AM 205: lecture 18

- Last time: optimization methods
- Today: conditions for optimality

Existence of Global Minimum

For example:

- $f(x, y) = x^2 + y^2$ is coercive on $\mathbb{R}^2$ (global min. at $(0, 0)$)
- $f(x) = x^3$ is not coercive on $\mathbb{R}$ ($f \to -\infty$ for $x \to -\infty$)
- $f(x) = e^x$ is not coercive on $\mathbb{R}$ ($f \to 0$ for $x \to -\infty$)

Convexity

An important concept for uniqueness is convexity

A set $S \subset \mathbb{R}^n$ is convex if it contains the line segment between any two of its points

That is, S is convex if for any $x, y \in S$, we have

$$\{\theta x + (1-\theta)y : \theta \in [0,1]\} \subset S$$

Convexity

Similarly, we define convexity of a function $f : S \subset \mathbb{R}^n \to \mathbb{R}$

f is convex if its graph along any line segment in S is on or below the chord connecting the function values

i.e. f is convex if for any $x, y \in S$ and any $\theta \in (0, 1)$, we have

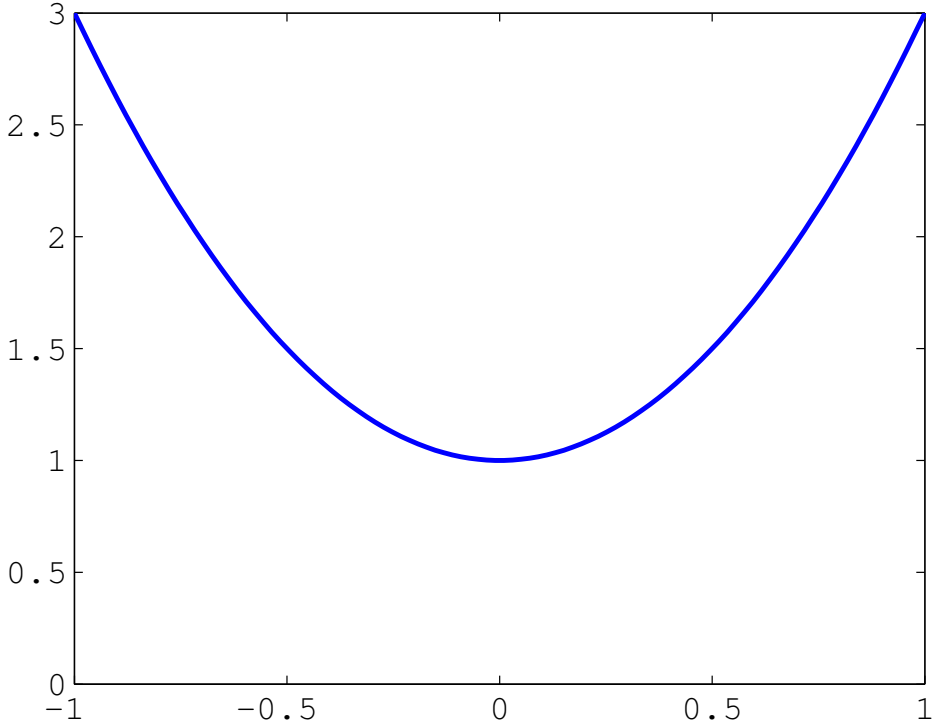
$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$$

Also, if

$$f(\theta x + (1 - \theta)y) < \theta f(x) + (1 - \theta)f(y)$$

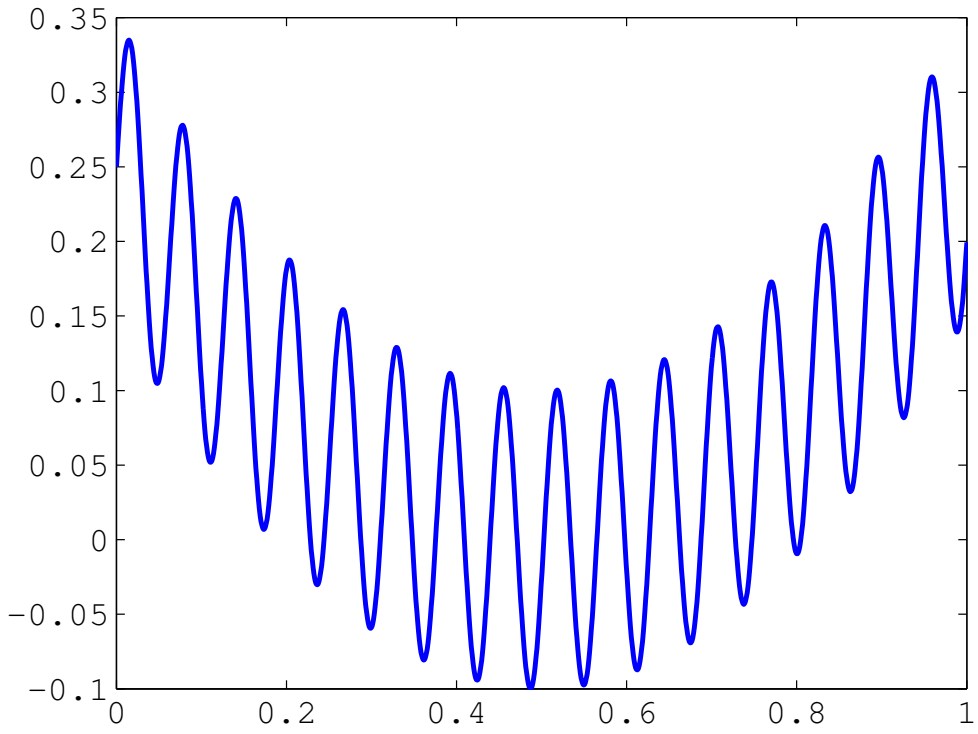
then f is strictly convex

Convexity



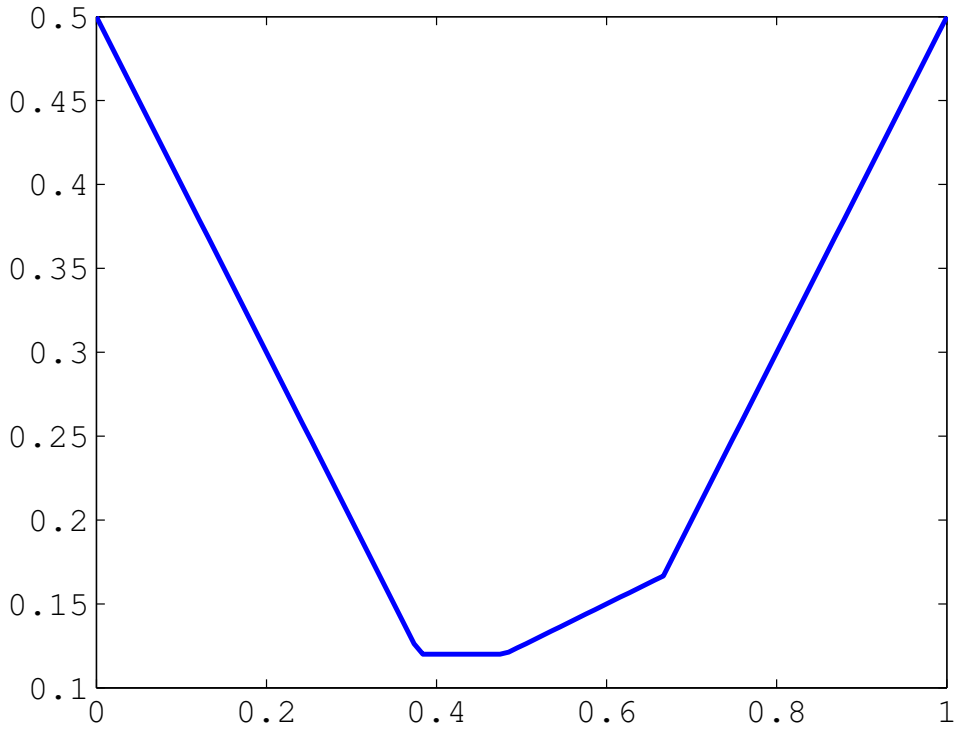
Strictly convex

Convexity



Non-convex

Convexity



Convex (not strictly convex)

Convexity

If f is a convex function on a convex set S, then any local minimum of f must be a global minimum[1]

Proof: Suppose x is a local minimum, *i.e.* $f(x) \leq f(y)$ for $y \in B(x, \epsilon)$ (where $B(x, \epsilon) \equiv \{y \in S : \|y - x\| \leq \epsilon\}$)

Suppose that x is not a global minimum, *i.e.* that there exists $w \in S$ such that $f(w) < f(x)$

(Then we will show that this gives a contradiction)

[1] A global minimum is defined as a point z such that $f(z) \leq f(x)$ for all $x \in S$. Note that a global minimum may not be unique, *e.g.* if $f(x) = -\cos x$ then 0 and 2π are both global minima.

Convexity

Proof (continued...):

For $\theta \in [0,1]$ we have $f(\theta w + (1-\theta)x) \leq \theta f(w) + (1-\theta)f(x)$

Let $\sigma \in (0,1]$ be sufficiently small so that

$$z \equiv \sigma w + (1-\sigma)\, x \in B(x, \epsilon)$$

Then

$$f(z) \leq \sigma f(w) + (1-\sigma)\, f(x) < \sigma f(x) + (1-\sigma)\, f(x) = f(x),$$

i.e. $f(z) < f(x)$, which contradicts that $f(x)$ is a local minimum!

Hence we cannot have $w \in S$ such that $f(w) < f(x)$ $\quad \square$

Convexity

Note that convexity does not guarantee uniqueness of global minimum

e.g. a convex function can clearly have a "horizontal" section (see earlier plot)

If f is a strictly convex function on a convex set S, then a local minimum of f is the unique global minimum

Optimization of convex functions over convex sets is called convex optimization, which is an important subfield of optimization

Optimality Conditions

We have discussed existence and uniqueness of minima, but haven't considered how to find a minimum

The familiar optimization idea from calculus in one dimension is: set derivative to zero, check the sign of the second derivative

This can be generalized to $\mathbb{R}^n$

Optimality Conditions

If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then the gradient vector $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is

$$\nabla f(x) \equiv \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

The importance of the gradient is that ∇f points "uphill," *i.e.* towards points with larger values than $f(x)$

And similarly $-\nabla f$ points "downhill"

Optimality Conditions

This follows from Taylor's theorem for $f : \mathbb{R}^n \to \mathbb{R}$

Recall that

$$f(x + \delta) = f(x) + \nabla f(x)^T \delta + \text{H.O.T.}$$

Let $\delta \equiv -\epsilon \nabla f(x)$ for $\epsilon > 0$ and suppose that $\nabla f(x) \neq 0$, then:

$$f(x - \epsilon \nabla f(x)) \approx f(x) - \epsilon \nabla f(x)^T \nabla f(x) < f(x)$$

Also, we see from Cauchy–Schwarz that $-\nabla f(x)$ is the steepest descent direction

Optimality Conditions

Similarly, we see that a necessary condition for a local minimum at $x^* \in S$ is that $\nabla f(x^*) = 0$

In this case there is no "downhill direction" at x^*

The condition $\nabla f(x^*) = 0$ is called a first-order necessary condition for optimality, since it only involves first derivatives

Optimality Conditions

$x^* \in S$ that satisfies the first-order optimality condition is called a critical point of f

But of course a critical point can be a local min., local max., or saddle point

(Recall that a saddle point is where some directions are "downhill" and others are "uphill", *e.g.* $(x, y) = (0, 0)$ for $f(x, y) = x^2 - y^2$)

Optimality Conditions

As in the one-dimensional case, we can look to second derivatives to classify critical points

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, then the Hessian is the matrix-valued function $H_f : \mathbb{R}^n \to \mathbb{R}^{n \times n}$

$$H_f(x) \equiv \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n x_1} & \frac{\partial^2 f(x)}{\partial x_n x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

The Hessian is the Jacobian matrix of the gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$

If the second partial derivatives of f are continuous, then $\partial^2 f / \partial x_i \partial x_j = \partial^2 f / \partial x_j \partial x_i$, and H_f is symmetric

Optimality Conditions

Suppose we have found a critical point x^*, so that $\nabla f(x^*) = 0$

From Taylor's Theorem, for $\delta \in \mathbb{R}^n$, we have

$$\begin{aligned} f(x^* + \delta) &= f(x^*) + \nabla f(x^*)^T \delta + \frac{1}{2}\delta^T H_f(x^* + \eta\delta)\delta \\ &= f(x^*) + \frac{1}{2}\delta^T H_f(x^* + \eta\delta)\delta \end{aligned}$$

for some $\eta \in (0, 1)$

Optimality Conditions

Recall positive definiteness: A is positive definite if $x^T Ax > 0$

Suppose $H_f(x^*)$ is positive definite

Then (by continuity) $H_f(x^* + \eta\delta)$ is also positive definite for $\|\delta\|$ sufficiently small, so that: $\delta^T H_f(x^* + \eta\delta)\delta > 0$

Hence, we have $f(x^* + \delta) > f(x^*)$ for $\|\delta\|$ sufficiently small, *i.e.* $f(x^*)$ is a local minimum

Hence, in general, positive definiteness of H_f at a critical point x^* is a second-order sufficient condition for a local minimum

Optimality Conditions

A matrix can also be negative definite: $x^T Ax < 0$ for all $x \neq 0$

Or indefinite: There exists x, y such that $x^T Ax < 0 < y^T Ay$

Then we can classify critical points as follows:

- $H_f(x^*)$ positive definite $\implies$ x^* is a local minimum
- $H_f(x^*)$ negative definite $\implies$ x^* is a local maximum
- $H_f(x^*)$ indefinite $\implies$ x^* is a saddle point

Optimality Conditions

Also, positive definiteness of the Hessian is closely related to convexity of f

If $H_f(x)$ is positive definite, then f is convex on some convex neighborhood of x

If $H_f(x)$ is positive definite for all $x \in S$, where S is a convex set, then f is convex on S

Question: How do we test for positive definiteness?

Optimality Conditions

Answer: A is positive (resp. negative) definite if and only if all eigenvalues of A are positive (resp. negative)[2]

Also, a matrix with positive and negative eigenvalues is indefinite

Hence we can compute all the eigenvalues of A and check their signs

[2] This is related to the Rayleigh quotient, see Unit V

Heath Example 6.5

Consider

$$f(x) = 2x_1^3 + 3x_1^2 + 12x_1x_2 + 3x_2^2 - 6x_2 + 6$$

Then

$$\nabla f(x) = \begin{bmatrix} 6x_1^2 + 6x_1 + 12x_2 \\ 12x_1 + 6x_2 - 6 \end{bmatrix}$$

We set $\nabla f(x) = 0$ to find critical points[3] $[1, -1]^T$ and $[2, -3]^T$

[3] In general solving $\nabla f(x) = 0$ requires an iterative method

Heath Example 6.5, continued...

The Hessian is

$$H_f(x) = \begin{bmatrix} 12x_1 + 6 & 12 \\ 12 & 6 \end{bmatrix}$$

and hence

$$H_f(1,-1) = \begin{bmatrix} 18 & 12 \\ 12 & 6 \end{bmatrix}, \text{ which has eigenvalues } 25.4, -1.4$$

$$H_f(2,-3) = \begin{bmatrix} 30 & 12 \\ 12 & 6 \end{bmatrix}, \text{ which has eigenvalues } 35.0, 1.0$$

Hence $[2,-3]^T$ is a local min. whereas $[1,-1]^T$ is a saddle point

Optimality Conditions: Equality Constrained Case

So far we have ignored constraints

Let us now consider equality constrained optimization

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad g(x) = 0,$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$, with $m \leq n$

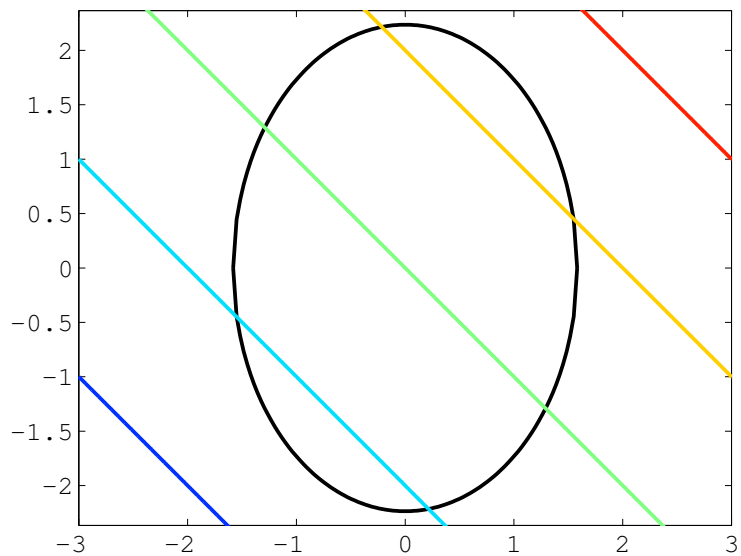
Since g maps to $\mathbb{R}^m$, we have m constraints

This situation is treated with Lagrange mutlipliers

Optimality Conditions: Equality Constrained Case

We illustrate the concept of Lagrange multipliers for $f, g : \mathbb{R}^2 \to \mathbb{R}$

Let $f(x, y) = x + y$ and $g(x, y) = 2x^2 + y^2 - 5$



∇g is normal to S:[4] at any $x \in S$ we must move in direction $(\nabla g(x))_\perp$ (tangent direction) to remain in S

[4] This follows from Taylor's Theorem: $g(x + \delta) \approx g(x) + \nabla g(x)^T \delta$

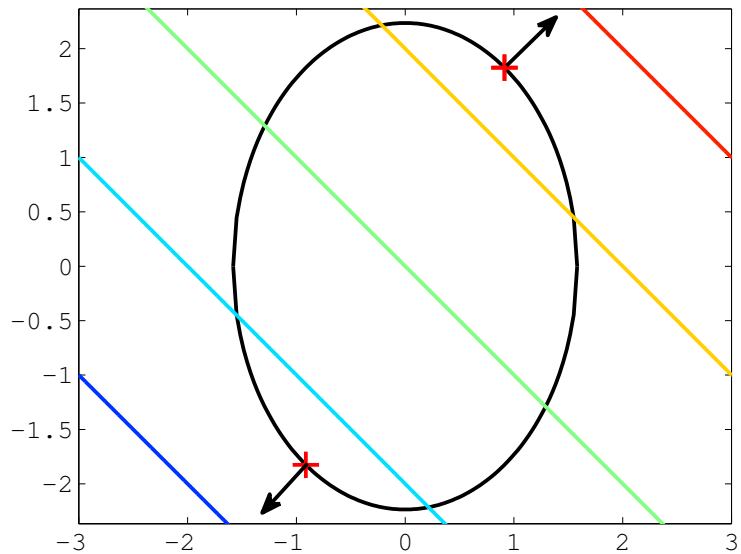
Optimality Conditions: Equality Constrained Case

Also, change in f due to infinitesimal step in direction $(\nabla g(x))_\perp$ is

$$f(x \pm \epsilon(\nabla g(x))_\perp) = f(x) \pm \epsilon \nabla f(x)^T (\nabla g(x))_\perp + \text{H.O.T.}$$

Hence stationary point $x^* \in S$ if $\nabla f(x^*)^T (\nabla g(x^*))_\perp = 0$, or

$$\nabla f(x^*) = \lambda^* \nabla g(x^*), \quad \text{for some } \lambda^* \in \mathbb{R}$$



Optimality Conditions: Equality Constrained Case

This shows that for a stationary point with $m = 1$ constraints, ∇f cannot have any component in the "tangent direction" to S

Now, consider the case with $m > 1$ equality constraints

Then $g : \mathbb{R}^n \to \mathbb{R}^m$ and we now have a set of constraint gradient vectors, ∇g_i, $i = 1, \ldots, m$

Then we have $S = \{x \in \mathbb{R}^n : g_i(x) = 0, i = 1, \ldots, m\}$

Any "tangent direction" at $x \in S$ must be orthogonal to all gradient vectors $\{\nabla g_i(x), i = 1, \ldots, m\}$ to remain in S

Optimality Conditions: Equality Constrained Case

Let $\mathcal{T}(x) \equiv \{v \in \mathbb{R}^n : \nabla g_i(x)^T v = 0, i = 1, 2, \ldots, m\}$ denote the orthogonal complement of $\{\nabla g_i(x), i = 1, \ldots, m\}$

Then, for $\delta \in \mathcal{T}(x)$ and $\epsilon \in \mathbb{R}_{>0}$, $\epsilon\delta$ is a step in a "tangent direction" of S at x

Since we have

$$f(x^* + \epsilon\delta) = f(x^*) + \epsilon \nabla f(x^*)^T \delta + \text{H.O.T.}$$

it follows that for a stationary point we need $\nabla f(x^*)^T \delta = 0$ for all $\delta \in \mathcal{T}(x^*)$

Optimality Conditions: Equality Constrained Case

Hence, we require that at a stationary point $x^* \in S$ we have

$$\nabla f(x^*) \in \text{span}\{\nabla g_i(x^*), i = 1, \ldots, m\}$$

This can be written succinctly as a linear system

$$\nabla f(x^*) = (J_g(x^*))^T \lambda^*$$

for some $\lambda^* \in \mathbb{R}^m$, where $(J_g(x^*))^T \in \mathbb{R}^{n \times m}$

This follows because the columns of $(J_g(x^*))^T$ are the vectors $\{\nabla g_i(x^*), i = 1, \ldots, m\}$

Optimality Conditions: Equality Constrained Case

We can write equality constrained optimization problems more succinctly by introducing the Lagrangian function, $\mathcal{L} : \mathbb{R}^{n+m} \to \mathbb{R}$,

$$\begin{aligned} \mathcal{L}(x, \lambda) &\equiv f(x) + \lambda^T g(x) \\ &= f(x) + \lambda_1 g_1(x) + \cdots + \lambda_m g_m(x) \end{aligned}$$

Then we have,

$$\frac{\partial \mathcal{L}(x,\lambda)}{\partial x_i} = \frac{\partial f(x)}{\partial x_i} + \lambda_1 \frac{\partial g_1(x)}{\partial x_i} + \cdots + \lambda_n \frac{\partial g_n(x)}{\partial x_i}, \quad i = 1, \ldots, n$$

$$\frac{\partial \mathcal{L}(x,\lambda)}{\partial \lambda_i} = g_i(x), \quad i = 1, \ldots, m$$

Optimality Conditions: Equality Constrained Case

Hence

$$\nabla \mathcal{L}(x, \lambda) = \begin{bmatrix} \nabla_x \mathcal{L}(x, \lambda) \\ \nabla_\lambda \mathcal{L}(x, \lambda) \end{bmatrix} = \begin{bmatrix} \nabla f(x) + J_g(x)^T \lambda \\ g(x) \end{bmatrix},$$

so that the first order necessary condition for optimality for the constrained problem can be written as a nonlinear system:[5]

$$\nabla \mathcal{L}(x, \lambda) = \begin{bmatrix} \nabla f(x) + J_g(x)^T \lambda \\ g(x) \end{bmatrix} = 0$$

(As before, stationary points can be classified by considering the Hessian, though we will not consider this here...)

[5] $n + m$ variables, $n + m$ equations

Optimality Conditions: Equality Constrained Case

See Lecture: Constrained optimization of cylinder surface area

Optimality Conditions: Equality Constrained Case

As another example of equality constrained optimization, recall our underdetermined linear least squares problem from I.3

$$\min_{b \in \mathbb{R}^n} f(b) \quad \text{subject to} \quad g(b) = 0,$$

where $f(b) \equiv b^T b$, $g(b) \equiv Ab - y$ and $A \in \mathbb{R}^{m \times n}$ with $m < n$

Optimality Conditions: Equality Constrained Case

Introducing Lagrange multipliers gives

$$\mathcal{L}(b, \lambda) \equiv b^T b + \lambda^T (Ab - y)$$

where $b \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m$

Hence $\nabla \mathcal{L}(b, \lambda) = 0$ implies

$$\begin{bmatrix} \nabla f(b) + J_g(b)^T \lambda \\ g(b) \end{bmatrix} = \begin{bmatrix} 2b + A^T \lambda \\ Ab - y \end{bmatrix} = 0 \in \mathbb{R}^{n+m}$$

Optimality Conditions: Equality Constrained Case

Hence, we obtain the $(n+m) \times (n+m)$ square linear system

$$\begin{bmatrix} 2\mathrm{I} & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} b \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}$$

which we can solve for $\begin{bmatrix} b \\ \lambda \end{bmatrix} \in \mathbb{R}^{n+m}$

Optimality Conditions: Equality Constrained Case

We have $b = -\frac{1}{2}A^T\lambda$ from the first "block row"

Subsituting into $Ab = y$ (the second "block row") yields $\lambda = -2(AA^T)^{-1}y$

And hence

$$b = -\frac{1}{2}A^T\lambda = A^T(AA^T)^{-1}y$$

which was the solution we introduced (but didn't derive) in $\mathrm{I}.3$

Optimality Conditions: Inequality Constrained Case

Similar Lagrange multiplier methods can be developed for the more difficult case of inequality constrained optimization

Steepest Descent

We first consider the simpler case of unconstrained optimization (as opposed to constrained optimization)

Perhaps the simplest method for unconstrained optimization is steepest descent

Key idea: The negative gradient $-\nabla f(x)$ points in the "steepest downhill" direction for f at x

Hence an iterative method for minimizing f is obtained by following $-\nabla f(x_k)$ at each step

Question: How far should we go in the direction of $-\nabla f(x_k)$?

Steepest Descent

We can try to find the best step size via a subsidiary (and easier!) optimization problem

For a direction $s \in \mathbb{R}^n$, let $\phi : \mathbb{R} \to \mathbb{R}$ be given by

$$\phi(\eta) = f(x + \eta s)$$

Then minimizing f along s corresponds to minimizing the one-dimensional function ϕ

This process of minimizing f along a line is called a line search[6]

[6] The line search can itself be performed via Newton's method, as described for $f : \mathbb{R}^n \to \mathbb{R}$ shortly, or via a built-in function

Steepest Descent

Putting these pieces together leads to the steepest descent method:

1: choose initial guess x_0
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad s_k = -\nabla f(x_k)$
4: $\quad$ choose η_k to minimize $f(x_k + \eta_k s_k)$
5: $\quad x_{k+1} = x_k + \eta_k s_k$
6: **end for**

However, steepest descent often converges very slowly

Convergence rate is linear, and scaling factor can be arbitrarily close to 1

(Steepest descent will be covered on Assignment 5)

Newton's Method

We can get faster convergence by using more information about f

Note that $\nabla f(x^*) = 0$ is a system of nonlinear equations, hence we can solve it with quadratic convergence via Newton's method[7]

The Jacobian matrix of $\nabla f(x)$ is $H_f(x)$ and hence Newton's method for unconstrained optimization is:

1: choose initial guess x_0
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: solve $H_f(x_k)s_k = -\nabla f(x_k)$
4: $x_{k+1} = x_k + s_k$
5: **end for**

[7] Note that in its simplest form this algorithm searches for stationary points, not necessarily minima

Newton's Method

We can also interpret Newton's method as seeking stationary point based on a sequence of local quadratic approximations

Recall that for small δ

$$f(x+\delta) \approx f(x) + \nabla f(x)^T \delta + \frac{1}{2}\delta^T H_f(x)\delta \equiv q(\delta)$$

where $q(\delta)$ is quadratic in δ (for a fixed x)

We find stationary point of q in the usual way:[8]

$$\nabla q(\delta) = \nabla f(x) + H_f(x)\delta = 0$$

This leads to $H_f(x)\delta = -\nabla f(x)$, as in the previous slide

[8] Recall I.4 for differentiation of $\delta^T H_f(x)\delta$