Notes for Lecture 3

Scribe: Tingkai Jia

1 The Convexity of $\|\mathbf{x}\|$

The function $\|\mathbf{x}\|$ of a norm of $\mathbf{x}$ is a convex function.

Proof. Consider any vector $\mathbf{x}, \mathbf{y}$, by the property of norm, we have

$$\|\theta\mathbf{x} + (1-\theta)\mathbf{y} \geq \|\theta\mathbf{x}\| + \|(1-\theta)\mathbf{y}\| = \theta\|\mathbf{x}\| + (1-\theta)\|\mathbf{y}\|.$$

Thus the function $\|\mathbf{x}\|$ is convex. □

2 First-order Condition of Convex Functions

Lemma 1. *Let $f(\mathbf{x})$ be a convex function. For any vector $\mathbf{h}$, we have*

$$\langle \nabla f(\mathbf{x}), \mathbf{h} \rangle = \lim_{t \to 0^+} \frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t}.$$

Proof. Let $q(t) = f(\mathbf{x} + t\mathbf{h})$. By applying the chain rule, we obtain

$$q'(t) = \langle \nabla f(\mathbf{x} + t\mathbf{h}), \mathbf{h} \rangle.$$

Consider when $t = 0$, we get

$$q'(0) = \lim_{t \to 0} \frac{q(t) - q(0)}{t} = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t},$$

also

$$q'(0) = \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t},$$

the proof is thus complete. □

Theorem 1 (first-order condition)**.** *Suppose f is differentiable and has convex domain, then f is convex if and only if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

holds for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom} f$.

Proof. Consider any $\mathbf{x}, \mathbf{y}$, we first obtain

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

$$\alpha[f(\mathbf{y}) - f(\mathbf{x})] \geq f(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})$$

By taking the limit as α approaches 0 in above last inequality, we obtain the following inequality

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \lim_{\alpha \to 0^+} \frac{f(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\alpha}$$

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

The last inequality uses Lemma 1, and we have thus proved the property. □

3 Second-order Condition of Convex Functions

Theorem 2 (second-order condition)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function. Suppose that $\nabla^2 f(\cdot)$ is continuous in an open neighborhood of $\mathbf{x}^* \in \mathbb{R}^d$.*

1. *If $\mathbf{x}^*$ is a local minimizer of $f(\cdot)$, then it holds that*

$$\nabla f(\mathbf{x}^*) = 0 \qquad \text{and} \qquad \nabla^2 f(\mathbf{x}^*) \succeq 0$$

2. *If it holds that*

$$\nabla f(\mathbf{x}^*) = 0 \qquad \text{and} \qquad \nabla^2 f(\mathbf{x}^*) \succ 0$$

then the point $\mathbf{x}^*$ is a strict local minimizer of $f(\cdot)$

Proof. **Part I:** Suppose $\nabla f(\mathbf{x}^*) \neq 0$. We define

$$\mathbf{p} = -\nabla f(\mathbf{x}^*),$$

which means $\langle \mathbf{p}, \nabla f(\mathbf{x}^*) \rangle < 0$. The continuity of $\nabla f(\cdot)$ means there exists some $T > 0$ such that

$$\langle \mathbf{p}, \nabla f(\mathbf{x}^* + t\mathbf{p}) \rangle < 0$$

for any $t \in (0, T]$. For any t, Taylor's theorem means there exist some $t \in (0, \hat{t})$ such that

$$f(\mathbf{x}^* + t\mathbf{p}) = f(\mathbf{x}^*) + t\langle \mathbf{p}, \nabla f(\mathbf{x}^* + t\mathbf{p}) \rangle < f(\mathbf{x}^*),$$

which leads to contradiction. Hence $\nabla f(\mathbf{x}^*) = 0$.

Suppose $\nabla f(\mathbf{x}^*)$ is not positive semi-definite. Then we can find some $\mathbf{p} \in \mathbb{R}^d$ such that $\langle \nabla^2 f(\mathbf{x}^*)\mathbf{p}, \mathbf{p} \rangle < 0$. The continuity of Hessian means there exists some $T > 0$ such that for any $t \in [0, T]$ holds that

$$\langle \nabla^2 f(\mathbf{x}^* + t\mathbf{p})\mathbf{p}, \mathbf{p} \rangle < 0.$$

Doing Taylor expansion around $\mathbf{x}^*$, we have for all $t \in (0, T]$, there exist some $t \in (0, \hat{t})$ such that

$$f(\mathbf{x}^* + \hat{t}\mathbf{p}) = f(\mathbf{x}^*) + t\langle \mathbf{p}, \nabla f(\mathbf{x}^*) \rangle + \frac{1}{2} t^2 \langle \nabla^2 f(\mathbf{x}^* + t\mathbf{p})\mathbf{p}, \mathbf{p} \rangle < f(\mathbf{x}^*),$$

which leads to contradiction. Hence $\nabla f(\mathbf{x}^*)$ should be positive semi-definite.

Part II: The continuity of Hessian means the positive definiteness of Hessian still hold in $\mathcal{B}(\mathbf{x}^*, \delta)$ for some $\delta > 0$. For any $\mathbf{p} \in \mathbb{R}^d$ with $\|\mathbf{p}\|_2 \leq \delta$, then we have

$$f(\mathbf{x}^* + \mathbf{p}) = f(\mathbf{x}^*) + \langle \mathbf{p}, \nabla f(\mathbf{x}^*) \rangle + \frac{1}{2} p^T \nabla^2 (\mathbf{x}^* + t\mathbf{p})\mathbf{p} > f(\mathbf{x}^*)$$

for some $t \in (0, 1)$, which leads to contradiction. The point $\mathbf{x}^*$ is a strict local minimizer. □

4 Epigraph Theorem of Convex Functions

Theorem 3. *A function $f(\mathbf{x})$ is convex if and only if its epigraph is a convex set.*

Proof. **Part I:** Suppose $f : \mathcal{C} \to \mathbb{R}$ is convex. Let $(\mathbf{x}_1, u_1)$ and $(\mathbf{x}_2, u_2)$ in

$$\text{epi} f \triangleq \{(\mathbf{x}, u) \in \mathcal{C} \times \mathbb{R} : f(\mathbf{x}) \leq u\}.$$

For any $\alpha \in [0, 1]$, the point

$$\alpha(\mathbf{x}_1, u_1) + (1 - \alpha)(\mathbf{x}_2, u_2) = (\alpha\mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2, \alpha u_1 + (1 - \alpha)u_2)$$

satisfies

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \le \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2) \le \alpha u_1 + (1-\alpha) u_2,$$

where the last inequality use the convexity of f and the second one is due to $(\mathbf{x}_1, u_1)$ and $(\mathbf{x}_2, u_2)$ in epi$f$. Hence, the point $\alpha(\mathbf{x}_1, u_1) + (1-\alpha)(\mathbf{x}_2, u_2)$ also in epi$f$, which means the epigraph is convex.
Part II: Suppose the epigraph

$$\text{epi} f \triangleq \{(\mathbf{x}, u) \in \mathcal{C} \times \mathbb{R} : f(\mathbf{x}) \le u\}.$$

is convex. It is easy to see $\mathcal{C}$ is convex by fixing some u. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$, $u_1 = f(\mathbf{x}_1)$ and $u_2 = f(\mathbf{x}_2)$. The convexity of epigraph means

$$\alpha(\mathbf{x}_1, u_1) + (1-\alpha)(\mathbf{x}_2, u_2) = (\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2, \alpha u_1 + (1-\alpha) u_2) \in \text{epi} f,$$

which leads to

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \le \alpha u_1 + (1-\alpha) u_2 = \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2).$$

□

5 A Property of Convex Functions

Lemma 2 (Supporting Hyperplane Theorem)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set and $\mathbf{x}_0$ belongs to its boundary. Then, there exists a nonzero vector $\mathbf{w} \in \mathbb{R}^d$ such that*

$$\langle \mathbf{w}, \mathbf{x} \rangle \le \langle \mathbf{w}, \mathbf{x}_0 \rangle.$$

Theorem 4. *The convex function has the following properties*

1. *If any $\mathbf{x} \in \text{dom} f$ satisfies $\partial f(\mathbf{x}) \neq \emptyset$, then f is convex.*
2. *If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and $\mathbf{x}$ belongs to the interior of* dom$f$*, then $\partial f(\mathbf{x}) \neq \emptyset$.*

Proof. **Part I:** We first check the sum rule (one side) of subgradient. For any $\mathbf{x}_1, \mathbf{x}_2 \in \text{dom} f \subseteq \mathbb{R}^d$ and $\alpha \in [0, 1]$, there exist $\mathbf{g}_1, \mathbf{g}_2 \in \mathbb{R}^d$ such that

$$f(z) \ge f(\mathbf{x}_1) + \langle \mathbf{g}_1, \mathbf{z} - \mathbf{x}_1 \rangle \quad \text{and} \quad f(\mathbf{z}) \le f(\mathbf{x}_2) + \langle \mathbf{g}_2, \mathbf{z} - \mathbf{x}_2 \rangle$$

for any $\mathbf{z} \in \mathbb{R}^d$, which leads to

$$f(\mathbf{z}) \ge \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2) + \langle \alpha \mathbf{g}_1 + (1-\alpha)\mathbf{g}_2, \mathbf{z} - (\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \rangle$$

$$\ge f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) + \langle \alpha \mathbf{g}_1 + (1-\alpha)\mathbf{g}_2, \mathbf{z} - (\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \rangle.$$

Let $\mathbf{g} = \alpha \mathbf{g}_1 + (1-\alpha)\mathbf{g}_2$, above inequality means $\mathbf{g} \in \partial f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2)$. The definition of subdifferential means for any $\mathbf{x}_1, \mathbf{x}_2$, we have

$$f(\mathbf{x}_1) \ge f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) + \langle \mathbf{g}, \mathbf{x}_1 - (\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \rangle$$

$$= f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) + \langle \mathbf{g}, (1-\alpha)(\mathbf{x}_1 - \mathbf{x}_2) \rangle$$

and

$$f(\mathbf{x}_2) \le f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) + \langle \mathbf{g}, \mathbf{x}_2 - (\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \rangle$$

$$= f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) + \langle \mathbf{g}, \alpha(\mathbf{x}_2 - \mathbf{x}_1) \rangle.$$

Weighted summing over the above inequality obtains

$$\alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2) \ge f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2).$$

Part II: Consider that $(\mathbf{x}, f(\mathbf{x}))$ is on the boundary of epi$f$. The hyperplane supporting theorem (Lemma 2) say there exists $(\mathbf{a}, b)$ with $\mathbf{a} \neq 0$ such that

$$\left\langle \begin{bmatrix} \mathbf{a} \\ b \end{bmatrix}, \begin{bmatrix} \mathbf{y} - \mathbf{x} \\ t - f(\mathbf{x}) \end{bmatrix} \right\rangle \leq 0$$

for any $(\mathbf{y}, t) \in$ epi$f$. That is

$$\langle \mathbf{a}, \mathbf{y} - \mathbf{x} \rangle + b(t - f(\mathbf{x})) \leq 0.$$

We can conclude $b \leq 0$. Otherwise, let $t \to +\infty$ (t can be arbitrary large for fixed $\mathbf{x}, \mathbf{y}$ and $\mathbf{a}$) leads to LHS tends to $+\infty$. Since $\mathbf{x}$ is in the interior, we can find some $\epsilon > 0$ such that $\mathbf{y} = \mathbf{x} + \epsilon \mathbf{a} \in \text{dom} f$, which leads to LHS be $-\epsilon \|\mathbf{a}\|_2^2$. It means $b \neq 0$. Hence we can say $b < 0$ and dividing by b obtains

$$\langle \frac{\mathbf{a}}{b}, \mathbf{y} - \mathbf{x} \rangle + (t - f(\mathbf{x})) \geq 0 \iff t \geq f(\mathbf{x}) + \langle -\frac{\mathbf{a}}{b}, \mathbf{y} - \mathbf{x} \rangle.$$

Taking $t = f(\mathbf{y})$ means $\mathbf{g} = -\frac{\mathbf{a}}{b}$ is a subgradient at $\mathbf{x}$. □

6 Some Examples of Convex Functions

Here are three different functions and the process of determining their convexity using the Hessian matrix.

Example 1. $\mathbf{x} \in \mathbb{R}^n, f : \mathbb{R}^n \to \mathbb{R}, f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$

$$\nabla^2 f(\mathbf{x}) = 2\mathbf{A}^\top \mathbf{A}$$
$$\mathbf{x}^\top \nabla^2 f(\mathbf{x}) \mathbf{x} = 2(\mathbf{A}\mathbf{x})^\top (\mathbf{A}\mathbf{x}) \succeq 0.$$

The Hessian matrix of $f(\mathbf{x})$ is a positive semi-definite matrix, which means it is a convex function.

Example 2. $x \in \mathbb{R}, y \in \mathbb{R}, f(x, y) = \frac{x^2}{y}$

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial y \partial x} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} \frac{2}{y} & -\frac{2x}{y^2} \\ -\frac{2x}{y^2} & \frac{2x^2}{y^3} \end{bmatrix} = \frac{2}{y^3} \begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix}$$
$$\begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix} = \begin{bmatrix} y \\ -x \end{bmatrix} (y, -x) \succeq 0.$$

The Hessian matrix of $f(\mathbf{x})$ is a positive semi-definite matrix, which means it is a convex function.

Example 3. $\mathbf{x} \in \mathbb{R}^m, f(\mathbf{x}) = \ln \left(\sum_{i=1}^n \exp(\mathbf{a}_i^\top \mathbf{x} + b_i) \right)$.

$$\nabla f(\mathbf{x}) = \mathbf{A}^\top \frac{1}{\sum_{i=1}^n \exp(\mathbf{a}_i^\top \mathbf{x} + b_i)} \begin{bmatrix} \exp(\mathbf{a}_1^\top \mathbf{x} + b_1) \\ \exp(\mathbf{a}_2^\top \mathbf{x} + b_2) \\ \vdots \\ \exp(\mathbf{a}_n^\top \mathbf{x} + b_n) \end{bmatrix} = \frac{1}{\mathbf{1}^\top \mathbf{z}} \mathbf{A}^\top \mathbf{z},$$

where $z_i = \exp(\mathbf{a}_i^\top \mathbf{x} + b_i)$.

$$\nabla^2 f(\mathbf{x}) = \frac{1}{\mathbf{1}^\top \mathbf{z}} \text{diag}(\mathbf{z}) - \frac{1}{(\mathbf{1}^\top \mathbf{z})^2} \mathbf{z} \cdot \mathbf{z}^\top$$
$$\mathbf{x}^\top \nabla^2 f(\mathbf{x}) \mathbf{x} = \frac{1}{\mathbf{1}^\top \mathbf{z}} \sum_{i=1}^n z_i x_i^2 - \frac{1}{(\mathbf{1}^\top \mathbf{z})^2} \sum_{i=1}^n (z_i x_i)^2$$

$$= \frac{1}{(\mathbf{1}^\top \mathbf{z})^2} \left(\sum_{i=1}^{n} z_i x_i^2 \sum_{i=1}^{n} z_i - \sum_{i=1}^{n} (z_i x_i)^2 \right) \geq 0,$$

the last inequality applies the Cauchy-Schwarz inequality, where

$$\mathbf{u} = \begin{bmatrix} z_1 x_1^2 \\ z_2 x_2^2 \\ \vdots \\ z_n x_n^2 \end{bmatrix}, \mathbf{v} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix},$$

$$\|\mathbf{u}\|_1 \|\mathbf{v}\|_1 - |\mathbf{u}^\top \mathbf{v}| = \sum_{i=1}^{n} z_i x_i^2 \sum_{i=1}^{n} z_i - \sum_{i=1}^{n} (z_i x_i)^2 \geq 0.$$

The Hessian matrix of $f(\mathbf{x})$ is a positive semi-definite matrix, which means it is a convex function.

7 Subgradient of Differentiable Functions

Theorem 5. *If a function is differentiable, the only subgradient at each point is the gradient,* i.e.*,*

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}.$$

Proof. For any $\mathbf{x}, \mathbf{y}$, let $\mathbf{g}$ be a subgradient of $\mathbf{x}$, we first obtain

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle.$$

Let $\mathbf{y} - \mathbf{x} = t\mathbf{h}$, we then get

$$f(\mathbf{x} + t\mathbf{h}) \geq f(\mathbf{x}) + \langle \mathbf{g}, t\mathbf{h} \rangle$$
$$\frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t} \geq \langle \mathbf{g}, \mathbf{h} \rangle,$$

By taking the limit as t approaches 0 in above last inequality, we obtain the following inequality

$$\lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t} = \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle \geq \langle \mathbf{g}, \mathbf{h} \rangle$$
$$\langle \nabla f(\mathbf{x}) - \mathbf{g}, \mathbf{h} \rangle \geq 0,$$

the first equation comes from Lemma 1. For the last inequality, we find that if and only if $\nabla f(\mathbf{x}) = \mathbf{g}$, the inequality can holds for any vector $\mathbf{h}$, thus we finish the proof. □

8 Basic Rules of Subgradient

In this part, we first show basic rules of subgradient operation, and then illustrate some functions to analyze its subdifferential, during which we may use these rules.

- Scaling: $\partial(\alpha f) = \alpha \partial f$ (for $\alpha > 0$)
- Summation: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$.
- Affine Transformation: if $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$, then

$$\partial h(\mathbf{x}) = \mathbf{A}^\top \partial f(\mathbf{A}\mathbf{x} + \mathbf{b}).$$

- Chain Rule: suppose f is convex, and g is differentiable, nondecreasing, and convex. Let $h = g \circ f$, then
$$\partial h(\mathbf{x}) = g'(f(\mathbf{x}))\partial f(\mathbf{x}).$$
- Composition: suppose $f(\mathbf{x}) = h(f_1(\mathbf{x}), \cdots, f_n(\mathbf{x}))$, where f_i's are convex, and h is differentiable, nondecreasing, and convex. Let $\mathbf{q} = \nabla h(\mathbf{y})|_{y=[f_1(\mathbf{x}),\cdots,f_n(\mathbf{x})]}$, and $\mathbf{g}_i \in \partial f_i(\mathbf{x})$. Then
$$q_1\mathbf{g}_1 + \cdots + q_n\mathbf{g}_n \in \partial f(\mathbf{x}).$$
- Pointwise Maximum: if $f(\mathbf{x}) = \max_{1\le i\le k} f_i(\mathbf{x})$, then
$$\partial f(\mathbf{x}) = \text{conv}\left\{\cup\left\{\partial f_i(\mathbf{x}) | f_i(\mathbf{x}) = f(\mathbf{x})\right\}\right\}.$$
- Pointwise Supremum: if $f(\mathbf{x}) = \sup_{\alpha\in\mathcal{F}} f_\alpha(\mathbf{x})$, then
$$\partial f(\mathbf{x}) = \text{closure}\left(\text{conv}\left\{\cup\left\{\partial f_\alpha(\mathbf{x}) | f_\alpha(\mathbf{x}) = f(\mathbf{x})\right\}\right\}\right).$$

Example 4 (the l_1 norm)**.** $\mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$.

Let $f_i(\mathbf{x}) = |\mathbf{x}^T \cdot \mathbf{e}_i| = |x_i|$, then we obtain
$$f(\mathbf{x}) = \sum_{i=1}^n |x_i| = \sum_{i=1}^n f_i(\mathbf{x}).$$
Consider the summation rule of subgradient, we can see $\partial f(\mathbf{x}) = \sum_{i=1}^n \partial f_i(\mathbf{x})$, then
$$\partial f_i(\mathbf{x}) = \begin{cases} \text{sgn}(x_i)\mathbf{e}_i, & \text{if } x_i \neq 0 \\ [-1,1]\cdot\mathbf{e}_i, & \text{if } x_i = 0 \end{cases}$$
now we have
$$\partial f(\mathbf{x}) = \left\{ \mathbf{g} \middle| \begin{matrix} g_i = \text{sgn}(x_i), & \text{if } x_i \neq 0 \\ g_i \in [-1,1], & \text{if } x_i = 0 \end{matrix} \right\}.$$

Example 5. $\mathbf{x} \in \mathbb{R}^n, h(\mathbf{x}) = \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_1 = \sum_{i=1}^n |\mathbf{a}_i^\top \mathbf{x} + b_i|$.

Let $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$, it follows that $f(\mathbf{x}) = \|\mathbf{x}\|_1$. Consider the affine transformation rule of subgradient, we can see $\partial h(\mathbf{x}) = \partial f(\mathbf{A}\mathbf{x} + \mathbf{b})$, then we have
$$\mathbf{g} = \sum_{i:\mathbf{a}_i^\top\mathbf{x}+b_i\neq 0} \text{sgn}(\mathbf{a}_i^\top\mathbf{x} + b_i)\mathbf{e}_i \in \partial f(\mathbf{A}\mathbf{x} + \mathbf{b}),$$
which means
$$\mathbf{A}^T\mathbf{g} = \sum_{i:\mathbf{a}_i^\top\mathbf{x}+b_i\neq 0} \text{sgn}(\mathbf{a}_i^\top\mathbf{x} + b_i)\mathbf{a}_i \in \partial h(\mathbf{x}).$$
It should be noted that the final set we get is actually part of the subdifferential, because we did not consider cases when $\mathbf{a}_i^\top\mathbf{x} + b_i = 0$ instead of setting a default 0, which still belongs to the subdifferential of $h_i(\mathbf{x})$. In the actual running of gradient descent algorithm, we have no need to obtain the complete subdifferential of function, a practical subgradient would be enough.

Example 6 (piece-wise linear functions)**.** $\mathbf{x} \in \mathbb{R}^n$, $f(\mathbf{x}) = \max_{1\le i\le n}\{\mathbf{a}_i^\top\mathbf{x} + b_i\}$.

Let $f_i(\mathbf{x}) = \mathbf{a}_i^\top \mathbf{x} + b_i$, then we obtain

$$f(\mathbf{x}) = \max_{1 \le i \le n} \{\mathbf{a}_i^\top \mathbf{x} + b_i\} = \max_{1 \le i \le n} f_i(\mathbf{x}).$$

Consider that it satisfies the pointwise maximum situation, then pick any $\mathbf{a}_j$ s.t. $\mathbf{a}_j^\top \mathbf{x} + b_j = \max_i \{\mathbf{a}_i^\top \mathbf{x} + b_i\}$, we have

$$\mathbf{a}_j \in \partial f(\mathbf{x}).$$

Example 7 (the l_∞ norm)**.** $\mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) = \|\mathbf{x}\|_\infty = \max_{1 \le i \le n} |x_i|$.

Let $f_i(\mathbf{x}) = |\mathbf{x}^\top \mathbf{e}_i| = |x_i|$, then we obtain

$$f(\mathbf{x}) = \max_{1 \le i \le n} |x_i| = \max_{1 \le i \le n} f_i(\mathbf{x}).$$

Consider that it satisfies the pointwise maximum situation, then if $\mathbf{x} \neq 0$, pick any x_j obeying $|x_j| = \max_i |x_i|$, we obtain

$$\operatorname{sgn}(x_j)\mathbf{e}_j \in \partial f(\mathbf{x}).$$