Optimization

1 Motivation

In data analysis, optimization methods are used to fit models to the available data. Model parameters are chosen by maximizing or minimizing a **cost function**, such as the likelihood of the data given the parameters or the error achieved by the model on a training dataset. In these notes we will introduce some of the basic methods that can be applied to **unconstrained** optimization problems of the form

$$\min_{\boldsymbol{x}} f\left(\boldsymbol{x}\right), \tag{1}$$

for a cost function $f : \mathbb{R}^n \to \mathbb{R}$. We will begin by introducing the main concepts in a 1D setting that is easier to visualize, before moving on to multiple dimensions.

2 Optimization in 1D

2.1 Derivatives and convexity

The optimization methods that we will describe in the following sections start from an arbitrary point and try to make progress towards the minimum of the function of interest by using the local characteristics of the function, such as the slope or the curvature. This local information is captured by the **derivatives** of the function.

Definition 2.1 (Derivative)**.** *The derivative of a real-valued function $f : \mathbb{R} \to \mathbb{R}$ at $x \in \mathbb{R}$ is defined as*

$$f'\left(x\right) := \lim_{h \to 0} \frac{f\left(x+h\right) - f\left(x\right)}{h}. \tag{2}$$

Higher order derivatives are defined recursively,

$$f''\left(x\right) := \left(f'\left(x\right)\right)', \tag{3}$$

$$f'''\left(x\right) := \left(f''\left(x\right)\right)', \tag{4}$$

i.e. the second derivative of a function is the derivative of the first derivative, the third derivative is the derivative of the second derivative and so on.

A function is said to be n-times continuously differentiable if all its derivatives up to the nth exist and are continuous.

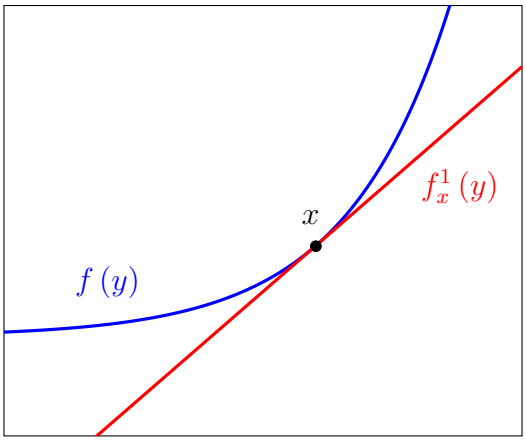

Figure 1: Linear approximation of a function f at x.

Geometrically, the derivative at a point x corresponds to the slope of the line that is tangent to the curve $(y, f(y))$ at x. This line is a **linear approximation** to the function, as illustrated in Figure 1.

Definition 2.2 (First-order approximation)**.** *The first-order or linear approximation of a differentiable function $f : \mathbb{R} \to \mathbb{R}$ at $x \in \mathbb{R}$ is*

$$f_x^1(y) := f(x) + f'(x)(y - x). \tag{5}$$

It follows immediately from the definition of derivative that the linear function $f_x^1(y)$ becomes an arbitrarily good approximation of f as we approach x, even if we divide the error by the distance to x.

Lemma 2.3. *The linear approximation $f_x^1 : \mathbb{R} \to \mathbb{R}$ at $x \in \mathbb{R}$ of a differentiable function $f : \mathbb{R} \to \mathbb{R}$ satisfies*

$$\lim_{y \to x} \frac{f(y) - f_x^1(y)}{x - y} = 0. \tag{6}$$

The value of the derivative provides valuable information about whether the value of the function is increasing or decreasing.

Lemma 2.4 (Local monotonicity)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function.*

- *If $f'(x) > 0$, f is increasing at x.*
- *If $f'(x) < 0$, f is decreasing at x.*
- *If $f'(x) = 0$, f reaches a local minimum or a local maximum at x.*

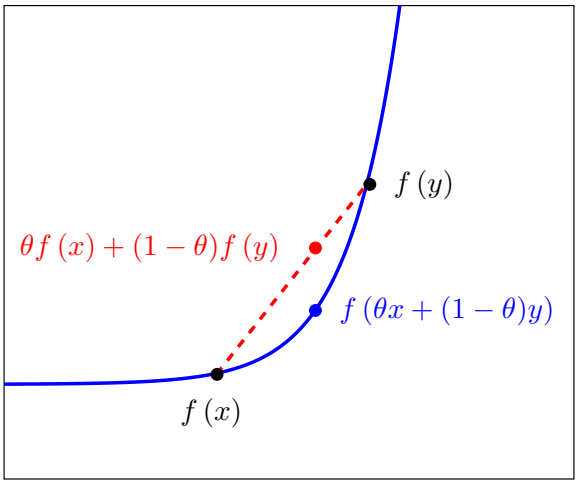

Figure 2: Illustration of condition (7) in Definition 2.5. The curve corresponding to the function must lie below any chord joining two of its points.

In general, information about the *local* behavior of a function does not necessarily reflect its *global* behavior. We could move in a direction in which the function decreases only to find a local minimum that is much larger than the global minimum of the function. This is *not* the case for **convex** functions.

Definition 2.5 (Convexity)**.** *A function is convex if for any* $x, y \in \mathbb{R}$ *and any* $\theta \in (0,1)$*,*

$$\theta f(x) + (1-\theta) f(y) \geq f(\theta x + (1-\theta) y). \tag{7}$$

The function is strictly convex if the inequality is strict

$$\theta f(x) + (1-\theta) f(y) > f(\theta x + (1-\theta) y). \tag{8}$$

A function f *is said to be concave is* $-f$ *is convex, and strictly concave if* $-f$ *is strictly convex.*

Condition (7) is illustrated in Figure 2. The curve corresponding to the function must lie below any chord joining two of its points. This is a general condition that applies to functions that are not differentiable. The following lemma, proved in Section A.1, provides an equivalent condition when the function is differentiable: its curve must lie above all of its linear approximations (for example, the function in Figure 1 is convex).

Lemma 2.6. *A differentiable function* f *is convex if and only if for all* $x, y \in \mathbb{R}$

$$f(y) \geq f_x^1(y) \tag{9}$$

and strictly convex if and only if for all $x, y \in \mathbb{R}$

$$f(y) > f_x^1(y). \tag{10}$$

An immediate corollary is that for a convex function, any point at which the derivative is zero is a **global minimum**.

Corollary 2.7. *If a differentiable function f is convex and $f'(x) = 0$, then for any $y \in \mathbb{R}$*

$$f(y) \geq f(x). \tag{11}$$

If the function is strictly convex, and $f'(x) = 0$, then for any $y \in \mathbb{R}$

$$f(y) > f(x), \tag{12}$$

i.e. x is the only minimizer.

This means that if we are trying to minimize a convex function, once we locate a point at which the derivative of the function is zero we are done!

For functions that are twice-differentiable, convexity is related to the curvature of the function. Curvature is quantified by the second derivative of the function. The following lemma, proved in Section A.2 of the appendix, establishes that such functions are convex if and only if their curvature is always nonnegative.

Lemma 2.8. *A 2-times continuously differentiable function f is convex if and only if $f''(x) \geq 0$ for all $x \in \mathbb{R}$. It is strictly convex if and only $f''(x) > 0$.*

A quadratic function is a second-order polynomial of the form $\frac{a}{2}x^2 + bx + c$. Its second derivative is equal to a, so if $a > 0$ the function is strictly convex, whereas if $a < 0$ it is strictly concave. If $a = 0$ then the function is linear. The **second-order approximation** of a function at a certain point is equal to the quadratic function that has the same value, slope and curvature at that point (these three parameters completely determine a quadratic function), as illustrated by Figure 3.

Definition 2.9 (Second-order approximation)**.** *The second-order or quadratic approximation of f at x_0 is*

$$f_x^2(x) := f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2. \tag{13}$$

The quadratic function $f_x^2(y)$ becomes an arbitrarily good approximation of f as we approach x, even if we divide the error by the squared distance between x and y. We omit the proof that follows from basic calculus.

Lemma 2.10. *The quadratic approximation $f_x^2 : \mathbb{R} \to \mathbb{R}$ at $x \in \mathbb{R}$ of a twice continuously differentiable function $f : \mathbb{R} \to \mathbb{R}$ satisfies*

$$\lim_{y \to x} \frac{f(y) - f_x^2(y)}{(x - y)^2} = 0. \tag{14}$$

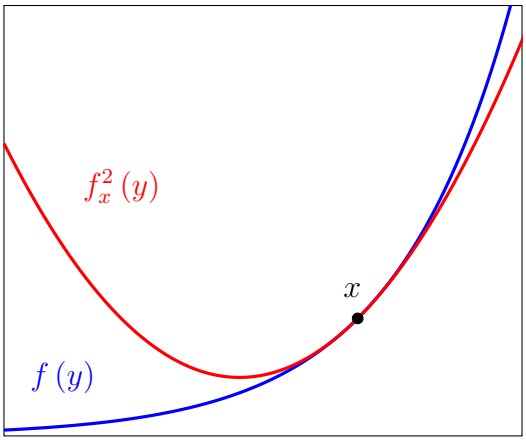

Figure 3: Quadratic approximation $f_x^2(y)$ of a function f at x.

2.2 Optimization algorithms

In order to explain how gradient descent works, we will first introduce a 1D-version which we call derivative descent (this is *not* a standard term). The idea is to start at an arbitrary point x_0 and move in the direction in which the function decreases.

Algorithm 2.11 (Derivative descent)**.**
Input*: A differentiable function $f : \mathbb{R} \to \mathbb{R}$, its derivative f', a step size α and a stopping threshold ϵ.*
Output*: An estimate of the minimum of the function.*

1. *Choose a random initialization $x_0 \in \mathbb{R}$.*
2. *For $i = 1, 2, \ldots$ compute*

$$x_i = x_{i-1} - \alpha\, f'(x_{i-1})\,. \tag{15}$$

until $|f'(x_i)| \leq \epsilon$.

Note that the size of the step we take in each iteration is proportional to the magnitude of the derivative at that point. This means that we will make rapid progress in regions where $|f'|$ is large and slower progress when $|f'|$ is small. For functions that are locally quadratic, the derivative gradually becomes smaller as we approach the minimum. For such functions, taking a step that is proportional to $|f'|$ decreases the magnitude of the steps and allows the algorithm to converge to the minimum. Figure 4 shows the progress of derivative descent when applied to a quadratic for two values of the constant α. If α is small, we make slower progress towards the minimum, but if it is very large we might repeatedly overshoot the minimum.

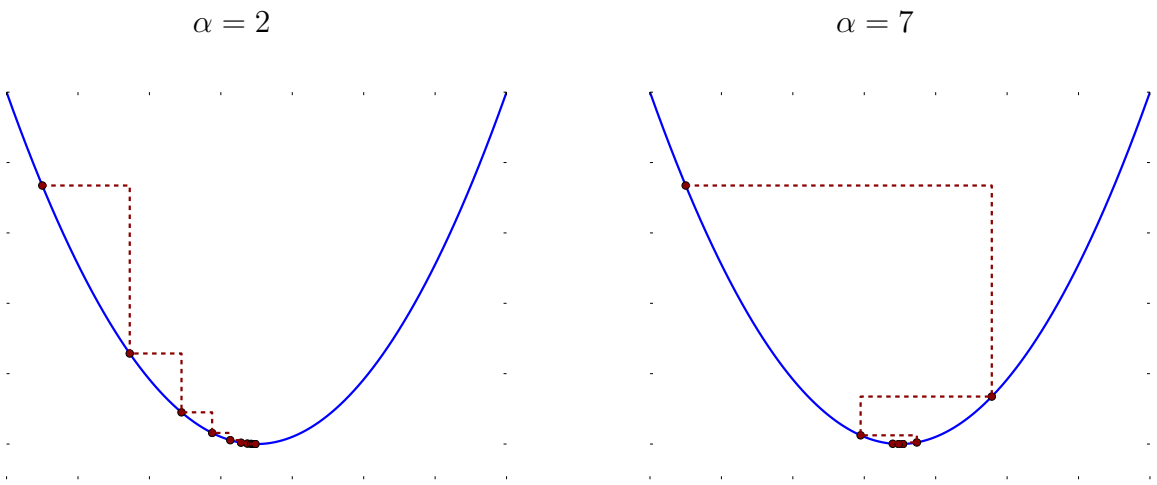

Figure 4: Derivative descent applied to a quadratic function for two different values of α. In both cases the iterations converge to the minimum.

Derivative descent only uses first-order information about the function. It does not take into account its curvature. If we know that the function is well approximated by a quadratic, we could adapt the step size that we take accordingly. This is the idea behind Newton's method. The following simple lemma shows that we always minimize a quadratic by taking a step that depends on the curvature.

Lemma 2.12. *Derivative descent finds the global minimum of a convex quadratic*

$$q(x) := \frac{a}{2}x^2 + bx + c \tag{16}$$

if we set

$$\alpha = \frac{1}{q''(x_0)} = \frac{1}{a}. \tag{17}$$

Proof. The global minimum of a quadratic function can be found by setting the derivative

$$q'(x) = ax + b \tag{18}$$

to zero, which yields that the minimum $x^* = -b/a$. The first step of derivative descent with a step size of $1/a$ always reaches x^*,

$$x_1 = x_0 - \alpha f'(x_0) \tag{19}$$

$$= x_0 - \frac{ax_0 + b}{a} \tag{20}$$

$$= x^*. \tag{21}$$

□

Corollary 2.13. *For a 2-times continuously differentiable function $f : \mathbb{R} \to \mathbb{R}$,*

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)} \tag{22}$$

minimizes $f^2_{x_0}$, the quadratic approximation of f at x_0.

Newton's method iteratively minimizes the quadratic approximation of a function.

Algorithm 2.14 (Newton's method)**.**
Input*: A twice-differentiable function $f : \mathbb{R} \to \mathbb{R}$, its first and second derivative f' and f'', a step size α and a stopping threshold ϵ.*
Output*: An estimate of the minimum of the function.*

1. *Choose a random initialization $x_0 \in \mathbb{R}$.*
2. *For $i = 1, 2, \ldots$ compute*

$$x_i = x_{i-1} - \frac{f'(x_{i-1})}{f''(x_{i-1})}. \tag{23}$$

   until $|f'(x_i)| \leq \epsilon$.

When the function is well approximated by a quadratic then Newton's method typically converges more rapidly than derivative descent. As expected, for a quadratic it converges in one step as shown in Figure 5. Figures 6 and 7 show the result of applying both algorithms to a convex function that is not a quadratic. Note how the step size of derivative descent depends on the slope of the function at each point. Figure 7 illustrates how Newton's method moves to the minimum of the quadratic approximation of the function in each step.

Both derivative descent and Newton's method will usually converge for *well-behaved* convex functions (this statement can be made rigorous, but proof of the convergence of these algorithms are beyond the scope of these notes). Figure 8 shows an example where they are applied to a nonconvex function that has local minima. Both algorithms converge to different local minima depending on the initialization.

3 Optimization in multiple dimensions

We now extend the ideas explored in the previous section to multiple dimensions.

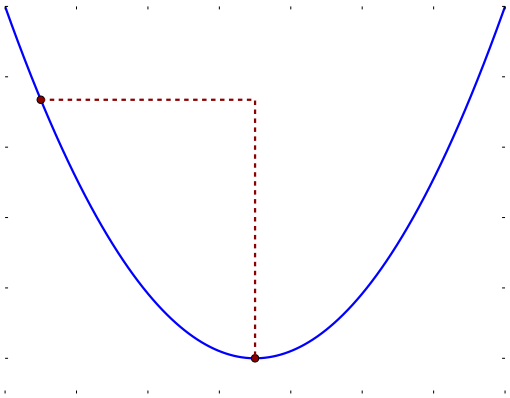

Figure 5: When applied to a quadratic function, Newton's method converges in one step.

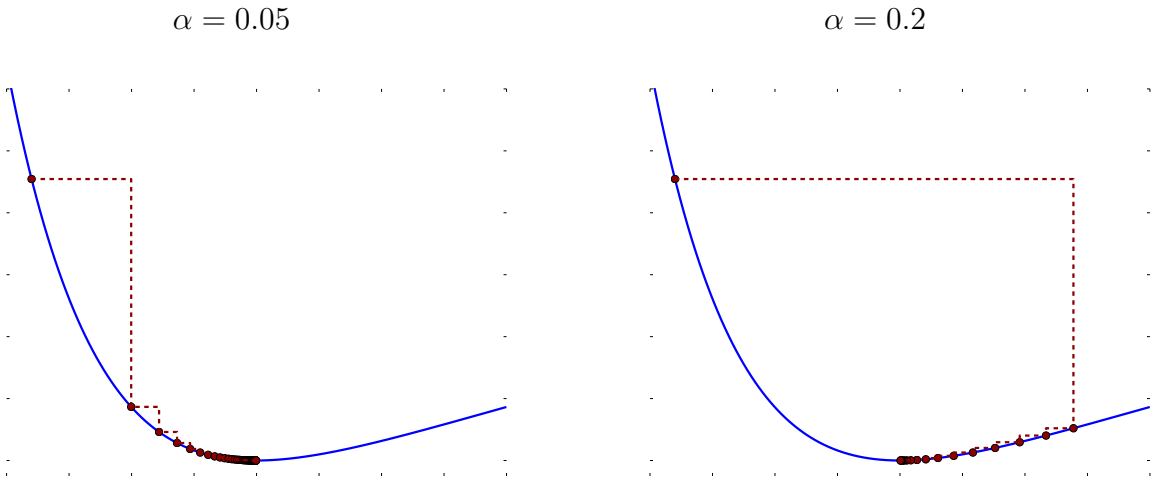

Figure 6: Derivative descent applied to a convex function for two different values of α. In both cases the iterations converge to the minimum.

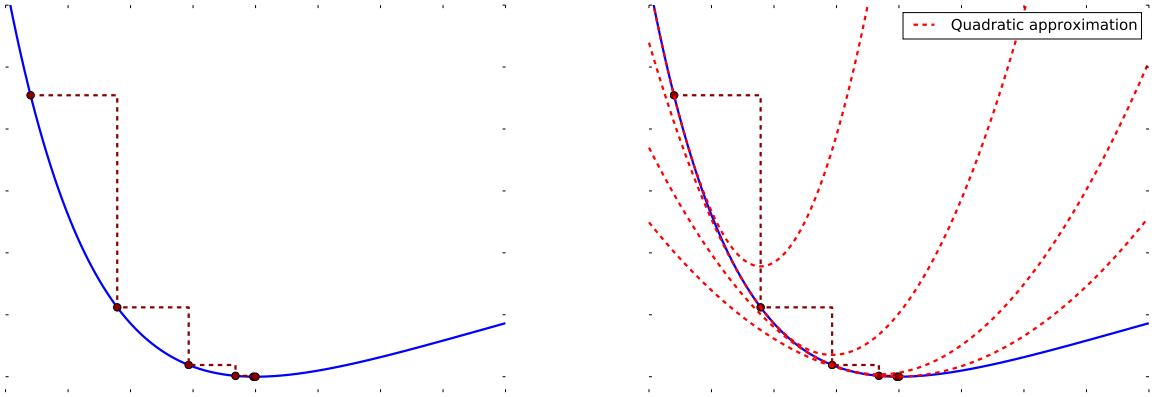

Figure 7: Newton's method applied to a convex function. The image on the right shows the quadratic approximations to the function at each iteration.

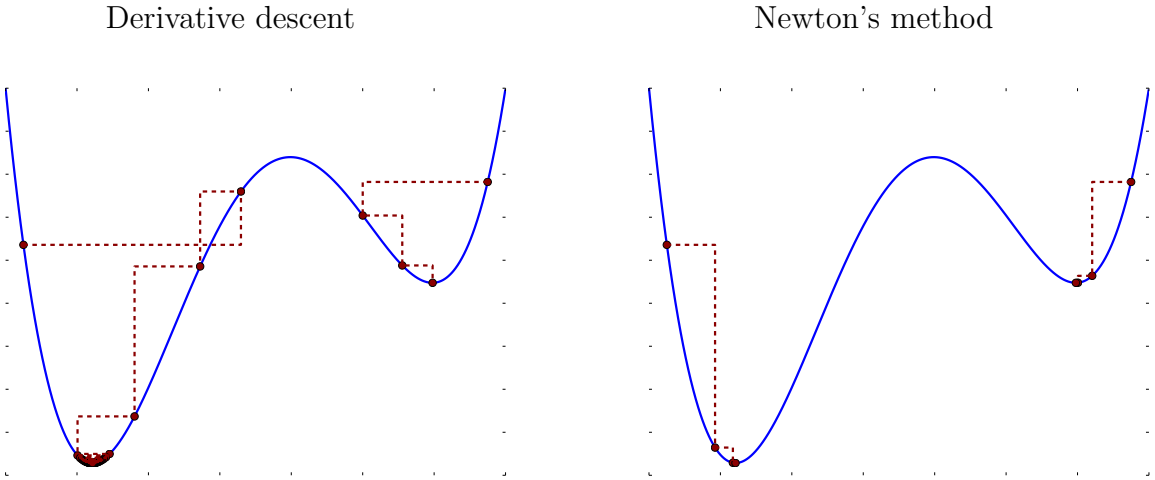

Figure 8: Derivative descent and Newton's method applied to a nonconvex function with two local minima. The algorithms converge to different minima depending on the initialization.

3.1 Gradient, Hessian and convexity

As we saw in the previous section, the derivative quantifies the variation of functions defined on $\mathbb{R}$. However, when the domain of a function is $\mathbb{R}^n$ instead, the function may vary differently in different directions. The **directional derivative** of the function quantifies the variation in a fixed direction.

Definition 3.1 (Directional derivative)**.** *The directional derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$ in the direction of a unit-norm vector $\boldsymbol{u}$ is*

$$f'_{\boldsymbol{u}}(\boldsymbol{x}) := \lim_{h \to 0} \frac{f(\boldsymbol{x} + h\boldsymbol{u}) - f(\boldsymbol{x})}{h}. \tag{24}$$

Higher-order directional derivatives in the same direction are computed recursively

$$f''_{\boldsymbol{u}}(\boldsymbol{x}) := \lim_{h \to 0} \frac{f'(\boldsymbol{x} + h\boldsymbol{u}) - f'(\boldsymbol{x})}{h} \tag{25}$$

The **partial derivatives** of a function are its directional derivatives in the direction of the axes.

Definition 3.2 (Partial derivative)**.** *The ith directional derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$ is*

$$\begin{aligned} \frac{\partial f(\boldsymbol{x})}{\partial x_i} &:= \lim_{h \to 0} \frac{f\left(\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_i + h \\ \cdots \\ x_n \end{bmatrix}\right) - f\left(\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_i \\ \cdots \\ x_n \end{bmatrix}\right)}{h} && (26) \\ &= f'_{\boldsymbol{e}_i}(\boldsymbol{x}). && (27) \end{aligned}$$

where $\boldsymbol{e}_i$ is the ith standard basis vector. Higher-order directional derivatives are computed recursively

$$\frac{\partial^2 f(\boldsymbol{x})}{\partial x_j \partial x_i} = \frac{\partial}{\partial x_j} \frac{\partial f(\boldsymbol{x})}{\partial x_i} \qquad \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i^2} = \frac{\partial}{\partial x_i} \frac{\partial f(\boldsymbol{x})}{\partial x_i} \tag{28}$$

The **gradient** is a vector that contains the partial derivatives of the function at a certain point.

Definition 3.3 (Gradient)**.** *The gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ is*

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \cdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}. \tag{29}$$

If the gradient exists at every point, the function is said to be differentiable.

More intuitively, the gradient encodes the variation of the function *in every direction*.

Lemma 3.4. *If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable,*

$$f'_{\boldsymbol{u}}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})^T \boldsymbol{u} \tag{30}$$

for any unit-norm vector $\boldsymbol{u} \in \mathbb{R}^n$.

We omit the proof of the lemma which uses basic multivariable calculus. An important corollary is that the gradient provides the direction of maximum positive and negative variation of the function.

Corollary 3.5. *The direction of the gradient ∇f of a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is the direction of maximum increase of the function. The opposite direction is the direction of maximum decrease.*

Proof. By the Cauchy-Schwarz inequality

$$\begin{aligned} |f'_{\boldsymbol{u}}(\boldsymbol{x})| &= \left| \nabla f(\boldsymbol{x})^T \boldsymbol{u} \right| && (31) \\ &\leq ||\nabla f(\boldsymbol{x})||_2 \, ||\boldsymbol{u}||_2 && (32) \\ &= ||\nabla f(\boldsymbol{x})||_2 && (33) \end{aligned}$$

with equality if and only if $\boldsymbol{u} = \pm \frac{\nabla f(\boldsymbol{x})}{||\nabla f(\boldsymbol{x})||_2}$. □

Figure 9 shows the gradient of a function $f : \mathbb{R}^2 \to \mathbb{R}$ at different locations. The gradient is orthogonal to the **contour lines** of the function. Indeed, since the function does not change value in those direction, the directional derivative is zero.

In one dimension, the linear approximation can be thought of as a line that is tangent to the curve $(x, f(x))$. In multiple dimensions, the linear approximation to the function at a point is a hyperplane that is tangent to the hypersurface $(\boldsymbol{x}, f(\boldsymbol{x}))$ at that point.

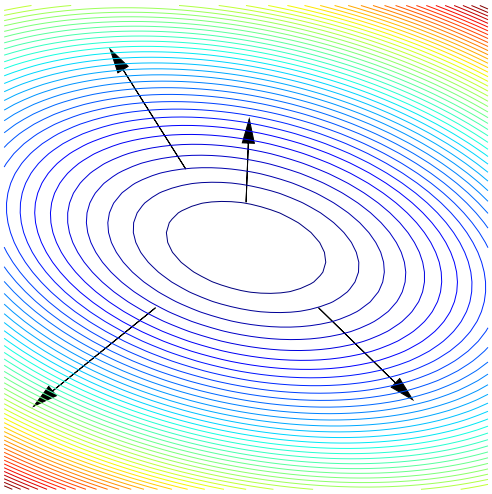

Figure 9: Contour lines of a function $f : \mathbb{R}^2 \to \mathbb{R}$. The gradients at different points are represented by black arrows, which are orthogonal to the contour lines.

Definition 3.6 (First-order approximation)**.** *The first-order or linear approximation of a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x}$ is*

$$f_{\boldsymbol{x}}^1(\boldsymbol{y}) := f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \tag{34}$$

As in the 1D case, this is an arbitrarily good approximation to the function as we approach $\boldsymbol{x}$ even if we divide the error by the distance to $\boldsymbol{x}$.

Lemma 3.7. *For any continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$*

$$\lim_{\boldsymbol{y} \to \boldsymbol{x}} \frac{f(\boldsymbol{y}) - f_{\boldsymbol{x}}^1(\boldsymbol{y})}{||\boldsymbol{y} - \boldsymbol{x}||_2} = 0 \tag{35}$$

We now extend the definition of convexity to multiple dimensions.

Definition 3.8 (Convexity)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and any $\theta \in (0, 1)$,*

$$\theta f(\boldsymbol{x}) + (1 - \theta) f(\boldsymbol{y}) \geq f(\theta \boldsymbol{x} + (1 - \theta) \boldsymbol{y}). \tag{36}$$

The function is strictly convex if the inequality is strict,

$$\theta f(\boldsymbol{x}) + (1 - \theta) f(\boldsymbol{y}) > f(\theta \boldsymbol{x} + (1 - \theta) \boldsymbol{y}). \tag{37}$$

The condition for convexity in higher dimensions is remarkably similar to the definition in 1D: the hypersurface $(\boldsymbol{z}, f(\boldsymbol{z}))$ must stay below any chord between two points $(\boldsymbol{x}, f(\boldsymbol{x}))$ and $(\boldsymbol{y}, f(\boldsymbol{y}))$. For differentiable functions, as in one dimension, an equivalent condition is for the linear approximation to remain below the hypersurface.

Lemma 3.9. *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}). \tag{38}$$

It is strictly convex if and only if

$$f(\boldsymbol{y}) > f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}). \tag{39}$$

We omit the proof, which is similar to that of the 1D case. An immediate corollary is that for a convex function, any point at which the gradient is zero is a **global minimum** and if the function is strictly convex, the point is the only minimum.

Corollary 3.10. *If a differentiable function f is convex and $\nabla f(\boldsymbol{x}) = 0$, then for any $\boldsymbol{y} \in \mathbb{R}$*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}). \tag{40}$$

If f is strictly convex then for any $\boldsymbol{y} \in \mathbb{R}$

$$f(\boldsymbol{y}) > f(\boldsymbol{x}). \tag{41}$$

The **Hessian matrix** of a function contains its second-order partial derivatives.

Definition 3.11 (Hessian matrix)**.** *The Hessian matrix of a function $f : \mathbb{R}^n \to \mathbb{R}$ is*

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ & & \cdots & \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}. \tag{42}$$

If a function has a Hessian matrix, we say that the function is twice differentiable.

The Hessian matrix encodes the curvature of the function in every direction. It follows from the definition (and some basic multivariable calculus) that

$$f''_{\boldsymbol{u}}(\boldsymbol{x}) = \boldsymbol{u}^T \nabla^2 f(\boldsymbol{x}) \boldsymbol{u}, \tag{43}$$

i.e. the second directional derivative of the function can be computed from the quadratic form induced by the Hessian matrix. The following lemma shows how to obtain the local directions of maximum and minimum curvature of the function.

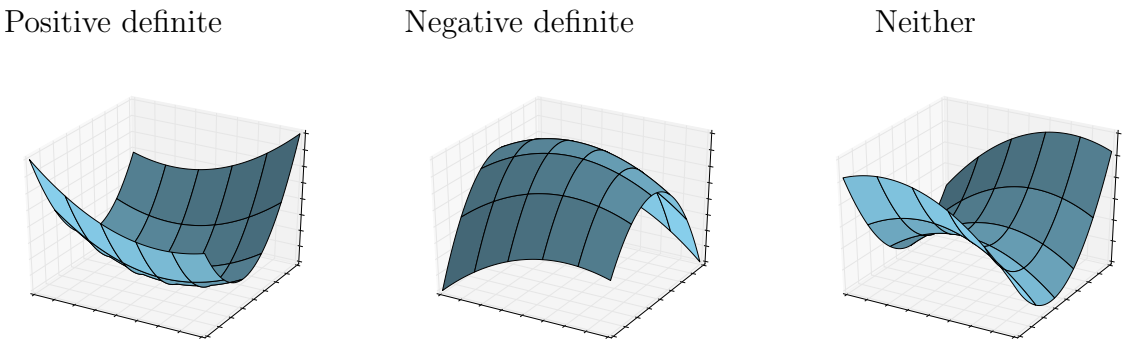

Figure 10: Quadratic forms induced by a positive definite matrix (left), a negative definite matrix (center) and a matrix that is neither positive nor negative definite (right).

Lemma 3.12. *Consider the eigendecomposition of the Hessian matrix of a twice-differentiable function f at a point $\boldsymbol{x}$. The maximum curvature of f at $\boldsymbol{x}$ is given by the largest eigenvalue of $\nabla^2 f(\boldsymbol{x})$ and is in the direction of the corresponding eigenvector. The smallest curvature, or the largest negative curvature, of f at $\boldsymbol{x}$ is given by the smallest eigenvalue of $\nabla^2 f(\boldsymbol{x})$ and is in the direction of the corresponding eigenvector.*

Proof. $\nabla^2 f(\boldsymbol{x})$ is symmetric, so the lemma follows from Theorem 6.2 in Lecture Notes 9 and (43). □

Example 3.13 (Quadratic forms)**.** If all the eigenvalues of a symmetric matrix $A \in \mathbb{R}^{n \times n}$ are nonnegative, the matrix is said to be **positive semidefinite**. If the eigenvalues are positive the matrix is **positive definite**. If the eigenvalues are all nonpositive, the matrix is negative semidefinite. If they are negative, the matrix is negative definite.

The quadratic forms

$$q(\boldsymbol{x}) := \boldsymbol{x}^T A \boldsymbol{x} + \boldsymbol{b}^T x + c \tag{44}$$

induced by A has nonnegative curvature in every direction if the matrix is positive semidefinite by Lemma 3.12 because A is the Hessian matrix of the quadratic form. Similarly, the curvature is always positive if A is positive definite and negative if A is negative definite. Figure 10 illustrates this with quadratic forms in two dimensions.

As in 1D, if a function has nonnegative curvature in every direction then it is convex.

Lemma 3.14. *A twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for every $\boldsymbol{x} \in \mathbb{R}^n$, the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ is positive semidefinite. The function is strictly convex if and only if for every $\boldsymbol{x} \in \mathbb{R}^n$, $\nabla^2 f(\boldsymbol{x})$ is positive definite.*

The proof of the lemma, which we omit, is similar to the one of the 1D case.

As in 1D we can define a quadratic approximation of a twice-differentiable function at each point.

Definition 3.15 (Second-order approximation)**.** *The second-order or quadratic approximation of f at $\boldsymbol{x}$ is*

$$f_{\boldsymbol{x}}^2(\boldsymbol{y}) := f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \tag{45}$$

The quadratic form $f_{\boldsymbol{x}}^2(\boldsymbol{y})$ becomes an arbitrarily good approximation of f as we approach x, even if we divide the error by the squared distance between x and y. We omit the proof that follows from multivariable calculus.

Lemma 3.16. *The quadratic approximation $f_x^2 : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}$ of a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies*

$$\lim_{\boldsymbol{y} \to \boldsymbol{x}} \frac{f(\boldsymbol{y}) - f_{\boldsymbol{x}}^2(\boldsymbol{y})}{||\boldsymbol{y} - \boldsymbol{x}||_2^2} = 0 \tag{46}$$

3.2 Optimization algorithms

Gradient descent is arguably the most important optimization algorithm. The idea is simple, and very similar to the 1D version that we dubbed derivative descent. By Corollary 3.5 the direction of $-\nabla f$ is the direction of steepest descent (gradient descent is also called steepest descent in the literature).

Algorithm 3.17 (Gradient descent)**.**
Input*: A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, its gradient ∇f, a step size α and a stopping threshold ϵ.*
Output*: An estimate of the minimum of the function.*

1. *Choose a random initialization $\boldsymbol{x_0} \in \mathbb{R}^n$.*
2. *For $i = 1, 2, \ldots$ compute*

$$\boldsymbol{x_i} = \boldsymbol{x_{i-1}} - \alpha \nabla f(\boldsymbol{x_{i-1}}). \tag{47}$$

until $||\nabla f(\boldsymbol{x_i})||_2 \leq \epsilon$.

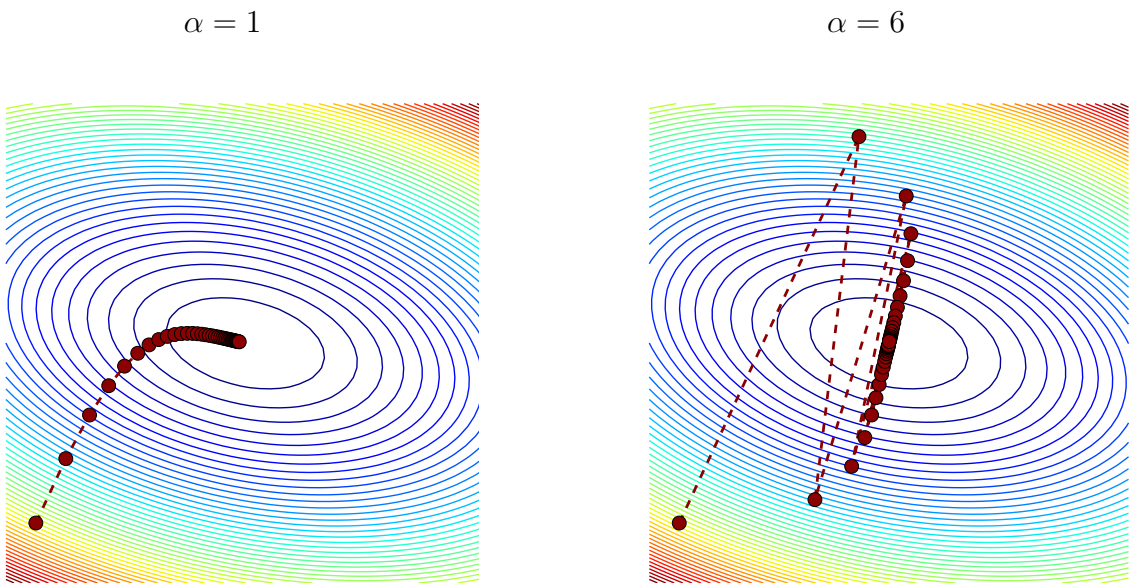

Figure 11: Gradient descent applied to a quadratic function for two different values of α. In both cases the iterations converge to the minimum.

Figure 11 shows the results of applying the method on a quadratic function. As in the 1D case, choosing the step size implies trying to reach a tradeoff between making slow progress or repeatedly overshooting the minimum.

Newton's method consists of modifying the step in gradient descent using curvature information. The global minimum $\boldsymbol{x}^*$of a convex quadratic form

$$q(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c \tag{48}$$

is $-A^{-1}\boldsymbol{b}$. As a result the minimum $\boldsymbol{x}^*$ can be found by computing

$$\boldsymbol{x}^* = \boldsymbol{x} - \nabla^2 f(\boldsymbol{x_{i-1}})^{-1} \nabla f(\boldsymbol{x_{i-1}}) = \boldsymbol{x} - A^{-1}(A\boldsymbol{x} + \boldsymbol{b}). \tag{49}$$

Note that multiplying by the inverse of the Hessian not only changes the magnitude of the step but also its direction.

Newton's method iteratively minimizes the quadratic approximation of a function.

Algorithm 3.18 (Newton's method)**.**
Input*: A twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, a way to compute its gradient ∇f and Hessian matrix $\nabla^2 f$, a step size α and a stopping threshold ϵ.*
Output*: An estimate of the minimum of the function.*

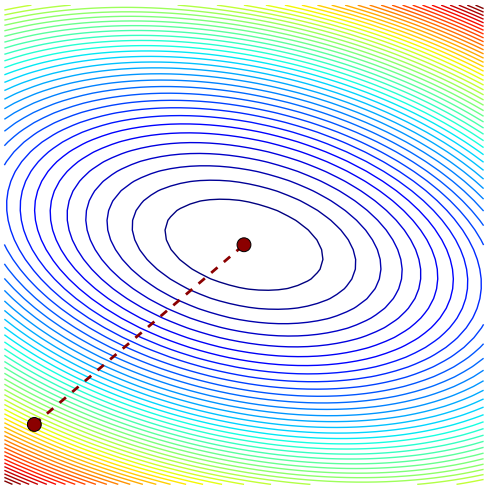

Figure 12: When applied to a quadratic function, Newton's method converges in one step.

1. *Choose a random initialization* $x_0 \in \mathbb{R}$.
2. *For* $i = 1, 2, \ldots$ *compute*

$$\boldsymbol{x_i} = \boldsymbol{x_{i-1}} - \nabla^2 f\left(\boldsymbol{x_{i-1}}\right)^{-1} \nabla f\left(\boldsymbol{x_{i-1}}\right) \tag{50}$$

until $||\nabla f\left(\boldsymbol{x_i}\right)||_2 \leq \epsilon$.

As expected, when applied to a quadratic function, Newton's method converges in one step. This is shown in Figure 12. Figure 13 shows the result of applying gradient descent and Newton's method to a convex function $f : \mathbb{R}^2 \to \mathbb{R}$.

3.3 Stochastic optimization

Let us assume that our cost function of interest can be decomposed into the sum of m cost functions

$$f\left(\boldsymbol{x}\right) = \frac{1}{m} \sum_{i=1}^{m} f_i\left(\boldsymbol{x}\right). \tag{51}$$

An important example is the squared ℓ_2 error achieved by a model on a training set. In this case, each f_i is the error for a single example in the training set. The aim of stochastic-optimization algorithms is to optimize the whole cost function as the different f_i are revealed.

Gradient descent

Newton's method

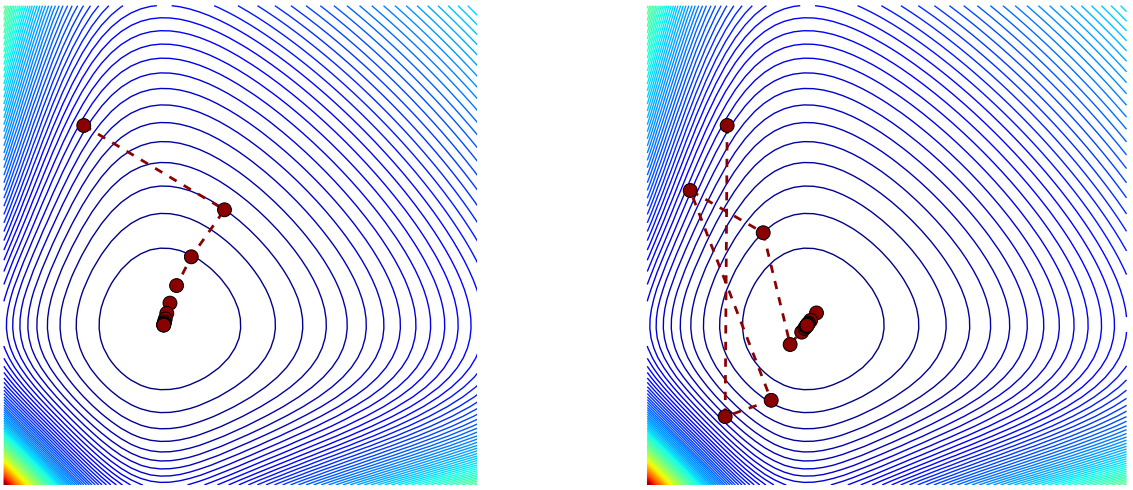

Figure 13: Gradient descent and Newton's method applied to a convex function. Both algorithms converge to the minimum.

Intuitively, this allows to refine the model parameters as more data become available. By far, the most popular algorithm for stochastic optimization is stochastic gradient descent, which consists of taking a gradient-descent step for each piece of data that becomes available.

Algorithm 3.19 (Stochastic gradient descent)**.**
Input*: A differentiable function* $f : \mathbb{R}^n \to \mathbb{R}$*, its gradient* ∇f*, a step size* α *and a stopping threshold* ϵ*.*
Output*: An estimate of the minimum of the function.*

1. *Choose a random initialization* $\boldsymbol{x_0} \in \mathbb{R}^n$*.*
2. *For* $i = 1, 2, \ldots, m$ *compute*

$$\boldsymbol{x_i} = \boldsymbol{x_{i-1}} - \alpha \nabla f_i \left(\boldsymbol{x_{i-1}} \right). \tag{52}$$

until $||\nabla f \left(\boldsymbol{x_i} \right)||_2 \leq \epsilon$*.*

A Proofs

A.1 Proof of Lemma 2.6

First we show that convexity implies that (9) holds. If f is convex then for any $x, y \in \mathbb{R}$ and any $0 \leq \theta \leq 1$

$$\theta\left(f\left(y\right)-f\left(x\right)\right)+f\left(x\right) \geq f\left(x+\theta\left(y-x\right)\right). \tag{53}$$

Rearranging the terms we have

$$f\left(y\right) \geq \frac{f\left(x+\theta\left(y-x\right)\right)-f\left(x\right)}{\theta}+f\left(x\right). \tag{54}$$

Setting $h=\theta\left(y-x\right)$, this implies

$$f\left(y\right) \geq \frac{f\left(x+h\right)-f\left(x\right)}{h}\left(y-x\right)+f\left(x\right). \tag{55}$$

Taking the limit when $h \to 0$ yields

$$f\left(y\right) \geq f'\left(x\right)\left(y-x\right). \tag{56}$$

To complete the proof, we show that if (9) holds, then the function is convex. Indeed, let $z=\theta x+\left(1-\theta\right)y$, then by (9)

$$\begin{aligned} f\left(x\right) &\geq f'\left(z\right)\left(x-z\right)+f\left(z\right) && (57)\\ &= f'\left(z\right)\left(1-\theta\right)\left(x-y\right)+f\left(z\right) && (58)\\ f\left(y\right) &\geq f'\left(z\right)\left(y-z\right)+f\left(z\right) && (59)\\ &= f'\left(z\right)\theta\left(y-x\right)+f\left(z\right) && (60) \end{aligned}$$

Multiplying (58) by θ, then (60) by $1-\theta$ and summing the inequalities, we obtain

$$\theta f\left(x\right)+\left(1-\theta\right)f\left(y\right) \geq f\left(\theta x+\left(1-\theta\right)y\right). \tag{61}$$

A.2 Proof of Lemma 2.8

The second derivative is nonnegative anywhere if and only if the first derivative is nondecreasing, because f'' is the derivative of f'.

First we prove that if the function is convex then the derivative is nondecreasing. By Lemma 2.6, if the function is convex then for any $x, y \in \mathbb{R}$ such that $y > x$

$$\begin{aligned} f\left(x\right) &\geq f'\left(y\right)\left(x-y\right)+f\left(y\right), && (62)\\ f\left(y\right) &\geq f'\left(x\right)\left(y-x\right)+f\left(x\right). && (63) \end{aligned}$$

Rearranging, we obtain

$$f'(y)(y-x) \geq f(y) - f(x) \geq f'(x)(y-x). \tag{64}$$

Since $y - x > 0$, we have $f'(y) \geq f'(x)$.

To complete the proof we show that if the derivative is nondecreasing, then the function is convex. For this we will use a basic result from real analysis.

Theorem A.1 (Mean-value theorem)**.** *If a function f is differentiable in the interval $[x, y]$ then there exists a point $y \leq \gamma y$ such that*

$$f'(\gamma) = \frac{f(y) - f(x)}{y - x}. \tag{65}$$

For arbitrary $x, y, \theta \in \mathbb{R}$, such that $y > x$ and $0 < \theta < 1$, let $z = \theta y + (1-\theta)x$. Since $y > z > x$, there exist $\gamma_1 \in [x, z]$ and $\gamma_2 \in [z, y]$ such that

$$f'(\gamma_1) = \frac{f(z) - f(x)}{z - x}, \tag{66}$$

$$f'(\gamma_2) = \frac{f(y) - f(z)}{y - z}. \tag{67}$$

Since $\gamma_1 < \gamma_2$, if f' is nondecreasing

$$\frac{f(y) - f(z)}{y - z} \geq \frac{f(z) - f(x)}{z - x}, \tag{68}$$

which implies

$$\frac{z-x}{y-x} f(y) + \frac{y-z}{y-x} f(x) \geq f(z). \tag{69}$$

Recall that $z = \theta y + (1-\theta)x$, so that $\theta = (z-x)/(y-x)$ and $\theta = (z-x)/(y-x)$ and $1 - \theta = (y-z)/(y-x)$. (69) is consequently equivalent to

$$\theta f(y) + (1-\theta) f(x) \geq f(\theta y + (1-\theta)x). \tag{70}$$