Exercise 1.1 Make sure you understand how to derive formula (1.7) using function $\varphi(t) = f(x+td)$.

▼ **Solution**

Similarly, we can prove the *mean value theorem*: for a differentiable f and any x, d there exists $z \in [x, x+d]$ such that

$$f(x+d) = f(x) + \langle \nabla f(z), d \rangle. \tag{1.7}$$

Mean value theorem.

For differentiable f and $(\forall x, d)(\exists z \in [x, x+d])$ s.t.

$$f(x+d) = f(x) + \langle \nabla f(z), d \rangle$$

Proof.
Consider the function $\varphi(t) = f(x+td)$:
φ is differentiable. We can find its derivative using the chain rule:
$\varphi'(t) = f'(x+td)d = \langle \nabla f(x+td), d \rangle$
Therefore, we can apply the mean value theorem to φ at $[0, 1]$:
$\varphi(1) - \varphi(0) = \varphi'(t_0)$, for some $t_0 \in (0, 1)$.
By setting $z = x + t_0 d$ and plugging f back, we get:
$f(x+d) = f(x) + \langle \nabla f(z), d \rangle$, for $z \in (x, x+d)$

Exercise 1.2 Let $f(x) = \frac{1}{2}\|Ax - b\|^2$ for $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Compute the gradient in two ways: using the definition (1.5) and using the chain rule.

Solution

Using the definition:

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called differentiable at $x \in \mathbb{R}^n$ if there is $u \in \mathbb{R}^n$ such that for all $d \in \mathbb{R}^n$ we have

$$f(x+d) = f(x) + \langle u, d\rangle + o(||d||)$$

Let's consider the difference:

$$f(x+d) - f(x)$$
$$= \frac{1}{2}(\|A(x+d) - b\|^2 - \|Ax - b\|^2)$$
$$= \frac{1}{2}(\|Ax - b + Ad\|^2 - \|Ax - b\|^2)$$
$$= \frac{1}{2}(\|Ax - b\|^2 + 2\langle Ax - b, Ad\rangle + \|Ad\|^2 - \|Ax - b\|^2)$$
$$= \frac{1}{2}(2\langle Ax - b, Ad\rangle + \|Ad\|^2)$$
$$= \frac{1}{2}(2\langle A^\top(Ax - b), d\rangle + \|Ad\|)$$
$$= \langle A^\top(Ax - b), d\rangle + \frac{||Ad||}{2}$$

$\Rightarrow u = \nabla f(x) = A^\top(Ax - b)$

Using Chain Rule:

Let $h(x) = \frac{1}{2}\|x\|^2$, $g(x) = Ax - b$, then $f(x) = h(g(x))$

By the Chain Rule have:

$$\nabla f(x) = g'(x)^\top \nabla h(g(x))$$

As $g(x)$ is an affine function: $g'(x) = A$

Since $h(x) = \frac{1}{2}\sum x_j^2$, $\nabla h(x) = x$

$$\nabla f(x) = A^\top(Ax - b)$$

Exercise 1.3 Find the gradient and the Hessian of $f(x) = \frac{1}{2}\langle Qx, x\rangle - \langle b, x\rangle + c$, where $x, b \in \mathbb{R}^n$, $Q \in \mathbb{R}^{n\times n}$, and $c \in \mathbb{R}$.

▼ **Solution**

Just like in the previous task:

$$f(x+d) = f(x) + \langle u, d\rangle + o(||d||)$$

Considering the difference:

$$\begin{aligned} f(x+d) - f(x) &= \frac{1}{2}((x+d)^\top Q(x+d) - x^\top Qx) - b^T d \\ &= \frac{1}{2}(x^\top Qd + d^\top Qx + d^\top Qd) - b^T d \\ &= \frac{1}{2}(\langle Q^\top x, d\rangle + \langle Qx, d\rangle) - \langle b, d\rangle + \frac{1}{2}\langle Qd, d\rangle \end{aligned}$$

$$\Rightarrow u = \nabla f(x) = Qx - b$$

Hessian:

$$\nabla^2 f(x) = [\nabla f(x)]' = Q$$

Exercise 1.4 Make sure you understand what equation (1.10) means and why that derivation is correct.

▼ Solution

1. Gradients define the direction of the local fastest increase of f:

$$\frac{\nabla f(x)}{\|\nabla f(x)\|} = \underset{\|d\|=1}{\operatorname{argmax}} \lim_{t\to 0} \frac{f(x+td)-f(x)}{t} = \underset{\|d\|=1}{\operatorname{argmax}} \langle \nabla f(x), d\rangle.$$

This follows directly from the Cauchy-Schwarz inequality.

By **Cauchy–Bunyakovsky–Schwarz** inequality $|\langle \nabla f(x), d\rangle| \leq ||\nabla f(x)||||d||$

Since it is sufficient to only choose direction from vectors of $||d|| = 1$, we can maximize $\langle \nabla f(x), d\rangle$ by setting $d = \dfrac{\nabla f(x)}{||\nabla f(x)||}$.

Exercise 1.5 When is the function defined in Exercise 1.3 (i) convex; (ii) strongly convex?

▼ Solution

Exercise 1.3 Find the gradient and the Hessian of $f(x) = \frac{1}{2}\langle Qx, x\rangle - \langle b, x\rangle + c$, where $x, b \in \mathbb{R}^n$, $Q \in \mathbb{R}^{n\times n}$, and $c \in \mathbb{R}$.

(i) $(Q + Q^\top)$ is positive semidefinite
(i) $(Q + Q^\top) \succeq \mu I$ for some positive μ

Exercise 1.6 Prove equivalence in (ii) in Lemma 1.2.

▼ Solution

- f is μ-strongly convex if and only if

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \qquad \text{for all } x, y. \tag{1.15}$$

By definition of strong convexity:

μ-strongly convex if

$$\lambda f(x) + (1-\lambda) f(y) \geqslant f(\lambda x + (1-\lambda)y) + \frac{\mu\lambda(1-\lambda)}{2}\|x - y\|^2 \qquad \forall x, y \quad \forall \lambda \in [0,1] \tag{1.13}$$

($\Rightarrow$) Assume that function is strongly convex and

$$f(x) - f(y) \geq \frac{f(y + \lambda(x-y)) - f(y)}{\lambda} + \frac{\mu(1-\lambda)}{2}||x - y||^2$$

Since $f(x)$ is differentiable, letting $\lambda \to 0$, got

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

Conversely ($\Leftarrow$)
Take $z = \lambda x + (1-\lambda)y$

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

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

Multiplying the first inequality by λ and the second one by $1-\lambda$ and adding them up leads to the definition of $\mu-$strongly convex functions. To see that:

$$x - z = x - \lambda x - (1-\lambda)y = (1-\lambda)(x-y)$$
$$y - z = y - \lambda x - (1-\lambda)y = \lambda(y - x)$$

yielding $\lambda(x-z)+(1-\lambda)(y-z)=0.$

💡

Exercise 1.7 One may expect that in Lemma 1.3, strict convexity is equivalent to $\nabla^2 f(x) \succ 0$. Show that this is not true. ▪

▼ Solution

Consider $f(x)=x^4$ It is strictly convex:

$$\forall x,y \in \mathbb{R}$$
$$(1-\alpha)x^4+\alpha y^4 > ((1-\alpha)x+\alpha y)^4$$

$$((1-\alpha)x+\alpha y)^4 = (((1-\alpha)x+\alpha y)^2)^2$$

Since $g(x)=x^2$ is strictly convex have

$$(((1-\alpha)x+\alpha y)^2)^2 < ((1-\alpha)x^2+\alpha y^2)^2 < (1-\alpha)x^4+\alpha y^4$$

Exercise 1.8 Prove Lemma 1.4.

Lemma 1.4 Let $f : \mathbb{R}^n \to \mathbb{R}$. The following holds:

(i) If f is convex, then every local minimum is a global one and the set of minima is convex.

(ii) If f is strictly convex, then it has at most one minimum.

(iii) If f is strongly convex, then minimum always exists.

Solution

(i) Proof. By contradiction, let x_1, x_2 be local minima such that $f(x_1) < f(x_2)$.
Take some neighborhood $B(x_2, \alpha_2), \forall y \in B(x_2, \alpha_2) : f(y) \geq f(x_2)$. Such neighborhood always exist by definition of local minimum. But then, for all small enough $t \in (0, \alpha_2]$:

$$f((1-t)x_2 + tx_1) \geq f(x_2) > (1-t)f(x_2) + tf(x_1)$$

Which is a contradiction of f being convex.
Now let us prove that the set of minima is convex:
by contradiction, let us assume that x, y - global minima, but $z = (1-t)x + ty, t \in (0,1)$ isn't one.
By definition of convexity, we have:
$(1-t)f(x) + tf(y) = f(x) \geq f(z)$
However, we have previously stated that $f(z) > f(x)$. Therefore, we get a contradiction, and thus the set of minima is convex.
(ii) Proof. Suppose there are more than one minimum. Take two arbitrary points x_1, x_2 that minimize the function f. By (i) $f(x_1) = f(x_2)$, as strict convexity implies convexity.
We can consider three possible cases:

1. f attains value greater than $f(x_1), f(x_2)$ on interval $[x_1, x_2]$. This would clearly contradict definition of strict convexity.
2. f attains the same values as $f(x_1), f(x_2)$ on interval $[x_1, x_2]$. This would also contradict strict inequality in definition of strict convexity.

3. f attains values smaller $f(x_1), f(x_2)$ on interval $[x_1, x_2]$. But then this contradicts part (i).

(iii) Proof.
By definition of strong convexity (by taking $y = 0$)

$$(1-\lambda)f(x) + \lambda f(0) \geq f((1-\lambda)x) + \frac{\mu\lambda(1-\lambda)}{2}||x||^2, \qquad \forall x, \forall\lambda \in [0,1]$$

Setting $\lambda = \frac{1}{2}$ we get

$$f(x) \geq 2f\left(\frac{x}{2}\right) - f(0) + \frac{\mu}{4}||x||^2$$

Since $f(x) \geq 2f\left(\frac{x}{2}\right) - f(0)$ by convexity, as $||x|| \to \infty$ whole RHS goes to infinity, thus $f(x) \to \infty$. Then $(\forall M \in \Re)(\exists k \in \Re_+)(\forall x : ||x|| \geq k)\{f(x) \geq M\}$.
Therefore, for arbitrary M, we can find such k so that the minima has to belong to a closed ball $B(0, k)$.
Then, by theorem of Weierstrass a continuous function on closed and bounded set attains its minimum and maximum $\Rightarrow$minimum always exist.

Exercise 2.1 Show that the update in (2.3) is indeed equivalent to the GD update.

▼ Solution

2. Minimizing f is hard, so let's minimize its Taylor's approximation around x_k. Consider the first-order approximation

$$f(x) \approx f(x_k) + \langle \nabla f(x_k), x - x_k \rangle.$$

This approximation is linear over x, so minimizing it doesn't make much sense — we always obtain $-\infty$. To prevent this, we can add some regularization and thus, we define x_{k+1} as

$$x_{k+1} = \operatorname*{argmin}_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}. \tag{2.3}$$

It is easy to check that this will lead to GD as in (2.2).

Dropping the terms not dependent on x, we have:

Let us consider function

$$g(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2a_k} \|x - x_k\|^2$$

It's gradient is:

$$\nabla g(x) = \nabla f(x_k) + \frac{1}{a_k}(x - x_k)$$

And Hessian is:

$$\nabla^2 g(x) = \frac{1}{a_k} I$$

Since $\nabla^2 g(x) \succeq \frac{1}{a_k}$. $g(x)$ is strongly convex and, thus, have one local minimum, which is also global. By optimality condition $\nabla g(x) = 0$.

$$\begin{gathered} 0 = a_k \nabla f(x_k) + x - x_k \\ x = x_k - a_k \nabla f(x_k) \end{gathered}$$

Then call this minimizer as $x_{k+1} := x$ Which gives GD iteration

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

Exercise 2.2 Show that $\alpha = \frac{1}{L}$ is indeed the maximum of $\alpha(2-\alpha L)$.

▼ Solution

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}(\alpha(2-\alpha L)) = 2-\alpha L-\alpha L = 0 \Rightarrow \alpha = \frac{1}{L}$$

Exercise 2.3 Prove that f in (2.7) is convex if and only if $Q \succcurlyeq 0$.

▼ Solution

$$f(x) = \frac{1}{2}\langle Qx, x\rangle - \langle b, x\rangle, \tag{2.7}$$

$\nabla^2 f(x) = \frac{1}{2}(Q+Q^\top)$,
$f(x)$ is convex if and only if $\nabla^2 f(x) \succeq 0$.
Under the implicit assumption that Q is symmetric, this indeed becomes equivalent to $Q \succeq 0$.

Exercise 2.4 Prove that f in (2.7) is L-smooth, with $L = \|Q\| = \lambda_{\max}(Q)$.

▼ Solution

$$f(x) = \frac{1}{2}\langle Qx, x\rangle - \langle b, x\rangle, \qquad (2.7)$$

$\nabla^2 f(x) = \frac{1}{2}(Q + Q^\top)$,
$f(x)$ is L-smooth if and only if $\nabla^2 f(x) \preceq LI$.
Under the implicit assumption that Q is symmetric, this becomes equivalent to $\lambda_{max}(Q) \leq L$. Therefore, statement proven.

Exercise 2.6 Prove that $1 - t \leqslant e^{-t}$ using only convexity arguments (one line proof).

▼ Solution

$f(t) = e^{-t}$ is differentiable and convex; let us use an alternative definition of convexity for a differentiable function, i.e., that
$f(t) \geq f(0) + tf'(0)$:
$e^{-t} \geq e^0 - e^0 t = 1 - t$