Optimization

Course Notes by Ohannes Karakashian
Transcribed and Annotated by Gregory Ƶitelli

Lagrange Multipliers

Suppose we want to minimize the function $2x^2 + 3y^2 + z^2$ subject to the constraints $x+y+z = 1$ and $2x-y+3z = 4$. We say this is a constrained optimization problem with equality constraints. The typical approach to a problem like this is to set up a Lagrange function $\Lambda(x, \lambda_1, \lambda_2) = f(x, y, z) + \lambda_1 g_2 + \lambda_2 g_2$, where $f(x, y, z)$ is the function to be minimized and g_1, g_2 are the constraint functions. Then the extremal points for our original function f will occur when the partials of Λ are all zero. This can be made precise in the following sense.

Theorem (Lagrange Multipliers)**.** *Let $\Omega = V_1 \times V_2$ be an open domain, with V_2 complete, $\varphi : \Omega \to V_2$ a $C^1(\Omega)$ constraint function. We let $U = \{(u_1, u_2) \in \Omega : \varphi(u_1, u_2) = 0\} \subseteq \Omega$ be the constraint set. We assume that $\partial_2\varphi$ is invertible and has bounded inverse. If $f : \Omega \to \mathbb{R}$ is a function differentiable at some point $u \in U$, and if f has a local minimum at u with respect to the set U, then there exists $\Lambda(u) \in \mathcal{L}(V_2, \mathbb{R})$ such that*

$$f'(u) + \Lambda(u)\varphi'(u) = 0$$

Proof. We require that $\partial_2\varphi$ be bounded and invertible in order to use the implicit function theorem. We let $g'(u_1) = -\left(\partial_2\varphi(u_1, u_2)\right)^{-1} \partial_1\varphi(u_1, u_2)$, the implicit function. Next, define $G(u_1) = f(u_1, g(u_1))$ so that $G : V_1 \to \mathbb{R}$. Now if f has a local minimum at $u = (u_1, u_2)$ with respect to U, then G must have a local minimum at u_1 and so $G'(u_1) = 0$.

$$\begin{aligned} G'(u_1) &= \partial_1 f(u_1, u_2) + \partial_2 f(u_1, u_2) g'(u_1) \\ &= \partial_1 f(u_1, u_2) - \partial_2 f(u_1, u_2)\Big(\partial_2\varphi(u_1, u_2)\Big)^{-1} \partial_1\varphi(u_1, u_2) = 0 \end{aligned}$$

It is fairly trivial that

$$0 = \partial_2 f(u_1, u_2) - \partial_2 f(u_1, u_2)\Big(\partial_2 \varphi(u_1, u_2)\Big)^{-1} \partial_2 \varphi(u_1, u_2)$$

and so combining these two we have

$$0 = \nabla f(u) + \Lambda(u) \nabla \varphi(u)$$

where $\Lambda(u) = -\partial_2 f(u)\big(\partial_2 \varphi(u)\big)^{-1}$. ■

In the example, we can construct $\varphi : \mathbb{R}^3 \to \mathbb{R}^2$ to be our constraint function by letting $\varphi_1(x, y, z) = x + y + z - 1$ and $\varphi_2(x, y, z) = 2x - y + 3z - 4$. Then $\Omega = V_1 \times V_2 = \mathbb{R} \times \mathbb{R}^2$, and the second partial of φ (where the second variable is thought of as (y, z)) is an element of $\mathcal{L}(\mathbb{R}^2, \mathbb{R}^2)$. In fact, we can directly express the second partial of φ as

$$\begin{pmatrix} 1 & 1 \\ -1 & 3 \end{pmatrix}$$

which is certainly an invertible, bounded linear operator from $\mathbb{R}^2$ to $\mathbb{R}^2$. By the theorem, there exists a Lagrange function $\Lambda(u) \in \mathcal{L}(\mathbb{R}^2, \mathbb{R})$, which we can express as (λ_1, λ_2).

$$\begin{pmatrix} 4u_1 & 6u_2 & 2u_3 \end{pmatrix} + \begin{pmatrix} \lambda_1 & \lambda_2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 2 & -1 & 3 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}$$

This produces a system of five equations (the three above and the two constraints) in five unknowns (the original three variables and λ_1, λ_2), which will indicate where the extremal points might be. Note that Lagrange multipliers do not guarantee that all points found will be extremal, it is simply a necessary condition if an extremal point exists.

Exercise. Let $f(x) = \frac{1}{2} x^T A x - b^T x$, where $f : \mathbb{R}^n \to \mathbb{R}$ and A is symmetric positive definite. We would like to minimize f subject to $Cx = d$, where C is a given $m \times n$ matrix and we assume $m < n$. If f has a local minimum on the set U where $Cx = d$ is satisfied, then there must be a Lagrange multiplier where $f'(u) + \Lambda(u)\varphi'(u) = 0$, where, $\varphi(x) = Cx - d$. We decompose our region $\mathbb{R}^n$ into $\mathbb{R}^{n-m} \times \mathbb{R}^m$, and then $\Lambda \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}) = \mathbb{R}^m$. Then our constraints become $Au - b + C^T \lambda = 0$ and $Cu - d = 0$. In block matrix form, this becomes.

$$\begin{pmatrix} A & C^T \\ C & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix}$$

We would now like to know when the block matrix is solvable. A necessary condition is that C must have full rank.

Second Derivative

Definition. We say that a function $f : \Omega \to \mathbb{R}$ has a strict local minimum at a point $u \in \Omega$ if there exists an open neighborhood O containing u such that $f(v) < f(u)$ for all $v \in O \backslash \{u\}$.

Theorem. *Let $f : \Omega \to \mathbb{R}$ be differentiable on $\Omega \subseteq V$ and twice differentiable at some $u \in \Omega$. If f has a local minimum at u, then $f''(u)(w, w) \geq 0$ for all $w \in V$.*

Proof. Fix some $w \in V$, and suppose that f has a local minimum at u. Let $|t| < \frac{r}{\|w\|}$, then $f(u + tw) \geq f(u)$. Then

$$f(u + tw) = f(u) + f'(u)tw + \frac{1}{2}f''(u)(tw, tw) + \epsilon\|tw\|^2$$

By Fermat, $f'(u) = 0$, and so

$$\begin{aligned} 0 \leq f(u_t w) - f(u) &= \frac{1}{2}t^2 f''(u)(w, w) + t^2\epsilon(tw)\|w\|^2 \\ &= t^2\left(\frac{1}{2}f(u)(w, w) + \epsilon(tw)\|w\|^2\right) \end{aligned}$$

Now if $\frac{1}{2}f(u)(w, w)$ is negative, then $\epsilon(tw)\|w\|^2$ will eventually become smaller than it in norm since $\epsilon(tw) \to 0$ as $t \to 0$. However, this would make the entire right hand side negative, which contradicts the fact that it is greater than or equal to zero. Therefore, $\frac{1}{2}f(u)(w, w)$ must be nonnegative. ■

An identical argument can be used to show that a local maximum at a twice differentiable point implies that the second derivative is always negative, rather than positive.

As an example, consider $f(x) = \frac{1}{2}x^T Ax - b^T x$ with A symmetric. Suppose that f has a local minimum at u, then by the theorem, $f''(u)(w,w) \geq 0$ for all $w \in \mathbb{R}^n$. However we have already computed $f''(u)(w,w)$ as $w^T Aw$, and so we know that symmetric matrices are positive semidefinite since $w^T Aw \geq 0$. Consequently, if A has a negative eigenvalue then it cannot have a local minimum at any point.

Theorem. *Let $f : \Omega \to \mathbb{R}$ be differentiable at $u \in \Omega \subseteq V$, and $f'(u) = 0$.*

(i) If $f''(u)$ exists and there is some $\alpha > 0$ such that $f''(u)(w,w) \geq \alpha\|w\|^2$ for all $w \in V$ then f has a local minimum at u.

(ii) If f'' exists on Ω and there exists an open ball $B \subseteq \Omega$ centered at u such that $f''(v)(w,w) \geq 0$ for all $v \in B$ and $w \in V$, then f has a local minimum at u.

Proof. For part (i), by Taylor's theorem we have that for $w \in V$, $\|w\|$ small,

$$\begin{aligned} f(u+w) - f(u) &= f'(u)w + \frac{1}{2}f''(u)(w,w) + \epsilon(w)\|w\|^2 \\ &\geq \frac{1}{2}\alpha\|w\|^2 + \epsilon(w)\|w\|^2 \\ &= \left(\frac{1}{2}\alpha + \epsilon(w)\right)\|w\|^2 \end{aligned}$$

Then for $\|w\|$ small enough, we have that $f(u+w) - f(u) > 0$. ■

If A is a symmetric, positive definite matrix, then the problem of minimizing $f(x) = \frac{1}{2}x^T Ax - b^T x$ reduces to finding location of the zeros of $f'(u) = Au - b$. This is because the first condition of the preceding theorem is always satisfied, since $x^T Ax \geq \lambda_{\min}\|x\|^2$ and symmetric, positive definite matrices only have positive eigenvalues.

Convexity

Definition. Let $U \subseteq V$, then we say that U is convex if for all $u, v \in U$, $\alpha u + (1-\alpha)v \in U$ for any $\alpha \in [0,1]$. This is to say, U is convex if any two points can be connected by a straight line lying in U.

Definition. A function $f : U \to \mathbb{R}$, where U is a convex set, is called convex if

$$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha)f(v)$$

for all $\alpha \in [0,1]$.

Theorem. *Let $f : \Omega \to \mathbb{R}$, U convex, and suppose that $f'(u)$ exists and f has a local minimum at u with respect to U. Then $f'(u)(v-u) \geq 0$ for all $v \in U$.*

Proof. Let $v \in U$, and write $w = v - u$. Then since U is convex, $u + \theta w \in U$ for $0 \leq \theta \leq 1$. The fact that $f'(u)$ exists implies that

$$0 \leq \frac{f(u+\theta w) - f(u)}{\theta} = \frac{f'(u)\theta w}{\theta} + \frac{\epsilon(\theta w)\|\theta w\|}{\theta} = f'(u)w \pm \epsilon(\theta w)\|w\|$$

Since $\epsilon(\theta w) \to 0$, we have that $f'(u)(v-u) \geq 0$ for all $v \in U$. ■

Theorem. *Let $f : \Omega \to \mathbb{R}$ be differentiable on $\Omega \subseteq V$, with a set $U \subseteq \Omega$ convex.*

(i) f is convex on U if and only if $f(v) \geq f(u) + f'(u)(v-u)$ for all $v, u \in U$.

(ii) f is strictly convex on U if and only if $f(v) > f(u) + f'(u)(v-u)$ for all $v, u \in U$, $v \neq u$.

Proof. For the first part, let $u \neq v$ and $\theta \in (0,1)$. Then if f is convex, then

$$f(u + \theta(v-u)) \leq (1-\theta)f(u) + \theta f(v)$$

$$\frac{f(u+\theta(v-u)) - f(u)}{\theta} \leq f(v) - f(u)$$

Taking $\theta \to 0^+$, the left hand side goes to $f'(u)(v-u)$, and so we have our result.

Next, suppose that $f(v) \geq f(u) + f'(u)(v-u)$ holds for all $u, v \in U$, and let $\theta \in (0,1)$. Then if we let $w = u + \theta(u-v)$,

$$\begin{aligned} f(v) &\geq f(w) + f'(w)(v-w) \\ &= f(v+\theta(u-v)) + f'(v+\theta(u-v))(-\theta(u-v)) \\ &= f(v+\theta(u-v)) - \theta f'(v+\theta(u-v))(u-v) \end{aligned}$$

Similarly,

$$\begin{aligned} f(u) &\geq f(w) + f'(w)(u-w) \\ &\geq f(v+\theta(u-v)) - (1-\theta)f'(v+\theta(u-v))(u-v) \end{aligned}$$

Now multiply the first intequality by $1-\theta$ and the second by θ, and add them together. Then

$$(1-\theta)f(v)+\theta f(u) \geq f(v+\theta(v-u))$$

The proof of the second part follows similarly. ■

Theorem. *Let $f:\Omega \to \mathbb{R}$ be twice differentiable, and let $U \subseteq \Omega$ be convex.*

(i) f is convex on U if and only if $f''(u)(v-u,v-u) \geq 0$ for all $v,u \in U$.
(ii) If $f''(u)(v-u,v-u) > 0$ for all $v,u \in U$ with $u \neq v$, then f is strictly convex.

Proof. From Taylor's theorem,

$$f(v)-f(u)-f'(u)(v-u)=\frac{1}{2}f''(w)(v-u,v-u)$$

where $w \in (u,v)$, and so $w=u+\theta(v-u)$ for $\theta \in (0,1)$. Then

$$f(v)-f(u)-f'(u)(v-u)=\frac{1}{2\theta^2}f''(w)(w-u,w-u)$$

Then if $f''(u)(v-u,v-u) \geq 0$ for all $u,v \in U$, then $f(v) \geq f(u)+f'(u)(v-u)$, which by our previous result happens if and only if f is convex.

Now if $f''(u)(v-u,v-u) > 0$ then we have the same implication in one direction, that f is strictly convex. For the other direction for the first part, the proof is straightforward. ■

As an example, let $f(x)=\frac{1}{2}x^TAx-b^Tx$ with A symmetric. Then f is convex on $\mathbb{R}^n$ if and only if A is positive-semidefinite, and strictly convex if it is positive-definite. Note that any matrix can be written as $A=\frac{1}{2}(A+A^T)+\frac{1}{2}(A-A^T)$, where the first term is symmetric and the second is called skew-symmetric.

Theorem. *Let $f:\Omega \to \mathbb{R}$ with $U \subseteq \Omega$ convex.*

(i) If f is convex on U and has a local minimum at $u \in U$, then u is also the global minimum with respect to U.
(ii) If f is strictly convex on U then it has at most one minimum, and that minimum is strict.
(iii) If f is convex on U, and if $f'(u)$ exists for some $u \in U$, then f has a minimum at u with respect to U if and only if $f'(u)(v-u) \geq 0$ for all $u,v \in U$.
(iv) If U is open then the preceding condition is equivalent to $f'(u)=0$.

Least Squares

One application of convexity is the least squares problem. Let $B \in \mathbb{R}^{m\times n}$ and $c \in \mathbb{R}^m$. Then we want to find a vector $u \in \mathbb{R}^n$ such that $\|Bv-c\|_2$ is minimized. Let $f(v) = \frac{1}{2}\|Bv - c\|^2 - \frac{1}{2}\|c\|^2$, then minimizing f is the same as solving our original problem. Note that the function can be rewritten as $f(v) = \frac{1}{2}(B^TBv, v) - (B^Tc, v)$, and that B^TB is positive semidefinite. Then $f''(v) = B^TB$, and so by our previous theorems we know that f is convex. Then by our last theorem, the set of solutions of the least squares problem coincides with the set of solutions of $f'(u) = 0$. This amounts to solving $B^TBu - B^Tc = 0$, called the normal equations.

If we assume that $m \geq n$ and the rank of B is n, then it can be shown that B^TB is positive definite and $B^TBu - B^Tc = 0$ has a unique solution. Therefore, we have solved the least squares problem in the full rank case. When the rank of B is smaller than n, we will eventually see that there are infinitely many solutions.

Newton's Method

Let $f : V \to \mathbb{R}$ be a function, and suppose we want to find a root of $f(u) = 0$. Newton's method begins with some given u_0, and iterates a method to produce a sequence u_{k+1} where $f'(u_k)(u_{k+1} - u_k) = -f(u_k)$.

Assuming that a root u does exist, and that $f'(u)$ is invertible, then we can show that Newton's method exhibits quadratic convergence where

$$\|u - u_{k+1}\| \leq c\|u - u_k\|^2$$

However, the method requires that u_0 begin sufficiently close to u in order to converge at all. Consider, for example, the arctangent function, which has a simple root at zero. There are large enough values such that Newton's method will not converge, and will instead diverge to $\pm\infty$. Significant work has been put into using Newton's method enhanced to provide global convergence.

Computing f' can be very expensive. If $f : \mathbb{R}^n \to R$ and n is large (say even 100) then f' requires over fifty thousand partials to be computed. This leads to quasi-Newton methods, for example taking $f'(u_0)(u_{k+1} - u_k) = -g(u_k)$ where the slope is fixed throughout.

Projection Theorem

Theorem. *Let V be a Hilbert space, and let $U \subseteq V$ be a closed, convex subset. Then for any $w \in V$, there exists a unique element $u \in V$ such that $\|w-u\| = \inf_{v\in U}\|w-v\|$. We write $u = Pw$ and say that P is the projection from V onto U. Additionally, $\langle Pw - w, v - Pw\rangle \geq 0$ for all $v \in U$ and P is a contraction.*

Proof. The quantity $\rho = \inf_{v\in U}\|w-v\|$ exists and is finite since it is bounded below by zero. By definition of the infimum, there must be a minimizing sequence $u_n \in U$ such that $\rho \leq \|w-u_n\| < \rho + \frac{1}{n}$. If $\|w-u_n\| = \rho$ for some $n \in \mathbb{N}$ then u_n is the minimizer. Otherwise, we look at the sequence u_n and would like to show that it will converge to something inside of V. By the parellelogram law,

$$\|u_n - u_m\|^2 + 4\left\|w - \frac{1}{2}(u_n + u_m)\right\|^2 = 2\|w-u_m\|^2 + 2\|w-u_n\|^2$$

Since U is convex, $\frac{(}{u}_n + u_m) \in U$, and so the second term is greater than or equal to ρ^2. For large n and m, the right hand side approaches $4\rho^2$, and so we have that $\|u_n - u_m\|^2 \to 0$. This implies that u_n is a Cauchy sequence, and so it converges to some u in V. Now we simply need to show that $u \in U$, however this follows from the fact that U is closed. The uniqueness is trivial. ■

For arbitrary closed, convex sets, the projection operator P is not linear. If U is a closed subspace, however, then P is linear. In fact, the converse also holds, that if P is linear then it projects onto a closed subspace of V.

For example, consider the closed convex set $R^2_+ = \{(x,y) : x,y \geq 0\}$ contained in $\mathbb{R}^2$. Then the projection onto R^2_+ can be easily seen to be $P(x,y) = (\max\{x,0\}, \max\{y,0\})$.

The projection operator is always a contraction, so that $\|Pw_1 - Pw_2\| \leq \|w_1 - w_2\|$. The space U is a closed subspace if and only if its projection P is linear, which further happens if and only if the angle condition becomes an orthogonality condition $P(w-w,v) = 0$ for all $v \in U$.

One application of these kinds of projections is the finite element method. Suppose we have the PDE give by $-u'' = f(x)$ for $0 < x < 1$, and $u(0) = u(1) = 0$. The weak formulation of the problem involves finding some u' such that $(u', v') = (f, v)$ for all $v \in H^1_0$. Since H^1_0 is infinite dimensional, we approximate our weak solution u' by elements in the span of linear bumps inside of $[0,1]$. This approximation is done by projecting the solution down into this subspace.

Theorem (Riesz-Representation Theorem)**.** *Let V be a Hilbert space. If f is a bounded linear functional on V, then there exists a unique element $v \in V$ such that $f(w) = (w, v)$.*

In finite dimensional real spaces, linear operators $A : V \to W$ as seen as matrices have the property that $(A^T w, v)_V = (w, Av)_W$ for all $v \in V$ and $w \in W$ (assuming the inner products are the typical ones). The operator represented by A^T is the adjoint of A, a dual operator with this property. What about for infinite dimensional spaces? Well, using the Riesz representation theorem, we can prove that such an operator exists by defining the map $v \mapsto (w, A_v)$.

Suppose that $f : V \to \mathbb{R}$ is differentiable at a, so that $f'(a) \in \mathcal{L}(V, \mathbb{R})$. If V is a Hilbert space, then there exists some element $w \in V$ such that $f'(a)v = (w, v)_V$. We call this element w the gradient of f at a, and write $\nabla f(a)$. If V is $\mathbb{R}^n$, then the gradient is simply the vector consisting of the various partial derivatives.

If A is an $m \times n$ matrix, then the rank nullity theorem tells us that $\mathbb{R}^n = \ker A \oplus \text{ran} A^T$ and $\mathbb{R}^m = \text{ran} A \oplus \ker A^T$, where the sums are orthogonal direct sums. Such a decomposition means that not only do the linear combinations of elements in the two sets add together to span the whole space, but their is a unique representation of any elements with one entry in each space. The orthogonal direct sum means that the two sets in consideration are orthogonal.

Definition. We say that V_1 is orthogonal to V_2 if $(v_1, v_2) = 0$ for all $v_1 \in V_1$ and $v_2 \in V_2$.

Definition. Let V be a Hilbert space. We say that $u, v \in V$ are orthogonal if $(u, v) = 0$.

Definition. If U is a subset of V, the orthogonal complement $U^\perp$ is the set of elements orthogonal to all elements in U.

Note that $U^\perp$ is always a closed subspace, regardless of what U is. Consequently, $U^{\perp\perp}$ is actually the closed linear span of the set U.

Theorem. *Let U be a closed subspace of a Hilbert space V. Then $V = U \oplus U^\perp$, that is to say every element of V can be written as the sum of two unique elements, one in U and one in U^+.*

Definition. A function $f : V \to \mathbb{R}$ is coercive if $\lim_{\|u\|\to\infty} f(u) = +\infty$.

Theorem. *Let U be a nonempty, closed subset of $\mathbb{R}^n$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function. If U is bounded, then f attains both its supremum and infimum values on U. If U is unbounded and f is coercive, then it attains its infimum.*

As an example, if $f(x) = \frac{1}{2}x^T Ax - b^T x$ is symmetric positive-definite then it is coercive.

Elliptic Functionals

Theorem. *Every separable Hilbert space has a countable basis.*

Definition. A functional $f : V \to \mathbb{R}$ on a Hilbert space is called elliptic if

(i) it is continuously differentiable on V.
(ii) there exists $\alpha > 0$ such that for all $u, v \in V$,

$$(\nabla f(v) - \nabla f(u), v - u) \geq \alpha \|v - u\|^2$$

Elliptic functionals generalize quadratic functions $f(u) = \frac{1}{2}u^T Au - b^T u$, where A is a symmetric positive definite matrix. In that situation, α can be taken to be the smallest eigenvalue of A.

Theorem. *An elliptic functional $f : V \to \mathbb{R}$ is strictly convex and coercive. Furthermore, for all $u, v \in V$,*

$$f(v) - f(u) \geq (\nabla f(u), v - u) + \frac{\alpha}{2}\|v - u\|^2$$

Theorem. *If U is a nonempty, closed, convex subset of V and f is elliptic, then there is a unique point $u \in V$ such that $f(u) = \inf_{v\in U} f(v)$.*

Theorem. *Suppose U is convex and f is elliptic. Then an element $u \in U$ achieves the infimum of f on U if and only if for all $v \in U$,*

$$(\nabla f(u), v - u) \geq 0$$

Theorem. *A function which is twice differentiable on V is elliptic if and only if $f''(u)(w, w) \geq \alpha\|w\|^2$ for all $w \in V$.*

Relaxation Method

The relaxation method can be described as follows. Suppose $f : \mathbb{R}^n \to R$ is a function we wish to minimize over some set. Minimizing over $\mathbb{R}$ is fairly straightforward to do, but minimizing a function $f(u_1, \ldots, u_n)$ will be much more difficult. To simplify the problem, we choose some initial point $(u_1^1, u_2^1, \ldots, u_n^1) \in \mathbb{R}^n$, and then minimize $f(\chi, u_2^1, \ldots, u_n^1)$ using single variable methods. The minimizer in the first component becomes u_1^2, and then the process repeats minimizing over $f(u_1^2, \chi, u_3^1, \ldots, u_n^1)$. Once this has been done n times, we have a new point $(u_1^2, \ldots, u_n^2)$, which should be closer to the minima than the original point $(u_1^1, \ldots, u_n^1)$. This is repeated, using the same method, to iterate from $(u_1^k, \ldots, u_n^k)$ to $(u_1^{k+1}, \ldots, u_n^{k+1})$.

Theorem. *If $f : \mathbb{R}^n \to \mathbb{R}$ is elliptic, then the relaxation method converges for any starting vector u^0.*

We've already said that $f(v) = \frac{1}{2} v^T A v - b^T v$ is elliptic for A symmetric, positive-definite. Because of this, finding a minima of f is equivalent to finding a solution to $Au = b$. It turns out that using the relaxation method to find a solution to f is equivalent to Gauss Seidel method applied to $Au = b$. Consequently, the preceding theorem says that the Gauss Seidel method converges for the problem $Au = b$ when A is symmetric, positive-definite.

Gradient Method

The gradient method is similar to the relaxation technique, however rather than moving along the axis it chooses to move in the direction opposite of the gradient (down the steeping slope).

$$u_{k+1} = u_k - \rho_k \nabla f(u) k)$$

$$f(u_{k+1}) = \inf_{\alpha > 0} f(u_k - \alpha \nabla f(u_k))$$

Theorem. *Suppose f is elliptic, then the gradient method converges.*

Proof. Assume that $\nabla f(u_k) \neq 0$. Consider $\varphi_k : \rho \mapsto \varphi_k(\rho) = f(u_k - \rho \nabla f(u_k))$. It can be shown that φ_k is coercive and strictly convex, so that it has a unique minimum such that $\varphi_k'(\rho_k) = 0$. Then $\varphi_k'(\rho) = -(\nabla f(u_k - \rho \nabla f(u_k)), \nabla f(u_k))$, so that

$$0 = \varphi_k'(\rho_k) = (\nabla f(u_k - \rho_k \nabla f(u_k)), \nabla f(u_k)) = (\nabla f(u_{k+1}), \nabla f(u_k))$$

This shows that the gradient of successive iterants are orthogonal. Furthermore,

$$(\nabla f(u_{k+1}), u_{k+1} - u_k) = (\nabla f(u_{k+1}), -\rho_k \nabla f(u_k)) = 0$$

Then by previous inequalities we've shown, $f(u_k) - f(u_{k+1}) \geq \frac{\alpha}{2}\|u_k - u_{k+1}\|^2$, so that $f(u_k)$ is a decreasing sequence. On the other hand, the fact that the iterants are orthogonal means that $\|\nabla f(u_k)\| \leq \|\nabla f(u_k) - \nabla f(u_{k+1})\|$. Then the sequence $f(u_k)$ is bounded, and f is assumed to be elliptic, ∇f is uniformly continuous where we are concerned, and so $\lim_{k\to\infty} \|\nabla f(u_k)\| = 0$. Now write $\alpha\|u_k - u\|^2 \leq (\nabla f(u_k) - \nabla f(u), u_k - u) = (\nabla f(u_k), u_k - u) \leq \|\nabla f(u_k)\|\|u_k - u\|$ by the ellipticity, so that $\|u_k - u\| \to 0$. ∎

Theorem. *Let V be a Hilbert space, $f : V \to \mathbb{R}$ differentiable on V. Suppose there exists some positive constants α, M such that*

$$(\nabla f(v) - \nabla f(u), v - u) \geq \alpha\|v - u\|^2$$

for all $u, v \in V$, and $\|\nabla f(u) - \nabla f(v)\| \leq M\|v - u\|^2$ (so that f is elliptic). Suppose that ρ_k satisfies $0 < a \leq \rho_k << \frac{2d}{M^2}$, then the gradient method with step sizes ρ_k will converge. Moreover, there exists $\beta = \beta(\alpha, M, a, b) < 1$ such that $\|u_{k+1} - u\| \leq \beta\|u_k - u\|$.

Suppose that $f : V \to \mathbb{R}$ is such that

$$(\nabla f(v) - \nabla f(u), v - u) \geq \alpha\|v - u\|^2$$

$$\|\nabla f(u) - \nabla f(v)\| \leq M\|v - u\|^2$$

for all $u, v \in V$. Let $u_{k+1} = u_k - \rho_k \nabla f(u_k)$, for the gradient method with parameter ρ_k.

Theorem. *If* $0 < a \leq \rho_k \leq b < \frac{2\alpha}{M^2}$*, then* $\|u_{k+1} - u\| \leq \beta \|u_k - u\|$*, with* $\beta < 1$*.*

Suppose $f(v) = \frac{1}{2} v^T A v - b^T v$, with A symmetric, positive-definite. Let $\alpha = \lambda_1$ be the smallest eigenvalue, and $M = \lambda_n$ the largest. Typically, $\lambda_1 << \lambda_n$, so that $\frac{2\alpha}{M^2} = \frac{2\lambda_1}{\lambda_n^2}$ is very small. This enlarges the range of ρ_k's that we can use, so that $0 < a \leq \rho_k \leq b < \frac{2}{\lambda_n} = \frac{2\lambda_1}{\lambda_n^2} \cdot \frac{\lambda_n}{\lambda_1}$.

For quadratic f, $u_{k+1} = u_k - \rho_k(Au_k - b)$ implies that $u_{k+1} - u = u_k - u - \rho(Au_k - Au)$, which further implies that $\|u_{k+1} - u\| = \|(I - \rho_k A)(u_k - u)\| = \|I - \rho_k A\| \|u_k - u\|$. We would like to choose ρ_k so that $\|I - \rho_k A\| < 1$. If A is symmetric, then $I - \rho_k A$ is symmetric, and A is normal (commutes with its transpose). Furthermore, $\rho(A)$, the spectral radius, is simply $\|A\|$. We would like to minimize $\|I - \rho_k A\| = \rho(I - \rho_k A) = \max_{\lambda \in \sigma(A)} |1 - \rho_k \lambda|$, which is generally an impossible problem. We can create some kinds of bounds, however, using the fact that $\max_{\lambda \in \sigma(A)} |1 - \rho_k \lambda| \leq \max_{\lambda_1 \leq \lambda \leq \lambda_n} |1 - \rho_k \lambda| = \max\{|1 - \rho_k \lambda_1|, |1 - \rho_k \lambda_n|\}$. Then $0 < \rho_k < \frac{2}{\lambda_n}$, which implies that $|1 - \rho_k \lambda_1|, |1 - \rho_k \lambda_n| < 1$. The minimizer to this problem can be found to be $\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}$.

Conjugate Gradient Method

This method is often studied as a method for solving $Ax = b$ where A is symmetric, positive-definite. If A is not s.p.d., then we employ some kind of GMRES, QMR, etc. CGM is also a minimization algorithm. Given $f : \mathbb{R}^n \to \mathbb{R}$ elliptic, there is a unique global minimum u when $\nabla f(u) = 0$. We begin with a starting vector u_0, and employ the method of steepest decent to go from u_0 to u_1, so that $f(u_1) = \inf_\alpha f(u_0 + \alpha \nabla f(u_0))$. Note that when f is quadratic, $\alpha = -\frac{\|\nabla f(u_0)\|^2}{\nabla f(u_0)^T A \nabla f(u_0)}$ is the optimal α.

Subsequence points u_k are constructed as follows. At each step, let

$$G_k = \text{span}\{\nabla f(u_0), \nabla f(u_1), \ldots, \nabla f(u_k)\}$$

the subspace spanned by the gradients at each of our previous points. Then we pick the new point u_{k+1} as the minimizer of $\inf_{v \in u_k + G_k} f(v)$.

Lemma. *Assume* $f(v) = \frac{1}{2} v A^T v - b^T v$ *for* A *symmetric, positive-definite. Define* $\Delta_l = v_{l+1} - v_l \in G_l$*, so that* $\Delta_l = \sum_{i=0}^{l} \delta_{il} \nabla f(u_i)$ *for some scalars* δ_{il}*,* $0 \leq i \leq l \leq k$*.Assume that* $\nabla f(u_0), \ldots, \nabla f(u_k)$ *are nonzero, then* $\Delta_0, \ldots, \Delta_k$ *are conjugate with respect to* A*.*