II. Calculus

II.05. Optimization and Convexity

Lecture based on

https://github.com/gwthomas/math4ml (Garrett Thomas, 2018)

and

https://web.stanford.edu/b̃oyd/cvxbook/bv_cvxbook.pdf (Boyd and Vandenberghe, 2004)

Prof. Dr. Christoph Lippert

Digital Health & Machine Learning

Unconstrained Minimization

$$\text{minimize}\, f(x)$$

- f twice continuously differentiable
- optimal value $p^\star = \inf_x f(x)$ is attained (and finite)

Unconstrained minimization methods

- produce a sequence of points $x^{(k)} \in \operatorname{dom} f, k = 0, 1, \ldots$ with
$$f(x^{(k)}) \to p^\star$$
- can be interpreted as iterative methods for solving the optimality condition
$$\nabla f(x^\star) = 0$$

Initial point

- $x^{(0)} \in \operatorname{dom} f$

Descent Methods

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)} \quad \text{with} \quad f(x^{(k+1)}) < f(x^{(k)})$$

- alternative notation: $x := x + t\Delta x$
- Δ is the **step direction**, or **search direction**; t is the **step size**, or **step length**
- typically, chose $\nabla f(x)^\top \Delta x < 0$ (i.e., Δx is a **descent direction**)

General Descent method:

given a starting point $x \in \operatorname{dom} f$.

repeat:

1. Determine **descent direction** Δx
2. **Line search**. Chose a step size $t > 0$.
3. **Update**. $x := x + t\Delta x$

until stopping criterion is satisfied.

Line Search Types

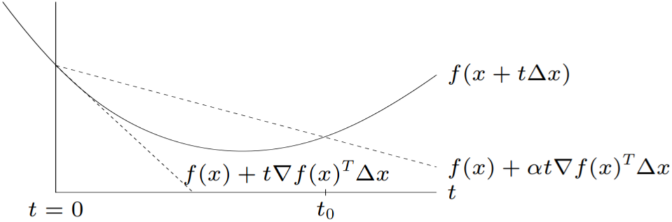
exact line search: $t = \arg\min_{t>0} f(x + t\Delta x)$

backtracking line search (with parameters $\alpha \in (0, 1/2), \beta \in (0, 1)$)

- starting at $t = 1$, repeat $t := \beta t$ until

$$f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^\top \Delta x$$

- graphical interpretation: backtrack until $t \leq t_0$


$f(x + t\Delta x)$
$f(x) + t\nabla f(x)^T \Delta x$
$f(x) + \alpha t\nabla f(x)^T \Delta x$
t
$t = 0$
t_0

No line search Set t to a small constant (e.g. 10^-3) and decrease t over iterations. Typically used in stochastic gradient descent.

Gradient Descent

general descent method with $\Delta x = -\nabla f(x)$

given a starting point $x \in \operatorname{dom} f$.

repeat:

1. $\Delta x := -\nabla f(x)$
2. **Line search**. Chose a step size t via exact or backtracking line search.
3. **Update**. $x := x + t\Delta x$

until stopping criterion is satisfied.

- stopping criterion usually in the form $\|\nabla f(x)\|_2 \leq \epsilon$
- very simple, but often very slow.

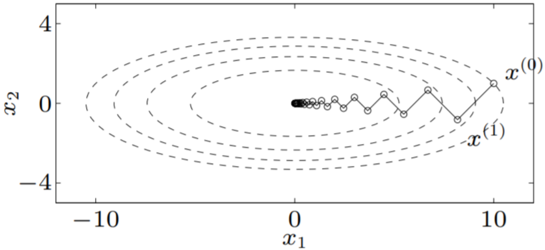
Quadratic problem in $\mathbb{R}^2$

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2) \qquad (\gamma > 0)$$

with exact line search, starting at $x^{(0)} = (\gamma, 1)$:

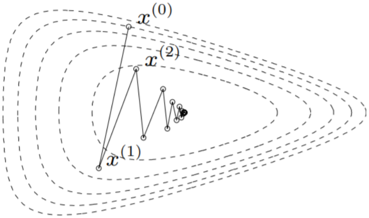
$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1}\right)^k, \quad x_2 = \left(-\frac{\gamma - 1}{\gamma + 1}\right)^k$$

- very slow if $\gamma \gg 1$ or $\gamma \ll 1$.
- example for $\gamma = 10$:

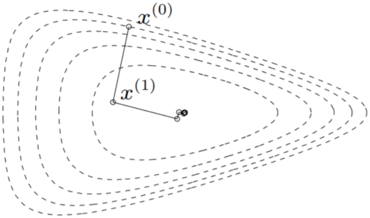


Nonquadratic example

$$f(x) = \exp(x_1 + 3x_2 - 0.1) + \exp(x_1 - 3x_2 - 0.1) + \exp(-x_1 - 0.1)$$



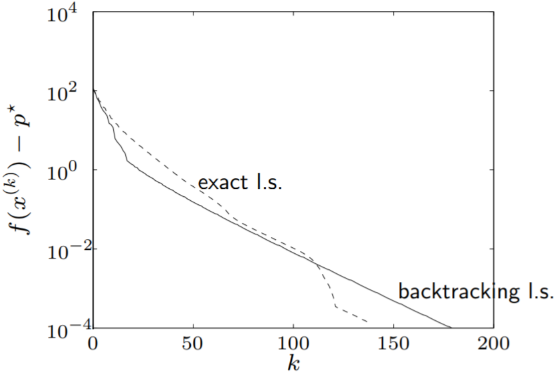
backtracking line search



exact line search

A problem in $\mathbb{R}^{100}$

$$f(x) = c^\top x - \sum_{i=1}^{500} \log\Big(b_i - a_i^\top x\Big)$$



'linear' convergence, i.e., a straight line on a semi-log plot

Steepest Descent

Normalized steepest descent direction (at $\mathbf{x}$, for norm $\|\cdot\|$):

$$\Delta\mathbf{x}_{\mathrm{nsd}} = \arg\min\{\nabla f(\mathbf{x})^\top \mathbf{v} \| \|\mathbf{v}\| = 1\}$$

interpretation: for small $\mathbf{v}$, $f(\mathbf{x}+\mathbf{v}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{v}$;

direction $\Delta\mathbf{x}_{\mathrm{nsd}}$ is unit-norm step with most negative directional derivative

(unnormalized) steepest descent direction

$$\Delta\mathbf{x}_{\mathrm{sd}} = \|\nabla f(\mathbf{x})\|_* \Delta\mathbf{x}_{\mathrm{nsd}}$$

satisfies $\nabla f(\mathbf{x})^\top \Delta\mathbf{x}_{\mathrm{sd}} = -\|\Delta f(\mathbf{x})\|_*^2$

steepest descent method

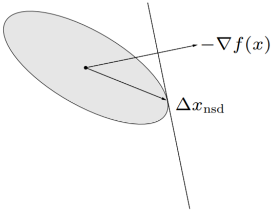
- general descent method with $\Delta\mathbf{x} = \Delta\mathbf{x}_{\mathrm{sd}}$
- Covergence properties similar to gradient descent

Examples

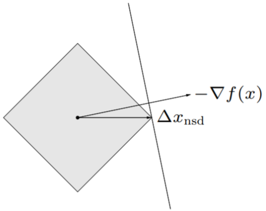
- Euclidean norm: $\Delta \mathbf{x}_{\mathrm{sd}} = -\nabla f(\mathbf{x})$
- quadratic norm $\|\mathbf{x}\|_{\mathbf{A}} = \left(\mathbf{x}^\top \mathbf{A}\mathbf{x}\right)^{1/2}$, with symmetric and positive definite n by n matrix $\mathbf{A}$: $\Delta x_{\mathrm{sd}} = -\mathbf{A}^{-1}\nabla f(x)$
- ℓ_1-norm: $\Delta \mathbf{x}_{\mathrm{sd}} = -\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i}\mathbf{e}_i$, where $|\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i}| = \|\nabla f(\mathbf{x})\|_\infty$

unit balls and normalized steepest descent directions for

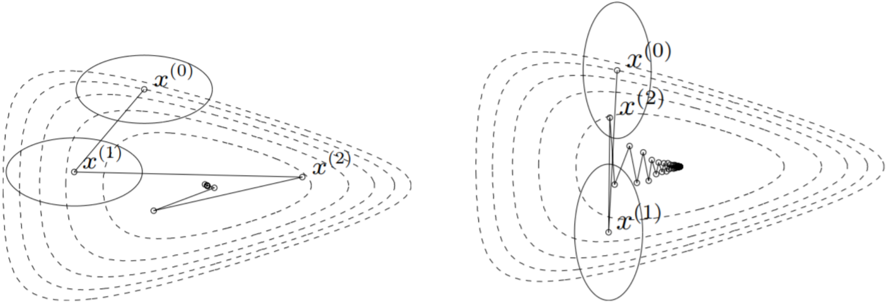
quadratic norm



ℓ_1-norm



Choice of norm for steepest descent



- steepest descent with backtracking line search for two quadratic norms
- ellipses show $\{x | \|x - x^{(k)}\|_P = 1\}$
- equivalent interpretation of steepest descent with quadratic norm $\|\dot{|}\|_P$: gradient descent after change of variables $\tilde{x} = P^{1/2}x$

shows that choice of P has strong effect on speed of convergence

Newton step

$$\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

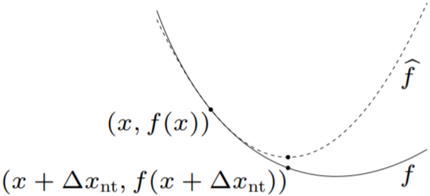
Interpretations:

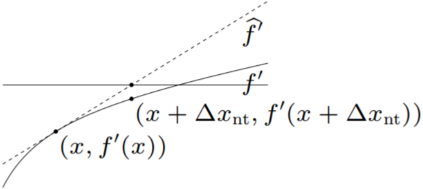
- $x + \Delta x_{\mathrm{nt}}$ minimizes second order approximation

$$\hat{f}(x+v) = f(x) + \nabla f(x)^\top v + \frac{1}{2} v^\top \nabla^2 f(x) v$$

- $x + \Delta x_{\mathrm{nt}}$ solves linearized optimality condition

$$\nabla f(x+v) \approx \nabla \hat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$





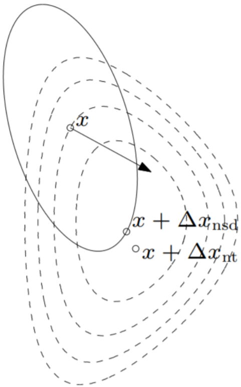
Newton step

$$\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

Interpretations:

- $x + \Delta x_{\mathrm{nt}}$ is the steepest descent direction at x in local Hessian norm

$$\|u\|_{\nabla^2 f(x)} = \left(u^\top \nabla^2 f(x) u\right)^{1/2}$$



- dashed lines are contour lines of f
- ellipse is $\{x + v | v^\top \nabla^2 f(x) v = 1\}$
- arrow shows $-\nabla f(x)$

Newton decrement

$$\lambda(x) = \sqrt{\nabla f(x)^\top \nabla^2 f(x)^{-1} \nabla f(x)}$$

is a measure of proximity of x to $x^\star$

Properties

- gives an estimation of $f(x) - f(x^\star)$, using the quadratic approximation $\hat{f}$:

$$f(x) - \inf_y \hat{f}(y) = \frac{1}{2}\lambda(x)^2$$

- equal to the norm of the Newton step in the quadratic Hessian norm

$$\lambda(x) = \Delta x_{\mathrm{nt}} \nabla^2 f(x) \Delta x_{\mathrm{nt}}$$

- directional derivative in the Newton direction:

$$\nabla f(x)^\top \Delta x_{\mathrm{nt}} = -\lambda(x)^2$$

Newton's method

given a starting point $x \in \operatorname{dom} f$, tolerance $\epsilon > 0$.

repeat:

1. Compute the **Newton step** and **Newton decrement**.

$$\Delta x_{\text{nt}} := -\nabla^2 f(x)^{-1} \nabla f(x); \;\; \lambda^2 := \nabla f(x) \nabla^2 f(x)^{-1} \nabla f(x)$$

2. **Stopping criterion**. *quit* if $\frac{\lambda^2}{2} \leq \epsilon$.
3. **Line search**. Choose step size t by backtracking line search
4. **Update**. $x := x + t\Delta x_{\text{nt}}$

affine invariant

- independent of linear changes of coordinates
- Newton iterates for $\tilde{f}(y) = f(Ty)$ with starting point $y^{(0)} = T^{-1}x^{(0)}$ are

$$y^{(k)} = T^{-1}x^{(k)}$$

Newton's method requires an invertible Hessian matrix.

- When $\nabla^2 f(x)$ is not p.d., Δx_{nt} may not even be defined.
- Even when it is defined, Δx_{nt} may not be a descent direction, if

$$\nabla f(x)^\top \Delta x_{\mathrm{nt}} < 0$$

- In this case, line search methods needs to modify the search direction.

Convexity is a term that pertains to both sets and functions.

For functions, there are different degrees of convexity, and how convex a function is tells us a lot about its minima:

- do they exist
- are they unique
- how quickly can we find them using optimization algorithms
- etc.

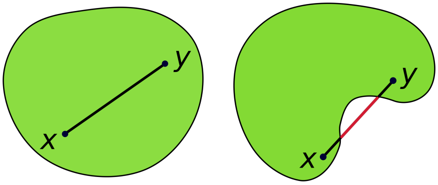
Here, we present basic results regarding

- convexity
- strict convexity
- strong convexity.

A set $\mathcal{X} \subseteq \mathbb{R}^d$ is **convex** if

$$t\mathbf{x} + (1-t)\mathbf{y} \in \mathcal{X}$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and all $t \in [0,1]$.
Geometrically, this means that all the points on the line segment between any two points in $\mathcal{X}$ are also in $\mathcal{X}$.



(a) A convex set (b) A non-convex set

Figure: What convex sets look like

Why do we care whether or not a set is convex?

- The nature of minima can depend greatly on whether or not the feasible set is convex.
- Undesirable pathological results can occur when we allow the feasible set to be arbitrary, so for proofs we will need to assume that it is convex.
- Fortunately, we often want to minimize over all of $\mathbb{R}^d$, which is easily seen to be a convex set.

Theorem

Assume that X and Y are convex sets.
Then $X \cap Y$ is also convex.

Proof.

Consider any $a, b \in X \cap Y$.
Since X and Y are convex, the line segments connecting a and b are contained in both X and Y.

Given that, they also need to be contained in $X \cap Y$. □

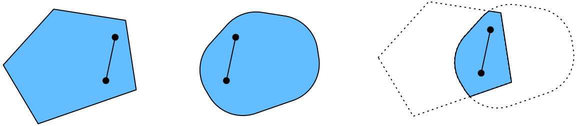
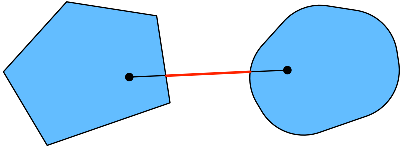


Figure: The intersection between two convex sets is convex

- Given convex sets X_i, their intersection $\cap_i X_i$ is convex.
- The converse is not true, consider two disjoint sets $X \cap Y = \emptyset$.
- Now pick $a \in X$ and $b \in Y$.
- The line segment connecting a and b needs to contain some part that is neither in X nor Y, since we assumed that $X \cap Y = \emptyset$.
- Hence the line segment isn't in $X \cup Y$ either, thus proving that in general unions of convex sets need not be convex.



The union of two convex sets need not be convex.

Typically the problems in machine learning are defined on convex domains.

- $\mathbb{R}^d$ is a convex set (after all, the line between any two points in $\mathbb{R}^d$ remains in $\mathbb{R}^d$).
- In some cases we work with variables of bounded length, such as balls of radius r as defined by

$$\{\mathbf{x}|\mathbf{x} \in \mathbb{R}^d \text{ and } \|\mathbf{x}\|_2 \leq r\}$$

.

In the remainder, assume $f : \mathbb{R}^d \to \mathbb{R}$ unless otherwise noted.

We'll start with the definitions and then give some results.

- A function f is **convex** if
$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$
for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and all $t \in [0,1]$.
- If the inequality holds strictly (i.e. $<$ rather than $\leq$) for all $t \in (0,1)$ and $\mathbf{x} \neq \mathbf{y}$, then we say that f is **strictly convex**.
- A function f is **strongly convex with parameter** m (or m-**strongly convex**) if the function
$$\mathbf{x} \mapsto f(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$$
is convex.

These conditions are given in increasing order of strength; strong convexity implies strict convexity which implies convexity.

Geometric interpretation:

- **Convexity** means that the line segment between two points on the graph of f lies on or above the graph itself.
- **Strict convexity** means that the graph of f lies strictly above the line segment, except at the segment endpoints.

(The function in the figure is strictly convex.)

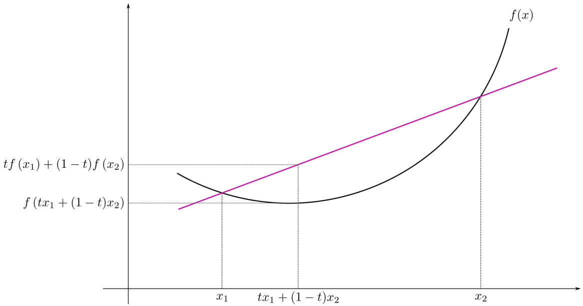


Figure: What convex functions look like

Consequences of convexity:

Why do we care if a function is (strictly/strongly) convex?

- The various notions of convexity have implications about the nature of minima.
- The stronger conditions tell us more about the minima.

Proposition

Let $\mathcal{X}$ be a convex set.
If f is convex, then any local minimum of f in $\mathcal{X}$ is also a global minimum.

Proof.

Suppose f is convex, and let $\mathbf{x}^*$ be a local minimum of f in $\mathcal{X}$.
Then for some neighborhood $N \subseteq \mathcal{X}$ about $\mathbf{x}^*$, we have $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x} \in N$.
Suppose towards a contradiction that there exists $\tilde{\mathbf{x}} \in \mathcal{X}$ such that $f(\tilde{\mathbf{x}}) < f(\mathbf{x}^*)$.
Consider the line segment $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$, noting that $\mathbf{x}(t) \in \mathcal{X}$ by the convexity of $\mathcal{X}$.
Then by the convexity of f, if follows for all $t \in (0,1)$:

$$f(\mathbf{x}(t)) \leq tf(\mathbf{x}^*) + (1-t)f(\tilde{\mathbf{x}}) < tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) = f(\mathbf{x}^*)$$

We can pick t to be sufficiently close to 1 that $\mathbf{x}(t) \in N$;
then $f(\mathbf{x}(t)) \geq f(\mathbf{x}^*)$ by the definition of N, but $f(\mathbf{x}(t)) < f(\mathbf{x}^*)$ by the above inequality, a contradiction.
It follows that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$.
So $\mathbf{x}^*$ is a global minimum of f in $\mathcal{X}$. □

Proposition

Let $\mathcal{X}$ be a convex set.
If f is strictly convex, then there exists at most one local minimum of f in $\mathcal{X}$.
Consequently, if it exists it is the unique global minimum of f in $\mathcal{X}$.

Proof.

The second sentence follows from the first, so all we must show is that if a local minimum exists in $\mathcal{X}$ then it is unique.
Suppose $\mathbf{x}^*$ is a local minimum of f in $\mathcal{X}$, and suppose towards a contradiction that there exists a local minimum $\tilde{\mathbf{x}} \in \mathcal{X}$ such that $\tilde{\mathbf{x}} \neq \mathbf{x}^*$.
Since f is strictly convex, it is convex, so $\mathbf{x}^*$ and $\tilde{\mathbf{x}}$ are both global minima of f in $\mathcal{X}$ by the previous result. Hence $f(\mathbf{x}^*) = f(\tilde{\mathbf{x}})$.
Consider the line segment $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$, which again must lie entirely in $\mathcal{X}$.
By the strict convexity of f, it follows for all $t \in (0,1)$ that

$$f(\mathbf{x}(t)) < tf(\mathbf{x}^*) + (1-t)f(\tilde{\mathbf{x}}) = tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) = f(\mathbf{x}^*)$$

But this contradicts the fact that $\mathbf{x}^*$ is a global minimum.
Therefore if $\tilde{\mathbf{x}}$ is a local minimum of f in $\mathcal{X}$, then $\tilde{\mathbf{x}} = \mathbf{x}^*$, so $\mathbf{x}^*$ is the unique minimum in $\mathcal{X}$. □

It is worthwhile to examine how the feasible set affects the optimization problem. We will see why the assumption that $\mathcal{X}$ is convex is needed in the results so far.

Consider the function $f(x) = x^2$, which is a strictly convex function.

The unique global minimum of this function in $\mathbb{R}$ is $x = 0$.

But let's see what happens when we change the feasible set $\mathcal{X}$.

i $\mathcal{X} = \{1\}$: This set is actually convex, so we still have a unique global minimum. But it is not the same as the unconstrained minimum!

ii $\mathcal{X} = \mathbb{R} \setminus \{0\}$: This set is non-convex, and we can see that f has no minima in $\mathcal{X}$.

For any point $x \in \mathcal{X}$, one can find another point $y \in \mathcal{X}$ such that $f(y) < f(x)$.

iii $\mathcal{X} = (-\infty, -1] \cup [0, \infty)$: This set is non-convex, and we can see that there is a local minimum $(x = -1)$ which is distinct from the global minimum $(x = 0)$.

iv $\mathcal{X} = (-\infty, -1] \cup [1, \infty)$:

This set is non-convex, and we can see that there are two global minima $(x = \pm 1)$.

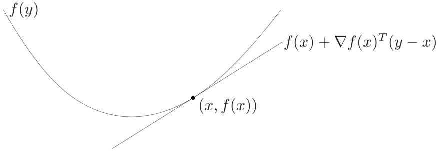
Proposition (First order Condition)

Suppose f is differentiable. Then f is convex if and only if, for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$

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

Moreover, f is strictly convex, if and only if for $\mathbf{x} \neq \mathbf{y}$

$$f(\mathbf{y}) > f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$$



f convex $\Rightarrow f(y) \geq f(x) + f'(x)(y-x)$ (one direction for the 1-dim case only).

Assume $f : \mathbb{R} \to \mathbb{R}$ is convex and $x, y \in \operatorname{dom} f$.
Since $\operatorname{dom} f$ is convex (i.e., an interval), it follows for all $0 < t \leq 1$ and $x + t(y-x)$:

$$f(x + t(y-x)) \leq (1-t)f(x) + tf(y)$$

$$tf(y) \geq f(x + t(y-x)) - (1-t)f(x) = f(x + t(y-x)) - f(x) + tf(x)$$

$$f(y) \geq f(x) + \frac{f(x + t(y-x)) - f(x)}{t}$$

By letting $t \to 0$, we get $f(y) \geq f(x) + f'(x)(y-x)$ □

Proposition (Second order conditions)

Suppose f is twice differentiable.
Then

i f *is convex if and only if* $\nabla^2 f(\mathbf{x}) \succeq 0$ *for all* $\mathbf{x} \in \operatorname{dom} f$.

ii *If* $\nabla^2 f(\mathbf{x}) \succ 0$ *for all* $\mathbf{x} \in \operatorname{dom} f$, *then* f *is strictly convex.*

iii f *is* m*-strongly convex if and only if* $\nabla^2 f(\mathbf{x}) \succeq mI$ *for all* $\mathbf{x} \in \operatorname{dom} f$.

Proposition

If f is convex and $\alpha \geq 0$, then αf is convex.

Proof.

Suppose f is convex and $\alpha \geq 0$.
Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(\alpha f) = \operatorname{dom} f$,

$$\begin{aligned}(\alpha f)(t\mathbf{x} + (1-t)\mathbf{y}) &= \alpha f(t\mathbf{x} + (1-t)\mathbf{y}) \\ &\leq \alpha \left(t f(\mathbf{x}) + (1-t) f(\mathbf{y})\right) \\ &= t(\alpha f(\mathbf{x})) + (1-t)(\alpha f(\mathbf{y})) \\ &= t(\alpha f)(\mathbf{x}) + (1-t)(\alpha f)(\mathbf{y})\end{aligned}$$

so αf is convex. □

How can we show that a function is (strictly/strongly) convex?

It is of course possible (in principle) to directly show that the condition in the definition holds, but this is usually not the easiest way.

Proposition

Norms are convex.

Proof.

Let $\|\cdot\|$ be a norm on a vector space V.
Then for all $\mathbf{x}, \mathbf{y} \in V$ and $t \in [0,1]$,

$$\|t\mathbf{x} + (1-t)\mathbf{y}\| \leq \|t\mathbf{x}\| + \|(1-t)\mathbf{y}\| = |t|\|\mathbf{x}\| + |1-t|\|\mathbf{y}\| = t\|\mathbf{x}\| + (1-t)\|\mathbf{y}\|$$

where we have used respectively the triangle inequality, the homogeneity of norms, and the fact that t and $1-t$ are nonnegative.
Hence $\|\cdot\|$ is convex. □

Proposition

If f and g are convex, then $f + g$ is convex.
Furthermore, if g is strictly convex, then $f + g$ is strictly convex, and if g is m-strongly convex, then $f + g$ is m-strongly convex.

Proof.

Suppose f and g are convex. Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(f + g) = \operatorname{dom} f \cap \operatorname{dom} g$,

$$\begin{aligned}
(f+g)(t\mathbf{x} + (1-t)\mathbf{y}) &= f(t\mathbf{x} + (1-t)\mathbf{y}) + g(t\mathbf{x} + (1-t)\mathbf{y}) \\
&\le tf(\mathbf{x}) + (1-t)f(\mathbf{y}) + g(t\mathbf{x} + (1-t)\mathbf{y}) && \text{convexity of } f \\
&\le tf(\mathbf{x}) + (1-t)f(\mathbf{y}) + tg(\mathbf{x}) + (1-t)g(\mathbf{y}) && \text{convexity of } g \\
&= t(f(\mathbf{x}) + g(\mathbf{x})) + (1-t)(f(\mathbf{y}) + g(\mathbf{y})) \\
&= t(f+g)(\mathbf{x}) + (1-t)(f+g)(\mathbf{y}) && \text{so } f+g \text{ is convex.}
\end{aligned}$$

If g is strictly convex, the second inequality above holds strictly for $\mathbf{x} \neq \mathbf{y}$ and $t \in (0, 1)$, so $f + g$ is strictly convex.
If g is m-strongly convex, then the function $h(\mathbf{x}) \equiv g(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$ is convex, so $f + h$ is convex.
But $f + g$ is m-strongly convex:

$$(f+h)(\mathbf{x}) \equiv f(\mathbf{x}) + h(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2 \equiv (f+g)(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$$

Proposition

If $f_1, \dots, f_n$ are convex and $\alpha_1, \dots, \alpha_n \geq 0$, then

$$\sum_{i=1}^{n} \alpha_i f_i$$

is convex.

Proof.

Follows from the previous two propositions by induction. □

Proposition

If f is convex, then $g(\mathbf{x}) \equiv f(\mathbf{A}\mathbf{x} + \mathbf{b})$ is convex for any appropriately-sized $\mathbf{A}$ and $\mathbf{b}$.

Proof.

Suppose f is convex and g is defined like so.
Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} g$,

$$\begin{aligned} g(t\mathbf{x} + (1-t)\mathbf{y}) &= f(\mathbf{A}(t\mathbf{x} + (1-t)\mathbf{y}) + \mathbf{b}) \\ &= f(t\mathbf{A}\mathbf{x} + (1-t)\mathbf{A}\mathbf{y} + \mathbf{b}) \\ &= f(t\mathbf{A}\mathbf{x} + (1-t)\mathbf{A}\mathbf{y} + t\mathbf{b} + (1-t)\mathbf{b}) \\ &= f(t(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-t)(\mathbf{A}\mathbf{y} + \mathbf{b})) \\ &\le tf(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-t)f(\mathbf{A}\mathbf{y} + \mathbf{b}) && \text{convexity of } f \\ &= tg(\mathbf{x}) + (1-t)g(\mathbf{y}) \end{aligned}$$

Thus g is convex. □

Proposition

If f and g are convex, then $h(\mathbf{x}) \equiv \max\{f(\mathbf{x}), g(\mathbf{x})\}$ is convex.

Proof.

Suppose f and g are convex and h is defined like so.
Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} h$,

$$\begin{aligned} h(t\mathbf{x} + (1-t)\mathbf{y}) &= \max\{f(t\mathbf{x} + (1-t)\mathbf{y}), g(t\mathbf{x} + (1-t)\mathbf{y})\} \\ &\le \max\{tf(\mathbf{x}) + (1-t)f(\mathbf{y}), tg(\mathbf{x}) + (1-t)g(\mathbf{y})\} \\ &\le \max\{tf(\mathbf{x}), tg(\mathbf{x})\} + \max\{(1-t)f(\mathbf{y}), (1-t)g(\mathbf{y})\} \\ &= t\max\{f(\mathbf{x}), g(\mathbf{x})\} + (1-t)\max\{f(\mathbf{y}), g(\mathbf{y})\} \\ &= th(\mathbf{x}) + (1-t)h(\mathbf{y}) \end{aligned}$$

Note that in the first inequality we have used convexity of f and g plus the fact that $a \le c, b \le d$ implies $\max\{a, b\} \le \max\{c, d\}$.
In the second inequality we have used the fact that $\max\{a + b, c + d\} \le \max\{a, c\} + \max\{b, d\}$.
Thus h is convex. □

Theorem (Jensen's Inequality)

Let f be a convex function.

$$\sum_i \alpha_i f(x_i) \geq f\left(\sum_i \alpha_i x_i\right)$$

$$\textit{and } \mathbf{E}_x[f(x)] \geq f\left(\mathbf{E}_x[x]\right)$$

Proof.

To prove the first inequality we repeatedly apply the definition of convexity to one term in the sum at a time.
The expectation can be proven by taking the limit over finite segments. □

The expectation of a convex function $(\mathbf{E}_x[f(x)])$ is larger than the convex function of an expectation $f(\mathbf{E}_x[x])$.

One of the common applications of Jensen's inequality is related to the log-likelihood of partially observed random variables.

$$\mathbf{E}_{y\sim p(y)}[-\log p(x|y)] \geq -\log \underbrace{p(x)}_{\mathbf{E}_{y\sim p(y)}[p(x|y)]} .$$

This is used in **variational methods**.

- y is typically the unobserved random variable
- $p(y)$ is the best guess of how it might be distributed
- $p(x)$ is the marginal distribution (with y integrated out).
- For instance, in clustering y might be the cluster labels and $p(x|y)$ is the generative model when applying cluster labels.

A good way to gain intuition about the distinction between convex, strictly convex, and strongly convex functions is to consider examples where the stronger property fails to hold.

Functions that are convex but not strictly convex:

i $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + \alpha$ for any $\mathbf{w} \in \mathbb{R}^d, \alpha \in \mathbb{R}$. Such a function is called an **affine function**, and it is both convex and concave. (In fact, a function is affine if and only if it is both convex and concave.) Note that linear functions and constant functions are special cases of affine functions.

ii $f(\mathbf{x}) = \|\mathbf{x}\|_1$

Functions that are strictly but not strongly convex:

1. $f(x) = x^4$. This example is interesting because it is strictly convex but you cannot show this fact via a second-order argument (since $f''(0) = 0$).
2. $f(x) = \exp(x)$. This example is interesting because it's bounded below but has no local minimum.
3. $f(x) = -\log x$. This example is interesting because it's strictly convex but not bounded below.

Functions that are strongly convex:

1. $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$