_t^T x_i)<1\}} + 2\lambda w$
Adaptive step size - $\eta_t = \frac{\lambda}{2t}$
Since f is G-Lipschitz and 2λ-strongly convex, the convergence rate is

$$\frac{G^2 \log T}{2\lambda T} = O\left(\frac{\log T}{T}\right)$$