Methods of Optimization Fall 2009
Prof. S. Ta'Asan*

Chris Almost†

Contents

*email?
†cdalmost@cmu.edu

1 Review and Definitions

1.1 Convex sets

The *line* through $x_1, x_2 \in \mathbb{R}^n$ is the set of points $\{(1-\theta)x_1 + \theta x_2 \mid \theta \in \mathbb{R}\}$.

An *affine set* is a set that contains the entire line through any pair of points in it, i.e. C is affine if $(1-\theta)x_1 + \theta x_2 \in C$ for all $x_1, x_2 \in C$ and $\theta \in \mathbb{R}$. Notice that if C is an affine set and $x_0 \in C$ is chosen, then $C - x_0 = \{x - x_0 \mid x \in C\}$ is a vector space. Indeed, $(x_1 - x_0) + (x_2 - x_0) = 2(\frac{1}{2}x_1 + \frac{1}{2}x_2) - x_0 \in C - x_0$ and $\theta(x - x_0) = \theta x + (1-\theta)x_0 - x_0 \in C - x_0$. If A is a matrix then the set of solutions x to $Ax = b$ is an affine set.

A *convex set* is a set that contains the line segment joining any pair of points in it, i.e. C is convex if $(1-\theta)x_1 + \theta x_2 \in C$ for all $x_1, x_2 \in C$ and $\theta \in [0,1]$. A *cone* is a set that contains the ray emanating from any point in it, i.e. K is a cone if $\theta x \in K$ for all $x \in K$. A *hyperplane* is $\{x \mid a^T x - b = 0\}$ and a *half space* is $\{a^T x - b \leq 0\}$, where a and b are vectors. The *norm cone* is $\{(x,t) \mid \|x\| \leq t\} \subseteq \mathbb{R}^{n+1}$, where $\|\cdot\|$ is any norm on $\mathbb{R}^n$.

The set of symmetric $n \times n$ matrices is denoted $\mathscr{S}^n$. Define

$$\mathscr{S}^n_+ = \{A \in \mathscr{S}^n \mid x^T A x \geq 0 \text{ for all } x \in \mathbb{R}^n\}$$

and

$$\mathscr{S}^n_{++} = \{A \in \mathscr{S}^n \mid x^T A x > 0 \text{ for all } x \neq 0\}.$$

Note that $\mathscr{S}^n_+$ and $\mathscr{S}^n_{++}$ are convex cones in the convex set $\mathscr{S}^n \subseteq \mathbb{R}^{n \times n}$.

1.2 Convex functions

A *convex function* is a function whose graph lies below the line segment joining any two points of it. More precisely, f is convex if its domain is convex and $f((1-\theta)x_1 + \theta x_2) \leq (1-\theta)f(x_1) + \theta f(x_2)$ for all $x_1, x_2 \in \mathrm{dom}(f)$ and $\theta \in [0,1]$. Any norm is a convex function with domain $\mathbb{R}^n$.

Recall from multi-variable calculus that if f is a differentiable convex function then $f(y) \geq f(x) + \nabla f(x)^T (y - x)$ for all $x, y \in \mathrm{dom}(f)$. In fact,

1.2.1 Proposition. *If f is differentiable and* $\mathrm{dom}(f)$ *is convex then f is convex if and only if $f(y) \geq f(x) + \nabla f(x)^T (y - x)$ for all $x, y \in \mathrm{dom}(f)$.*

PROOF: Let f be differentiable. On $\mathbb{R}$, if f is convex then $f(\theta y + (1-\theta)x) \leq \theta f(y) + (1-\theta)f(x)$, so we get

$$f(y) - f(x) \geq \frac{1}{\theta}(f(x + \theta(y-x)) - f(x)) \to f'(x)(y-x)$$

as $\theta \to 0$, which is the required inequality.

Conversely, still on $\mathbb{R}$, let $x, y \in \mathrm{dom}(f)$, $\theta \in [0,1]$, and $z = \theta x + (1-\theta)y$. We have $f(x) \geq f(z) + f'(z)(x - z)$ and $f(y) \geq f(z) + f'(z)(y - z)$, so multiplying

and adding we get

$$\theta f(x) + (1-\theta)f(y) \geq f(z) + f'(z)(\theta x + (1-\theta)y - z) = f(z).$$

In $\mathbb{R}^n$, let $g(t) := f(ty + (1-t)x)$. Then g is a convex function on its domain, a subset of $\mathbb{R}$. Note that $g'(t) = \nabla f(ty+(1-t)x)^T(y-x)$, so the one dimensional case implies the n dimensional case. □

1.2.2 Proposition. *If f is twice differentiable and* $\mathrm{dom}(f)$ *is convex then f is convex if and only if $\nabla^2 f(x) \succeq 0$.*

PROOF: Let f be twice differentiable. We have seen that f is convex if and only if $f(y) \geq f(x) + \nabla f(x)^T(y-x)$ for all $x, y \in \mathrm{dom}(f)$. On $\mathbb{R}$, if f is convex, for $x, y \in \mathrm{dom}(f)$ we have

$$f(y) \geq f(x) + f'(x)(y-x) \text{ and } f(x) \geq f(y) + f'(y)(x-y).$$

Adding and dividing by $(x-y)^2$,

$$0 \leq \frac{f'(x) - f'(y)}{x-y} \to f''(x)$$

as $y \to x$.

Conversely, still on $\mathbb{R}$, let $x, y \in \mathrm{dom}(f)$ and note that the following is non-negative.

$$\begin{aligned}\int_x^y f''(z)(z-y)dz &= f'(z)(y-z)\Big|_x^y + \int_x^y f'(z)dz \\ &= -f'(x)(y-x) + f(y) - f(x)\end{aligned}$$

In $\mathbb{R}^n$, for $x, y \in \mathrm{dom}(f)$ fixed, define $g(t) = g_{x,y} := f(x + t(y-x))$. Then $g'(t) = \nabla f(x + t(y-x))^T(y-x)$, and

$$g''(t) = t^2(y-x)^T \nabla^2 f(x + t(y-x))(y-x).$$

Now $g = g_{x,y}$ is convex for every $x, y \in \mathrm{dom}(f)$ if and only if f is convex, and $g''(t) \geq 0$ for all $t > 0$ for all $x, y \in \mathrm{dom}(f)$ if and only if

$$(y-x)^T \nabla^2 f(x + t(y-x))(y-x) \geq 0$$

for all $t > 0$ and all $x, y, \in \mathrm{dom}(f)$. It follows that g is convex if and only if $\nabla^2 f(x) \succeq 0$ for all $x \in \mathrm{dom}(f)$. □

Let f be convex in its first variable for every value of its second variable in some set $\mathscr{A}$ (not necessarily convex). Let $g(x) := \sup_{y \in \mathscr{A}} f(x, y)$. Then g is

convex. Indeed,

$$\begin{aligned} g(\theta x_1 + (1-\theta)x_2) &= \sup_{y \in \mathscr{A}} f(\theta x_1 + (1-\theta)x_2, y) \\ &\leq \sup_{y \in \mathscr{A}} \theta f(x_1, y) + (1-\theta) f(x_2, y) \\ &\leq \sup_{y \in \mathscr{A}} \theta f(x_1, y) + \sup_{y \in \mathscr{A}} (1-\theta) f(x_2, y) \\ &= \theta \sup_{y \in \mathscr{A}} f(x_1, y) + (1-\theta) \sup_{y \in \mathscr{A}} f(x_2, y) \\ &= \theta g(x_1) + (1-\theta) g(x_2) \end{aligned}$$

Let $C \subseteq \mathbb{R}^n$ be a convex set and let $f : \mathbb{R}^m \times C \to \mathbb{R}$ be convex. Then $g(x) := \inf_{y \in C} f(x, y)$ is convex. Indeed, for $x_1, x_2 \in \mathbb{R}^m$ and $\theta \in [0, 1]$,

$$\begin{aligned} g(\theta x_1 + (1-\theta)x_2) &= \inf_{y \in C} f(\theta x_1 + (1-\theta)x_2, y) \\ &= \inf_{y_1, y_2 \in C} f(\theta x_1 + (1-\theta)x_2, \theta y_1 + (1-\theta) y_2) \\ &\leq \inf_{y_1, y_2 \in C} (\theta f(x_1, y_1) + (1-\theta) f(x_2, y_2)) \\ &\leq \theta \inf_{y_1 \in C} f(x_1, y_1) + (1-\theta) \inf_{y_2 \in C} f(x_2, y_2)) \\ &= \theta g(x_1) + (1-\theta) g(x_2) \end{aligned}$$

since the sets C and $\theta C + (1-\theta)C$ are equal.

1.3 Convex optimization

The *constrained optimization problem* we consider is that of minimizing a function $f_0 : \mathbb{R}^n \to \mathbb{R}$ subject to *inequality constraints* $f_i(x) \leq 0$ for $i = 1, \ldots, m$, and *equality constraints* $h_j(x) = 0$ for $j = 1, \ldots, p$, where the f_i are convex and the h_j are affine. Call the set of points satisfying the constraints X.

If f_0 is convex and differentiable then $f_0(y) \geq f_0(x) + \nabla f_0(x)^T (y - x)$ for all $x, y \in X$. If $x^* \in X$ is a minimizer then $\nabla f_0(x^*)^T (y - x^*) \geq 0$ for all $y \in X$. These are the *first order conditions*. Note that for the *unconstrained problem* $X = \mathbb{R}^n$, the first order conditions are $\nabla f_0(x^*) = 0$, and they are necessary and sufficient.

Here is a useful trick. If you can write $f(x + tv) = f(x) + tw^T v + O(t^2)$ then $w = \nabla f(x)$. This follows from Taylor's theorem. For example, if $P \in \mathscr{S}^n_+$ and $f(x) := \frac{1}{2} x^T P x + q^T x + r$ then

$$\begin{aligned} f(x + tv) &= \tfrac{1}{2}(x + tv)^T P (x + tv) + q^T (x + tv) + r \\ &= f(x) + t(q + Px)^T v + O(t^2) \\ \text{so } \nabla f(x) &= Px + q \end{aligned} \tag{Q}$$

1.3.1 Example. Solve the unconstrained problem $\min_x \|Ax - b\|_2^2$. This is a general form of the problem of finding a least-squares approximation. Note that

$$\|Ax - b\|_2^2 = x^T (A^T A) x - 2(A^T b)^T x + b^T b,$$

so the gradient is $2(A^TAx - A^Tb)$ by (Q). Setting the gradient equal to zero gives the *normal equations*, $A^TAx = A^Tb$.

2 Unconstrained Optimization

2.1 Inequalities

Consider the unconstrained problem $\min_x f(x)$, where f is a convex function. Let $x^* = \text{argmin}_x f(x)$. We assume that x^* is unknown, $p^* = f(x^*)$ is unknown, but $\nabla f(x)$ exists and is known. We would like to find x (our guess at the minimum) such that $\|p^* - f(x)\|$ is small or $\|x - x^*\|$ is small.

Assume that f is twice differentiable. By Taylor's theorem,

$$f(y) = f(x) + \nabla f(x)^T(y-x) + \tfrac{1}{2}(y-x)^T\nabla^2 f(z)(y-x)$$

for some z on the line segment between x and y. Assume that, for some $m > 0$, $\nabla^2 f(x) \succeq mI$ for all $x \in \mathbb{R}^n$. Then

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \tfrac{m}{2}\|y-x\|^2.$$

This expression is a quadratic function of y, so it has a minimum. Differentiating with respect to y, for the minimizer y^* we get $\nabla f(x) + my^* - mx = 0$, i.e. $y^* - x = -\frac{1}{m}\nabla f(x)$. Then

$$f(y) \geq f(x) - \tfrac{1}{m}\nabla f(x)^T\nabla f(x) + \tfrac{m}{2}\|\tfrac{1}{m}\nabla f(x)\|^2 = f(x) - \tfrac{1}{2m}\|\nabla f(x)\|^2.$$

Whence $p^* = \min_y f(y) \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2$, so

$$0 \leq f(x) - p^* \leq \tfrac{1}{2m}\|\nabla f(x)\|^2. \tag{1}$$

Under the same assumptions, for any $x, y \in \mathbb{R}^n$, as before,

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T(y-x) + \tfrac{1}{2}(y-x)^T\nabla^2 f(z)(y-x) \\ &\geq f(x) + \nabla f(x)^T(y-x) + \tfrac{m}{2}\|y-x\|^2 \\ \text{so } p^* = f(x^*) &\geq f(x) + \nabla f(x)^T(x^*-x) + \tfrac{m}{2}\|x^*-x\|^2 \\ &\geq f(x) - \|\nabla f(x)\|\|x^*-x\| + \tfrac{m}{2}\|x^*-x\|^2 \end{aligned}$$

Therefore $0 \geq p^* - f(x) \geq -\|\nabla f(x)\|\|x^*-x\| + \frac{m}{2}\|x^*-x\|^2$, so

$$\|x^* - x\| \leq \tfrac{2}{m}\|\nabla f(x)\| \tag{2}$$

Now assume that, for some $M > 0$, $MI \succeq \nabla^2 f(x)$ for all $x \in \mathbb{R}^n$. Then

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T(y-x) + \tfrac{1}{2}(y-x)^T\nabla^2 f(z)(y-x) \\ &\leq f(x) + \nabla f(x)^T(y-x) + \tfrac{M}{2}\|y-x\|^2 \\ \text{so } p^* = \min_y f(y) &\leq \min_y\{f(x) + \nabla f(x)^T(y-x) + \tfrac{M}{2}\|y-x\|^2\} \end{aligned}$$

We can solve the minimization problem on the right hand side, as above, by setting the derivative with respect to y to zero. $\nabla f(x) + My^* - Mx = 0$, so $y^* - x = -\frac{1}{M}\nabla f(x)$. Plugging this into the equations,

$$p^* - f(x) \leq -\tfrac{1}{M}\nabla f(x)^T \nabla f(x) + \tfrac{1}{2M}\|\nabla f(x)\|^2 = -\tfrac{1}{2M}\|\nabla f(x)\|^2,$$

so

$$\tfrac{1}{2M}\|\nabla f(x)\|^2 \leq f(x) - p^* \tag{3}$$

2.2 Descent methods

Let $\{x^{(n)}\}_{n=1}^{\infty}$ be a sequence of vectors. If $x^{(n+1)} = \rho x^{(n)}$ for some $0 < \rho < 1$ then $x^{(n)} \to 0$, for any starting value $x^{(0)}$, and we say that $x^{(n)}$ has *linear convergence* with constant ρ.

The general *descent method* is as follows.

```
def descent(f, dx, x0):
    x = x0
    while abs(df(x)) > delta:
        dx = compute_dir(f, df, x)
        t = line_search(f, df, x, dx)
        x = x + t * dx
    return x
```

Gradient descent

The *gradient descent method* chooses $\Delta x = -\nabla f(x)$. The line search for this method is either the *exact line search*, (i.e. solve $\min_t f(x + t\Delta x)$ exactly for t) or the *backtracking line search*, as follows. The function $g(t) := f(x + t\Delta x)$ is convex, and the line

$$\{(t, f(x) + t\nabla f(x)^T \Delta x) \mid t \in \mathbb{R}\}$$

supports the graph of g at the point $t = 0$. Choose an $\alpha < \frac{1}{2}$ and form the reference line

$$\{(t, f(x) + \alpha t \nabla f(x)^T \Delta x) \mid t \in \mathbb{R}\},$$

which intersects the graph of g at $t = 0$ and $t = t_0 > 0$. See Figure 1. Choose a reduction factor $\beta < 1$ and run the following.

```
t = 1
while f(x + t * dx) > f(x) + alpha * t * df * dx
    t = beta * t
return t
```

Observe that this algorithm terminates with $t \in [\beta t_0, t_0 \wedge 1]$, where t_0 is the other point at which the reference line intersects the graph of f.

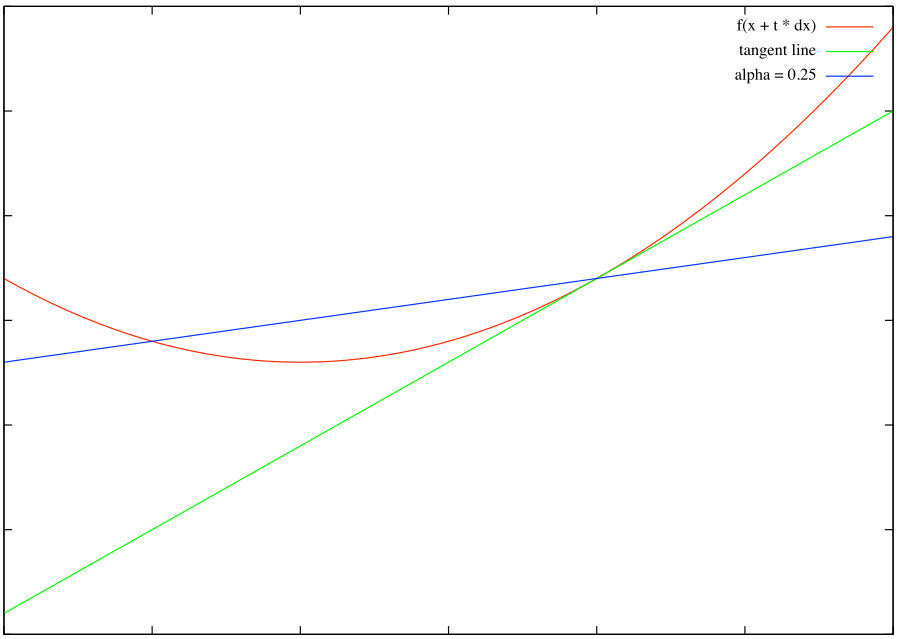

Figure 1: Backtracking line search

2.2.1 Theorem. *Let f be twice continuously differentiable and such that $mI \preceq \nabla^2 f(x) \preceq MI$ for all $x \in \mathbb{R}^n$. Assume that $m > 0$. Then the gradient descent method applied to f converges linearly with constant $1 - \frac{m}{M}$ using exact line search and constant $1 - 2\alpha\beta\frac{m}{M}$ using backtracking line search.*

PROOF: Let $\{x^{(n)}\}_{n=1}^{\infty}$ be the sequence of points encountered by the gradient descent method.

For exact line search, let $t_{\text{exact}} := \operatorname{argmin}_t f(x - t\|\nabla f(x)\|^2)$.

$$\begin{aligned} f(x - t\nabla f) &\leq f(x) - t\|\nabla f(x)\|^2 + t^2\tfrac{M}{2}\|\nabla f(x)\|^2 \\ f(x - t_{\text{exact}}\nabla f(x)) &\leq f(x) - \tfrac{1}{2M}\|\nabla f(x)\|^2 \\ \text{so } f(x^{(n+1)})) - p^* &\leq f(x^{(n)}) - p^* - \tfrac{1}{2M}\|\nabla f(x^{(n)})\|^2 \\ &\leq (1 - \tfrac{m}{M})(f(x^{(n)}) - p^*) \end{aligned}$$

The third line follows from minimizing the right hand side of the second line with respect to t, and the last line follows from the inequality (1), written as $-\|\nabla f(x)\|^2 \leq -2m(f(x) - p^*)$. This proves convergence since $m < M$.

For the backtracking method, as above,

$$\begin{aligned} f(x - t\nabla f) &\leq f(x) - t\|\nabla f(x)\|^2 + t^2\tfrac{M}{2}\|\nabla f(x)\|^2 \\ &= f(x) + (-t + \tfrac{M}{2}t^2)\|\nabla f(x)\|^2 \end{aligned}$$

$$\leq f(x) - \tfrac{1}{2}t\|\nabla f(x)\|^2$$
$$\leq f(x) - \alpha t\|\nabla f(x)\|^2$$

where the third line holds for all $0 \leq t \leq \frac{1}{M}$. It follows from this that the backtracking step always terminates with a non-zero t, and further, $\frac{\beta}{M} \leq t \leq 1$. Now, either $t = 1$ and $f(x - \nabla f(x)) \leq f(x) - \alpha\|\nabla f(x)\|^2$, or

$$f(x - t\nabla f(x)) \leq f(x) - \alpha t\|\nabla f(x)\|^2 \leq f(x) - \tfrac{\alpha\beta}{M}\|\nabla f(x)\|^2$$

Therefore, for the iteration,

$$\begin{aligned} f(x^{(n+1)}) - p^* &\leq f(x^{(n)}) - p^* - \min\{1, \tfrac{\alpha\beta}{M}\}\|\nabla f(x)\|^2 \\ &\leq (f(x^{(n)}) - p^*)(1 - \min\{2m, 2\alpha\beta\tfrac{m}{M}\}) \\ &= (f(x^{(n)}) - p^*)(1 - 2\alpha\beta\tfrac{m}{M}) \end{aligned}$$

□

Data processing

Here is our first application. Given data $\{u_j^*\}_{j=1}^N$, interpreted as coming from some "rough" or "noisy" source, find $\{u_j\}_{j=1}^N$ such that u is "nice" and u is "close to" u^*.

If by "close to" we mean that $\|u - u^*\|_2^2$ is small, and if by "nice" we mean that u has a reasonably sized quadratic variation, then the problem becomes

$$\min_u \sum_j (u_j - u_j^*)^2 + \beta \sum_j (u_{j+1} - u_j)^2.$$

The term on the right is the *roughness penalty* and it involves an arbitrary parameter controlling the relative importance of the two goals. Call this objective function $f(u)$. It is convex (as a sum of convex functions) and twice continuously differentiable.

$$\frac{\partial f}{\partial u_k} = 2(u_k - u_k^*) + 2\beta(2u_k - u_{k-1} - u_{k+1})$$

for $1 \leq k \leq N$ (where $u_k := 0$ if $k = 0$ or $N + 1$). Since $\nabla f(u) = 0$, we need to solve the following linear system of equations.

$$\begin{bmatrix} 2+4\beta & -2\beta & 0 & \dots & 0 \\ -2\beta & 2+4\beta & -2\beta & & \vdots \\ 0 & -2\beta & 2+4\beta & \ddots & 0 \\ \vdots & & \ddots & 2+4\beta & -2\beta \\ 0 & \dots & 0 & -2\beta & 2+4\beta \end{bmatrix} u = 2u^*.$$

We can of course do the same thing when u^* is two-dimensional (e.g. an image). Then the problem becomes

$$\min_u \sum_{ij} (u_{ij}^* - u_{ij})^2 + \beta\left(\sum_{ij} (u_{i+1,j} - u_{i,j})^2 + \sum_{ij} (u_{i,j+1} - u_{i,j})^2\right).$$

Steepest Descent

Given $z \in \mathbb{R}^n$, $z^* : \mathbb{R}^n \to \mathbb{R} : x \mapsto z^T x$ defines a linear functional. We write $\|z\|_* := \max_{\|x\|=1} |z^T x|$, and call $\|\cdot\|_*$ the *dual norm* of $\|\cdot\|$. The dual norm of the Euclidean norm is itself. Recall that the operator norm of a matrix A acting on $(\mathbb{R}^n, \|\cdot\|_1)$ is the largest of the L^1-norms of the columns of A. Also recall that the operator norm of a matrix A acting on $(\mathbb{R}^n, \|\cdot\|_\infty)$ is the largest of the L^1-norms of the rows of A. It follows that $\|z\|_{1,*} = \|z\|_\infty$ and $\|z\|_{\infty,*} = \|z\|_1$.

The *steepest descent method* chooses Δx so that $f(x + t\Delta x)$ decreases in t most rapidly, for some definition of "rapidly". Recall from Taylor's theorem that $f(x + t\Delta x) \approx f(x) + t\nabla f(x)^T \Delta x$. Given a norm $\|\cdot\|$ on $\mathbb{R}^n$, we would like to find a direction vector Δx_{nsd} (i.e. $\|\Delta x_{nsd}\| = 1$) that minimizes $\nabla f(x)^T \Delta x$ (this minimum will be negative). That is to say, for steepest descent,

$$\Delta x_{nsd} := \operatorname{argmin}\{\nabla f(x)^T \Delta x : \|\Delta x\| = 1\}.$$

Let $P \in \mathscr{S}^n_{++}$ and consider that

$$\|x\|_P^2 = x^T P x = x^T P^{\frac{1}{2}} P^{\frac{1}{2}} x = (P^{\frac{1}{2}} x)^T (P^{\frac{1}{2}} x) = \|P^{\frac{1}{2}} x\|_2^2.$$

Therefore we need to solve

$$\begin{aligned} \min\{\nabla f(x)^T \Delta x : \|P^{\frac{1}{2}} \Delta x\|_2 = 1\} & \\ &= \min\{\nabla f(x)^T P^{-\frac{1}{2}} (P^{\frac{1}{2}} \Delta x) : \|P^{\frac{1}{2}} \Delta x\|_2 = 1\} \\ &= \min\{(P^{-\frac{1}{2}} \nabla f(x))^T (P^{\frac{1}{2}} \Delta x) : \|P^{\frac{1}{2}} \Delta x\|_2 = 1\} \end{aligned}$$

The solution to the final problem is known to be $P^{\frac{1}{2}} \Delta x = -P^{\frac{1}{2}} \nabla f(x)^T$, so it follows that the steepest descent direction for $\|\cdot\|_P$ is $\Delta x = -P^{-1} \nabla f(x)^T$.

2.2.2 Example. Let $f(x, y) = \varepsilon x^2 + y^2$. Then f is twice continuously differentiable, $\nabla f(x, y) = \left[\begin{smallmatrix} 2\varepsilon x \\ 2y \end{smallmatrix}\right]$, and $\nabla^2 f(x, y) = \left[\begin{smallmatrix} 2\varepsilon & 0 \\ 0 & 2 \end{smallmatrix}\right]$. If we take $P = \left[\begin{smallmatrix} 2\varepsilon & 0 \\ 0 & 2 \end{smallmatrix}\right]$ then the steepest descent direction is

$$\Delta x = -\begin{bmatrix} 2\varepsilon & 0 \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 2\varepsilon x \\ 2y \end{bmatrix} = -\begin{bmatrix} x \\ y \end{bmatrix}.$$

2.2.3 Example. For the norm $\|\cdot\|_1$, the steepest descent direction is

$$\Delta x = -\operatorname{sign}(\nabla f(x)^T) e_i,$$

where $i = \operatorname{argmax}_i |\nabla f(x)_i^T|$.

2.2.4 Theorem. *Let f be twice continuously differentiable and such that $mI \preceq \nabla^2 f(x) \preceq MI$ for all $x \in \mathbb{R}^n$. Assume that $m > 0$. Fix a norm $\|\cdot\|$ on $\mathbb{R}^n$ and choose $\gamma \in (0, 1)$ such that $\|x\| \geq \gamma \|x\|_2$ for all $x \in \mathbb{R}^n$.*

The steepest descent method with backtracking line search applied to f converges linearly with constant $1 - 2m\alpha\gamma^2 \min\{1, \beta\gamma^2/M\}$.

PROOF: Define

$$\Delta x_{nsd} := \min\{\nabla f(x)^T \Delta x : \|\Delta x\| = 1\} \quad \text{and} \quad \Delta x_{sd} := \|\nabla f(x)\|_* \Delta x_{nsd}.$$

Then

$$\begin{aligned}
f(x + t\Delta x_{sd}) &\le f(x) + t\nabla f(x)^T \Delta x_{sd} + t^2 \tfrac{M}{2} \|\Delta x_{sd}\|_2^2 && \text{Taylor's theorem} \\
&\le f(x) + t\nabla f(x)^T \Delta x_{sd} + t^2 \tfrac{M}{2\gamma^2} \|\Delta x_{sd}\|_*^2 && \text{by choice of } \gamma \\
&\le f(x) - t\|\nabla f(x)\|_*^2 + t^2 \tfrac{M}{2\gamma^2} \|\nabla f(x)\|_*^2 && \text{definition of } \Delta x_{sd} \\
&= f(x) - t(1 - t\tfrac{M}{2\gamma^2}) \|\nabla f(x)\|_*^2 \\
&\le f(x) - \tfrac{1}{2} t \|\nabla f(x)\|_*^2 && \text{see below} \\
&\le f(x) - \alpha t \|\nabla f(x)\|_*^2 && \alpha < \tfrac{1}{2}
\end{aligned}$$

The second last line follows because the minimum value of $-t + t^2 \frac{M}{2\gamma^2}$ is $-\frac{\gamma^2}{2M}$ and is attained at $t = \frac{\gamma^2}{M}$. The line joining the vertex of this parabola and the origin has slope $-\frac{1}{2}$ and lies above the parabola for $t \in (0, \frac{\gamma^2}{M}]$.

It follows that the result of the backtracking line search satisfies the inequality $t \ge \min\{1, \frac{\beta\gamma^2}{M}\}$, so by (1),

$$\begin{aligned}
f(x^{(n+1)}) - p^* &\le (f(x^{(n)}) - p^*) - \alpha \min\{1, \tfrac{\beta\gamma^2}{M}\} \|\nabla f(x)\|_*^2 \\
&\le (f(x^{(n)}) - p^*) - \alpha\gamma^2 \min\{1, \tfrac{\beta\gamma^2}{M}\} \|\nabla f(x)\|_2^2 \\
&\le (f(x^{(n)}) - p^*) \Big(1 - 2m\alpha\gamma^2 \min\{1, \tfrac{\beta\gamma^2}{M}\}\Big). \qquad \square
\end{aligned}$$

2.3 Newton's method

Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of real numbers. If $x_{n+1} = K x_n^2$ then, if $x_n \to 0$, we say that x_n has *quadratic convergence*. Unfortunately, the conditions under which such a sequence converges depend on both K and x_0.

Newton's method is steepest descent where, at each step, we move in the steepest descent direction for the norm $\|\cdot\|_{\nabla^2 f(x^{(n)})}$. We will see that Newton's method has quadratic convergence.

Another approach to Newton's method is as follows. Recall

$$f(x + tv) \approx f(x) + t\nabla f(x)^T v + \tfrac{1}{2} t^2 v^T \nabla^2 f(x) v.$$

The minimum value of $\min_v \nabla f^T v + \frac{1}{2} v^T \nabla^2 f\, v$ occurs when $\nabla^2 f(x) v = -\nabla f(x)$, by (Q). Newton's method takes $\Delta x = -\nabla^2 f(x)^{-1} \nabla f(x)$.

A third approach is to note $f(x^*) = \min_x f(x)$ if and only if $\nabla f(x^*) = 0$. We can approximate the latter by solving $\nabla f(x + v) = 0$, which is approximately solved by v satisfying $\nabla f(x) + \nabla^2 f(x) v = 0$, by Taylor's theorem.

2.3.1 Theorem. *Let f be twice continuously differentiable and such that $mI \preceq \nabla^2 f(x) \preceq MI$ for all $x \in \mathbb{R}^n$. Assume that $m > 0$ and there is L (the Lipschitz constant) such that*

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x - y\|_2$$

for all $x, y \in \mathbb{R}^n$. Newton's method with backtracking line search applied to f satisfies the following. There is $0 < \eta \leq m^2/L$ and $\gamma = \gamma(\eta) > 0$ such that

(i) if $\|\nabla f(x)\| \geq \eta$ then $f(x^{(n+1)}) - f(x^{(n)}) \leq -\gamma$;

(ii) if $\|\nabla f(x)\| \leq \eta$ then $\|\frac{L}{2m^2}\nabla f(x^{(n+1)})\|_2 \leq \|\frac{L}{2m^2}\nabla f(x^{(n)})\|_2^2$ and the line search terminates with $t = 1$.

In fact, $\eta = \max\{3(1-2\alpha), 1\}m^2/L$.

Remark. It follows from (i) that the number of iterations needed, from any starting point $x^{(0)}$, to get to case (ii) is about $(f(x^{(0)}) - p^*)/\gamma$. If we are in case (ii) then $\|\nabla f(x^{(n)})\| \leq \eta$ implies $\|\nabla f(x^{(n+1)})\| \leq \eta$, so we remain in case (ii). Further, for any $\ell > k$,

$$\left\|\frac{L}{2m^2}\nabla f(x^{(\ell)})\right\| \leq \left\|\frac{L}{2m^2}\nabla f(x^{(k)})\right\|^{2^{\ell-k}} \leq \left(\frac{1}{2}\right)^{2^{\ell-k}}$$

By (1),

$$f(x^{(n)}) - p^* \leq \tfrac{1}{2m}\|\nabla f(x^{(n)})\|_2^2 \leq \frac{1}{2m}\left(\frac{2m^2}{L}\right)^2\left(\frac{1}{2}\right)^{2^{n+1}} = \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^n}$$

Note that in this case the number of steps required so that $f(x^{(n)}) - p^* < \varepsilon$ is on the order of $\log_2 \log_2 \frac{1}{\varepsilon}$. For all practical purposes, at most 6 steps will be required!

PROOF: Define $\Delta x_{nt} := -(\nabla^2 f(x))^{-1}\nabla f(x)$. Let

$$\lambda^2(x) := \nabla f(x)^T(\nabla^2 f(x))^{-1}\nabla f(x) > 0.$$

Note that $\lambda^2(x) = -\nabla f(x)^T \Delta x_{nt}$. From the linear approximation, this is the amount we decrease our approximation with a step of unit size. Call it *Newton's decrement*. Further,

$$\|\Delta x_{nt}\|_2^2 = \nabla f(x)^T(\nabla^2 f(x))^{-1}(\nabla^2 f(x))^{-1}\nabla f(x) \leq \tfrac{1}{m}\lambda^2(x).$$

As usual,

$$\begin{aligned} f(x + t\Delta x_{nt}) &\leq f(x) + t\nabla f(x)^T \Delta x_{nt} + t^2 \tfrac{M}{2}\|\Delta x_{nt}\|_2^2 \\ &\leq f(x) - t\lambda^2(x) + t^2 \tfrac{M}{2m}\lambda^2(x) \\ &\leq f(x) + (-t + t^2 \tfrac{M}{2m})\lambda^2(x) \end{aligned}$$

The parabola is minimized at $\hat{t} = \frac{m}{M}$ and the slope of the line joining $(0,0)$ with the vertex $(\frac{m}{M}, -\frac{m}{2M})$ is $-\frac{1}{2}$. So when $0 < t \leq \frac{m}{M}$,

$$\begin{aligned} f(x + t\Delta x_{nt}) &\leq f(x) - \tfrac{1}{2} t\lambda^2(x) \\ &\leq f(x) - \alpha t\lambda^2(x) && \alpha < \tfrac{1}{2} \\ &= f(x) + \alpha t \nabla f(x)^T \Delta x_{nt} \end{aligned}$$

Therefore any $t \in (0, \frac{m}{M}]$ satisfies the backtracking condition. It follows that the t returned by the backtracking algorithm is at least $\hat{t} = \beta \frac{m}{M}$. What is the reduction in f?

$$\begin{aligned} f(x + \hat{t}\Delta x_{nt}) &\leq f(x) - \alpha\beta \tfrac{m}{M} \lambda^2(x) \\ \text{so } f(x^{(n+1)}) - f(x^{(n)}) &\leq -\alpha\beta \tfrac{m}{M} \lambda^2(x) \end{aligned}$$

But $\lambda^2(x) = \nabla f(x)^T (\nabla^2 f(x))^{-1} \nabla f(x)$, so

$$\tfrac{1}{M} \|\nabla f(x)\|_2^2 \leq \lambda^2(x) \leq \tfrac{1}{m} \|\nabla f(x)\|_2^2.$$

Whence

$$f(x^{(n+1)}) - f(x^{(n)}) \leq -\alpha\beta \tfrac{m}{M^2} \|\nabla f(x)\|_2^2 \leq -\alpha\beta \tfrac{m}{M^2} \eta^2 =: -\gamma.$$

Now for the second, more difficult, part. Let $\tilde{f}(t) = f(x + t\Delta x_{nt})$. The idea is to derive an estimate for $\tilde{f}''$ and then $\tilde{f}'$, and then $\tilde{f}$.

$$\begin{aligned} \tilde{f}'(t) &= \nabla f(x + t\Delta x_{nt})^T \Delta x_{nt} \\ \tilde{f}''(t) &= \Delta x_{nt}^T \nabla^2 f(x + t\Delta x_{nt}) \Delta x_{nt} \\ \tilde{f}'(0) &= \nabla f(x)^T \Delta x_{nt} = -\lambda^2(x) \\ \tilde{f}''(0) &= \Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt} = \lambda^2(x) \end{aligned}$$

It follows from the Lipschitz condition that

$$|\tilde{f}''(t) - \tilde{f}''(0)| = \Delta x_{nt}^T (\nabla^2 f(x + t\Delta x_{nt}) - \nabla^2 f(x)) \Delta x_{nt} \leq tL \|\Delta x_{nt}\|_2^3.$$

Whence,

$$\tilde{f}''(t) \leq \tilde{f}''(0) + tL\|\Delta x_{nt}\|_2^3 \leq \lambda^2(x) + tL\|\Delta x_{nt}\|_2^3 \leq \lambda^2(x) + t\frac{L}{m^{3/2}} \lambda^3(x),$$

so, integrating,

$$\tilde{f}'(x) \leq \tilde{f}'(0) + \lambda^2(x)t + t^2 \frac{L}{2m^{3/2}} \lambda^3(x) = -\lambda^2(x) + \lambda^2(x)t + t^2 \frac{L}{2m^{3/2}} \lambda^3(x).$$

Integrating again,

$$\begin{aligned} \tilde{f}(t) &\leq \tilde{f}(0) - t\lambda^2(x) + \frac{1}{2} t^2 \lambda^2(x) + t^3 \frac{L}{6m^{3/2}} \lambda^3(x) \\ f(x + t\Delta x_{nt}) - f(x) &\leq -t\lambda^2(x) + \frac{1}{2} t^2 \lambda^2(x) + t^3 \frac{L}{6m^{3/2}} \lambda^3(x). \end{aligned}$$

We need to check the exit condition for the backtracking line search.

$$\begin{aligned} f(x+t\Delta x_{nt}) &\leq f(x)+\alpha t\nabla f(x)^T\Delta x_{nt} \\ \iff \tilde{f}(t) &\leq \tilde{f}(0)-\alpha t\lambda^2(x) \end{aligned}$$

We want to exit the line search algorithm on the first step, with $t=1$. The following steps are reversible.

$$\begin{aligned} -\lambda^2(x)+\tfrac{1}{2}\lambda^2(x)+\frac{L}{6m^{3/2}}\lambda^3(x) &\leq \alpha\lambda^2(x) \\ -\frac{1}{2}+\frac{L}{6m^{3/2}}\lambda(x) &\leq -\alpha \\ 2\alpha-1 &\leq \frac{L\lambda(x)}{3m^{3/2}} \\ \lambda(x) &\leq \frac{3(1-2\alpha)m^{3/2}}{L} \end{aligned}$$

We have the inequality $\sqrt{m}\lambda(x)\leq\|\nabla f(x)\|_2$, so if we choose $\eta=\frac{3(1-2\alpha)m^2}{L}$ then, since $\|\nabla f(x)\|\leq\eta$ in case (ii), the inequality is satisfied and line search terminates with $t=1$. It remains to show that

$$\left\|\frac{L}{2m^2}\nabla f(x^{(n+1)})\right\|_2 \leq \left\|\frac{L}{2m^2}\nabla f(x^{(n)})\right\|_2^2$$

Consider the following.

$$\begin{aligned} \|\nabla f(x^{n+1})\|_2 &= \|\nabla f(x^{(n)}+\Delta x_{nt})-\nabla f(x^{(n)})-\nabla^2 f(x^{(n)})\Delta x_{nt})\|_2 \\ &= \left\|\int_0^1\left(\nabla^2 f(x^{(n)}+t\Delta x_{nt})-\nabla f(x^{(n)})\right)\Delta x_{nt}dt\right\|_2 \\ &\leq \int_0^1\|\nabla^2 f(x^{(n)}+t\Delta x_{nt})-\nabla f(x^{(n)})\|\|\Delta x_{nt}\|_2 dt \\ &\leq L\int_0^1 t\|\Delta x_{nt}\|_2^2 dt=\frac{L}{2}\|\Delta x_{nt}\|_2^2\leq\frac{L}{2m^2}\|\nabla f(x)\|_2^2 \end{aligned}$$

This proves the inequality because norms are homogeneous. The last inequality follows from

$$\begin{aligned} \|\Delta x_{nt}\|_2 &= \|(\nabla^2 f)^{-1}\nabla f(x)\|_2 \\ &\leq \|(\nabla^2 f)^{-1}\|_2\|\nabla f(x)\|_2\leq\tfrac{1}{m}\|\nabla f(x)\|_2. \end{aligned}$$

□

2.4 Non-differentiable optimization

In applications the objective functions, while convex, may not always be differentiable. Whether or not f is not differentiable at a point x, define the *sub-gradient*

of f at x to be

$$\partial f(x) = \{g \in \mathbb{R}^n \mid f(y) \geq f(x) + g^T(y-x) \text{ for all } y \in \mathbb{R}^n\}.$$

For example, $\partial|0| = [-1,1]$, and

$$\partial(|x|+|y|) = \begin{cases} (\text{sign}(x), \text{sign}(y)) & (x,y) \neq (0,0) \\ & x = 0, y \neq 0 \\ & y = 0, x \neq 0 \\ & (x,y) = (0,0) \end{cases}$$

Note that the sub-gradient is always convex.

2.4.1 Lemma. *$x^* = \text{argmin}_x f(x)$ if and only if $0 \in \partial f(x^*)$.*

PROOF: If $0 \in \partial f(x)$ then by definition, $f(y) \geq f(x) + 0^T(y-x) = f(x)$ for all $y \in \mathbb{R}^n$, so x is a minimizer. Conversely, if $0 \notin \partial f(x)$ then there is $y \in \mathbb{R}^n$ such that $f(y) < f(x) + 0^T(y-x) = f(x)$, so x is not the minimizer. □

The sub-gradient method is described below. Choose a step size α_k and a direction $g^{(k)} \in \partial f(x^{(k)})$ and let $x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$.

```
x = xbest = x0
while :
    choose g in sub-gradient of f(x)
    calculate alpha
    x = x - alpha g
    if f(x) < f(xbest): xbest = x
```

Since $g^{(k)} \in \partial f(x^{(k)})$, it follows that $f(x^*) \geq f(x^{(k)}) + g^{(k)T}(x^* - x^{(k)})$, so

$$\begin{aligned} \|x^{(k+1)} - x^*\|_2^2 &= \|x^{(k)} - \alpha_k g^{(k)} - x^*\|_2^2 \\ &= \|x^{(k)} - x^*\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^*) + \alpha_k^2 \|g^{(k)}\|_2^2 \\ &\leq \|x^{(k)} - x^*\|_2^2 - 2\alpha_k (f(x^{(k)}) - f(x^*)) + \alpha_k^2 \|g^{(k)}\|_2^2 \\ &\leq \|x^{(1)} - x^*\|_2^2 - 2\sum_{j=1}^k \alpha_j (f(x^{(j)}) - f(x^*)) + \sum_{j=1}^k \alpha_j^2 \|g^{(j)}\|_2^2 \end{aligned}$$

In particular we conclude the following.

$$2\sum_{j=1}^k \alpha_j (f(x^{(j)}) - f(x^*)) \leq \|x^{(1)} - x^*\|_2^2 + \sum_{j=1}^k \alpha_j^2 \|g^{(j)}\|_2^2$$

$$2(f(x_{\text{best}}) - f(x^*)) \sum_{j=1}^k \alpha_j \leq \|x^{(1)} - x^*\|_2^2 + \sum_{j=1}^k \alpha_j^2 \|g^{(j)}\|_2^2$$

$$f(x_{\text{best}}) - f(x^*) \leq \frac{\|x^{(1)} - x^*\|_2^2 + \sum_{j=1}^k \alpha_j^2 \|g^{(j)}\|_2^2}{2\sum_{j=1}^k \alpha_j}$$

There are several methods for choosing the sequence $\{\alpha_k\}_{k=1}^{\infty}$.

(i) Constant step size: $\alpha_k = h$ for all $k \geq 1$.
(ii) Variable step size, scaled by $g^{(k)}$: $\alpha_k = h/\|g^{(k)}\|_2$ for all $k \geq 1$.
(iii) ℓ^2 sequence: $\sum_k \alpha_k^2 < \infty$ but $\sum_k \alpha_k = \infty$.
(iv) c_0 sequence: $\lim_{k\to\infty} \alpha_k = 0$ but $\sum_k \alpha_k = \infty$.

We assume that $\partial f(x)$ is bounded for all $x \in \mathbb{R}^n$ by the same radius G. We analyze each case below.

(i)

$$f(x_{\text{best}}) - f(x^*) \leq \frac{\|x^{(1)} - x^*\|_2^2 + 2kh^2G^2}{2kh} \to \frac{hG^2}{2} \text{ as } k \to \infty$$

(ii)

$$f(x_{\text{best}}) - f(x^*) \leq \frac{\|x^{(1)} - x^*\|_2^2 + kh^2}{2h\sum_{j=1}^{k} \frac{1}{\|g^{(j)}\|}} \leq \frac{\|x^{(1)} - x^*\|_2^2 G + kh^2 G}{2hk} \to \frac{hG}{2}$$

(iii)

$$f(x_{\text{best}}) - f(x^*) \leq \frac{\|x^{(1)} - x^*\|_2^2 + \sum_{j=1}^{k} \alpha_j^2 \|g^{(j)}\|_2^2}{2\sum_{j=1}^{k} \alpha_j} \to 0$$

(iv) Exercise.

2.4.2 Example. In a data fitting problem we can choose our error function. Let's compare a sum-of-squares smoothness error

$$\sum_j (u_j - u_j^*)^2 + \beta \sum_j (u_{j+1} - u_j)^2$$

with the non-differentiable

$$\sum_j (u_j - u_j^*)^2 + \beta \sum_j |u_{j+1} - u_j|.$$

Let $h = 1/N$ and consider the two possible fits $\{u_j = jh\}$ and $\{v_j = \delta_{j,N}\}$ on the grid $\{jh \mid j = 0, \ldots, N\}$. Then $\sum_j |u_{j+1} - u_j| = \sum_j h = Nh = 1$ and $\sum_j |v_{j+1} - v_j| = 1$. But $\sum_j (u_{j+1} - u_j)^2 = \sum_j h^2 = Nh^2 = 1/N$ while $\sum_j (v_{j+1} - v_j)^2$ is still 1. If the "true" curve is expected to have jumps then we would prefer the non-differentiable error function, because the sum-of-squares error function gives too much preference to smoother fits, i.e. we lose "sharpness."

3 Constrained Optimization

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function. Recall that $\min_x f(x)$ has a solution at x^* if and only if $\nabla f(x^*) = 0$. This is the unconstrained case. Consider

the problem $\min_x f(x)$ subject to constraints such as $Ax = b$ or $x \geq 0$. The minimizer may occur at some x^* such that $\nabla f(x^*) \neq 0$.

Recall that $f(y) \geq f(x) + \nabla f(x)^T(y - x)$ for all $x, y \in \mathbb{R}^n$ and that x^* is a minimizer if and only if $\nabla f(x^*)^T(y - x^*)$ for all y satisfying the constraints. For an *equality constraint* $Ax = b$ we require that $Ay = b$. The minimizer must of course satisfy the constraints, so we obtain $A(y - x^*) = 0$. Furthermore, every element of $\ker(A)$ is obtained in this way for some y. Therefore, for any $v \in \ker(A)$, $\nabla f(x^*)^T \geq 0$. Again, this is a linear functional bounded below on a linear space, so it must be the case that $\nabla f(x^*)^T \equiv 0$ on $\ker(A)$. It is a fact that $\mathbb{R}^n = \ker(A) \oplus \text{range}(A^T)$, so $\nabla f(x^*)^T \in \text{range}(A^T)$. Write $\nabla f(x^*)^T = -A^T v$. It follows that x^* is a minimizer if and only if $Ax^* = b$ and $\nabla f(x^*)^T + A^T v = 0$ for some $v \in \mathbb{R}^n$.

3.0.3 Example. We solve $\min_x \frac{1}{2}\|x - x^0\|_2^2$ subject to $Ax = b$. Then $\nabla f(x) = x - x^0$ so the optimality condition becomes $Ax^* = b$ and $A^T v + x^* - x^0 = 0$. We can write this as the system

$$\begin{bmatrix} A & 0 \\ I & A^T \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} = \begin{bmatrix} b \\ x^0 \end{bmatrix}.$$

For the *non-negativity constraint*, recall that $\nabla f(x^*)^T \cdot (y - x^*) \geq 0$ for all $y \in X$. Then $\nabla f^T y \geq \nabla f^T x^*$, so at $y = 0$ we get $\nabla f^T(x^*) \cdot x^* \leq 0$. Further, $\nabla f^T(x^*)$ is a linear form bounded below on the positive cone, so $\nabla f(x^*) \geq 0$. Combining these, we find that $x_i^*(\nabla f(x^*))_i = 0$ for all $i = 1, \ldots, n$. This is also called *complementary slackness*.

3.1 Separating hyperplanes

3.1.1 Theorem. *Let $C, D \subseteq \mathbb{R}^n$ be disjoint convex sets. Then there is $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that $a^T c \leq b$ for all $c \in C$ and $a^T d \geq b$ for all $d \in D$. If C is closed and D compact then the inequalities may be taken to be strict. If C and D are both open then the inequalities may be taken to be strict.*

Consider the inequalities $Ax \leq b$. This is equivalent to $b - Ax \geq 0$, and these inequalities have a solution if and only if $C = \{b - Ax \mid x \in \mathbb{R}^n\}$ and $D = \{y \mid y \geq 0\}$ have a non-empty intersection. If $C \cap D = \varnothing$ the by the separating hyperplane theorem there are $\lambda \in \mathbb{R}^n$ and $\mu \in \mathbb{R}$ such that $\lambda^T z + \mu \leq 0$ for all $z \in C$ and $\lambda^T z + \mu \geq 0$ for all $z \in D$.

$\lambda^T(b - Ax) + \mu \leq 0$ implies that $-\lambda^T Ax \leq -\lambda^T b$ for all $x \in \mathbb{R}^n$. A linear functional is bounded above if and only if it is constant, so $-\lambda^T A = 0$, or $A^T\lambda = 0$. Further, $\lambda^T b + \mu \leq 0$.

$\lambda^T y + \mu \geq 0$ for all $y \geq 0$ implies in particular that $\mu \geq 0$ (taking $y = 0$). It follows that $\lambda^T b \leq -\mu \leq 0$. Further, if λ had a negative coordinate then the inequality wouldn't be satisfied for all $y \geq 0$, so $\lambda > 0$.

Therefore $Ax \leq b$ does not have a solution if and only if there is $\lambda > 0$ such that $A^T\lambda = 0$ and $\lambda^T b \leq 0$.

3.2 Conjugate functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be any function and let $f^*(y) = \sup_{x\in\mathrm{dom}(f)}(y^T x - f(x))$ for all y for which this supremum is not ∞. We say that f^* is the *conjugate function* of f.

3.2.1 Examples.

(i) If $f(x) = ax + b$ then

$$f^*(y) = \sup_{x\in\mathbb{R}}(yx - ax - b) = \sup_x(x(y-a) - b) = \infty \text{ if } y \neq a.$$

Therefore $\mathrm{dom}(f^*) = \{a\}$ and $f^*(a) = -b$.

(ii) If $f(x) = -\log(x)$ then $\mathrm{dom}(f) = \mathbb{R}_{++}$.

$$f^*(y) = \sup_{x\in\mathbb{R}_{++}}(yx - \log x) = \begin{cases}\infty & y \geq 0\\ -1 - \log(-y) & y < 0.\end{cases}$$

(iii) Let $Q \in \mathscr{S}^n_{++}$ and let $f(x) = \frac{1}{2}x^T Qx$.

$$f^*(y) = \sup_{x\in\mathbb{R}^n}(y^T x - \tfrac{1}{2}x^T Qx) = \tfrac{1}{2}y^T Q^{-1} y.$$

The maximum occurs at the x for which $y + Qx = 0$ by (Q).

(iv) Let $f(x) = \|x\|$, where $\|\cdot\|$ is some norm. If $\|y\|_* > 1$ then there is $z \in \mathbb{R}^n$ with $\|z\| = 1$ such that $|y^T z| > 1$. Notice that in this case

$$y^T(tz) - \|tz\| = (y^T z - 1)t$$

is unbounded. Suppose that $\|y\|_* \leq 1$. Then

$$y^T x - \|x\| \leq \|y\|_*\|x\| - \|x\| \leq 0,$$

with a maximum of 0 attained at $x = 0$. Therefore

$$f^*(y) = \sup_{x\in\mathbb{R}^n}(y^T x - \|x\|) = \begin{cases}0 & \|y\|_* \leq 1\\ \infty & \|y\|_* > 1.\end{cases}$$

(v) If $f(x) = \frac{1}{2}\|x\|^2$ then

$$f^*(y) = \sup_{x\in\mathbb{R}^n}(y^T x - \tfrac{1}{2}\|x\|^2) \leq \sup_{x\in\mathbb{R}^n}(\|y\|_*\|x\| - \tfrac{1}{2}\|x\|^2) = \tfrac{1}{2}\|y\|_*^2,$$

where the supremum is attained at $x = \mathrm{argmax}\{y^T x \mid \|x\| = \|y^*\|\}$.

3.3 Duality

The general problem is $\min_x f_0(x)$ subject to $f_i(x) \leq 0$, $i = 1,\ldots,m$ and $h_j(x) = 0$, $j = 1,\ldots,p$, where the f_i are convex and the h_j are affine. Let X be the set

of points satisfying the constraints, the *feasible* points, and $D = \bigcap_i \mathrm{dom}(f_i)$. We assume $\nabla^2 f \geq mI$ and let p^* be the optimal value. Define the *Lagrangian* to be

$$\mathscr{L}(x,\lambda,\nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)$$

and $g(\lambda,\nu) := \inf_{x\in D} \mathscr{L}(x,\lambda,\nu)$. We show that $g(\lambda,\nu) \leq f_0(x)$ for $\lambda \geq 0$ and any feasible point x. Indeed,

$$g(\lambda,\nu) = \inf_{x\in D} \mathscr{L}(\lambda,\nu) \leq \mathscr{L}(\tilde{x},\lambda,\mu) = f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{j=1}^{p} \nu_j h_j(\tilde{x}) \leq f_0(\tilde{x}).$$

It follows that $g(\lambda,\nu) \leq p^*$ for all $\lambda \geq 0$ and ν arbitrary. The problem $\max_{\lambda\geq 0,\nu} g(\lambda,\nu)$ is the *dual problem*.

3.3.1 Examples.

(i) For the problem $\min_x x^T x$ subject to $Ax = b$,

$$\begin{aligned} \mathscr{L}(x,\nu) &= x^T x + \nu^T(Ax-b) \\ g(\nu) &= \inf_x (x^T x + \nu^T(Ax-b)) \\ &= -\nu^T b - \tfrac{1}{4}\nu^T A A^T \nu \end{aligned}$$

(ii) For the problem $\min_x c^T x$ subject to $Ax = b$ and $x \geq 0$,

$$\begin{aligned} \mathscr{L}(x,\lambda,\nu) &= c^T x - \lambda^T x + \nu^T(Ax-b) \\ g(\lambda,\nu) &= \inf_x (c^T x - \lambda^T x + \nu^T(Ax-b)) \\ &= -\nu^T b + \inf_x ((c-\lambda+A^T\nu)^T x) \\ &= \begin{cases} -\nu^T b & c-\lambda+A^T\nu = 0 \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

(iii) For the problem $\min_x f_0(x)$ subject to $Ax = b$ and $Cx \leq d$,

$$\begin{aligned} \mathscr{L}(x,\lambda,\nu) &= f_0(x) + \lambda^T(Cx-d) + \nu^T(Ax-b) \\ g(\lambda,\nu) &= \inf_x (f_0(x) + \lambda^T(Cx-d) + \nu^T(Ax-b)) \\ &= -\nu^T b - \lambda^T d + \inf_x (f_0(x) + (C^T\lambda + A^T\nu)^T x) \\ &= -\nu^T b - \lambda^T d - \sup_x ((-A^T\nu - C^T\lambda)^T x - f_0(x)) \\ &= -\nu^T b - \lambda^T d - f_0^*(-A^T\nu - C^T\lambda) \end{aligned}$$

(iv) To solve the problem $\min_x \|x\|$ subject to $Ax = b$, note that

$$g(\nu) = -\nu^T b - f_0^*(A^T\nu) = \begin{cases} -\nu^T b & \|A^T b\| \leq 1 \\ -\infty & \text{otherwise.} \end{cases}$$

We have the following relationships.

$$\begin{aligned} g(\lambda,\nu) &\le f_0(x) \text{ for any } x \in X \\ g(\lambda,\nu) &\le \inf_{x\in X} f_0(x) = p^* \\ \sup_{\lambda\ge 0,\nu} g(\lambda,\nu) &\le \inf_{x\in X} f_0(x) \end{aligned}$$

This is *weak duality*. The problem $\sup_{\lambda\ge 0,\nu} g(\lambda,\nu)$ is called the *dual problem* to the *primal problem* $\inf_{x\in X} f_0(x)$.

3.3.2 Examples.

(i) For the problem $\min_x c^T x$ subject to $Ax = b$ and $x \ge 0$,

$$g(\lambda,\nu) = \begin{cases} -\nu^T b & c - \lambda + A^T\nu = 0 \\ -\infty & \text{otherwise} \end{cases}$$

so the dual problem is $\max_\nu -b^T\nu$ subject to $A^T\nu + c \ge 0$.

(ii) For the problem $\min_x c^T x$ subject to $Cx \le d$,

$$g(\lambda) = \begin{cases} -\lambda^T d & c + C^T\lambda = 0 \\ -\infty & \text{otherwise} \end{cases}$$

so the dual problem is $\max_\lambda -d^T\lambda$ subject to $C^T\lambda + c = 0$.

Under some conditions

$$d^* = \sup_{\lambda\ge 0,\nu} g(\lambda,\nu) = \inf_{x\in X} f_0(x) = p^*.$$

This is referred to as *strong duality*.

Suppose for now that there are no equality constraints. Then $g(\lambda) = \inf_x \mathscr{L}(x,\lambda)$ and

$$\sup_{\lambda\ge 0} \mathscr{L}(x,\lambda) = \sup_{\lambda\ge 0}(f_0(x) + \sum_{j=1}^{p} \lambda_j f_j(x)) \le f_0(x)$$

Therefore

$$\inf_x \sup_{\lambda\ge 0} \mathscr{L}(x,\lambda) \le \inf_x f_0(x) = p^*.$$

On the other hand,

$$d^* = \sup_{\lambda\ge 0} g(\lambda) = \sup_{\lambda\ge 0} \inf_x \mathscr{L}(x,\lambda).$$

Finally,

$$d^* = \sup_{\lambda\ge 0} \inf_x \mathscr{L}(x,\lambda) \le \inf_x \sup_{\lambda\ge 0} \mathscr{L}(x,\lambda) \le p^*.$$

3.3.3 Proposition (Slater condition). *If the following two conditions hold then $d^* = p^*$, i.e. strong duality holds.*

(i) There is $\tilde{x} \in int(D)$ such that $f_i(\tilde{x}) < 0$ for all $i = 1, \ldots, m$ and $h_j(\tilde{x}) = 0$ for all $j = 1, \ldots, p$; and
(ii) $rank(A) = p$, where $Ax = b$ is the other representation of the equality constraints.

PROOF: Define the sets

$$\mathscr{G} := \{(f_1(x), \ldots, f_m(x), h_1(x), \ldots, h_p(x), f_0(x)) \mid x \in D\}$$
$$\mathscr{A} := \{(u, v, t) \in \mathbb{R}^{m+p+1} \mid \exists\, x \in D \forall\, i, j (f(x) \leq u_i \wedge h_j(x) = v_j \wedge f_0(x) \leq t)\}$$
$$\mathscr{B} := \{(0, 0, t) \in \mathbb{R}^{m+p+1} \mid t < p^*\}$$

Then $\mathscr{A}$ is convex and $\mathscr{B}$ is convex, and $p^* = \inf\{t \mid (0, 0, t) \in \mathscr{A}\}$. It follows that $\mathscr{A}$ and $\mathscr{B}$ are disjoint. If $p^* = -\infty$ then $d^* = p^*$, so we may assume that $\mathscr{B}$ is non-empty, and by assumption $\mathscr{A}$ is non-empty.

By the separating hyperplane theorem there is $(\lambda, \nu, \mu) \neq 0$ and $\alpha \in \mathbb{R}$ such that $(\lambda, \nu, \mu)^T (u, v, t) \geq \alpha$ for all $(u, v, t) \in \mathscr{A}$ and $(\lambda, \nu, \mu)^T (u, v, t) \leq \alpha$ for all $(u, v, t) \in \mathscr{B}$. From the latter, $\mu t \leq \alpha$ for all $(0, 0, t) \in \mathscr{B}$, so $\mu p^* \leq \alpha$. From the former, $\lambda, \mu \geq 0$ since u and t can be taken to $+\infty$ in $\mathscr{A}$.

If $\mu > 0$ then we may take $\mu = 1$ without loss of generality, and for any feasible x,

$$d^* \geq \sum_i \lambda_i f_i(x) + \sum_j \nu_j h_j(x) + \mu f_0(x) \geq \alpha \geq p^*.$$

Assume that $\mu = 0$. For any $x \in D$,

$$\sum_i \lambda_i f_i(x) + \nu^T (Ax - b) \geq \alpha \geq 0.$$

For the particular point $\tilde{x}$, $\sum_i \lambda_i f_i(\tilde{x}) \geq 0$, so $\lambda = 0$. Therefore $\nu^T (Ax - b) \geq 0$ for all $x \in D$, and $\nu \neq 0$. Since $\tilde{x}$ is in the interior of D, if $\nu^T A \neq 0$ then there is a point $x_- \in \text{int}(D)$ such that $\nu^T (Ax_- - b) < 0$. This contradiction implies that $A^T \nu = 0$. But this too is a contradiction since the rank of A is p. Therefore $\mu > 0$.□

We know that $g(\lambda, \nu) \leq p^*$ for any $\lambda \geq 0$ and ν. Therefore

$$f_0(x) - p^* \leq f_0(x) - g(\lambda, \nu). \tag{4}$$

This estimate is very important in applications.

3.4 Complementary slackness

Assume for now that strong duality holds. Let x^* be primal optimal and (λ^*, ν^*) be dual optimal. Recall that

$$f(x^*) = g(\lambda^*, \nu^*) = \inf_{x \in D} \{f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{j=1}^{p} \nu_j^* h_j(x)\}$$

$$\begin{aligned}&\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{j=1}^p \nu_j^* h_j(x^*)\\ &\leq f_0(x^*)\end{aligned}$$

It follows that we have equalities on every line, and in particular $\lambda_i^* f_i(x^*) = 0$ for all $i = 1, \ldots, m$. If $f_i(x^*) < 0$ then $\lambda_i = 0$ is zero, and we say that this constraint is an *inactive constraint*.

3.4.1 Proposition (Karish-Kuhn-Tucker conditions).
Assume that f_0 and f_i, $i = 1, \ldots, m$, are convex and h_j, $j = 1, \ldots, p$, are affine, and all functions are differentiable. Then x^* is primal optimal, (λ^*, ν^*) is dual optimal, and strong duality holds and only if

$$\begin{aligned} f_i(x^*) &\leq 0 && i = 1, \ldots, m\\ h_j(x^*) &= 0 && j = 1, \ldots, p\\ \lambda_i^* &\geq 0 && i = 1, \ldots, m\\ \lambda_i^* f_i(x^*) &= 0 && i = 1, \ldots, m\\ \nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(x^*) &= 0 \end{aligned}$$

PROOF: Note that he function

$$x \mapsto \mathscr{L}(x, \lambda^*, \nu^*) = f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{j=1}^p \nu_j^* h_j(x)$$

is convex. If $x^* = \operatorname{argmin}_x \mathscr{L}(x, \lambda^*, \nu^*)$. Therefore $\nabla_x \mathscr{L}(x, \lambda^*, \nu^*) = 0$.
Finish this proof. □

3.4.2 Example. Solve $\min_x \frac{1}{2} x^T P x + q^T x + r$ subject to $Ax = b$. Here

$$\mathscr{L}(x, \nu) = \tfrac{1}{2} x^T P x + q^T x + r + \nu^T (Ax - b).$$

Therefore $0 = \nabla \mathscr{L}$ implies $Px + q + A^T \nu = 0$ and $Ax = b$, so the original problem is equivalent to

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \nu \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}.$$

3.4.3 Example. Solve $\min_x f_0(x) = \sum_{i=1}^n f_i(x_i)$ subject to $a^T x = b$, that x lies on a hyperplane. The dual function is

$$\begin{aligned} g(\nu) &= \inf_x \mathscr{L}(x,\nu) \\ &= \inf_x \sum_{i=1}^n f_i(x_i) + \nu(a^T x - b) \\ &= -\nu b + \inf_x \sum_{i=1}^n (f_i(x_i) + \nu a_i x_i) \\ &= -\nu b - \sum_{i=1}^n \sup_{x_i}(-\nu a_i x_i - f_i(x_i)) \\ &= -\nu b - \sum_{i=1}^n f_i^*(-\nu a_i). \end{aligned}$$

The dual problem is to maximize $g(\nu)$ over $\nu \in \mathbb{R}$. This should be easy, so say the maximum occurs at ν^*. Now to find the optimizer for the primal problem, we must solve the unconstrained problem $\min_x \sum_{i=1}^n (f_i(x_i) + \nu^* a_i x_i)$.

3.4.4 Example (Optimal transport). Given a list of locations with given amounts of material (x_i, p_i), we would like to find a scheme to move it to locations with given capacities (y_j, q_j), minimizing some distance quantity.

Let the flows be $\mu_{i,j} \geq 0$. So $\sum_j \mu_{i,j} = p_i$, and $\sum_i \mu_{i,j} = q_j$. The distance quantity is $\sum_{i,j} \mu_{i,j} |x_i - y_j|^2$ (*Wasserstein distance*). This problem has the $\mu_{i,j}$ as the quantities to be found, and there are on the order of N^2 unknowns. The dual problem is found as follows. Notice that

$$\sum_{i,j} \mu_{i,j} |x_i - y_j|^2 = \sum_i |x_i|^2 p_i + \sum_j |y_j|^2 q_j - 2 \sum_{i,j} \mu_{i,j} x_i \cdot y_j,$$

so it suffices to minimize $-\sum_{i,j} \mu_{i,j} x_i \cdot y_j$.

$$\begin{aligned} \mathscr{L}(\mu, a, b, \lambda) &= -\sum_{i,j} \mu_{i,j} x_i \cdot y_j + \sum_i a_i \left(\sum_j \mu_{i,j} - p_i \right) + \sum_j b_j \left(\sum_i \mu_{i,j} - q_j \right) - \sum_{i,j} \lambda_{i,j} \mu_{i,j} \\ &= \sum_{i,j} \mu_{i,j} (-x_i \cdot y_j + a_i + b_j - \lambda_{i,j}) - \sum_i a_i p_i - \sum_j b_j q_j \\ g(a, b, \lambda) &= \inf_\mu \mathscr{L}(\mu, a, b, \lambda) \\ &= \begin{cases} -\sum_i a_i p_i - \sum_j b_j q_j & -x_i y_j + a_i + b_j - \lambda_{i,j} = 0 \text{ for all } i, j \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

The supremum of $g(a, b, \lambda)$ over $\lambda \geq 0$ will occur when $-x_i y_j + a_i + b_j - \lambda_{i,j} = 0$ for all i, j. Then $0 \leq \lambda_{i,j} = a_i + b_j - x_i \cdot y_j$. The dual problem becomes $-\min_{a,b} \sum_i a_i p_i + \sum_j b_j q_j$ subject to $a_i + b_j \geq x_i \cdot y_j$. But consider the constraints: $b_j \geq x_i \cdot y_j - a_i$ for all i, so we may take $b_j = \max_i (x_i \cdot y_j - a_i)$. In

some sense $b = a^*$, and the dual problem reduces to the unconstrained problem $\min_a \sum_i a_i p_i + \sum_j a_j^* q_j$. If strong duality holds (which it likely does), the minimum distance is $2d^* + \sum_i |x_i|^2 p_i + \sum_j |y_i|^2 q_j$.

To actually implement the problem we need an algorithm for calculating a^* given a, and we need some descent method. From the KKT conditions, $\mu_{i,j}\lambda_{i,j} = 0$, so it $\lambda_{i,j} > 0$ then $\mu_{i,j} = 0$, and it can be seen that many of the primal variables are actually zero.

3.5 Alternatives

Either there is a solution to $Ax = b$ with $x \gg 0$ or there is λ such that $A^T\lambda \geq 0$, $A^T\lambda \neq 0$, and $b^T\lambda \leq 0$. Let $C = \{x \gg 0 \mid x \in \mathbb{R}^n\}$ and $D = \{x \in \mathbb{R}^n \mid Ax = b\}$. $C \cap D = \varnothing$ if and only if there is no solution $x \gg 0$. C and D are convex and non-empty, so there is $c \in \mathbb{R}^n$ and $d \in \mathbb{R}$ such that $c^T d \geq d$ on C and $c^T x \leq d$ on D. The fact that $c^T x \geq d$ implies $c \geq 0$, since the coördinates of $x \in C$ may be taken to positive infinity. Taking $x \to 0$, it is seen that $d \leq 0$. D is an affine space, so $c^T x$ is constant on D since it is bounded above. Rename d, if necessary, to be equal to this constant. Then $c^T x = d$ on D. Write $D = \{x = x_0 + Fy \mid y \in \mathbb{R}^r\}$, where $r = \dim\ker A$ and F is a matrix whose columns are a basis of $\ker A$. Then $c^T(x_0 + Fy) = d$ for any $y \in \mathbb{R}^r$, which is only possible if $c^T F = 0$, so $c \perp \ker A$. But $\ker A \oplus \text{range} A^T = \mathbb{R}^n$, so $c \in \text{range} A^T$, i.e. $c = A^T\lambda$ for some λ. Finally, $\lambda^T A = c \geq 0$, $\lambda^T A = c \neq 0$, and $\lambda^T A x_0 = \lambda^T b = d \leq 0$. This was problem 2.20 in the textbook.

3.5.1 Example. $f_i(x) \leq 0$ for $i = 1, \ldots, m$ and $h_j(x) = 0$ for $j = 1, \ldots, p$. We would like to find an *alternative* for this collection of inequalities. The idea is to consider the problem $\min_x 0$ subject to $f_i(x) \leq 0$ and $h_j(x) = 0$. If there is any feasible x then the minimum is $p^* = 0$, otherwise $p^* = \infty$. The dual function is

$$g(\nu, \lambda) = \inf_x 0 + \sum_{j=1}^{p} \nu_j h_j(x) \sum_{i=1}^{m} \lambda_i f_i(x).$$

If $\alpha > 0$ then $g(\alpha\nu, \alpha\lambda) = \alpha g(\nu, \lambda)$. Whence, if there is (ν, λ) with $\lambda \geq 0$ such that $g(\nu, \lambda) > 0$ then $p^* \geq d^* = \infty$ by weak duality. If there is no feasible solution to $\lambda \geq 0$, $g(\nu, \lambda) > 0$ then $d^* = 0$. Summarizing, if the dual is feasible (i.e. there is (ν, λ) such that $\lambda \geq 0$ and $g(\nu, \lambda) > 0$) then the primal is not feasible. This is a *weak alternative* because it could be the case that neither are feasible.

3.5.2 Example. $f_i(x) < 0$ for $i = 1, \ldots, m$ and $h_j(x) = 0$ for $j = 1, \ldots, p$. We would like to show that $\lambda \geq 0$, $\lambda \neq 0$, $g(\nu, \lambda) \geq 0$ is an alternative. Consider the problem $\min_x 0$ subject to $f_i(x) \leq 0$ and $h_j(x) = 0$. The dual function is

$$g(\nu, \lambda) = \inf_x \sum_{j=1}^{p} \nu_j h_j(x) \sum_{i=1}^{m} \lambda_i f_i(x).$$

If the primal is feasible then $g(\nu, \lambda) < 0$ for all $\lambda \geq 0$, $\lambda \neq 0$, so the proposed alternative cannot hold. (Not quite right.)

If strong duality holds then $g(\lambda^*, \nu^*) = d^* = p^* = f_0(x^*)$. We assume for now that strong duality holds.

3.5.3 Example. For $f_i(x) < 0$, $i = 1, \ldots, m$, and $Ax = b$, the alternative is $\lambda \geq 0$, $\lambda \neq 0$, and $g(\lambda, \nu) \geq 0$.

Consider the problem $\min_{x,s} s$ subject to $f_i(x) - s \leq 0$ and $Ax = b$. Then $p^* < 0$ if and only if there is x such that $f_i(x) < 0$ for all i and $Ax = b$.

The dual function is

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x,s} s + \sum_{i=1}^{m} \lambda_i (f_i(x) - s) + \nu^T (Ax - b) \\ &= -\nu^T b + \inf_{x,s} s \left(1 - \sum_{i=1}^{m} \lambda_i\right) + \sum_{i=1}^{m} \lambda_i f_i(x) + \nu^T Ax \end{aligned}$$

This is $-\infty$ unless $\mathbf{1}^T \lambda = 1$. The dual problem hence is

$$\max_{\substack{\mathbf{1}^T \lambda = 1 \\ \lambda \geq 0}} g(\lambda, \nu).$$

By strong duality there at λ^* and ν^* optimizing the dual. Now, $p^* < 0$ if and only if there is x such that $Ax = b$ and $f_i(x) < 0$ for all i. In this case $g(\lambda^*, \nu^*) = d^* = p^* < 0$, so the alternative does not hold, in particular $g(\lambda, \nu) < 0$ for all $\lambda \geq 0$ and ν. Similarly, if there is no such x then the alternative holds.

4 Algorithms

Now we move to chapters 10 and 11, concerning problems of the from $\min f_0(x)$ subject to $Ax = b$, and $\min f_0(x)$ subject to $Ax = b$ and $f_i(x) \leq 0$, respectively.

One way to solve the problem with only equality constraints is to use the dual problem.

$$\begin{aligned} g(\nu) := \inf_x (f_0(x) + \nu^T (Ax - b)) &= -\nu^T b + \inf_x (f_0(x) + (A^T \nu)^T x) \\ &= -\nu^T b - \sup_x (-(A^T \nu)^T x - f_0(x)) \\ &= -\nu^T b - f_0^*(-A^T \nu) \end{aligned}$$

Solving the primal is equivalent (under strong duality) to solving the unconstrained problem $\sup_\nu (-\nu^T b - f_0^*(-A^T \nu))$. Suppose ν^* is the solution. Then plugging this into the Lagrangian, we can solve the unconstrained problem $\inf_x (f_0(x) + \nu^{*T}(Ax - b)$ for x^*, the solution to the original problem.

Another is way to use Newton's method. Suppose x approximates the solution and we wish to improve it. Approximate $f_0(x + v)$ up to the quadratic term,

$$f_0(x + v) \approx f_0(x) + \nabla f_0(x)^T v + \tfrac{1}{2} v^T \nabla^2 f(x) v$$

and use this approximation to find the direction to move to improve x. This is also known as *sequential quadratic programming*. We solve the following problem for the best direction v.

$$\min_v f_0(x) + \nabla f_0(x)^T v + \tfrac{1}{2} v^T \nabla^2 f(x) v \text{ subject to } Av = Ax - b$$

The optimality conditions are as follows. The Lagrangian is

$$\mathscr{L} = \tfrac{1}{2} v^T \nabla^2 f(x) v + \nabla f_0(x)^T v + \nu^T (Ax - b - Av)$$

so the KKT condition is that $\nabla^2 f_0(x) v + \nabla f_0(x) + A^T \nu = 0$ and of course the constraint $Av = Ax - b$ must be satisfied. We get the following system of equations for the best direction v and the dual variable ν.

$$\begin{bmatrix} \nabla^2 f_0(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v_{\text{nt}} \\ \nu_{\text{nt}} \end{bmatrix} = - \begin{bmatrix} \nabla f_0(x) \\ Ax - b \end{bmatrix}.$$

A second derivation of this system comes from the optimality conditions for the original problem. They are $\nabla f_0(x) + A^T \nu = 0$ and $Ax = b$. We wish to find $x + \Delta x$ so that the optimality conditions still hold (but at which some other quantity is improved). Linearizing,

$$\begin{aligned} \nabla f_0(x) + \nabla^2 f_x(x) \Delta x + A^T \Delta \nu + A^T \nu &= 0 \\ Ax + A\Delta x &= b \\ \begin{bmatrix} \nabla^2 f_0(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\text{nt}} \\ \Delta \nu_{\text{nt}} \end{bmatrix} &= - \begin{bmatrix} \nabla f_0(x) + A^T \nu \\ Ax - b \end{bmatrix} \end{aligned}$$

In the unconstrained Newton's method we want $\nabla f_0(x + v) = 0$ so we pick v so that $\nabla^2 f_0(x) v = -\nabla f_0(x)$.

$$\begin{aligned} f_0(x + tv) &= f_0(x) + t \nabla f_0(x)^T v + O(t^2) \\ &= f_0(x) - t v^T \nabla^2 f_0(x)^T v + O(t^2) \\ \frac{d}{dt} f(x + tv)\Big|_{t=0} &= -\lambda^2(x) < 0 \end{aligned}$$

since $\nabla^2 f_0(x)$ is positive semidefinite. With the backtracking method $f_0(x)$ decreases with every iteration.

With equality constraints, however, $f_0(x)$ will not necessarily decrease if $Ax \neq b$. We need to find some other quantity that will go down with every iteration to use as the backtracking quantity. From Newton's step,

$$\begin{aligned} \frac{d}{dt} f_0(x + t\Delta x)\Big|_{t=0} &= \nabla f_0(x)^T \Delta x_{\text{nt}} \\ &= -\Delta x_{\text{nt}}^T (\nabla^2 f_0(x) \Delta x_{\text{nt}}^T + A^T (\nu + \Delta \nu_{\text{nt}})) \\ &= -\Delta x_{\text{nt}}^T \nabla^2 f_0(x) \Delta x_{\text{nt}}^T - (\nu + \Delta \nu_{\text{nt}})^T (A \Delta x_{\text{nt}}) \\ &= -\Delta x_{\text{nt}}^T \nabla^2 f_0(x) \Delta x_{\text{nt}}^T - (\nu + \Delta \nu_{\text{nt}})^T (b - Ax) \end{aligned}$$

We show now that

$$r = r(x,v) = \begin{bmatrix} r_{dual} \\ r_{primal} \end{bmatrix} = \begin{bmatrix} \nabla f_0(x) + A^T v \\ Ax - b \end{bmatrix}$$

strictly decreases, and that $r = 0$ at optimum. Consider the new variable $y = (x,v)$ and do Newton's method on $r(y)$. The linear approximation is $r(y+z) = r(y) + Dr(y)z$, so we solve this equation for z. r is called the *residual*. This material is in §10.3.

$$\begin{aligned} &\frac{d}{dt}\|r(y+t\Delta y_{\text{nt}})\|_2^2\Big|_{t=0} \\ &\quad = \frac{d}{dt}(r(y)+tDr(y)\Delta y_{\text{nt}}), r(y)+tDr(y)\Delta y_{\text{nt}}) + O(t^2)\Big|_{t=0} \\ &\quad = 2(Dr(y)\Delta y_{\text{nt}}, r(y)) = -2r(y)^T r(y) = -2\|r(y)\|_2^2 \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{d}{dt}\|r(y+t\Delta y_{\text{nt}})\|_2\Big|_{t=0} &= \frac{d}{dt}\sqrt{\|r(y+t\Delta y_{\text{nt}})\|_2^2}\Big|_{t=0} \\ &= \frac{-2\|r(y)\|_2^2}{2\sqrt{\|r(y)\|_2^2}} = -\|r\|_2 \end{aligned}$$

Hence the tangent line to the graph of $\|r(y+tv)\|_2$ at $t=0$ is $(1-t)\|r(y)\|_2$. We build our line search as follows. Find t such that

$$\|r(y+t\Delta y_{\text{nt}})\|_2 \le (1-\alpha t)\|r\|_2.$$

The *infeasible start Newton's method* is as follows. (See §10.3.)

```
input: (x0, v0), e > 0, 0 < a < 0.5, 0 < b < 1
x, v = x0, v0
while |r(x, v)| > e:
    dxnt, dvnt = compute_dir(x, v, etc.)
    t = 1.0
    while |r(x + t dxnt, v + t dvnt)| > (1 - a t) |r(x, v)|
        t = b * t
    x, v = x0 + t dxnt, v0 + t dvnt
```

Some observations. If we accept the Newton step, i.e. if $t = 1$, then the solution becomes feasible, i.e. $A(x+\Delta x_{\text{nt}}) = b$. Otherwise, if we step by $t < 1$ then we have $b - A(x+t\Delta x_{\text{nt}}) = (1-t)(b-Ax)$, so we "improve the feasibility" by a factor of $1-t$.

Now we move on to the problem with inequality constraints. Consider the problem $\min f_0(x)$ subject to $Ax = b$ and $f_i(x) \le 0$, $i = 1,\dots,m$. Introduce the functional $\varphi(x) = -\sum_{i=1}^m \log(-f_i(x))$ and solve

$$\min_x t f_0(x) + \varphi(x) \text{ subject to } Ax = b.$$

We hope that the solutions get close to the solution of the original problem for large t. Compute

$$\nabla\varphi(x) = -\sum_{i=1}^{m} \frac{1}{f_i(x)} \nabla f_i(x)$$
$$\nabla^2\varphi(x) = -\sum_{i=1}^{m} \left(\frac{1}{f_i(x)^2} \nabla f_i(x) \nabla f_i(x)^T - \frac{1}{f_i(x)} \nabla^2 f_i(x) \right)$$

Combining with the dual, the optimality conditions are

$$t\nabla f_0(x^*) - \sum_{i=1}^{m} \frac{1}{f_i(x^*)} \nabla f_i(x^*) + A^T \nu^* = 0$$
$$Ax^* - b = 0$$

where the first line is $t\nabla f_0(x) + \nabla\varphi + A^T\nu = 0$. These are the equations for the *barrier method*. We known that $-f_i(x^*) > 0$, so write $\lambda_i^* := \frac{1}{-tf_i(x^*)} > 0$. Rescaling ν^*, we can write the above equations as

$$\nabla f_0(x^*) + \sum \lambda_i^* \nabla f_i(x^*) + A^T \nu^* = 0$$
$$Ax^* - b = 0$$
$$-\lambda_i^* f_i = \tfrac{1}{t}$$

These are the equations for the *primal-dual method*. The Lagrangian for the problem is

$$\mathscr{L}(x, \nu, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \nu^T (Ax - b)$$
$$\nabla\mathscr{L}(x, \nu, \lambda) = \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + A^T \nu$$

Note that $\nabla\mathscr{L}(x^*, \nu^*, \lambda^*) = 0$. With the dual function g we have

$$g(\lambda^*, \nu^*) = f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \nu^{*T}(Ax^* - b)$$
$$\text{so } f_0(x^*) - p^* \le f_0(x^*) - g(\lambda^*, \nu^*)$$
$$= \sum_{i=1}^{m} \frac{1}{tf(x^*)} f_i(x^*) = \frac{m}{t} \to 0 \text{ as } t \to \infty$$

For the barrier method we have the variables (x, ν). Use Newton's method with line search on the residuals to find Δx and $\Delta\nu$.

$$\begin{bmatrix} t\nabla^2 f_0 + \nabla^2\varphi & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\nu \end{bmatrix} = \begin{bmatrix} t\nabla f_0 + \nabla\varphi + A^T\nu \\ Ax - b \end{bmatrix}$$

How do we start the barrier method? We need to start with a feasible point because otherwise $\varphi(x)$ is not defined. The first phase is to solve the auxiliary problem

$$\min_{x,s} s \text{ subject to } f_i(x) \le s \text{ and } Ax = b.$$

This problem is always feasible: choose x^0 such that $Ax^{(0)} = b$ and set $s^{(0)} = \max_i f_i(x^{(0)})$. Use any method to get a sequence of approximate solutions to this problem. If we find $s^{(n)} < 0$ for some n then we know $p^* < 0$ for the auxiliary problem, so the original problem is feasible and we have obtained a feasible point $x^{(n)}$. However, if find that p^* is probably positive then the original problem is probably not feasible. In this case we hope to find (λ, ν) for the dual to the auxiliary problem for which $g(\lambda, \nu) > 0$, proving infeasibility of the original problem. Another way to get started is to instead consider the problem

$$\min_{x,s} \mathbf{1}^T s \text{ subject to } f_i(x) \le s_i,\ i = 1, \dots, m, \text{ and } Ax = b.$$

For the primal-dual method we apply Newton's method with line search on the residuals to the modified KKT equations

$$\begin{aligned} r_{\text{dual}} &= \nabla f_0(x) + \sum \lambda_i \nabla f_i(x) + A^T \nu = 0 \\ r_{\text{cent}} &= -\lambda_i f_i(x) - \tfrac{1}{t} = 0 \\ r_{\text{prim}} &= Ax - b = 0 \end{aligned}$$

to find Δx $\Delta\lambda$, and $\Delta\nu$.

$$\begin{bmatrix} \nabla^2 f_0 + \sum \lambda_i \nabla^2 f_i & [\nabla f_1 \mid \cdots \mid \nabla f_m] & A^T \\ -[\lambda_1 \nabla f_1 \mid \cdots \mid \lambda_m \nabla f_m]^T & -\operatorname{diag}(f_1, \dots, f_m) & 0 \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\text{nt}} \\ \Delta\lambda_{\text{nt}} \\ \Delta\nu_{\text{nt}} \end{bmatrix} = - \begin{bmatrix} r_{\text{dual}} \\ r_{\text{cent}} \\ r_{\text{prim}} \end{bmatrix}$$

$$Dr(y) = \Delta y_{\text{nt}}$$

Recall the inequality $f_0(x) - p^* \le \frac{m}{t}$. Define $\hat{\eta} = -\sum_i \lambda_i f_i(x) \approx \frac{m}{t}$. Since $\lambda_i > 0$ and $-\lambda_i f_i(x) = \frac{1}{t} > 0$, we must have $f_i(x) < 0$.

Write $y^+ = y + s\Delta y_{\text{nt}}$. The line search consists of the following steps:

(i) Choose $s_1 \le 1$ so that $\lambda^+ \ge 0$.

(ii) Starting with $s = s_1$, $s \leftarrow \beta s$ until $f_i(x^+) < 0$. Call the result s_2.

(iii) Starting with $s = s_2$, $s \leftarrow \beta s$ until $\|r(y^+)\|_2 \le (1 - \alpha t)\|r(y)\|_2$.

What follows are fragments of the proof that this all works.

Let $t_{\max} = \inf\{t > 0 : \|r(y + t\Delta y_{\text{nt}})\|_2 > \|r(y)\|_2\}$. This is the first t for which moving along the Newton direction by this much will not improve the value of $\|r(y)\|_2$. Let

$$S := \{t : \|r(y + t\Delta y_{\text{nt}})\|_2 \le \|r(y)\|_2\} \supseteq [0, t_{\max}],$$

and assume that S is closed. Note that we have $Dr(y)\Delta y = -r(y)$ because it is the definition of the Newton increment.

$$\begin{aligned} r(y+t\Delta y) &= r(y) + \int_0^1 tDr(y+t\tau\Delta y)\Delta y d\tau \\ &= r(y) + tDr(y)\Delta y + \int_0^1 [Dr(y+t\tau\Delta y) - Dr(y)]t\Delta y d\tau \\ &= (1-t)r(y) + e \\ \|r(y+t\Delta y)\|_2 &\le (1-t)\|r(y)\|_2 + \|e\|_2 \end{aligned}$$

Now, if Dr is Lipschitz, say $\|Dr(y) - Dr(y)\|_2 \le L\|y - y'\|_2$, then

$$\begin{aligned} \|e\|_2 = \left\| \int_0^1 [Dr(y+t\tau\Delta y) - Dr(y)]t\Delta y d\tau \right\|_2 \\ &\le t\int_0^1 \|Dr(y+t\tau\Delta y) - Dr(y)\|_2 \|\Delta y\|_2 d\tau \\ &\le Lt^2 \int_0^1 \|\Delta y\|_2^2 \tau d\tau = \tfrac{1}{2}Lt^2\|\Delta y\|_2^2 \\ &= \tfrac{1}{2}Lt^2\|[Dr(y)^{-1}r(y)\|_2^2 \end{aligned}$$

Suppose as well that $\|[Dr(y)]^{-1}\|_2 \le K$ for all y. Then for all $t \in [0, 1 \wedge t_{\max}]$,

$$\|r(y+t\Delta y)\|_2 \le (1-t)\|r(y)\|_2 + \tfrac{1}{2}t^2LK^2\|r(y)\|_2^2. \tag{5}$$

This parabola in t is minimized at $\bar{t} = 1/(LK^2\|r(y)\|_2)$. It is clear that $1/LK^2$ is a critical value of $\|r(y)\|_2$. There are hence two cases: (i) $\|r\| > 1/(K^2L)$ and (ii) $\|r\| \le 1/(K^2L)$.

In the first case $\bar{t} < 1 \wedge t_{max}$, so $\bar{t}$ is a valid step size and we get

$$\begin{aligned} \|r(y+t\Delta y)\|_2 &\le (1-\bar{t})\|r(y)\|_2 + \tfrac{1}{2}\bar{t}^2LK^2\|r(y)\|_2^2 \\ &= \|r(y)\|_2 - \frac{1}{2K^2L} \\ &\le \|r(y)\|_2 - \alpha\bar{t}\frac{1}{K^2L} \\ &\le (1-\alpha\bar{t})\|r(y)\|_2 \end{aligned}$$

since $\bar{t} < 1$, $\alpha < \frac{1}{2}$, and $1/(K^2L) < \|r(y)\|_2$. Therefore the line search terminates for some $t \ge \beta\bar{t}$.

$$\|r(y+t\Delta y)\|_2 \le (1-\alpha\bar{t})\|r(y)\|_2 \le \|r(y)\|_2 - \frac{\alpha\beta}{K^2L}$$

so the reduction after the line search as at least $\frac{\alpha\beta}{K^2L}$. Therefore, after finitely many steps, we will end up in case (ii).

In the second case $\|r(y)\|_2 \leq 1/(K^2L)$ so

$$\|r(y + t\Delta y)\|_2 \leq (1-t)\|r(y)\|_2 + \tfrac{1}{2}t^2LK^2\|r(y)\|_2^2 \leq (1 - t + \tfrac{1}{2}t^2)\|r(y)\|_2.$$

From this it follows that $t_{\max} > 1$, and for $t = 1$ we have

$$\|r(y + t\Delta y)\|_2 \leq \tfrac{1}{2}\|r(y)\|_2 \leq (1 - \alpha t)\|r(y)\|_2$$

so the line search exits with $t = 1$. Finally, plugging $t = 1$ into the original inequality,

$$\|r(y + \Delta y)\|_2 \leq \frac{LK^2}{2}\|r(y)\|_2^2$$
$$\frac{K^2L}{2}\|r(y + \Delta y)\|_2 \leq \left(\frac{LK^2}{2}\|r(y)\|_2\right)^2$$

so we have quadratic convergence in case (ii).

Index