ECE 273 Convex Optimization and Applications Instructor: Jun-Kun Wang
Scribe: Eric Bressinger, Reya Sadhu, Subhadeep Chatterjee April 4, 2024
Editor/TA: Marialena Sfyraki

Lecture 2: Gradient Flow and Convex Analysis I

1 Gradient Descent and Gradient Flow

A formal specification of the Gradient Descent (GD) algorithm follows.

Algorithm 1 Gradient Descent

```
Input: an initial point $\mathbf{x}_0 \in \mathbf{dom}\ f$ and step size $\eta$.
for $k = 1$ to $K$ do
    $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \eta \nabla f(\mathbf{x_k})$
end for
Return $\mathbf{x}_{k+1}$.
```

Remark: The parameter η is called the *step size* or *learning rate.*

In order to better understand gradient descent, let's consider the curve that at each instant proceeds in the direction of steepest descent of f. For this method, let's consider a function $f : X \to \mathbb{R}$, the method of gradient flow starts at some initial point $x_0 \in X$ and seek to find the optimum of f by following the integral curve defined by the following differential equations.

Definition 1. ***(Gradient Flow)****: Let $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. Gradient flow is a smooth curve $\mathbf{x} : \mathbb{R} \to \mathbb{R}^d$ such that*

$$\frac{d\mathbf{x}(t)}{dt} = -\nabla f\big(\mathbf{x}(t)\big)$$

1.1 Insights into the Algorithm

Gradient Flow is Gradient Descent as $\eta \to 0$. Consider the update step

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k),$$

then

$$\lim_{\eta \to 0} \frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\eta} = \lim_{\eta \to 0} -\nabla f(\mathbf{x}_k)$$

$$\Leftrightarrow \frac{d\mathbf{x}}{dt} = -\nabla f(\mathbf{x}).$$

Consider applying Gradient Flow to $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, that is

$$\frac{d\mathbf{x}(t)}{dt} = -\nabla f\left(\mathbf{x}(t)\right).$$

Let

$$\mathbf{x} \triangleq \begin{bmatrix} x[1] \\ x[2] \\ \vdots \\ x[i] \\ \vdots \\ x[d] \end{bmatrix} \in \mathbb{R}^d.$$

Then,

$$\begin{aligned}
\frac{df}{dt} &= \sum_i^d \frac{\partial f}{\partial x[i]} \frac{\partial x[i]}{\partial t} && \text{, by Chain rule} \\
&= \left\langle \nabla f(\mathbf{x}), \frac{d\mathbf{x}(t)}{dt} \right\rangle \\
&= \langle \nabla f(\mathbf{x}), -\nabla f(\mathbf{x}) \rangle && \text{, by Gradient Flow} \\
&= -||\nabla f(\mathbf{x})||_2^2 \\
&\leq 0.
\end{aligned}$$

Remarks:

- Thus, as long as $\nabla f(\mathbf{x}) \neq \mathbf{0}$, the function is always decreasing. This means gradient flow is always making progress as long as it is not stationary. This does not necessarily imply that it finds the optimal point.
- We are using the differential equation of gradient flow as as a continuous analog for the gradient descent update process. The gradient descent algorithm is used to simulate the dynamics of gradient flow.

1.2 Gradient Dominant Condition

Definition 2. *(**Gradient Dominant** or **Polyak-Lojasiewicz (PL) Condition**): We say a function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the "Gradient Dominance" condition if* $\forall \mathbf{x} \in \mathbb{R}^d$

$$||\nabla f(\mathbf{x})||_2^2 \geq 2\mu \left(f(\mathbf{x}) - \min_{\mathbf{x}} f(\mathbf{x}) \right), \quad \textit{for some } \mu > 0.$$

We say that f is μ-gradient dominant.

Definition 3. ***(Stationary Point)**: Given a differentiable function f such that $f : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$, a stationary point is a point such that*

$$\nabla f(\mathbf{x}) = \mathbf{0} \in \mathbb{R}^d.$$

Remark: For any function satisfying the P.L. condition, every stationary point is a global optimum point.

Proof. By definition,

$$f(\mathbf{x}) - \min_{\mathbf{x}} f(\mathbf{x}) \geq 0.$$

For a stationary point $\mathbf{x} \in \mathbb{R}^d$,

$$\nabla f(\mathbf{x}) = \mathbf{0} \Rightarrow ||\nabla f(\mathbf{x})||_2^2 = \mathbf{0}.$$

Thus, for any function f satisfying the P.L condition with $\mathbf{x}$ as a stationary point,

$$\mathbf{0} \geq 2\mu \left(f(\mathbf{x}) - \min_{\mathbf{x}} f(\mathbf{x}) \right) \geq \mathbf{0}, \mu > 0,$$

which is only true when equality holds, by squeeze theorem. Thus,

$$f(\mathbf{x}) - \min_{\mathbf{x}} f(\mathbf{x}) = \mathbf{0} \Leftrightarrow \mathbf{x} \text{ is a global optimum point.}$$

□

Example 1: All strongly convex functions

Example 2: $f(x) = x^2 + 2\sin^2(x)$

Remark: For simplicity, we can define $f_* := \min_{\mathbf{x}} f(\mathbf{x})$. Thus, we can rewrite the optimality gap as $f(\mathbf{x}_t) - f_*$.

Consequence: Suppose that f is additionally μ-gradient dominant. Then, taking the derivative of an optimality gap we get

$$\begin{aligned} \frac{d(f(\mathbf{x}_t) - f_*)}{dt} &= \frac{df(\mathbf{x}_t)}{dt} && \text{, as } f_* \text{ is a constant} \\ &= -||\nabla f(\mathbf{x}_t)||_2^2 && \text{, by Gradient Flow} \\ &\leq -2\mu \left(f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x}) \right) && \text{, since } f \text{ is } \mu\text{-gradient dominant} \end{aligned} \qquad (1)$$

which implies that

$$f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x}) \leq e^{-2\mu t} \left(f(\mathbf{x}_0) - \min_{\mathbf{x}} f(\mathbf{x}) \right) \tag{2}$$

for μ-gradient dominant functions, where $\mathbf{x}_0$ is the initial point.

Remark: As $f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x})$ is the gap at t and $f(\mathbf{x}_0) - \min_{\mathbf{x}} f(\mathbf{x})$ is the initial gap, the above inequality implies that the optimality gap decays exponentially with time.

Why does (6) imply (7)? Let

$$\delta_t := f(\mathbf{x}_t) - f_*.$$

Then, inequality (6) can be expressed as

$$\begin{aligned}
&\frac{d\delta_t}{dt} \leq -2\mu\delta_t \\
&\Leftrightarrow \frac{d\delta_t}{\delta_t} \leq -2\mu dt \\
&\Rightarrow \int_{\delta_0}^{\delta_t} \frac{d\delta_t}{\delta_t} \leq \int_0^t -2\mu dt \\
&\Leftrightarrow \ln(\delta_t) - \ln(\delta_0) \leq -2\mu t \qquad\qquad , \text{ since } \frac{d}{dx}\ln x = \frac{1}{x}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{\delta_t}{\delta_0} \leq exp(-2\mu t) \\
&\Leftrightarrow \delta_t \leq \delta_0 exp(-2\mu t)
\end{aligned}$$

Plugging back in, we get

$$f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x}) \leq \exp(-2\mu t) \left(f(\mathbf{x}_0) - \min_{\mathbf{x}} f(\mathbf{x}) \right)$$

2 Convex Sets

Definition 4. *(**Convex Combination**): A linear combination is called convex if the coefficients of the variables are non-negative and sum to 1. In other words, given any finite number of points $x_i, i \in [1...n]$ in a real vector space, a convex combination of these points has the form:*

$$\alpha_1 x_1 + \alpha_2 x_2 ... + \alpha_n x_n,$$

where the coefficients α_i satisfy $\alpha_i \geq 0, \forall i$ and $\sum_{i=1}^n \alpha_i = 1$.

Definition 5. *(**Convex Set**): A set $C \subseteq \mathbb{R}^d$ is called convex if for any $\mathbf{x}, \mathbf{y} \in C$ and any $\alpha \in [0,1]$, we have*

$$\alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in C.$$

Here the above expression is defined as the Convex combination of x and y.

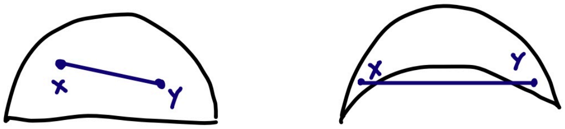

Figure 1: Difference between convex and non convex set

Remark: Let $x_\alpha = \alpha\mathbf{x} + (1-\alpha)\mathbf{y}$. Observe that x_α is a parametrization of the points of the line segment between $\mathbf{x}$ and $\mathbf{y}$. We can see from the above figure that the point x_α always lies on the line formed between $\mathbf{x}$ and $\mathbf{y}$. Hence, if x_α lies inside the set as in the left figure, the set is convex. Otherwise, if any point on the line formed between $\mathbf{x}$ and $\mathbf{y}$ lies outside the set as in the right figure, then $\exists\alpha \in [0,1]$ such that $x_\alpha \notin C$ and hence the set is non-convex.

3 Convex Functions

Definition 6. *(**Zero Order Characterization of Convex Functions**): A function $f : C \to \mathbb{R}$ defined over a convex set C is called convex if, for any $\mathbf{x}, \mathbf{y} \in C$ and any $\alpha \in [0,1]$, the following inequality holds*

$$f\left(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\right) \le \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}).$$

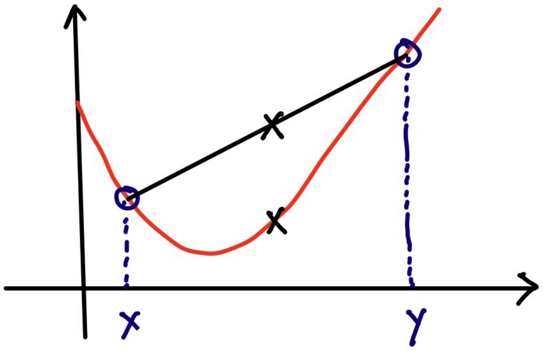

Figure 2: Zero Order Characterization

Remark: Let $x_\alpha = \alpha\mathbf{x} + (1-\alpha)\mathbf{y}$ be the parametrization of the points of the line segment between $\mathbf{x}$ and $\mathbf{y}$. Similarly, $\alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$ is the parametrization of the points on line segment between $f(\mathbf{x})$ and $f(\mathbf{y})$. The zero-order characterization implies that a function is convex if the function value at any convex combination of $\mathbf{x}$ and $\mathbf{y}$ is always less equal than the convex combination of the function values at $\mathbf{x}$ and $\mathbf{y}$, as can be seen in Figure 2. In other words, the line segment connecting $f(\mathbf{x})$ and $f(\mathbf{y})$ lies always above the function curve defined between $\mathbf{x}$ and $\mathbf{y}$.

Definition 7. *(**First Order Characterization of Convex Functions**): A differentiable function $f : C \to \mathbb{R}$ defined over a convex set C is called convex **if and only if**, for any $\mathbf{x}, \mathbf{y} \in C$*

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

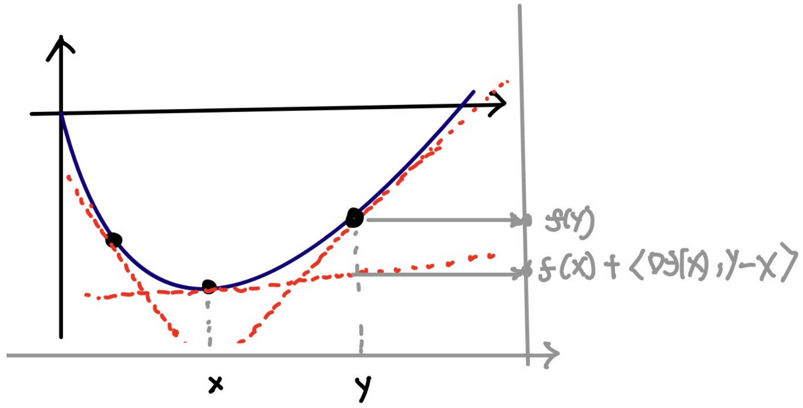

Figure 3: First Order Characterization

Remark: Observe that if we take a tangent at the point $\mathbf{x}$, we get the line equation $f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), (\mathbf{y}-\mathbf{x}) \rangle$, for some $\mathbf{y} \in C$. The first-order characterization implies that a function is convex, if the function always dominates its first order (linear) Taylor approximation.

Definition 8. *(**Second Order Characterization of Convex Functions**): A twice-differentiable function $f : C \to \mathbb{R}$ defined over a convex set C is convex **if and only if**, for any $\mathbf{x} \in C$, the Hessian matrix evaluated at $\mathbf{x}$ is positive semi-definite, i.e.*

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

3.1 Examples of Convex Functions:

Example 1: Linear Functions

$$f\left(\alpha\mathbf{x}+(1-\alpha)\mathbf{y}\right)=\alpha f(\mathbf{x})+(1-\alpha)f(\mathbf{y})$$

A linear function is also a concave function, i.e.

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

Example 2: Quadratic Functions

$$f(\mathbf{x})=\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}-\mathbf{b}^T\mathbf{x},\ \lambda_{min}(\mathbf{A})\geq 0$$

Example 3: Negative Entropy

$$F(\mathbf{x})=\sum_{i=1}^{d}x_i\log x_i,$$

where $x\in\mathbb{R}^d_{>0}$ (i.e., each element x_i of the vector $x\in\mathbb{R}^d$ satisfies $x_i>0$ for all $i\in[d]$).

Example 4: Non-negative weighted sum of convex functions

$$f(\mathbf{x})=\sum_{i=1}^{n}\alpha_i f_i(\mathbf{x}),\ \alpha_i\geq 0,\forall i$$

Example 5: Sum of squared loss

$$F(\mathbf{x})=\sum_{i=1}^{n}\frac{1}{2}\left(y_i-\mathbf{x}^\top\mathbf{z}_i\right)^2$$

4 Strongly Convex Functions

Definition 9. ***(Zero Order Characterization of μ-Strongly Convex Functions)****: A function $f:C\to\mathbb{R}$ defined over a convex set C is μ-strongly convex w.r.t. a norm $||\cdot||$ if, for any $\mathbf{x},\mathbf{y}\in C$ and any $\alpha\in[0,1]$ we have*

$$f\left((1-\alpha)\mathbf{x}+\alpha\mathbf{y}\right)\leq(1-\alpha)f(\mathbf{x})+\alpha f(\mathbf{y})-\frac{\mu}{2}\alpha(1-\alpha)||\mathbf{y}-\mathbf{x}||^2,$$

for some $\mu>0$.

Remark 1: We know that $\mu > 0, \alpha \in [0,1], ||\mathbf{y}-\mathbf{x}||^2 \geq 0, \forall \mathbf{x}, \mathbf{y} \in C$. Hence, we have

$$(1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \frac{\mu}{2}\alpha(1-\alpha)||\mathbf{y}-\mathbf{x}||^2 \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$

Thus, when f is strongly convex we get

$$f\left((1-\alpha)\mathbf{x} + \alpha\mathbf{y}\right) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$

Therefore, strongly convexity implies convexity.

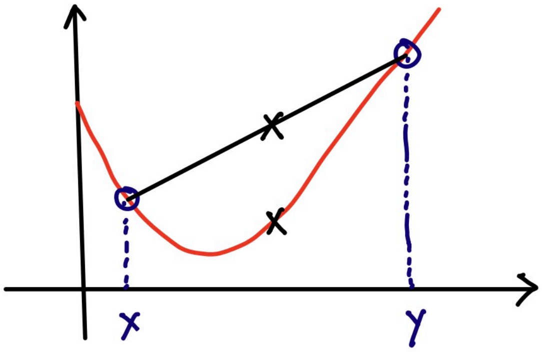

Figure 4: Zero Order Characterization

Remark 2: From the graph we can see that strong convexity just suggests that there will always be a difference of at least $\frac{\mu}{2}\alpha(1-\alpha)||\mathbf{y}-\mathbf{x}||^2$ thus imposing a restriction and hence ensuring strong convexity. This also implies regular convexity. When $\alpha = 0$ or $\alpha = 1$ we get the intersection points and thus the μ term will not impact the value. However, for any value between them, there will always be at least a difference of $\frac{\mu}{2}\alpha(1-\alpha)||\mathbf{y}-\mathbf{x}||^2$. This is why when we apply the same for a linear function, it fails because a linear function adheres to the inequality of normal convexity. The minimum gap is not followed by the linear function and hence a linear function is not strongly convex.

Definition 10. ***(First Order Characterization of μ-Strongly Convex Functions)****: A differentiable function $f : C \to \mathbb{R}$ defined over a convex set C is μ-strongly convex w.r.t. a norm $||\cdot||$* ***if and only if*** *for any $\mathbf{x}, \mathbf{y} \in C$ we have*

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

for some $\mu > 0$.

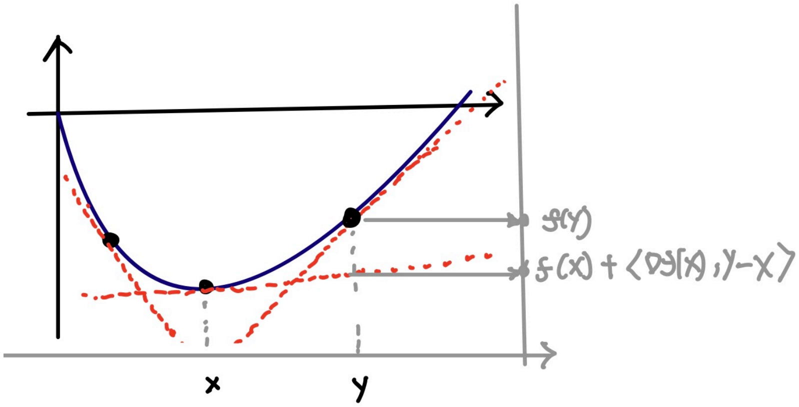

Figure 5: First Order Characterization

Remark: *We have the line $f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$, hence at $\mathbf{y}$ we for sure know that $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$ but strong convexity is defined when there is an extra amount of gap that is needed. That extra gap is at y and is equal to $\frac{\mu}{2}||\mathbf{y} - \mathbf{x}||^2$. This implies an extra parameter that ensures that it always is convex. It ensures a strong convexity for any $\mu > 0$.*

Definition 11. ***(Second Order Characterization of μ-Strongly Convex Functions)****: A twice differentiable function $f : C \to \mathbb{R}$ defined over a convex set C is μ-strongly convex w.r.t. a norm $||\cdot||$* ***if and only if*** *for any $\mathbf{x} \in C$ we have*

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

for some $\mu > 0$ and any $\mathbf{y} \in \mathbb{R}^d$.

Remarks:

1. Here $||\cdot||$ can be any norm and not restricted to l_2 norm. There are benefits of using non-Euclidean norm that will be more apparent in the later lectures.

2. Using the l_2 norm, the second-order characterization implies that $\lambda_{min}(\nabla^2 f(\mathbf{x})) \geq \mu > 0$.

3. Any strongly convex function is also always convex but the inverse is generally not true.

4. We generally use the second order characterization due to its relative ease in use as well as computation.

Theorem 1. *The μ-strong convexity implies the μ-Gradient Dominant condition, i.e., $\forall \mathbf{x} \in \mathbb{R}^d$,*

$$||\nabla f(\mathbf{x})||_2^2 \geq 2\mu \left(f(\mathbf{x}) - \min_{\mathbf{x}} f(\mathbf{x}) \right) , \quad \textit{for some} \ \ \mu > 0.$$

4.1 Equivalency of the strong convexity characterizations

For any $\mathbf{x} \in C \subseteq \mathbb{R}^d$, $\mathbf{y} \in C \subseteq \mathbb{R}^d$, $\mathbf{z} \in \mathbb{R}^d$, and any $\alpha \in [0,1]$, the three expressions below are equivalent (assuming that f is twice continuously differentiable):

$$f\big((1-\alpha)\mathbf{x} + \alpha\mathbf{y}\big) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \frac{\mu}{2}\alpha(1-\alpha)||\mathbf{y}-\mathbf{x}||^2 \quad (zero-order) \tag{3}$$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{\mu}{2}||\mathbf{y}-\mathbf{x}||^2 \qquad (first-order) \tag{4}$$

$$\mathbf{z}^\top \nabla^2 f(\mathbf{x})\mathbf{z} \geq \mu\|\mathbf{z}\|^2 \qquad (C \text{ is open}) \qquad (second-order) \tag{5}$$

for some $\mu > 0$, while convexity is when $\mu = 0$.

4.2 Proof of equivalency of the strong convexity

4.2.1 First Order Definition → Zero Order Definition

Denote $\mathbf{x}_\alpha := \mathbf{x} + \alpha(\mathbf{y}-\mathbf{x})$. Then,

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}_\alpha) + \langle \nabla f(\mathbf{x}_\alpha), \mathbf{y}-\mathbf{x}_\alpha\rangle + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}_\alpha\|^2 \\ &= f(\mathbf{x}_\alpha) + (1-\alpha)\langle \nabla f(\mathbf{x}_\alpha), \mathbf{y}-\mathbf{x}\rangle + \frac{\mu}{2}(1-\alpha)^2\|\mathbf{y}-\mathbf{x}\|^2 \end{aligned}$$

and

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{x}_\alpha) + \langle \nabla f(\mathbf{x}_\alpha), \mathbf{x}-\mathbf{x}_\alpha\rangle + \frac{\mu}{2}\|\mathbf{x}-\mathbf{x}_\alpha\|^2 \\ &= f(\mathbf{x}_\alpha) - \alpha\langle \nabla f(\mathbf{x}_\alpha), \mathbf{y}-\mathbf{x}\rangle + \frac{\mu}{2}(1-\alpha)^2\|\mathbf{y}-\mathbf{x}\|^2 \end{aligned}$$

Since $\alpha(1-\alpha)^2 + \alpha^2(1-\alpha) = \alpha(1-\alpha)$, adding a α multiple of the first equation to the $1-\alpha$ multiple of the second equation completes the proof.

4.2.2 Zero Order Definition → First Order Definition

Denote $\mathbf{x}_\alpha := \mathbf{x} + \alpha(\mathbf{y}-\mathbf{x})$. From $f(\mathbf{x}_\alpha) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \frac{\mu}{2}\alpha(1-\alpha)||\mathbf{y}-\mathbf{x}||^2$, we have

$$\begin{aligned} f(\mathbf{y}) &\geq \frac{f(\mathbf{x}_\alpha) - (1-\alpha)f(\mathbf{x}) + \frac{\mu}{2}\alpha(1-\alpha)||\mathbf{y}-\mathbf{x}||^2}{\alpha} \\ &= f(\mathbf{x}) + \frac{\mu}{2}(1-\alpha)\|\mathbf{y}-\mathbf{x}\|^2 + \frac{f(\mathbf{x}_\alpha) - f(\mathbf{x})}{\alpha} \end{aligned}$$

Let $\alpha \to 0$, then

$$\frac{f(\mathbf{x}_\alpha) - f(\mathbf{x})}{\alpha} = \frac{\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle}{1}$$

by the L'Hopital's rule and the chain rule we get the following:

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

4.2.3 Second Order Definition → First Order Definition

Denote $\mathbf{x}_\alpha := \mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})$. We know that:

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \int_0^\theta (\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{x}_\alpha)(\mathbf{y} - \mathbf{x}) d\alpha d\theta.$$

Starting from

$$\mathbf{z}^\top \nabla^2 f(\mathbf{x}) \mathbf{z} \geq \mu \|\mathbf{z}\|^2, \forall \mathbf{z} \in \mathbb{R}^d.$$

Now taking $\mathbf{z} := \mathbf{y} - \mathbf{x}$ and $\mathbf{x} := \mathbf{x}_\alpha$ we get

$$\begin{aligned}
f(\mathbf{y}) &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \int_0^\theta (\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{x}_\alpha)(\mathbf{y} - \mathbf{x}) d\alpha d\theta \\
&\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \int_0^\theta \mu ||y - x||^2 d\alpha d\theta \\
&= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \mu ||y - x||^2 \theta d\theta
\end{aligned}$$

and since we know $\int_0^1 \theta d\theta = \frac{1}{2}$ hence,

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

4.2.4 First Order Definition → Second Order Definition

Denote $\mathbf{x}_\alpha := \mathbf{x} + \alpha \mathbf{z}$, and denote $g(\alpha) := f(\mathbf{x}_\alpha)$.
Then, by chain rule, we have

$$\begin{aligned}
g^{'}(\alpha) &= \frac{\partial f(\mathbf{x}_\alpha)}{\partial \alpha} \\
&= \sum_{i=1}^d \frac{\partial f(\mathbf{x}_\alpha)}{\partial \mathbf{x}_\alpha[i]} \frac{\partial \mathbf{x}_\alpha[i]}{\partial \alpha} \\
&= \sum_{i=1}^d \frac{\partial f(\mathbf{x}_\alpha)}{\partial \mathbf{x}_\alpha[i]} z[i] \\
&= \langle \nabla f(\mathbf{x}_\alpha), \mathbf{z} \rangle
\end{aligned}$$

and

$$
\begin{aligned}
g^{''}(\alpha) &= \sum_{i=1}^{d} \frac{\partial}{\partial \alpha}\left(\frac{\partial f(\mathbf{x}_\alpha)}{\partial \mathbf{x}_\alpha[i]}\right) z[i] \\
&= \sum_{i=1}^{d}\left(\sum_{i=1}^{d} \frac{\partial^2 f(\mathbf{x}_\alpha)}{\partial \mathbf{x}_\alpha[i]\partial \mathbf{x}_\alpha[i]}\ \frac{\partial \mathbf{x}_\alpha[i]}{\partial \alpha}\right) z[i] \\
&= \sum_{i=1}^{d}\left(\sum_{i=1}^{d} \nabla^2 f(\mathbf{x}_\alpha)[i,j]\ z[i]\right) z[i] \\
&= \mathbf{z}^\top \nabla^2 f(\mathbf{x}_\alpha)\mathbf{z}.
\end{aligned}
$$

Additionally,

$$
\begin{aligned}
g^{''}(0) &= \lim_{\alpha\to 0} \frac{g^{'}(\alpha) - g^{'}(0)}{\alpha} \\
&= \lim_{\alpha\to 0} \frac{\langle \nabla f(\mathbf{x}_\alpha) - \nabla f(\mathbf{x}), \mathbf{z}\rangle}{\alpha} \\
&= \lim_{\alpha\to 0} \frac{\langle \nabla f(\mathbf{x}_\alpha) - \nabla f(\mathbf{x}), \mathbf{x}_\alpha - \mathbf{x}\rangle}{\alpha^2}
\end{aligned}
$$

We have that $\mathbf{z} = \frac{\mathbf{x}_\alpha - \mathbf{x}}{\alpha}$. We further lower-bound $\langle \nabla f(\mathbf{x}_\alpha) - \nabla f(\mathbf{x}), \mathbf{x}_\alpha - \mathbf{x}\rangle$ as follows: By strong convexity:

$$
f(\mathbf{x}_\alpha) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}_\alpha - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{x}_\alpha - \mathbf{x}\|^2 \tag{6}
$$

$$
f(\mathbf{x}) \geq f(\mathbf{x}_\alpha) + \langle \nabla f(\mathbf{x}_\alpha), \mathbf{x} - \mathbf{x}_\alpha\rangle + \frac{\mu}{2}\|\mathbf{x}_\alpha - \mathbf{x}\|^2 \tag{7}
$$

Adding the above two, we get

$$
\langle \nabla f(\mathbf{x}_\alpha) - \nabla f(\mathbf{x}), \mathbf{x}_\alpha - \mathbf{x}\rangle \geq \mu\|\mathbf{x}_\alpha - \mathbf{x}\|^2 = \mu\alpha^2\|\mathbf{z}\|^2. \tag{8}
$$

Combining the last three inequalities, we get the following:

$$
\mathbf{z}^\top \nabla^2 f(\mathbf{x})\mathbf{z} \geq \mu\|\mathbf{z}\|^2.
$$

We can see that we proved all the possible pairs for the equivalencies.

Bibliographic notes

More prelimiaries of calculus and linear algebra can be found in Chapter 1 of [Drusvyatskiy (2020)], Chapter 2 of [Vishnoi (2021)] and Chapter 3 and Chapter 4 of [Aaron Sidford (2024)].

References

[Aaron Sidford (2024)] Aaron Sidford. Optimization Algorithms. 2024.

[Drusvyatskiy (2020)] Dmitriy Drusvyatskiy. Convex Analysis and Nonsmooth Optimization. 2020.

[Vishnoi (2021)] Nisheeth K. Vishnoi. Algorithms for Convex Optimization. Cambridge University Press, 2021