CMO 1: Preliminaries

Eklavya Sharma

Contents

1 Central problem and algorithm template

Central Problem of the course 'Computational Methods of Optimization': Given an objective function $f : \mathbb{R}^d \mapsto \mathbb{R}$ and a constraint set $S \subseteq \mathbb{R}^d$, find $x^* = \operatorname{argmin}_{x \in S} f(x)$ and $f^* = f(x^*)$.

Example: for $\min_{x \in \mathbb{R}} (x - t)^2$, $x^* = t$ and $f^* = 0$.

All algorithms we develop to find x^* will follow this template:

Pick $x \in S$.
while x is not optimal **do**
    Pick another $x \in S$ such that $f(x)$ decreases.
end while
return x

---

2 Metric space

For any set S (we'll usually consider $S = \mathbb{R}^d$), $D : S \times S \mapsto \mathbb{R}$ is a distance function iff all of the following are true:

- $D(x, y) = 0 \iff x = y$.
- $D(x, y) \geq 0$.
- Symmetry: $D(x, y) = D(y, x)$.

- Triangle inequality: $D(x,y) + D(y,z) \geq D(x,z)$.

Theorem 1. *$D(x,y) = \|x-y\|$ is a distance function. Here*

$$\|x\| = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{d} x_i^2}$$

Theorem 2. *$D(x,y) = \sum_{i=1}^{d} |x_i - y_i|$ is a distance function.*

3 Neighborhood function and Open sets

Definition 1. *For $r > 0$ and $x \in \mathbb{R}^d$, $N_r(x) = \{z : D(x,z) < r\}$ is called a neighborhood of x of radius r.*

Definition 2. *$x \in \mathbb{R}^d$ is an interior point of S iff $\exists r > 0, N_r(x) \subseteq S$.*

Definition 3. *Let $x, y \in \mathbb{R}$.*

- $(x,y) = \{z : x < z < y\}$.
- $(x,y] = \{z : x < z \leq y\}$.
- $[x,y) = \{z : x \leq z < y\}$.
- $[x,y] = \{z : x \leq z \leq y\}$.

Definition 4. *S is an open set iff $\forall x \in S$, x is an interior point of S.*

Example 1. *$(0,1)$ is an open set but $[0,1)$ is not.*

Definition 5. *$x \in \mathbb{R}^d$ is a limit point of S iff $N_r(x) \cap S \neq \phi$.*

Example 2. *$0, \frac{1}{2}, 1$ are 3 of the limit points of $(0,1]$.*

Definition 6. *Closure of a set S is the set of all limit points of S.*

Definition 7. *A set S is closed iff all limit points of S lie in S.*

Example 3. *$[0,1]$ is a closed set.*

4 Limit and Bounds

Definition 8. *Let $[x_i]_{i \in \mathbb{N}}$ be an infinite sequence where $x \in \mathbb{R}^d$. Then*

$$\lim_{i \to \infty} x_i = x \iff \forall \epsilon > 0, \exists n, \forall i \geq n, \|x - x_i\| < \epsilon$$

Definition 9. *$S \subseteq \mathbb{R}^d$ is a bounded set iff $\exists M, \forall x \in S, \|x\| \leq M$.*

Definition 10. *For $x_i \in \mathbb{R}$, M is an upper bound of $[x_i]_{i \in \mathbb{N}}$ iff $\forall i, x_i \leq M$. A sequence with an upper bound is called an upper-bounded sequence.*

Definition 11. *g is a least upper bound (LUB) (of $[x_i]_{i \in \mathbb{N}}$) iff g is an upper bound and for every upper bound h, $g \leq h$.*

Example 4. *For $x_i = 1 - \frac{1}{i}$, LUB is 1.*

Theorem 3. *A monotonic bounded sequence has a limit.*

5 Continuity

Definition 12.

$$\lim_{x \to p} f(x) = q \iff \forall \epsilon > 0, \exists \delta > 0, \forall x \in N_\delta(p), f(x) \in N_\epsilon(q)$$

Definition 13. *f is continuous at x $\iff$ $\lim_{x \to p} f(x) = f(p)$. f is continuous over S $\iff$ f is continuous at all points $x \in S$.*

Theorem 4. *Let $S \subseteq \mathbb{R}^d$ be closed and bounded. Let $f(S) = \{f(x) : x \in S\}$. Let f be continuous over S. Then $f(S)$ is closed and bounded.*

For optimization problems, x^* is guaranteed to exist iff f is continuous and S is closed and bounded. Henceforth, we will assume S to be closed and bounded and assume functions to be continuous.

6 Asymptotics

$$a(x) \in o(b(x)) \iff \lim_{x \to x_0} \left| \frac{a(x)}{b(x)} \right| = 0$$

For example, at $x = 0$, $x^3 \in o(x^2)$.

If f is continuous at $x = p$, $f(x) = f(p) + o(1)$.

CMO 2: Taylor Series

Eklavya Sharma

Contents

1 Univariate Taylor Series

Let $f : [a, b] \mapsto \mathbb{R}$. Let $x, y \in [a, b]$.

Suppose f is differentiable k times. Then for some $z \in (x, y)$,

$$f(y) = \sum_{i=0}^{k-1} f^{(i)}(x) \frac{(y-x)^i}{i!} + f^{(k)}(z) \frac{(y-x)^k}{k!}$$

C^k is the set of all functions which are k-times differentiable and whose k^{th} derivative is continuous.

When $f^{(k)} \in C^k$,

$$f(y) = \sum_{i=0}^{k} f^{(i)}(x) \frac{(y-x)^i}{i!} + o(1) \frac{(y-x)^k}{k!}$$

Therefore, we can ignore the last term if x is close to y.

2 Multivariate Calculus

Definition 1. *Let $f : \mathbb{R}^m \mapsto \mathbb{R}^n$ be a function and $y = f(x)$. Then the Jacobian of y w.r.t. x is an n by m matrix where*

$$\left(\frac{\partial y}{\partial x}\right)_{i,j} = \frac{\partial y_i}{\partial x_j}$$

Theorem 1 (Chain rule)**.** *Let $y = f(x)$ and $z = g(y)$. Then*

$$\frac{\partial z}{\partial x} = \left(\frac{\partial z}{\partial y}\right)\left(\frac{\partial y}{\partial x}\right)$$

Definition 2. *For $f : \mathbb{R}^d \mapsto \mathbb{R}$, the gradient of f, denoted as ∇_f, is a d-dimensional vector defined as*

$$\nabla_f(x) = \left[\frac{\partial f(x)}{\partial x_i}\right]_{i=1}^{d}$$

For multivariate functions, $f \in C^1$ iff ∇_f exists and all components are continuous. Note that differentiability does not imply C_1.

Definition 3. *For $f : \mathbb{R}^d \mapsto \mathbb{R}$, the hessian of f, denoted as H_f, is a d by d matrix defined as*

$$\mathrm{H}_f(x)_{i,j} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

For multivariate function, $f \in C^2$ iff H_f exists and all its entries are continuous.

Theorem 2 (Proof omitted)**.** *When $f \in C^2$, H_f is symmetric.*

3 Multivariate Taylor Series

Let $g(t) = f(x + tu)$, where $t \in \mathbb{R}$ and $x, u \in \mathbb{R}^d$.

Theorem 3.

$$g'(t) = \nabla_f(x + tu)^T u \qquad \text{(when } f \in C^1\text{, by chain rule)}$$

$$g''(t) = u^T \mathrm{H}_f(x + tu)u \qquad \text{(when } f \in C^2\text{)}$$

Theorem 4. $f \in C^1 \implies g \in C^1$

Theorem 5. *If $f \in C^1$ and y is close to x,*

$$f(y) = f(x) + \nabla_f(x)^T(y - x) + o(\|y - x\|)$$

Proof. Let $g(t) = f(x + tu)$ and let $u = y - x$ be small. By applying univariate Taylor series on g at 0, we get

$$\begin{aligned}
&g(1) = g(0) + g'(\alpha), \text{ where } \alpha \in [0, 1] \\
&\Rightarrow f(x + u) = f(x) + \nabla_f(x + \alpha u)^T u \\
&\Rightarrow f(x + u) = f(x) + (\nabla_f(x) + o(1))^T u \qquad (\nabla_f \text{ is continuous and } u \text{ is small}) \\
&\Rightarrow f(y) = f(x) + \nabla_f(x)^T(y - x) + o(\|y - x\|)
\end{aligned}$$

□

Theorem 6. *If $f \in C^2$ and y is close to x,*

$$f(y) = f(x) + \nabla_f(x)^T(y - x) + \frac{1}{2}(y - x)^T \mathrm{H}_f(x)(y - x) + o(\|y - x\|^2)$$

Proof. Similar to previous theorem. □

CMO: Existence and Characterization of Minimum

Eklavya Sharma

Let $S \subseteq \mathbb{R}^d$ and $f : S \mapsto \mathbb{R}$.
x^* is a local minimum $\iff \exists r > 0, \forall x \in N_r(x^*) \cap S, f(x^*) \leq f(x)$.

We'll restrict our analysis in 2 ways:

- We'll only consider functions for which a global minimum exists. Here we'll discuss a sufficient condition for that.
- We'll only try to find a local minimum, since finding global minimum is difficult.

1 Necessary condition for local minimum of univariate function

Theorem 1. *If $f : \mathbb{R} \mapsto \mathbb{R}$ is differentiable, then x^* is the local minimum of $f \implies f'(x^*) = 0$.*

Proof. Let

$$h(t) = \frac{f(t) - f(x^*)}{t - x^*}$$

Then $f'(x^*) = \lim_{t \to x^*} h(t)$.

Suppose x^* is a local minimum in $(x - r, x + r)$. Then for $t \in (x - r, x)$, $h(t) \leq 0$ and for $t \in (x, x + r)$, $h(t) \geq 0$. Therefore, left derivative of f at x^* is non-positive and right derivative of f at x^* is non-negative. Since f is differentiable, left and right derivatives are equal. Therefore, $f'(x^*) = 0$. □

Theorem 2. *Let f be a C^2 function and x^* be a local minimum. Then $f''(x^*) \geq 0$.*

Proof. Using Taylor series near x^*, we get

$$f(x) = f(x^*) + (x - x^*)f'(x^*) + \frac{1}{2}(x - x^*)^2 f''(x^*) + o((x - x^*)^2)$$

$$\implies 0 \leq f(x) - f(x^*) = \frac{1}{2}(x - x^*)^2 f''(x^*) + o((x - x^*)^2)$$

For this to hold true for all x near x^*, $f''(x^*) \geq 0$. □

2 Characterization of functions which have a minimum

Consider a function from $\mathbb{R}^d$ to $\mathbb{R}$. Global minimum exists iff f is lower-bounded.

Definition 1.

$$\lim_{\|x\|\to\infty} f(x) = \infty \iff \forall F > 0, \exists M > 0, \forall x \in \mathbb{R}^d, (\|x\| > M \implies f(x) \geq F)$$

If $\lim_{\|x\|\to\infty} f(x) = \infty$*, then* f *is called a* ***coercive*** *function.*

Theorem 3 (Weierstrass' theorem)**.** *If a continuous function's domain is closed and bounded, the function has a global minimum and maximum.*

Theorem 4.

$$\lim_{\|x\|\to\infty} f(x) = \infty \wedge f \text{ is continuous} \implies f \text{ has global minimum}$$

Proof. Consider $F = f(0)$. Let $S_1 = \{x : \|x\| > M\}$ and $S_2 = \{x : \|x\| \leq M\}$.

Since f is coercive, $\forall x \in S_1, f(0) \leq f(x)$. By Weierstrass' theorem, a global minimum exists in S_2. Let it be x^*. Therefore, $f(x^*) \leq f(0)$. Therefore, x^* is a global minimum of $\mathbb{R}^d$. □

3 Sufficient condition for local minimum of univariate function

Theorem 5. $f'(x_0) = 0 \wedge f''(x_0) > 0 \implies x_0$ *is local minimum.*

Proof.

$$f(x) - f(x_0) = \frac{1}{2}(x - x^*)^2 f''(x^*) + o((x - x^*)^2)$$

In the neighborhood of x_0, the small-o term is negligible, so the $f''(x^*)$ makes $f(x) - f(x_0)$ positive. Therefore, x_0 is a local minimum in that neighborhood. □

4 Necessary condition for local minimum of multivariate function

Theorem 6. *Let* $f : \mathbb{R}^d \mapsto \mathbb{R}$ *be a differentiable function. Let* x^* *be a local minimum of* f*. Then* $\nabla_f(x^*) = 0$*.*

Proof. Let $u \in \mathbb{R}^d$ and $t \in \mathbb{R}$.

$$x^* + tu \in N_r(x^*) \iff \|tu\| < r \iff |t| \leq \frac{r}{\|u\|} \iff t \in N_{\frac{r}{\|u\|}}(0)$$

Let $g(t) = f(x^* + tu)$.

$$
\begin{aligned}
& x^* \text{ is local minimum of } f \\
& \Rightarrow \forall x \in N_r(x^*), f(x^*) \leq f(x) \\
& \Rightarrow \forall t \in N_{\frac{r}{\|u\|}}(0), f(x^*) \leq f(x^* + tu) \\
& \Rightarrow \forall t \in N_{\frac{r}{\|u\|}}(0), g(0) \leq g(t) \\
& \Rightarrow g \text{ has local minimum at } 0 \\
& \Rightarrow g'(0) = 0 \\
& \Rightarrow \nabla_f(x^*)^T u = 0
\end{aligned}
$$

Since this is true for all $u \in \mathbb{R}^d$, $\nabla_f(x^*) = 0$. □

Theorem 7. *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a differentiable function. Let x^* be a local minimum of f. Then $\mathrm{H}_f(x^*)$ is positive semi-definite.*

Proof. Similar to above proof. Use the fact that if g has a local minimum at 0, then $g''(0) \geq 0$. □

5 Sufficient condition for local minimum of multivariate function

Theorem 8. *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a differentiable function. Let $\nabla_f(x_0) = 0$ and $\mathrm{H}_f(x_0)$ be positive definite. Then x_0 is a local minimum of f.*

Proof. Proof follows directly from Taylor series. □

CMO: Convex functions

Eklavya Sharma

Definition 1 (Convex set)**.** *Let $S \subseteq \mathbb{R}^d$, S is convex iff*

$$\forall u, v \in S, \forall \alpha \in (0,1), (1-\alpha)u + \alpha v \in S$$

Definition 2 (Convex function)**.** *Let $f : S \mapsto \mathbb{R}$, where S is convex. f is convex iff*

$$\forall \alpha \in (0,1), \forall u, v \in S, f((1-\alpha)u + \alpha v) \leq (1-\alpha)f(u) + \alpha f(v)$$

Theorem 1. *f is convex* $\iff \big(\forall u, v \in S, \nabla_f(u)^T(v-u) \leq f(v) - f(u)\big)$

Proof.

f is convex
$\Rightarrow f((1-\alpha)u + \alpha v) \leq (1-\alpha)f(u) + \alpha(v)$
$\Rightarrow f(u + \alpha(v-u)) \leq f(u) + \alpha(f(v) - f(u))$
$\Rightarrow \dfrac{1}{\alpha}(f(u + \alpha(v-u)) - f(u)) \leq f(v) - f(u)$

Let $g(\alpha) = f(u + \alpha(v-u))$.

f is convex
$\Rightarrow \dfrac{g(\alpha) - g(0)}{\alpha} \leq g(1) - g(0)$
$\Rightarrow \lim\limits_{\alpha \to 0} \dfrac{g(\alpha) - g(0)}{\alpha} \leq \lim\limits_{\alpha \to 0}(g(1) - g(0))$
$\Rightarrow g'(0) \leq g(1) - g(0)$
$\Rightarrow \nabla_f(u)^T(v-u) \leq f(v) - f(u)$

Suppose $\forall u, v \in S, \nabla_f(u)^T(v-u) \leq f(v) - f(u)$.

For any arbitrarily chosen x_1 and x_2 ($x_1 \neq x_2$), let $x = (1-\alpha)x_1 + \alpha x_2$. Then $x_1 - x = \alpha(x_1 - x_2)$ and $x_2 - x = (1-\alpha)(x_2 - x_1)$.

Setting $u = x$ and $v = x_1$, we get

$$\nabla_f(x)^T \alpha(x_1 - x_2) \leq f(x_1) - f(x)$$

Setting $u = x$ and $v = x_2$, we get

$$-\nabla_f(x)^T(1-\alpha)(x_1 - x_2) \leq f(x_2) - f(x)$$

Adding these equations with weights $1-\alpha$ and α, we get

$$\begin{aligned}
&(1-\alpha)\,\nabla_f(x)^T\alpha(x_1-x_2) - \alpha\,\nabla_f(x)^T(1-\alpha)(x_1-x_2)\\
&\quad\leq (1-\alpha)(f(x_1)-f(x)) + \alpha(f(x_2)-f(x))\\
&\Rightarrow 0 \leq (1-\alpha)f(x_1) + \alpha f(x_2) - f(x)\\
&\Rightarrow f((1-\alpha)x_1 + \alpha x_2) \leq (1-\alpha)f(x_1) + \alpha f(x_2)\\
&\Rightarrow f \text{ is convex}
\end{aligned}$$

□

Theorem 2. *If f is convex, and x^* is a local minimum, then x^* is also a global minimum.*

Proof. For all $x \in \mathbb{R}^d$,

$$0 = f(x^*)^T(x-x^*) \leq f(x) - f(x^*)$$

□

Theorem 3 (Proof omitted)**.** *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ and $f \in C^2$. Then f is convex iff H_f is positive semi-definite.*

CMO: Minimizing a quadratic function

Eklavya Sharma

In many search algorithms, given the current point x, we choose the next point as $x+\alpha u$, where u is a descent direction (i.e. $\nabla_f(x)^T u \leq 0$) and $\alpha > 0$.

The strategy of choosing α as $\text{argmin}_{\alpha>0} f(x+\alpha u)$, is called **exact line search**.

1 Quadratic function

$$f(x) = \frac{1}{2}x^T Q x - d^T x$$

where Q is symmetric and positive definite.

$$\nabla_f(x) = Qx - d$$

$$\mathrm{H}_f(x) = Q$$

Since the hessian is positive definite, f is convex. So a local minimum is also a global minimum.

Define $x^* = Q^{-1}d$ (Q^{-1} exists because Q is positive definite). We find that x^* is a local minimum because it satisfies the sufficient conditions for it.

$$f(x^*) = -\frac{1}{2}{x^*}^T Q x^*$$

Although we have a closed form solution for x^*, this is sometimes not usable, since finding Q^{-1} takes $O(d^3)$ time, which can be too much if Q is large.

We will therefore explore descent-based methods to compute x^*.

2 Descent-based minimization of quadratic function

Let $u = \nabla_f(x) \neq 0$. Therefore, $u = Q(x - x^*)$.

Let $g(\alpha) = f(x - \alpha u)$.

$$g'(\alpha) = -u^T \nabla_f(x - \alpha u) = -u^T Q(x - \alpha u - x^*) = u^T(\alpha Q u - u)$$

Setting $g'(\alpha)$ to 0, we get

$$\alpha^* = \frac{\|u\|^2}{u^T Q u}$$

Since Q is positive definite, $\alpha^* > 0$.

$g''(\alpha) = u^T Q u > 0$, so α^* is a local minimum of g. Since $g''(\alpha) > 0$ for all α, g is convex, so α^* is a global minimum of g.

Apply Taylor series to find $f(x - \alpha^* u)$ around x,

$$\begin{aligned}
&f(x - \alpha^* u) = f(x) + \nabla_f(x)^T(-\alpha^* u) + \frac{1}{2}(-\alpha^* u)^T \,\mathrm{H}_f(x)(-\alpha^* u) \\
&\implies f(x) - f(x - \alpha^* u) = \alpha^* \,\nabla_f(x)^T u - \frac{(\alpha^*)^2}{2} u^T Q u \\
&= \left(\frac{\|u\|^2}{u^T Q u}\right)\|u\|^2 - \frac{1}{2}\left(\frac{\|u\|^2}{u^T Q u}\right)^2 u^T Q u \\
&= \frac{1}{2}\frac{\|u\|^4}{u^T Q u}
\end{aligned}$$

Apply Taylor series to find $f(x)$ around x^*,

$$\begin{aligned}
&f(x) = f(x^*) + \nabla_f(x^*)^T(x - x^*) + \frac{1}{2}(x - x^*)^T \,\mathrm{H}_f(x - x^*) \\
&\implies f(x) - f(x^*) = \frac{1}{2}(x - x^*)^T Q (x - x^*) = \frac{u^T Q^{-1} u}{2}
\end{aligned}$$

Before we can analyze the convergence of a descent-based algorithm to minimize f, we must look at an important result – Kantorovich's inequality.

Theorem 1 (Kantorovich's inequality)**.** *Let Q be a symmetric positive definite matrix. Let λ_1 and λ_d be its maximum and minimum eigenvalues respectively. Then*

$$\frac{\|u\|^4}{(u^T Q u)(u^T Q^{-1} u)} \geq \frac{4\lambda_1 \lambda_d}{(\lambda_1 + \lambda_d)^2}$$

Let $x^{(k+1)} = x^{(k)} - \alpha u$. Let $E(x) = f(x) - f(x^*)$. Then

$$\begin{aligned}
&\frac{E(x^{(k+1)})}{E(x^{(k)})} \\
&= 1 - \frac{f(x^{(i)}) - f(x^{(i+1)})}{f(x^{(i)}) - f(x^*)} \\
&= 1 - \frac{\|u\|^4}{(u^T Q u)(u^T Q^{-1} u)} \\
&\leq 1 - \frac{4\lambda_1 \lambda_d}{(\lambda_1 + \lambda_d)^2} && \text{(by Kantorovich's inequality)} \\
&\leq \left(\frac{\lambda_1 - \lambda_d}{\lambda_1 + \lambda_d}\right)^2
\end{aligned}$$

Therefore, E linearly converges to 0. We know that linear convergence is very fast, so this is a good descent method.

CMO: Minimizing a function with bounded hessian

Eklavya Sharma

Objective: Minimize a C^2 function $f : \mathbb{R}^d \mapsto \mathbb{R}$ for which $AI - \mathrm{H}_f(x)$ and $\mathrm{H}_f(x) - aI$ are positive semi-definite for all $x \in \mathbb{R}^d$ $(0 < a \leq A)$.

The trick we'll use is to lower-bound and upper-bound f.

Let $u = \nabla_f(x^{(i)})$. Let

$$f_l(x) = f(x^{(i)}) + u^T(x - x^{(i)}) + \frac{a}{2}\|x - x^{(i)}\|^2$$

$$f_h(x) = f(x^{(i)}) + u^T(x - x^{(i)}) + \frac{A}{2}\|x - x^{(i)}\|^2$$

By using Taylor series on f at $x^{(i)}$, we get that $\forall x \in \mathbb{R}^d, f_l(x) \leq f(x) \leq f_h(x)$.

Lemma 1.

$$f_l^* = \min_x f_l(x) = f(x^{(i)}) - \frac{\|u\|^2}{2a}$$

Proof sketch. Set $\nabla_{f_l}(x) = 0$ and solve for x. □

Lemma 2. *Let h_1 and h_2 be 2 functions such that $\forall x \in \mathbb{R}^d, h_1(x) \leq h_2(x)$. Let $h_1^* = \min_x h_1(x)$ and $h_2^* = \min_x h_2(x)$. Then $h_1* \leq h_2^*$.*

Proof. Let $x_2 = \mathrm{argmin}_x\, h_2(x)$. Then $h_1^* \leq h_1(x_2) \leq h_2(x_2) = h_2^*$. □

Let $x^* = \mathrm{argmin}_x\, f(x)$. Let $E(x) = f(x) - f(x^*)$.

Lemma 3.

$$E(x^{(i)}) \leq \frac{\|u\|^2}{2a}$$

Proof sketch. By lemma 2, $f_l^* \leq f(x^*)$. Now use lemma 1 to substitute f_l^*. □

Let $x^{(i+1)} = x^{(i)} - \frac{u}{A}$. (It can be proven that $x^{(i+1)}$ minimizes f_h, but we're not interested in that fact.)

Lemma 4.

$$E(x^{(i)}) - E(x^{(i+1)}) \geq \frac{\|u\|^2}{2A}$$

Proof.

$$
\begin{aligned}
f(x^{(i+1)}) &\leq f_h(x^{(i+1)}) && (f_h \text{ upper-bounds } f) \\
&= f(x^{(i)}) + u^T(x^{(i+1)} - x^{(i)}) + \frac{A}{2}\|x^{(i+1)} - x^{(i)}\|^2 \\
&= f(x^{(i)}) - \frac{\|u\|^2}{A} + \frac{A}{2}\frac{\|u\|^2}{A^2} \\
&= f(x^{(i)}) - \frac{\|u\|^2}{2A}
\end{aligned}
$$

$$
\implies E(x^{(i)}) - E(x^{(i+1)}) = f(x^{(i)}) - f(x^{(i+1)}) \geq \frac{\|u\|^2}{2A}
$$

□

Therefore,

$$
\frac{E(x^{(i+1)})}{E(x^{(i)})} = 1 - \frac{E(x^{(i)}) - E(x^{(i+1)})}{E(x^{(i)})} \leq 1 - \frac{a}{A}
$$

This proves the convergence of our algorithm.

CMO: Goldstein and Wolfe optimization

Eklavya Sharma

Definition 1. *C^1_L is the subset of C^1 functions for which*

$$\| \nabla_f(x) - \nabla_f(z)\| \leq L\|x - z\|$$

This is called the Lipschitz condition.

Objective: Minimize a lower-bounded C^1_L function $f : \mathbb{R}^d \mapsto \mathbb{R}$.

Contents

1 Goldstein and Wolfe conditions

Let u be a direction of decrease at $x^{(i)}$ (i.e. $\nabla_f(x^{(i)})^T u < 0$). Our descent algorithm will repeatedly choose a direction of descent (not necessarily steepest descent) and move in that direction with magnitude α.

Unlike the previous algorithms we saw, we'll not necessarily pick α as $\text{argmin}_{\alpha>0} f(x+\alpha u)$. This is called **inexact line search**. But this doesn't mean we can pick α arbitrarily. We still have to be smart about picking α to guarantee (quick) convergence. There are 2 famous ways of picking α: by the Goldstein conditions and the Wolfe conditions.

Let $g(\alpha) = f(x^{(i)} + \alpha u)$. Therefore, $g'(0) = \nabla_f(x^{(i)})^T u < 0$. Also, g is lower bounded because f is lower-bounded.

Draw a line which passes through $(0, g(0))$ with slope $m_1 g'(0)$, where $0 < m_1 < 1$ (note that the slope is negative). Let $h_1(\alpha) = g(0) + m_1 g'(0)\alpha$ be that line. Let $t(\alpha) = h_1(\alpha) - g(\alpha)$.

Lemma 1. *t has a positive zero. Let $\overline{\alpha}_1$ be the smallest positive zero. Then t is positive in the interval $(0, \overline{\alpha}_1)$. Formally,*

$$\exists \overline{\alpha}_1 > 0, (t(\overline{\alpha}_1) = 0 \wedge (\forall \alpha \in (0, \overline{\alpha}_1), t(\alpha) > 0))$$

Proof. Let f^* be the minimum value of f.

$$h_1(\alpha) - g(\alpha) < 0 \Leftarrow h_1(\alpha) - f^* < 0 \iff \alpha > \frac{f^* - g(0)}{m_1 g'(0)} > 0$$

Therefore, there is an α for which $t(\alpha) < 0$.

Since $g \in C^1$, by Taylor series, we get that for very small positive α,

$$g(\alpha) = g(0) + g'(0)\alpha + \alpha o(1)$$

$$\implies t(\alpha) = \alpha((1 - m_1)(-g'(0)) + o(1)) > 0$$

Therefore, there is an α for which $t(\alpha) > 0$.

Since g is continuous, by the intermediate value theorem, there must be an $\overline{\alpha}_1 > 0$ for which $t(\overline{\alpha}_1) = 0$. Without loss of generality, assume that $\overline{\alpha}_1$ is the smallest positive zero of t. Since $t(\alpha) > 0$ for small positive α, $t(\alpha) > 0$ for all $\alpha \in (0, \overline{\alpha}_1)$. □

In our descent algorithm, if we choose α from the interval $(0, \overline{\alpha}_1)$, then $g(0) = h_1(0) > h_1(\alpha) > g(\alpha)$. This means that $f(x^{(i)}) > f(x^{(i)} + \alpha u)$, which is what we required.

However, the decrease may be too small, especially if α is very close to 0. To counteract this, we'll impose another condition on α. We have 2 choices here.

1.1 Goldstein condition

Let $h_2(\alpha) = g(0) + m_2 g'(0)\alpha$, where $0 < m_1 < m_2 < 1$. Therefore, $h_2 - g$ has a smallest positive zero $\overline{\alpha}_2$. Also, $\overline{\alpha}_2 < \overline{\alpha}_1$. We'll chose α from the interval $(\overline{\alpha}_2, \overline{\alpha_1})$. This is called the Goldstein condition for choosing α.

1.2 Wolfe condition

Choose an $\alpha \in (0, \overline{\alpha}_1)$ such that $g'(\alpha) \geq m_3 g'(0)$, where $m_3 \in (0, 1)$. This is called the Wolfe condition.

Theorem 2. *If $m_3 \geq m_1$, it's possible to satisfy the Wolfe condition.*

Proof. Suppose we choose $\widehat{\alpha} \in (0, \overline{\alpha}_1)$. Since g is differentiable, by mean value theorem, we get

$$\exists \alpha \in [\widehat{\alpha}, \overline{\alpha}_1], g'(\alpha)(\overline{\alpha}_1 - \widehat{\alpha}) = g(\overline{\alpha}_1) - g(\widehat{\alpha})$$

Combine the above result with $g(\widehat{\alpha}) < h_1(\widehat{\alpha})$ and $g(\overline{\alpha}) = h_1(\overline{\alpha})$ to get $g'(\alpha) > g'(0)m_1$.

If we choose $m_3 \geq m_1$, then $g'(0)m_3 \leq g'(0)m_1 < g'(\alpha)$. Therefore, the Wolfe condition is satisfied for some $\alpha \in (0, \overline{\alpha}_1)$. □

2 Convergence of Wolfe condition

$$
\begin{aligned}
& g'(\alpha) \geq m_3 g'(0) && \text{(by Wolfe condition)} \\
& \Rightarrow \nabla_f(x^{(i)} + \alpha u)^T u \geq m_3 \nabla_f(x^{(i)})^T u \\
& \Rightarrow \left(\nabla_f(x^{(i)} + \alpha u) - \nabla_f(x^{(i)})\right)^T u \geq -(1 - m_3) \nabla_f(x^{(i)})^T u \\
& && \text{(subtract } \nabla_f(x^{(i)})^T u \text{ from both sides)} \\
& \Rightarrow \left\|\nabla_f(x^{(i)} + \alpha u) - \nabla_f(x^{(i)})\right\| \geq -(1 - m_3) \nabla_f(x^{(i)})^T u \\
& && \text{(both sides were +ve. Apply Cauchy-Schwarz inequality)} \\
& \Rightarrow L\alpha \|u\|^2 \geq -(1 - m_3) \nabla_f(x^{(i)})^T u && \text{(Lipschitz condition)} \\
& \Rightarrow \alpha \geq \frac{-(1 - m_3) \nabla_f(x^{(i)})^T u}{L\|u\|^2}
\end{aligned}
$$

$$
\begin{aligned}
& g(\alpha) < h_1(\alpha) = g(0) + m_1 g'(0)\alpha \\
& \Rightarrow f(x^{(i+1)}) < f(x^{(i)}) + m_1 \nabla_f(x^{(i)})^T u \alpha \\
& \Rightarrow f(x^{(i)}) - f(x^{(i+1)}) > \frac{m_1(1 - m_3)}{L} \left(\frac{\nabla_f(x^{(i)})^T u}{\|u\|}\right)^2
\end{aligned}
$$

Let $\nabla_f(x^{(i)})^T u = -\cos\theta_i \left\|\nabla_f(x^{(i)})\right\| \|u\|$. We'll impose another constraint: we'll choose u to not just be the descent direction, but also in a way that $\cos\theta_i$ is lower-bounded by a positive constant.

$$
f(x^{(i)}) - f(x^{(i+1)}) \geq \frac{m_1(1 - m_3)}{L} \cos^2\theta_i \| \nabla_f(x^{(i)})\|^2
$$

Summing i from 0 to $T - 1$, we get

$$
\forall T, f(x^{(i)}) - f^* \geq f(x^{(0)}) - f(x^{(T)}) \geq \frac{m_1(1 - m_3)}{L} \sum_{i=0}^{T-1} \cos^2\theta_i \left\|\nabla_f(x^{(i)})\right\|^2
$$

$\therefore \sum_{i=0}^{\infty} \cos^2\theta_i \left\|\nabla_f(x^{(i)})\right\|^2$ is a convergent series. So for $i \to \infty$, $\nabla_f(x^{(i)}) \to 0$.

Therefore, for $i \to \infty$, $x^{(i)}$ approaches a stationary point. Therefore, the descent algorithm which uses Wolfe condition converges to a stationary point, which would hopefully be a local minimum.

3 Alternate Characterization of C_L^1

Let $f \in C_L^1$. Let $g(\alpha) = f(x + \alpha(y - x))$. Then $g'(\alpha) = \nabla_f(x + \alpha(y - x))^T(y - x)$. Therefore, $g(0) = f(x)$, $g(1) = f(y)$ and $g'(0) = \nabla_f(x)^T(y - x)$.

$$
\int_0^1 (g'(\alpha) - g'(0))d\alpha = f(y) - f(x) - \nabla_f(x)^T(y - x)
$$

$$
\begin{aligned}
&\left|f(y) - f(x) - \nabla_f(x)^T(y-x)\right| \\
&= \left|\int_0^1 (g'(\alpha) - g'(0))d\alpha\right| \\
&\leq \int_0^1 |g'(\alpha) - g'(0)|d\alpha \\
&= \int_0^1 \left|\left(\nabla_f(x+\alpha(y-x)) - \nabla_f(x)\right)^T(y-x)\right| d\alpha \\
&\leq \int_0^1 \|\nabla_f(x+\alpha(y-x)) - \nabla_f(x)\|\,\|y-x\|d\alpha \qquad \text{(Cauchy-Schwarz inequality)} \\
&\leq \int_0^1 L\alpha\|y-x\|^2 d\alpha \qquad \text{(Lipschitz condition)} \\
&= \frac{L}{2}\|y-x\|^2
\end{aligned}
$$

4 Convergence of Goldstein condition

Let $u = \nabla_f(x^{(i)})$ and $x^{(i+1)} = x^{(i)} - \alpha u$.

Let $g(\alpha) = f(x^{(i)} - \alpha u)$. Then $g'(0) = -\nabla_f(x^{(i)})^T u = -\|u\|^2$.

$h_1(\alpha) = g(0) + m_1 g'(0)\alpha = f(x^{(i)}) - \alpha m_1\|u\|^2$. Similarly $h_2(\alpha) = f(x^{(i)}) - \alpha m_2\|u\|^2$.

$$
\begin{aligned}
&h_2(\alpha) \leq g(\alpha) \leq h_1(\alpha) \\
&\Rightarrow f(x^{(i)}) - m_2\alpha\|u\|^2 \leq f(x^{(i+1)}) \leq f(x^{(i)}) - m_1\alpha\|u\|^2 \\
&\Rightarrow m_1\alpha\|u\|^2 \leq f(x^{(i)}) - f(x^{(i+1)}) \leq m_2\alpha\|u\|^2
\end{aligned}
$$

$$
\begin{aligned}
&f(x^{(i)}) - f(x^{(i+1)}) + \nabla_f(x^{(i)})^T(x^{(i+1)} - x^{(i)}) \\
&\leq m_2\alpha\|u\|^2 + \nabla_f(x^{(i)})^T(x^{(i+1)} - x^{(i)}) \\
&= m_2\alpha\|u\|^2 - \alpha\|u\|^2 \\
&= -(1-m_2)\alpha\|u\|^2
\end{aligned}
$$

Therefore, by Lipschitz condition,

$$
\begin{aligned}
&(1-m_2)\alpha\|u\|^2 \leq \frac{L}{2}\|x^{(i+1)}-x^{(i)}\|^2 = \frac{L\alpha^2\|u\|^2}{2} \\
&\implies \alpha \geq \frac{2(1-m_2)}{L} \\
&\implies \frac{2(1-m_2)m_1}{L}\|u\|^2 \leq m_1\alpha\|u\|^2 \leq f(x^{(i)}) - f(x^{(i+1)}) \\
&\implies \forall T, \frac{2(1-m_2)m_1}{L}\sum_{i=0}^{T-1}\left\|\nabla_f(x^{(i)})\right\|^2 \leq f(x^{(0)}) - f(x^{(T)}) \leq f(x^{(0)}) - f^* \\
&\implies \forall T, \sum_{i=0}^{T-1}\left\|\nabla_f(x^{(i)})\right\|^2 \leq \frac{(f(x^{(0)})-f^*)L}{2m_1(1-m_2)}
\end{aligned}
$$

$\therefore \sum_{i=0}^{\infty}\left\|\nabla_f(x^{(i)})\right\|^2$ is a convergent series. So for $i \to \infty$, $\nabla_f(x^{(i)}) \to 0$.

Therefore, for $i \to \infty$, $x^{(i)}$ approaches a stationary point. Therefore, the descent algorithm which uses Goldstein condition converges to a stationary point, which would hopefully be a local minimum.

5 Rate of convergence

When descent direction is $-\nabla_f(x^{(i)})$, for both the Wolfe condition and the Goldstein condition, the sum $\sum_{i=0}^{T-1}\left\|\nabla_f(x^{(i)})\right\|^2$ is upper-bounded. Denote the upper bound by N.

Let $\delta = \min_i \|\nabla_f(x^{(i)})\|$. Then $T\delta^2 \leq N$. Therefore, $\delta \leq \sqrt{\frac{N}{T}}$. This tells us how fast $x^{(i)}$ converges to a stationary point.

CMO: Coordinate Descent

Eklavya Sharma

Objective: Minimize $f(x) = \frac{1}{2}x^TQx - b^Tx$, where Q is symmetric and positive definite. Q also has a large size. So large that it's stored on secondary/network storage.

Let $x^* = \text{argmin}_x f(x)$.

$$\nabla_f(x) = Qx - b = Q(x - x^*)$$

$$f(x^*) = -\frac{x^{*T}Qx^*}{2}$$

Let $g(\alpha) = f(x^{(i)} + \alpha u)$ where u is a descent direction. i.e. $\nabla_f(x^{(i)})^T u < 0$.

$$g'(0) = \nabla_f(x^{(i)})^T u$$

$$g''(0) = u^T Q u$$

Now we'll use exact line search to find out the step size.

Theorem 1. *Let $\alpha^* = \text{argmin}_\alpha g(\alpha)$. Then*

$$\alpha^* = -\frac{g'(0)}{g''(0)} > 0$$

$$g(\alpha^*) = f(x^{(i)}) - \frac{g'(0)^2}{2g''(0)} = f(x^{(i)}) - \frac{(\nabla_f(x^{(i)})^T u)^2}{2u^TQu}$$

Proof. By Taylor series,

$$g(\alpha) = g(0) + \alpha g'(0) + \frac{\alpha^2}{2}g''(0)$$

$$g'(\alpha) = g'(0) + \alpha g''(0)$$

By the necessary condition for local minimum,

$$g(\alpha^*) = 0 \implies \alpha^* = -\frac{g'(0)}{g''(0)}$$

□

Theorem 2.

$$f(x^{(i)}) - f(x^*) = \frac{\nabla_f(x^{(i)})^T Q^{-1} \nabla_f(x^{(i)})}{2}$$

Proof sketch. Let $v = x^{(i)} - x^*$. Replace $x^{(i)}$ by $x^* + v$ in $f(x^{(i)}) - f(x^*)$. The rest is algebraic manipulation. □

Let $E(x) = f(x) - f(x^*)$. Let $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$ be the eigenvalues of Q.

$$\begin{aligned}
\Delta &= \frac{E(x^{(i)}) - E(x^{(i+1)})}{E(x^{(i)})} \\
&= \frac{f(x^{(i)}) - f(x^{(i+1)})}{f(x^{(i)}) - f(x^*)} \\
&= \frac{g(0) - g(\alpha^*)}{f(x^{(i)}) - f(x^*)} \\
&= \frac{(\nabla_f(x^{(i)})^T u)^2}{2u^T Q u} \frac{2}{\nabla_f(x^{(i)})^T Q^{-1} \nabla_f(x^{(i)})} \\
&= \frac{(\nabla_f(x^{(i)})^T u)^2}{(u^T Q u)\left(\nabla_f(x^{(i)})^T Q^{-1} \nabla_f(x^{(i)})\right)} \\
&= \frac{(\nabla_f(x^{(i)})^T u)^2}{([\lambda_d, \lambda_1]\|u\|^2)\left(\left[\frac{1}{\lambda_1}, \frac{1}{\lambda_d}\right] \| \nabla_f(x^{(i)})\|^2\right)} \\
&\geq \frac{\lambda_d}{\lambda_1} \frac{(\nabla_f(x^{(i)})^T u)^2}{\|u\|^2 \| \nabla_f(x^{(i)})\|^2}
\end{aligned}$$

To prove linear convergence, we must come up with u such that Δ is lower-bounded by a positive constant (because $\frac{E(x^{(i+1)})}{E(x^{(i)})} = 1 - \Delta$).

Let e_j be the j^{th} column of the identity matrix. In coordinate-descent, we choose u to be e_j or $-e_j$ for some j. This has the advantage of being computationally lightweight. For example, $u^T Q u = Q_{j,j}$, which takes $O(1)$ time instead of $O(d^2)$.

Let $g = \nabla_f(x^{(i)})$. Let g_j be the j^{th} coordinate of g. For u to be a descent direction, we'll choose $u = -\operatorname{sgn}(g_j) e_j$. Therefore, $g^T u = -\operatorname{sgn}(g_j) g^T e_j = -|g_j| < 0$.

Also, we'll choose the j which has the highest value of $|g_j|$. Therefore,

$$\|g\|^2 = \sum_{k=1}^{d} |g_k|^2 \leq d|g_j|^2$$

$$\Delta \geq \frac{\lambda_d}{\lambda_1} \frac{(\nabla_f(x^{(i)})^T u)^2}{\|u\|^2 \| \nabla_f(x^{(i)})\|^2} = \frac{\lambda_d}{\lambda_1} \frac{|g_j|^2}{\|g\|^2} \geq \frac{\lambda_d}{d\lambda_1}$$

CMO: Conjugate Descent

Eklavya Sharma

Objective: Minimize $f(x) = \frac{1}{2}x^TQx - b^Tx$, where Q is symmetric and positive definite.

Contents

1 Q-conjugate vectors

Definition 1. *A set of d-dimensional non-0 vectors $U = \{u_0, u_1, \ldots, u_{k-1}\}$ is Q-conjugate iff $\forall i \neq j, u_i^TQu_j = 0$.*

Theorem 1. *If $U = \{u_0, \ldots, u_{d-1}\}$ is Q-conjugate, then U is a basis of $\mathbb{R}^d$.*

Proof. Assume U is linearly dependent. Then one of the vectors in U can be represented as a linear combination of the other (proof). Without loss of generality, assume $u_{d-1} = \sum_{i=0}^{d-2} \alpha_i u_i$.

$\forall i \neq d-1$,

$$0 = u_i^TQu_{d-1} = u_i^TQ\left(\sum_{j=0}^{d-2}\alpha_j u_j\right) = \sum_{j=0}^{d-2}\alpha_j u_i^TQu_j = \alpha_i u_i^TQu_i \implies \alpha_i = 0$$

Hence, $u_{d-1} = 0 \Rightarrow \perp$.

On assuming U to be linearly dependent, we got a contradiction. Therefore, U is linearly independent.

Since $|U| = d = \dim(\mathbb{R}^d)$, U is a basis of $\mathbb{R}^d$ (proof). □

Since Q is positive definite, $u_i^T Q u_i > 0$ for all i.

2 Descent algorithm using Q-conjugate vectors

We'll develop a descent algorithm which uses u_k in the k^{th} iteration with exact line search. The name of this algorithm will be 'Conjugate Gradient Algorithm'.

Let $g(\alpha) = f(x_k + \alpha u_k)$ and $g_k = \nabla_f(x_k)^T$ (sorry for overloading variables; the subscript will help distinguish them though). Therefore, $g'(0) = \nabla_f(x_k) = g_k$ and $g''(0) = u_k^T Q u_k$.

By univariate Taylor series, we get

$$g(\alpha) = g(0) + \alpha g'(0) + \frac{\alpha^2}{2} g''(0)$$

Let $\alpha_k^* = \text{argmin}_\alpha f(x_k + \alpha u_k)$. Therefore,

$$\alpha_k^* = -\frac{g'(0)}{g''(0)} = -\frac{g_k^T u_k}{u_k^T Q u_k}$$

We'll choose $x_{k+1} = x_k + \alpha_k^* u_k$. Therefore, $x_k = x_0 + \sum_{i=0}^{k-1} \alpha_i^* u_i$.

3 Proof of convergence

Theorem 2.

$$u_j^T g_k = \begin{cases} 0 & \text{if } j < k \\ u_j^T g_0 & \text{if } j \geq k \end{cases}$$

Proof.

$$\begin{aligned} g_k &= \nabla_f(x_k) = Q x_k - b \\ &= Q\left(x_0 + \sum_{i=0}^{k-1} \alpha_i^* u_i\right) - b \\ &= (Q x_0 - b) + \sum_{i=0}^{k-1} \alpha_i^* Q u_i \\ &= g_0 + \sum_{i=0}^{k-1} \alpha_i^* Q u_i \end{aligned}$$

$$
\begin{aligned}
u_j^T g_k &= u_j^T \left(g_0 + \sum_{i=0}^{k-1} \alpha_i^* Q u_i \right) \\
&= u_j^T g_0 + \sum_{i=0}^{k-1} \alpha_i^* u_j^T Q u_i \\
&= u_j^T g_0 + \sum_{i=0}^{k-1} \alpha_i^* \begin{Bmatrix} u_j^T Q u_j & i = j \\ 0 & i \neq j \end{Bmatrix} \\
&= u_j^T g_0 + \begin{Bmatrix} \alpha_j^* u_j^T Q u_j & j < k \\ 0 & j \geq k \end{Bmatrix} \\
&= u_j^T g_0 - \begin{Bmatrix} u_j^T g_j & j < k \\ 0 & j \geq k \end{Bmatrix}
\end{aligned}
$$

When $j = k$, we get $u_k^T g_k = u_k^T g_0$. Therefore,

$$
\begin{aligned}
u_j^T g_k &= u_j^T g_0 - \begin{Bmatrix} u_j^T g_j & j < k \\ 0 & j \geq k \end{Bmatrix} \\
&= u_j^T g_0 - \begin{Bmatrix} u_j^T g_0 & j < k \\ 0 & j \geq k \end{Bmatrix} \\
&= \begin{Bmatrix} 0 & j < k \\ u_j^T g_0 & j \geq k \end{Bmatrix}
\end{aligned}
$$

□

Corollary 2.1. $g_d = 0$. *This means that the conjugate descent algorithm converges in d iterations.*

Proof. By the previous theorem (2), $\forall 0 \leq j \leq d-1, u_j^T g_d = 0$. Since $U = \{u_0, u_1, \ldots, u_{d-1}\}$ forms a basis of $\mathbb{R}^d$, we get that $\forall x \in \mathbb{R}^d, x^T g_d = 0$. Therefore, $g_d^T g_d = 0 \implies g_d = 0$. □

We'll now look at an alternative way of proving convergence which will give us more insight.

Let $B_k = \{x_0 + \sum_{i=0}^{k-1} \beta_i u_i : \beta_i \in \mathbb{R}\}$. Since U is a basis of $\mathbb{R}^d$, $B_d = \mathbb{R}^d$. Therefore, to prove convergence of this algorithm, we'll prove the following theorem.

Theorem 3 (Expanding subspace theorem)**.** $\forall k, x_k = \operatorname{argmin}_{x \in B_k} f(x)$.

$x_k = x_0 + \sum_{i=0}^{k-1} \alpha_i^* u_i$. Let $\alpha^* = [\alpha_0^*, \ldots, \alpha_{k-1}^*]$. Let $h(\beta) = f(x_0 + \sum_{i=0}^{k-1} \beta_i u_i)$. Then $\min_{x \in B_k} f(x) = \min_{\beta \in \mathbb{R}^k} h(\beta)$. Since $h(\alpha^*) = f(x_k)$, if we prove that $\alpha^* = \operatorname{argmin}_{\beta \in \mathbb{R}^k} h(\beta)$, then $x_k = \operatorname{argmin}_{x \in B_k} f(x)$.

Lemma 4. *$h(\beta)$ is a convex function.*

Proof. Let $U = [u_0, u_1, \ldots, u_{k-1}]$ be a d by k matrix. Then

$$
(U\beta)_j = \sum_{i=0}^{k-1} U[j, i]\beta_i = \sum_{i=0}^{k-1} (u_i)_j \beta_i = \left(\sum_{i=0}^{k-1} u_i \beta_i \right)_j
$$

$$\implies h(\beta) = f\left(x_0 + \sum_{i=0}^{k-1} \beta_i u_i\right) = f(x_0 + U\beta)$$

$$\begin{aligned} h(\beta) &= f(x_0 + U\beta) \\ &= f(x_0) + \nabla_f(x_0)^T(U\beta) + \frac{1}{2}(U\beta)^T Q(U\beta) && \text{(by Taylor series)} \\ &= f(x_0) + (\nabla_f(x_0)^T U)\beta + \frac{1}{2}\beta^T(U^T QU)\beta \end{aligned}$$

This is a quadratic function in β. It is convex iff U^TQU is positive definite.

By the rules for multiplying stacked matrices, we get that $(U^TQU)_{i,j} = u_i^TQu_j$. Since vectors in U are Q-conjugate, $u_i^TQu_j = 0$ when $i \neq j$. Therefore, U^TQU is a diagonal matrix. Also, $\forall i, u_i^TQu_i > 0$ because Q is positive definite. Therefore, all diagonal entries of U^TQU are positive. Therefore, U^TQU is positive definite. □

Since $h(\beta)$ is convex, $\nabla_h(\beta) = 0$ is a necessary and sufficient condition for minimum.

For all $j \in [0, k-1]$

$$h(\beta)_j = \frac{\partial f(x_0 + \sum_{i=0}^{k-1} \beta_i u_i)}{\partial \beta_j} = u_j^T \nabla_f \left(x_0 + \sum_{i=0}^{k-1} \beta_i u_i\right)$$

$$h(\alpha^*)_j = u_j^T \nabla_f \left(x_0 + \sum_{i=0}^{k-1} \alpha_i^* u_i\right) = u_j^T \nabla_f(x_k) = u_j^T g_k = 0 \qquad \text{(by theorem 2)}$$

Therefore, α^* minimizes h, so x_d minimizes f.

4 Rate of convergence

Unlike the previous algorithms, this algorithm:

- Converges exactly (instead of only 'approaching' the solution).
- Converges very fast – in exactly d steps.

5 Choosing Q-conjugate pairs

We will find U as follows: $u_0 = -g_0$ and $u_{k+1} = -g_{k+1} + \beta_k u_k$. We'll choose β_k such that $u_k^TQu_{k+1} = 0$.

$$0 = u_k^TQu_{k+1} = -u_k^TQg_{k+1} + \beta_k u_k^TQu_k \implies \beta_k = \frac{u_k^TQg_{k+1}}{u_k^TQu_k}$$

Algorithm 1 CGA(x_0): Conjugate Gradient Algorithm for $f(x) = \frac{1}{2}x^TQx - b^Tx$. Takes starting point as input.

$g_0 = Qx_0 - b$
if g_0 == 0 **then**
    return x_0
end if
$u_0 = -g_0$
for $i \in [0, \infty)$ **do**
    $\alpha_i = \dfrac{-g_i^T u_i}{u_i^T Q u_i}$
    $x_{i+1} = x_i + \alpha_i u_i$
    $g_{i+1} = Qx_{i+1} - b$
    if g_{i+1} == 0 **then**
        return x_{i+1}
    end if
    $\beta_i = \dfrac{u_i^T Q g_{i+1}}{u_i^T Q u_i}$
    $u_{i+1} = -g_{i+1} + \beta_i u_i$
end for

Theorem 5. *U is Q-conjugate.*

Proof. Proof can be found in the lecture notes for the course 'Optimization II - Numerical Methods for Nonlinear Continuous Optimization' by A. Nemirovski, in Theorem 5.4.1, page 95. □

Proof sketch. First induct on k to prove that for all k,

$$\text{span}(\{g_0, g_1, \ldots, g_k\}) = \text{span}(\{g_0, Qg_0, \ldots, Q^k g_0\}) = \text{span}(\{u_0, u_1, \ldots, u_k\})$$

This can be done using the facts that $g_{k+1} - g_k = Q(x_{k+1} - x_k) = \alpha_k Q u_k$ and that $v_{k+1} = -g_{k+1} + \beta_k v_k$.

Then induct on k to prove that

$$\forall k, \forall i < k, u_k^T Q u_i = 0$$

To do this, express v_{k+1} as $-g_{k+1} + \beta_k v_k$, write Qv_i as a linear combination of $\{v_0, v_1, \ldots, v_{i+1}\}$ and carefully invoke theorem 2. □

6 Faster convergence for structured eigenvalues

When the eigenvalues of Q have certain properties, we can guarantee faster convergence.

$B_{k+1} = x_0 + \text{span}(u_0, \ldots, u_k)$. Therefore, any vector $x \in B_{k+1}$ can be expressed as $x_0 + \sum_{i=0}^{k} \gamma_i u_i$. Since $\text{span}(u_0, \ldots, u_k) = \text{span}(g_0, \ldots, Q^k g_0)$, $x = x_0 + \left(\sum_{i=0}^{k} \delta_i Q^i\right) g_0$.

Let Poly^k be the set of univariate polynomials of degree at most k where the coefficients are from $\mathbb{R}$ and the variable is an n by n matrix over $\mathbb{R}$. Therefore,

$$x \in B_{k+1} \implies \left(\exists P_k \in \text{Poly}^k, x = x_0 + P_k(Q) g_0\right)$$

$$x - x^* = (x_0 - x^*) + P_k(Q)g_0 = (x_0 - x^*) + P_k(Q)Q(x_0 - x^*)$$
$$= (I + QP_k(Q))(x_0 - x^*)$$

Define $E(x) = f(x) - f(x^*)$. By Taylor series,

$$E(x) = \frac{1}{2}(x - x^*)^T Q(x - x^*)$$
$$= \frac{1}{2}(x_0 - x^*)^T (I + QP_k(Q))^T Q(I + QP_k(Q))(x_0 - x^*)$$
$$= \frac{1}{2}(x_0 - x^*)^T Q(I + QP_k(Q))^2(x_0 - x^*)$$

Let $R = \{e_1, e_2, \ldots, e_d\}$ be the set of orthonormal eigenvectors of Q. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ be the corresponding eigenvalues. Since R forms a basis of $\mathbb{R}^d$, $x_0 - x^*$ can be represented as a linear combination of R. Let $x_0 - x^* = \sum_{i=1}^d \zeta_i e_i = \zeta_i$.

Lemma 6. $E(x_0) = \frac{1}{2}\sum_{i=1}^d \zeta_i^2 \lambda_i$

Proof. Let R be a matrix whose i^{th} column is e_i. Since the eigenvectors are orthonormal, $RR^T = R^T R = I$. Let $\zeta = [\zeta_1, \ldots, \zeta_d]^T$. Then

$$R\zeta = \sum_{i=1}^d \zeta_i e_i = x_0 - x^*$$

Since Q is symmetric, $Q = RDR^T$, Where D is a diagonal matrix whose i^{th} entry is λ_i. Therefore,

$$2E(x_0) = (x_0 - x^*)^T Q(x_0 - x^*) = (R\zeta)^T (RDR^T)(R\zeta)$$
$$= \zeta^T (R^T R) D (R^T R)\zeta = \zeta^T D \zeta = \sum_{i=1}^d \zeta_i^2 \lambda_i$$

□

Lemma 7 (Homework)**.** *Let T be a polynomial where $T(X) = X(I + XP_k(X))^2$. Then $E(x) = \frac{1}{2}\sum_{i=1}^d \zeta_i^2 T(\lambda_i)$.*

Hint. Use the fact that for all $j \in \mathbb{N}$, R is also the set of eigenvectors of Q^j and the corresponding eigenvalues are $\lambda_1^j, \ldots, \lambda_d^j$. □

Lemma 8. *For any polynomial $P_k \in Poly^k$,*

$$\frac{E(x_{k+1})}{E(x_0)} \leq \max_{i=0}^d (1 + \lambda_i P_k(\lambda_i))^2$$

Proof.

$$\begin{aligned}
E(x_{k+1}) &= \min_{x \in B_{k+1}} E(x) && \text{(Expanding subspace theorem)} \\
&= \min_{P_k \in \text{Poly}^k} \frac{1}{2} \sum_{i=1}^{d} \zeta_i^2 \lambda_i (1 + \lambda_i P_k(\lambda_i))^2 \\
&\leq \min_{P_k \in \text{Poly}^k} \frac{1}{2} \sum_{i=1}^{d} \left(\zeta_i^2 \lambda_i \left(\max_{i=0}^{d} (1 + \lambda_i P_k(\lambda_i))^2 \right) \right) \\
&= \min_{P_k \in \text{Poly}^k} \left(\frac{1}{2} \sum_{i=1}^{d} \zeta_i^2 \lambda_i \right) \left(\max_{i=0}^{d} (1 + \lambda_i P_k(\lambda_i))^2 \right) \\
&= E(x_0) \min_{P_k \in \text{Poly}^k} \max_{i=0}^{d} (1 + \lambda_i P_k(\lambda_i))^2
\end{aligned}$$

□

Therefore, by cleverly choosing a polynomial, we can prove useful bounds on convergence.

6.1 Q has r distinct eigenvalues

Suppose Q has r distinct eigenvalues $\mu_1 > \mu_2 > \ldots > \mu_r$. Let $\overline{P}_r(x) = 1 + xP_{r-1}(x)$.

We'll construct P_{r-1} such that $\overline{P}_r(x) = 0$ for all $1 \leq i \leq r$. This would mean that $\frac{E(x_r)}{E(x_0)} = 0$, so the conjugate gradient algorithm will converge in r iterations.

Define $\overline{P}_r$ and P_{r-1} as follows:

$$\overline{P}_r(x) = \prod_{j=1}^{r} \left(1 - \frac{x}{\mu_j}\right) \qquad\qquad P_{r-1}(x) = \frac{\overline{P}_r(x) - 1}{x}$$

Lemma 9. *P_{r-1} is a polynomial of degree $r-1$ such that $\forall 0 \leq i \leq r, \overline{P}_r(\mu_i) = 0$.*

Proof. Clearly, $\overline{P}_r(\mu_i) = 0$ for all i. Also, the degree of $\overline{P}$ is r.

Next, we must prove that P_{r-1} is a polynomial. Note that $\overline{P}_r(0) = 1$, so 0 is a root of $\overline{P}_r(x) - 1$. Therefore, x is a factor of $\overline{P}_r(x) - 1$ and hence P_{r-1} is a polynomial.

Since the degree of $\overline{P}_r$ is r, the degree of P_{r-1} is $r-1$. □

6.2 Theorem for a polynomial

In this section, we'll prove a theorem for a certain polynomial which we'll use in the next section.

Theorem 10. *Let $n \geq 2$. Let $0 < a_1 < a_2 < \ldots < a_n$. Let $p_1, p_2, \ldots, p_n$ be positive integers and let $p_1 = 1$.*

$$f(x) = \prod_{i=1}^{n} \left(1 - \frac{x}{a_i}\right)^{p_i} \qquad\qquad g(x) = f(x) - 1 + \frac{x}{a_1}$$

Then

1. *f is positive in $(-\infty, a_1)$, negative in (a_1, a_2) and 0 at a_1 and a_2.*

2. *$g(x) \leq 0$ for $x \in [0, a_1]$ and $g(x) \geq 0$ for $x \in [a_1, a_2]$.*

Proof. Since a_1 and a_2 are zeros of f, $f(a_1) = f(a_2) = 0$. Since a_1 is the leftmost zero of f, f has the same sign in $(-\infty, a_1)$ (by intermediate value theorem). Since $f(0) = 1$, f is positive in $(-\infty, a_1)$.

$$\frac{f'(x)}{f(x)} = \sum_{i=1}^{n} \frac{p_i}{x - a_i}$$

Let

$$h_1(x) = \prod_{i=1}^{n} (x - a_i)^{p_i - 1}$$

Then $h_1(x)$ divides $f'(x)$.

By Rolle's theorem, there must be points $b_1 < b_2 < \ldots < b_{n-1}$ such that for all i, $f'(b_i) = 0$ and $b_i \in (a_i, a_{i+1})$. Let

$$h_2(x) = \prod_{i=1}^{n-1} (x - b_i)$$

So $h_2(x)$ divides $f'(x)$.

Let $N = \sum_{i=1}^{n} p_i$. Then $\deg(f) = N$. Also

$$\deg(h_1 h_2) = \deg(h_1) + \deg(h_2) = (N - n) + (n - 1) = N - 1 = \deg(f')$$

Therefore, $f'(x) = \gamma h_1(x) h_2(x)$ for some $\gamma \in \mathbb{R}$.

Since $p_1 = 1$, b_1 is the leftmost zero of f' and it is the only zero in $(-\infty, a_2)$. Therefore, $f'(x)$ has the same sign for $x \in (-\infty, b_1)$. Since $f(0) = 1$, $f'(0) = -\sum_{i=1}^{n} \frac{1}{a_i} < 0$. Therefore, $f'(x) < 0$ for $x \in (-\infty, b_1)$.

Since $f(a_1) = 0$ and $f'(a_1) < 0$, $f(a_1 + \epsilon) < 0$ for all very small ϵ. Also, f has the same sign in (a_1, a_2), otherwise it would have a root in (a_1, a_2), which we know is false. Therefore, $f(x) < 0$ for $x \in (a_1, a_2)$. This completes the proof of part 1 of this theorem.

Applying Rolle's theorem to $f'(x)$ and by a similar argument (todo: expand this), we get that $f''(x)$ must have its leftmost root in (b_1, a_2). Therefore, $f''(x)$ has the same sign in $(-\infty, b_1]$.

$$\frac{f''(x)}{f(x)} = \left(\sum_{i=1}^{n} \frac{p_i}{a_i - x} \right)^2 - \sum_{i=1}^{n} \frac{p_i}{(a_i - x)^2}$$

$$\implies f''(0) = \left(\sum_{i=1}^{n} \frac{p_i}{a_i} \right)^2 - \sum_{i=1}^{n} \frac{p_i}{a_i^2} > 0$$

Therefore, $f''(x) > 0$ for $x \in (-\infty, b_1]$.

$f'(b_1) = 0$ and $f''(b_1) > 0$. Therefore, $f'(b_1 + \epsilon) > 0$ for all very small ϵ. $f'(x)$ has the same sign in (b_1, a_2) because b_1 is the only root of $f'(x)$ in $[b_1, a_2)$. Therefore, $f'(x) > 0$ for $x \in (b_1, a_2)$.

Since f is convex in $(-\infty, b_1]$, for $\alpha \in [0, 1]$,

$$f(\alpha a_1) = f((1-\alpha)0 + \alpha a_1) \leq (1-\alpha)f(0) + \alpha f(a_1) = (1-\alpha)$$

Setting α to x/a_1, we get that for $x \in [0, a_1]$, $f(x) \leq 1 - \frac{x}{a_1} \Rightarrow g(x) \leq 0$.

$g(0) = g(a_1) = 0$. By Rolle's theorem, $\exists x_0 \in (0, a_1), g'(x_0) = 0$. Since $g''(x) = f''(x) > 0$ for $x \in (-\infty, b_1]$, $g'(x) > 0$ for $x \in (x_0, b_1]$.

$g'(x) = f'(x) + \frac{1}{a_1}$. For $x \in (b_1, a_2)$, $f'(x) > 0 \Rightarrow g'(x) > 0$. Therefore, $g'(x) > 0$ for $x \in [a_1, b_1)$.

Since $g(a_1) = 0$ and $g'(x) > 0$ for $x \in [a_1, b_1)$, $g(x) > 0$ for $x \in (a_1, b_1)$. □

6.3 Q has some clustered eigenvalues

Suppose Q has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$, where for some constants a and b,

$$0 < a \leq \lambda_d \leq \ldots \leq \lambda_{r+1} < b < \lambda_r \leq \ldots \leq \lambda_1$$

Let $\mu_i = \lambda_i$ for i from 1 to r. Let $\mu_{r+1} = \frac{a+b}{2}$.

$$\overline{P}_{r+1}(x) = \prod_{j=1}^{r+1} \left(1 - \frac{x}{\mu_j}\right) \qquad P_r(x) = \frac{P_{r+1}(x) - 1}{x} \qquad h(x) = 1 - \frac{x}{\mu_{r+1}}$$

It's easy to prove (similar to lemma 9) that P_r is a polynomial and has degree r.

Since $\overline{P}_{r+1}$ is of the right form, we can apply theorem 10.

By part 1 of theorem 10, we get that for $x \in [a, \frac{a+b}{2}]$, $\overline{P}_{r+1}(x) \geq 0$. By part 2 of theorem 10, we get that for $x \in [a, \frac{a+b}{2}]$,

$$\overline{P}_{r+1}(x) \leq h(x) \leq h(a) = \frac{b-a}{b+a}$$

By part 1 of theorem 10, we get that for $x \in [\frac{a+b}{2}, b]$, $\overline{P}_{r+1}(x) \leq 0$. By part 2 of theorem 10, we get that for $x \in [\frac{a+b}{2}, b]$,

$$\overline{P}_{r+1}(x) \geq h(x) \geq h(b) = -\frac{b-a}{b+a}$$

Therefore, for $x \in [a, b]$, $\left|\overline{P}_{r+1}(x)\right| \leq \frac{b-a}{b+a}$. Therefore,

$$\frac{E(x_{r+1})}{E(x_0)} \leq \left(\frac{b-a}{b+a}\right)^2$$

We can use the above fact to design an algorithm called the 'partial conjugate gradient' algorithm. In this algorithm, we'll start at the point z_0 and run the conjugate gradient algorithm for $r+1$ steps to reach the point z_1. Then we'll rerun the conjugate gradient algorithm for $r+1$ steps from z_1 to reach a point z_2, then we'll rerun the conjugate gradient algorithm for $r+1$ steps from z_2 to reach a point z_3, and so on. We'll do this l times. After l iterations $\frac{E(z_l)}{E(z_0)} = \left(\frac{b-a}{b+a}\right)^{2l}$. This will give us linear convergence.

CMO: Newton's method

Eklavya Sharma

By second-order Taylor series, we get

$$f(y) = f(x) + \nabla_f(x)^T(y-x) + \frac{1}{2}(y-x)^T \, \mathrm{H}_f(x)(y-x) + \|y-x\|^2 o(1)$$

In Newton's method, we choose an update rule which minimizes the resulting quadratic function, assuming that $\mathrm{H}_f(x)$ is positive definite.

$$x^{i+1} = x^i - \mathrm{H}_f(x^i)^{-1} \, \nabla_f(x^i)$$

1 Matrix norm

Definition 1 (Spectral Norm)**.** *Let A be a d by d matrix. Then*

$$\|A\| = \max_{i=1}^{d} |\lambda_i(A)|$$

Theorem 1 (Homework)**.** *The spectral norm is a norm (i.e. $\|A\| = 0 \iff A = 0$ and $\|A\| \geq 0$).*

Theorem 2 (Homework)**.** $D(A, B) = \|A - B\| \implies D$ *is a distance metric.*

Proof hint for triangle inequality. Get a bound on eigenvalues of sum of matrices using the facts that the sum of positive semidefinite matrices is also positive semidefinite and that if (λ, v) is an eigenpair of a matrix A then $(\lambda - k, v)$ is an eigenpair of the matrix $A - kI$. □

2 Region of positive-definite hessian

Intuitively, Newton's method would work when the hessian is positive definite. Unfortunately, that need not be true for most real problems. However, at a local minimum, the hessian is guaranteed to be positive semidefinite. We hope that, due to f being in C^2, hessian would be positive semidefinite or definite *near* the local minimum too.

Therefore, if we somehow reach close to a local minimum, we can start using Newton's method. We'll now try to quantify how close do we need to get.

Definition 2. *Let $\lambda(A)$ denote the set of eigenvalues of matrix A.*

Definition 3. *For matrices A and B, $A \leq B$ iff $B - A$ is positive semidefinite. For matrices A and B, $A < B$ iff $B - A$ is positive definite. Similarly define $\geq$ and $>$.*

Theorem 3 (Transitivity of $\leq$ (Homework))**.** $A \leq B \wedge B \leq C \implies A \leq C$

Definition 4.

$$f \in C_M^2 \iff (\forall x, y \in \mathbb{R}^d, \| \mathrm{H}_f(y) - \mathrm{H}_f(x)\| \leq M\|y - x\|)$$

Theorem 4 (Homework)**.**

$$f \in C_M^2 \implies (\forall x, y \in \mathbb{R}^d, \mathrm{H}_f(x) - MrI \leq \mathrm{H}_f(y) \leq \mathrm{H}_f(x) + MrI)$$

where $r = \|y - x\|$.

Theorem 5. *Let $\mathrm{H}_f(x^*) \geq aI$, where $a > 0$. Then $r = \|x - x^*\| < \frac{a}{M} \implies \mathrm{H}_f(x)$ is positive definite.*

Proof.

$$r < \frac{a}{M} \implies 0 < (a - Mr)I \leq \mathrm{H}_f(x^*) - MrI$$

$$f \in C_M^2 \implies \mathrm{H}_f(x) \geq \mathrm{H}_f(x^*) - MrI > 0$$

□

We now have a region where $\mathrm{H}_f(x)$ is known to be positive definite. However, this is still not suitable for Newton's method, since the point in the next iteration may fall outside this region. We therefore impose another condition, that distance from x^* should reduce.

3 Newton's region and convergence

Lemma 6 (Proof omitted (beyond the scope of course?))**.** $\forall x, y \in \mathbb{R}^d$,

$$\nabla_f(y) - \nabla_f(x) = \int_0^1 \mathrm{H}_f(x + \alpha(y - x))(y - x)d\alpha$$

Lemma 7. *Let A be a symmetric matrix and u be a vector. Then $\|Au\| \leq \|A\|\|u\|$.*

Proof.

$$\frac{\|Au\|^2}{\|u\|^2} = \frac{u^T A^2 u}{\|u\|^2} \in [0, \|A\|^2] \implies \frac{\|Au\|}{\|u\|} \leq \|A\|$$

□

Theorem 8. *Let $\mathrm{H}_f(x^*) \geq a$, where $a > 0$. Let $r_k = \|x^k - x^*\|$. Then $r_k < \frac{2a}{3M} \implies r_{k+1} < r_k$.*

Proof. By theorem 6, we get

$$\nabla_f(x^k) = \int_0^1 \mathrm{H}_f(x^* + \alpha(x^k - x^*))(x^k - x^*)d\alpha$$

$$\begin{aligned} x^{k+1} - x^* &= (x^k - x^*) - \mathrm{H}_f^{-1}(x^k)\,\nabla_f(x^k) \\ &= \mathrm{H}_f^{-1}(x^k)(\mathrm{H}_f(x^k)(x^k - x^*) - \nabla_f(x^k)) \\ &= \mathrm{H}_f^{-1}(x^k)\int_0^1 \left(\mathrm{H}_f(x^k) - \mathrm{H}_f(x^k + \alpha(x^k - x^*))\right)(x^k - x^*)d\alpha \end{aligned}$$

$$\begin{aligned} r_{k+1} &= \left\| \mathrm{H}_f^{-1}(x^k)\int_0^1 \left(\mathrm{H}_f(x^k) - \mathrm{H}_f(x^k + \alpha(x^k - x^*))\right)(x^k - x^*)d\alpha \right\| \\ &\leq \left\|\mathrm{H}_f^{-1}(x^k)\right\| \left\| \int_0^1 \left(\mathrm{H}_f(x^k) - \mathrm{H}_f(x^k + \alpha(x^k - x^*))\right)(x^k - x^*)d\alpha \right\| \\ &\qquad\qquad \text{(by lemma 7)} \\ &\leq \left\|\mathrm{H}_f^{-1}(x^k)\right\| \int_0^1 \left\| \left(\mathrm{H}_f(x^k) - \mathrm{H}_f(x^k + \alpha(x^k - x^*))\right)(x^k - x^*) \right\| d\alpha \\ &\qquad\qquad \text{(by triangle inequality)} \\ &\leq \left\|\mathrm{H}_f^{-1}(x^k)\right\| \int_0^1 \left\|\mathrm{H}_f(x^k) - \mathrm{H}_f(x^k + \alpha(x^k - x^*))\right\| \left\|x^k - x^*\right\| d\alpha \quad \text{(by lemma 7)} \\ &\leq \left\|\mathrm{H}_f^{-1}(x^k)\right\| \int_0^1 \left(M \left\|(1-\alpha)(x^k - x^*)\right\|\right) \left\|x^k - x^*\right\| d\alpha \qquad (f \in C_M^2) \\ &= \left\|\mathrm{H}_f^{-1}(x^k)\right\| \left(\frac{Mr_k^2}{2}\right) \end{aligned}$$

$$\mathrm{H}_f(x^k) \geq (a - Mr_k)I \implies \mathrm{H}_f^{-1}(x^k) \leq \frac{1}{a - Mr_k} I \implies \left\|\mathrm{H}_f^{-1}(x^k)\right\| \leq \frac{1}{a - Mr_k}$$

$$\implies r_{k+1} \leq \frac{Mr_k^2}{2(a - Mr_k)}$$

$$\frac{Mr_k}{2(a - Mr_k)} < 1 \iff r_k < \frac{2a}{3M}$$

Therefore,

$$r_k \leq \frac{2a}{3M} \implies \frac{r_{k+1}}{r_k} \leq \frac{Mr_k}{2(a - Mr_k)} < 1 \implies r_{k+1} < r_k$$

□

The $\frac{2a}{3M}$-neighborhood of x^* is called the Newton region. In this region, Newton's method will always be applicable. Furthermore,

$$r_k < \frac{2a}{3M} \implies \frac{a}{M} - r_k > \frac{a}{3M} \implies r_{k+1} = \frac{r_k^2}{2(\frac{a}{M} - r_k)} < \frac{3M}{2a} r_k^2$$

which shows that Newton's method gives us quadratic convergence.

4 Quadratic function

Let $f(x) = \frac{1}{2}x^T Q x - b^T x$, where Q is symmetric and positive definite. $\nabla_f(x) = Qx - b = Q(x - x^*)$.

$$x_1 = x_0 - \mathrm{H}_f(x_0)^{-1} \nabla_f(x_0) = x_0 - Q^{-1}Q(x - x^*) = x^*$$

Therefore, Newton's method converges to the minimum in a single iteration.

CMO: Quasi-Newton Methods

Eklavya Sharma

Contents

1 Quasi-Newton method template

Newton's method's update rule:

$$x_{k+1} = x_k - \mathrm{H}_f^{-1}(x_k)\,\nabla_f(x_k)$$

This method is not useful, because it requires inverting the hessian, which can be prohibitively computationally expensive for high-dimensional data.

We will therefore try to model the change in the hessian's inverse, and approximate the hessian's inverse instead of calculating it exactly.

Let $g_k = \nabla_f(x_k)$, $\delta_k = x_{k+1} - x_k$ and $\gamma_k = g_{k+1} - g_k$.

$$\nabla_f(x_{k+1}) \approx \nabla_f(x_k) + \mathrm{H}_f(x_k)(x_{k+1} - x_k) \qquad \text{(by differentiating Taylor series)}$$

$$\implies \delta_k \approx \mathrm{H}_f^{-1}(x_k)\gamma_k$$

This inspires us to use an update rule of this form:

$$x_{k+1} = x_k - A_k g_k$$

and apply the following constraint on A_k:

$$\delta_k = A_{k+1}\gamma_k \tag{1}$$

This constraint is called the 'Quasi-Newton condition'.

Also, we must ensure that A_k is symmetric and positive (semi)definite.

Note that the Quasi-Newton condition is d equations, whereas there are d^2 entries in A_k. We therefore have a lot of slack in terms of how to update A_k.

In all Quasi-Newton methods described next, we choose A_0 as any matrix which is symmetric and positive (semi)definite. Generally, the identity matrix is used. Then we use A_k, δ_k and γ_k to obtain A_{k+1} via an update rule, like 'rank-1 update', 'rank-2 update' or 'BFGS'.

2 Rank-1 update

Here we impose a condition of the form $A_{k+1} = A_k + cuu^T$, where $c \in \mathbb{R}$ and $u \in \mathbb{R}^d$ (Note that $\text{rank}(uu^T) = 1$).

It's easy to see that A_{k+1} is symmetric for all c and positive definite for $c \geq 0$.

To get concrete values of c and u, we'll plug the rank-1 update condition into the Quasi-Newton condition (1).

$$\delta_k = (A_k + cuu^T)\gamma_k \implies (cu^T\gamma_k)u = \delta_k - A_k\gamma_k$$

Therefore, u is parallel to $\delta_k - A_k\gamma_k$. Let $u = \delta_k - A_k\gamma_k$. Then

$$u = (cu^T\gamma_k)u \implies cu^T\gamma_k = 1 \implies c = \frac{1}{u^T\gamma_k} = \frac{1}{\delta_k^T\gamma_k - \gamma_k^T A_k\gamma_k}$$

With these specific values of u and c, the rank-1 update condition will satisfy all required conditions (symmetry, positive definiteness and Quasi-Newton condition) if $c \geq 0$.

Unfortunately, it has not yet been proved or disproved whether $c \geq 0$.

2.1 Analysis for quadratic function

Let $f(x) = \frac{1}{2}x^TQx - b^Tx$, where Q is symmetric and positive definite. Then $\nabla_f(x) = Qx - b \implies \gamma_k = Q\delta_k$.

Lemma 1.

$$\forall i \in [0, k], A_{k+1}\gamma_i = \delta_i$$

Proof by induction on k.

$$P(l) : \forall i \in [0, l-1], A_l\gamma_i = \delta_i$$

We have to prove $P(l)$ for all $l \geq 1$.

Base case: Since A_1 was constructed to follow the Quasi-Newton condition, $\delta_0 = A_1\gamma_0 \implies P(1)$.

Inductive step: Assume $P(l)$ is true. We'll prove $P(l+1)$.

Let $i \in [0, l-1]$.

$$\begin{aligned} A_{l+1}\gamma_i &= \left(A_l + \frac{uu^T}{u^T\gamma_l}\right)\gamma_i && \text{(here } u = \delta_l - A_l\gamma_l\text{)} \\ &= \delta_i + \frac{u^T\gamma_i}{u^T\gamma_l}u && \text{($A_l\gamma_i = \delta_i$ by induction hypothesis)} \end{aligned}$$

$$\begin{aligned} u^T\gamma_i &= (\delta_l - A_l\gamma_l)^T\gamma_i \\ &= \delta_l^T\gamma_i - \gamma_l^T A_l\gamma_i \\ &= \delta_l^T\gamma_i - \gamma_l^T\delta_i && \text{(by induction hypothesis)} \\ &= \delta_l^T Q\delta_i - \delta_l^T Q\delta_i && (\forall j, \gamma_j = Q\delta_j) \\ &= 0 \end{aligned}$$

Therefore, $A_{l+1}\gamma_i = \delta_i$ for all $i \in [0, l-1]$. Since A_{l+1} was constructed to follow the Quasi-Newton condition, $A_{l+1}\gamma_l = \delta_l$. Therefore, $P(l+1)$ holds true. □

Lemma 2. *If all δ_i were orthonormal, then $A_d = Q^{-1}$.*

Proof. By lemma 1,

$$\forall i \in [0, d-1], \delta_i = A_d\gamma_i = A_dQ\delta_i$$

Therefore, $(1, \delta_i)$ is an eigenpair for A_dQ.

Let P be the matrix whose i^{th} columns is δ_i. P exists because real symmetric matrices are orthogonally diagonalizable and A_dQ is real and symmetric. Then $A_dQ = PIP^T = I \implies A_d = Q^{-1}$. □

Lemma 3. *If all δ_i are linearly independent, then $A_d = Q^{-1}$.*

Proof. Let $\Delta = \{\delta_0, \ldots, \delta_{d-1}\}$. Since $\Delta \subseteq \mathbb{R}^d$, $|\Delta| = d = \dim(\mathbb{R}^d)$ and Δ is linearly independent, Δ is a basis of $\mathbb{R}^d$.

Let $x \in \mathbb{R}^d$. Let $x = \sum_{i=0}^{d-1} c_i\delta_i$. Then

$$A_dQx = \sum_{i=0}^{d-1} A_dQ(c_i\delta_i) = \sum_{i=0}^{d-1} c_i(A_d\gamma_i) = \sum_{i=0}^{d-1} c_i\delta_i = x$$

Therefore, $\forall x \in \mathbb{R}^d, (A_dQ)x = x$, so $A_dQ = I$.

Note that the proof is not specific to rank-1 updates. Its correctness relies only on the Quasi-Newton condition and f being quadratic. □

Since $A_d = Q^{-1}$, the $(d+1)^{\text{th}}$ iteration would be identical to Newton's method. So the rank-1 update method will converge to the minimum in at most $d+1$ iterations.

2.2 Unresolved questions

- A_k is positive definite when $c \geq 0$. Is $c \geq 0$?
- Is $\{\delta_0, \delta_1, \ldots\}$ linearly independent?

3 Rank-2 update

$$A_{k+1} = A_k + cuu^T + bvv^T$$

It's easy to see that A_{k+1} is symmetric iff A_k is symmetric.

By Quasi-Newton condition, we get

$$\delta_k = A_{k+1}\gamma_k \implies (cu^T\gamma_k)u + (bv^T\gamma_k)v = \delta_k - A_k\gamma_k$$

Let $u = \delta_k$ and $v = A_k\gamma_k$. Then

$$c = \frac{1}{u^T\gamma_k} = \frac{1}{\delta_k^T\gamma_k} \qquad b = \frac{-1}{v^T\gamma_k} = \frac{-1}{\gamma_k^T A_k\gamma_k}$$

$$A_{k+1} = A_k + \frac{\delta_k\delta_k^T}{\delta_k^T\gamma_k} - \frac{A_k\gamma_k\gamma_k^T A_k}{\gamma_k^T A_k\gamma_k}$$

3.1 Analysis for quadratic function

Let $f(x) = \frac{1}{2}x^TQx - b^Tx$. Then $\gamma_k = Q\delta_k$.

Lemma 4 (Symmetric square root of a matrix)**.** *If A is a symmetric and positive definite matrix, then $\exists L$ such that $A = L^2$ and L is symmetric, positive semidefinite and invertible.*

Proof. Since A is real and symmetric, it is orthogonally diagonalizable. So there is a matrix P and a diagonal matrix D such that $A = PDP^T$ and $PP^T = P^TP = I$. Since A is positive definite, all diagonal entries of D are positive. Therefore, $\sqrt{D}$ exists. Also, all entries of $\sqrt{D}$ are positive, so $\sqrt{D}^{-1}$ exists. Let $L = P\sqrt{D}P^T$. Then L is symmetric and $L^2 = A$.

$$u^TL^u = u^T(P\sqrt{D}P^T)u = (P^Tu)^T\sqrt{D}(P^Tu) \geq 0$$

Therefore, L is also positive semidefinite. Also,

$$L(P\sqrt{D}^{-1}P^T) = P\sqrt{D}P^TP\sqrt{D}^{-1}P^T = I$$

Therefore, $L^{-1} = P\sqrt{D}^{-1}P^T$. □

Theorem 5. *Let A_k be symmetric and positive definite. Then A_{k+1} is positive definite.*

Proof.

$$c = \frac{1}{\delta_k^T\gamma_k} = \frac{1}{\delta_k^TQ\delta_k} > 0 \tag{2}$$

We'll now prove that $A_{k+1} - cuu^T$ is positive semidefinite. Let $w \in \mathbb{R}^d - \{0\}$.

$$\begin{aligned} &w^T(A_{k+1} - cuu^T)w \\ &= w^T(A_k + bvv^T)w \\ &= w^TA_kw - \frac{(w^TA_k\gamma_k)^2}{\gamma_k^TA_k\gamma_k} \end{aligned}$$

Since A_k is symmetric and positive definite, it has a symmetric and invertible square root L.

$$\begin{aligned}
&w^T(A_{k+1} - cuu^T)w \\
&= w^T L^T L w - \frac{(w^T L^T L \gamma_k)^2}{\gamma_k^T L^T L \gamma_k} \\
&= \|Lw\|^2 - \frac{((Lw)^T(L\gamma_k))^2}{\|L\gamma_k\|^2} \\
&\geq 0 && \text{(by Cauchy-Schwarz inequality)}
\end{aligned}$$

Therefore, $A_{k+1} - cuu^T$ is positive semidefinite. Since cuu^T is also positive semidefinite, A_{k+1} is also positive semidefinite.

The Cauchy-Schwarz inequality is tight iff the vectors are parallel or anti-parallel. Therefore, $A_{k+1} - cuu^T = 0 \iff Lw = \alpha L\gamma_k$ for some $\alpha \in \mathbb{R}$. Since L is invertible, this is equivalent to $w = \alpha\gamma_k$.

Assume A_{k+1} is not positive definite. $\exists w \in \mathbb{R}^d - \{0\}, w^T A_{k+1} w = 0$.

$$\begin{aligned}
&w^T A_{k+1} w = 0 \\
&\implies w^T(A_{k+1} - cuu^T)w + w^T(cuu^T)w = 0 \\
&\implies w^T(A_{k+1} - cuu^T)w = 0 \wedge w^T(cuu^T)w = 0 \\
&\implies (\alpha\gamma_k)^T(cuu^T)(\alpha\gamma_k) = 0 \\
&\implies c\alpha^2(\gamma_k^T\delta_k)^2 = 0 && (u = \delta_k) \\
&\implies \alpha^2(\delta_k^T Q \delta_k) = 0 && (\gamma_k = Q\delta_k \text{ and } 2)
\end{aligned}$$

This is not possible because $\delta_k^T Q \delta_k > 0$ (because Q is positive definite) and $\alpha \neq 0$ (because $w \neq 0$). Therefore, we have a contradiction. Therefore, A_{k+1} is positive definite. □

Lemma 6 (Proof omitted (probably beyond scope of course))**.**

$$\forall k \geq 1, \forall i \in [0, k-1], A_k\gamma_i = \delta_i \wedge \delta_k^T Q \delta_i = 0$$

Let $\Delta = \{\delta_0, \delta_1, \ldots\}$. Lemma 6 states that Δ is Q-conjugate. This implies that Δ is linearly independent. By lemma 3, we get that rank-2 updates converge to minimum in $d+1$ iterations.

4 BFGS

Instead of modeling the change in hessian's inverse, we'll now model the change in the hessian. But we need to do it in a way such that the change in the inverse is also easy to compute.

Let B_k be an approximation to the hessian and A_k be an approximation to the inverse of the hessian. Then $\gamma_k = B_{k+1}\delta_k$ and $\delta_k = A_{k+1}\gamma_k$.

We'll chose the update rule as

$$B_{k+1} = B_k + cuu^T + bvv^T$$

This will make sure that B_k is symmetric implies B_{k+1} is symmetric.

Applying the Quasi-Newton condition, we get

$$\gamma_k = B_{k+1}\delta_k \implies \gamma_k - B_k\delta_k = (cu^T\delta_k)u + (bv^T\delta_k)v$$

Let $u = \gamma_k$ and $v = B_k\delta_k$.

$$c = \frac{1}{u^T\delta_k} = \frac{1}{\gamma_k^T\delta_k} \qquad\qquad d = \frac{-1}{v^T\delta_k} = \frac{-1}{\delta_k^T B_k\delta_k}$$

$$B_{k+1} = B_k + \frac{\gamma_k^T\gamma_k}{\gamma_k^T\delta_k} - \frac{B_k\delta_k\delta_k^T B_k}{\delta_k^T B_k\delta_k}$$

Similar to theorem 5, we can prove that B_{k+1} is positive definite for quadratic functions. This implies that A_{k+1} is also symmetric and positive definite for quadratic functions.

To invert B_{k+1}, we'll use the Sherman-Morrison formula.

Theorem 7 (Sherman-Morrison formula)**.** *Let A be an invertible matrix. Then $A + uv^T$ is invertible iff $1 + v^TA^{-1}u \neq 0$. Also,*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

Applying the formula twice, we get

$$A_{k+1} = A_k + \frac{\delta_k\delta_k^T}{\delta_k^T\gamma_k}\left(1 + \frac{\gamma_k^TA_k\gamma_k}{\delta_k^T\gamma_k}\right) - \frac{A_k\gamma_k\delta_k^T + \delta_k\gamma_k^TA_k}{\delta_k^T\gamma_k}$$

5 Broyden Family

Let's explore this update rule:

$$A_{k+1} = A_k + a\frac{\delta_k\delta_k^T}{\delta^T\gamma_k} + c\frac{A_k\gamma_k\gamma_k^TA_k}{\gamma_k^TA\gamma_k} - b\frac{A_k\gamma_k\delta_k^T + \delta_k\gamma_k^TA_k}{\delta_k^T\gamma_k}$$

Applying the Quasi-Newton condition, we get

$$\delta_k - A_k\gamma_k = \left(a - b\frac{\gamma_k^TA_k\gamma_k}{\delta_k^T\gamma_k}\right)\delta_k + (c - b)A_k\gamma_k$$

Equating coefficients of δ_k and γ_k, we get

$$a = 1 + b\frac{\gamma_k^TA_k\gamma_k}{\delta_k^T\gamma_k} \qquad\qquad c = b - 1$$

On rearranging, we get

$$A_{k+1} = \left(A_k + \frac{\delta_k\delta_k^T}{\delta_k^T\gamma_k} - \frac{A_k\gamma_k\gamma_k^TA_k}{\gamma_k^TA_k\gamma_k}\right) + b(\gamma_k^TA_k\gamma_k)w_kw_k^T$$

where

$$w = \frac{\delta_k}{\delta_k^T \gamma_k} - \frac{A_k \gamma_k}{\gamma_k^T A_k \gamma_k}$$

This update rule is called the Broyden Family. Note that the first term is the same as the rank-2 update.

Define the following 2 functions:

$$\text{rank-2}(A, \delta, \gamma) = A + \frac{\delta \delta^T}{\delta^T \gamma} - \frac{A \gamma \gamma^T A}{\gamma^T A \gamma}$$

$$\text{bfgs}(A, \delta, \gamma) = A + \frac{\delta \delta^T}{\delta^T \gamma} \left(1 + \frac{\gamma^T A \gamma}{\delta^T \gamma}\right) - \frac{A \gamma \delta^T + \delta \gamma^T A}{\delta^T \gamma}$$

The Broyden family can also be rewritten as

$$A_{k+1} = (1 - b)\, \text{rank-2}(A_k, \delta_k, \gamma_k) + b\, \text{bfgs}(A_k, \delta_k, \gamma_k)$$

CMO: Constrained Optimization

Eklavya Sharma

In constrained optimization, we have to find

$$x^* = \underset{x \in C}{\operatorname{argmin}}\, f(x)$$

where $C \in \mathbb{R}^d$ is a closed set. C is called the feasible region. We say that x is feasible iff $x \in C$.

The methods which we developed for unconstrained optimization often don't work for constrained optimization because properties of optimal solutions are different here. For example, if x^* is an unconstrained minimum of f, then $\nabla_f(x^*) = 0$. This doesn't hold for constrained minima. $\min_{x \in [1,2]} x^2$ is an example.

We'll consider several special cases of constrained optimization.

Contents

1 Introduction

Definition 1 (Feasible directions)**.** *$u \in \mathbb{R}^d$ is feasible direction at $x \in C$ iff*

$$\exists \overline{\alpha} > 0, \forall \alpha \in [0, \overline{\alpha}], x + \alpha u \in C$$

The set of feasible directions at x is denoted by $\mathrm{FS}(x)$*.*

Theorem 1. *If x is a local minimum of f, then there is no feasible descent direction. Formally,*

$$\forall u \in \mathrm{FS}(x), \nabla_f(x)^T u \geq 0$$

Proof Sketch. If there is a feasible descent direction u at x, then for any arbitrarily small α, we can decrease f by moving α distance towards u. So f is not a local minimum. □

Note that the converse need not be true. Let x be a saddle point of f and let there be no constraints. Then every direction is not a descent direction (and not an ascent direction) but x is not a local minimum.

2 Projection onto a convex set

Theorem 2. *Let C be a convex set. Let $x^* = \operatorname{argmin}_{x \in C} f(x)$. Then*

$$\forall x \in C, x - x^* \in \mathrm{FS}(x^*)$$

In this section, we'll now fix the objective function to be $f(x) = \frac{1}{2}\|x - z\|^2$ and consider the feasible region C to be convex. Also, assume that $z \notin C$.

Definition 2 (Projection)**.** *Let $x^* = \operatorname{argmin}_{x \in C} f(x)$. Then x^* is called the projection of z onto C.*

Theorem 3.

$$x^* = \operatorname*{argmin}_{x \in C} f(x) \iff (\forall x \in C, (x^* - z)^T (x - x^*) \geq 0)$$

Proof. Let $x^* = \operatorname{argmin}_{x \in C}$. By theorem 2, we get that

$$\forall x \in C, x - x^* \in \mathrm{FS}(x^*)$$

By theorem 1, we get that

$$\begin{aligned} &\forall x \in C, \nabla_f(x^*)^T (x - x^*) \geq 0 \\ &\implies \forall x \in C, (x^* - z)^T (x - x^*) \geq 0 \end{aligned}$$

Now assume that $\forall x \in C, (x^* - z)^T (x - x^*) \geq 0$.

$$\begin{aligned} f(x) &= \frac{1}{2}\|(x - x^*) + (x^* - z)\|^2 \\ &= \frac{1}{2}\|x - x^*\|^2 + \frac{1}{2}\|x^* - z\|^2 + (x^* - z)^T (x - x^*) \\ &\geq f(x^*) \end{aligned}$$

Therefore, $x^* = \operatorname{argmin}_{x \in C} f(x)$. □

Theorem 4. *There is a half-space which separates C and z. Formally,*

$$\forall x \in C, w^T x > w^T z$$

where $w = x^* - z$.

Proof.

$$\begin{aligned} &(x^* - z)^T (x - x^*) \geq 0 && \text{(by theorem 3)} \\ &\implies (x^* - z)^T x && \geq (x^* - z)^T x^* \\ & && \geq (x^* - z)^T (x^* - z + z) \\ & && \geq \|x^* - z\|^2 + (x^* - z)^T z \\ & && > (x^* - z)^T z \end{aligned}$$

□

3 Inequality constraints

Define the feasible region as

$$C = \{x : (\forall i \in I, c_i(x) \geq 0) \wedge (\forall i \in E, h_i(x) = 0)$$

Here $\{c_i : i \in I\}$ is the set of inequality constraints and $\{h_i : i \in I\}$ is the set of equality constraints. Since we can write the constraint $h_i(x) = 0$ as the 2 constraints $h_i(x) \geq 0$ and $-h_i(x) \geq 0$, we'll ignore equality constraints for now.

Our minimization algorithm will iteratively choose a feasible descent direction and make a small step in that direction.

By the definition of feasible direction, we get

$$u \in \mathrm{FS}(x) \iff \exists \overline{\alpha} > 0, \forall \alpha \in [0, \overline{\alpha}], c_i(x + \alpha u) \geq 0$$

Also, for $x \in C$, define LFS (called linearized feasible directions) as

$$\mathrm{LFS}(x) = \bigcap_{i \in I} \begin{cases} \mathbb{R}^d & \text{if } c_i(x) > 0 \\ \{u : \nabla_{c_i}(x)^T u \geq 0\} & \text{if } c_i(x) = 0 \end{cases}$$

Intuitively, LFS should be the same as FS. Unfortunately, they need not be the same.

Define descent directions (DS) as

$$u \in \mathrm{DS}(x) \iff \nabla_f(x)^T u < 0$$

When $\mathrm{FS}(x) \cap \mathrm{DS}(x) = \mathrm{LFS}(x) \cap \mathrm{DS}(x)$, we say that x is regular. Regularity always holds when the constraints are linear.

At a point x, a constraint c_i is said to be active iff $c_i(x) = 0$.

Theorem 5 (Farkas' Lemma)**.** *Let A be a d by m matrix and $b \in \mathbb{R}^d$. For a vector x, let $x \geq 0$ mean that all components of x are non-negative. Let $T = \{u \mid b^T u < 0 \wedge A^T u \geq 0\}$. Let $L = \{\lambda \mid b = A\lambda \wedge \lambda \geq 0\}$. Then $T = \{\} \iff L \neq \{\}$.*

Let I' be the set of active constraints at x^*. Let $|I'| = m$. Let A be the matrix whose i^{th} column is $\nabla_{c_i}(x^*)$. Then A is a d by m matrix. Let $b = \nabla_f(x^*)$. Then

$$u \in \mathrm{LFS}(x^*) \iff A^T u \geq 0 \qquad\qquad u \in \mathrm{DS}(x^*) \iff b^T u < 0$$

Then by Farkas' lemma, we get that

$$\mathrm{LFS}(x^*) \cap \mathrm{DS}(x^*) = \{\} \iff (\exists \lambda \geq 0, b = A\lambda)$$

For such a λ, we have

$$\nabla_f(x^*) = A\lambda = \sum_{i \in I'} \lambda_i \nabla_{c_i}(x^*)$$

This is equivalent to saying that

$$\nabla_f(x^*) = \sum_{i \in I} \lambda_i \nabla_{c_i}(x^*) \quad \text{where } \lambda_i c_i(x^*) = 0$$

If x^* is a local minimum and a regular point, then $\mathrm{LFS}(x^*) \cap \mathrm{DS}(x^*) = \{\}$. So there exists $\lambda \in \mathbb{R}^m$ such that

- (Primal feasibility) $\forall i \in I, c_i(x^*) \geq 0$.
- (Stationarity) $\nabla_f(x^*) = \sum_{i \in I} \lambda_i \nabla_{c_i}(x^*)$.
- (Dual feasibility) $\forall i \in I, \lambda_i \geq 0$.
- (Complementary slackness) $\forall i \in I, \lambda_i c_i(x^*) = 0$.

These 4 conditions are called 'KKT conditions'. When these conditions hold for x and λ, (x, λ) is said to be a KKT point.

This is generally stated using the Lagrangian function (we're also going to consider the equality constraints now):

$$L(x, \lambda, \mu) = f(x) - \lambda^T c(x) - \mu^T h(x)$$

- (Primal feasibility) $c(x^*) \geq 0$ and $h(x^*) = 0$.
- (Stationarity) $\nabla_x L(x, \lambda, \mu) = 0$.
- (Dual feasibility) $\lambda \geq 0$.
- (Complementary slackness) $\forall i \in I, \lambda_i c_i(x^*) = 0$.

CMO: Constrained optimization for convex functions

Eklavya Sharma

1 Convex function and convex constraints

Let's analyze the following problem:

$$\begin{array}{ll} \min_x & f(x) \\ \text{where} & c_i(x) \leq 0 \quad \forall i \in I \\ & h_j(x) = 0 \quad \forall j \in I \end{array}$$

Here f and c_i are convex and C^1 and h_j is linear, i.e. $h_j(x) = a_j^T x - b_j$.

1.1 Feasible region is a convex set

Lemma 1 (Homework)**.** *The set $S_i = \{x : c_i(x) \leq 0\}$ is convex.*

Lemma 2 (Homework)**.** *The set $S_j = \{x : h_j(x) = 0\}$ is convex.*

Lemma 3 (Homework)**.** *The intersection of convex sets is convex.*

1.2 KKT point gives global minimum

Define the Lagrangian as

$$L(x, \lambda, \mu) = f(x) + \lambda^T c(x) + \mu^T h(x)$$

Lemma 4. *If $\lambda_i \geq 0$ and x is a feasible point, then $f(x) \geq L(x, \mu, \lambda)$.*

Proof. Since x is a feasible point,

$$\begin{aligned} & c_i(x) \leq 0 \wedge h_j(x) = 0 \\ & \implies \lambda^T c(x) \leq 0 \wedge \mu^T h(x) = 0 \\ & \implies f(x) + \lambda^T c(x) + \mu^T h(x) \leq f(x) \\ & \implies L(x, \lambda, \mu) \leq f(x) \end{aligned}$$

□

Lemma 5. *Let (x^*, λ^*, μ^*) be a KKT point. Then $f(x^*) = L(x^*, \mu^*, \lambda^*)$.*

Proof.

$$\begin{aligned}
&\lambda_i^* c_i(x^*) = 0 \wedge h_j(x^*) = 0 && \text{(complementary slackness and primal feasibility)}\\
&\implies \lambda^{*T} c(x^*) = 0 \wedge \mu^{*T} h(x^*) = 0\\
&\implies f(x^*) + \lambda^{*T} c(x^*) + \mu^{*T} h(x^*) = f(x^*)\\
&\implies L(x^*, \lambda^*, \mu^*) = f(x)
\end{aligned}$$

□

Theorem 6 (Proved previously)**.** *Let f be C^1 and convex. Then*

$$\forall u, v \in \mathbb{R}^d, f(v) \geq f(u) + \nabla_f(u)^T (v - u)$$

Theorem 7. *Let (x^*, λ^*, μ^*) be a KKT point. Then x^* is a constrained global minimum of f.*

Proof. Let x be a feasible point.

$$\begin{aligned}
f(x) &\geq L(x, \lambda^*, \mu^*) && \text{(by lemma 4)}\\
&= f(x) + \sum_i \lambda_i^* c_i(x) + \sum_j \mu_j^* (a_j^T x - b_j)\\
&\geq \left(f(x) + \nabla_f(x^*)^T (x - x^*)\right)\\
&\quad + \sum_i \lambda_i^* \left(c_i(x^*) + \nabla_{c_i}(x^*)^T (x - x^*)\right)\\
&\quad + \sum_j \mu_j^* \left(a_j^T (x - x^*) + (a_j^T x^* - b_j)\right) && \text{(by theorem 6)}\\
&= \left(f(x^*) + \sum_i \lambda_i^* c_i(x^*) + \sum_j \mu_j^* (a_j^T x^* - b_j)\right)\\
&\quad + (x - x^*)^T \left(\nabla_f(x^*) + \sum_i \lambda_i^* \nabla_{c_i}(x^*) + \sum_j \mu_j^* a_j\right) && \text{(rearrange terms)}\\
&= L(x^*, \lambda^*, \mu^*) + (x - x^*)^T (\nabla_x L)(x^*, \lambda^*, \mu^*)\\
&= f(x^*) && \text{(by lemma 5 and stationarity)}
\end{aligned}$$

Since for all feasible points $f(x) \geq f(x^*)$, x^* is a constrained global minimum of f. □

Note that unlike the necessary conditions for local minimum, here we do not require regularity.

1.3 Example: Projection over ball

Consider the optimization problem:

$$\min_x \frac{1}{2} \|x - z\|^2 \text{ where } \|x\|^2 \leq r^2$$

Here z lies outside the feasible region.

$\|x - z\|^2$ and $\|x\|^2$ are convex functions (because their hessian is $2I$, which is positive definite), so this is a convex optimization problem.

$$L(x, \lambda) = \frac{1}{2}\|x - z\|^2 + \lambda(\|x\|^2 - r^2)$$

Applying the KKT conditions, we get

- Stationarity: $z = (2\lambda + 1)x$.
- Primal feasibility: $\|x\|^2 \leq r^2$.
- Dual feasibility: $\lambda \geq 0$.
- Complementary slackness: $\lambda(\|x\|^2 - r^2) = 0$.

If we take $\lambda = 0$, then stationarity gives us $x = z$. This violates feasibility, so this is not possible. Therefore, complementary slackness gives us $\|x\|^2 = r^2$. On simplifying, we get

$$x = \frac{r}{\|z\|}z \qquad \lambda = \frac{1}{2}\left(\frac{\|z\|}{r} - 1\right) \qquad f(x) = \frac{1}{2}(r - \|z\|)^2$$

CMO: Projected Gradient Descent

Eklavya Sharma

Projected gradient descent is an algorithm for convex constrained optimization that is similar to gradient descent. In this algorithm, we use gradient descent and if we ever move out of the feasible set, we project the point back to the feasible set.

Algorithm 1 `proj-grad-desc`(f, C, x_0): Minimize $f : \mathbb{R}^d \mapsto \mathbb{R}$ (in C^1, not necessarily convex) over the convex feasible set C. $x_0 \in C$ is the initial point.

$x^{(\min)} = x^{(0)}$
for t from 0 to ∞ **do**
  Choose step size α_t.
  $x^{(t+1)} = \text{proj}_C(x^{(t)} - \alpha_t \nabla_f(x^{(t)}))$
  $x^{(\min)} = \text{argmin}_{x \in \{x^{(\min)}, x^{(t+1)}\}} f(x)$
  if (stopping criterion) **then**
    return $x^{(\min)}$
  end if
end for

1 Finding projection: Example

Projected gradient descent requires a subroutine for finding projection. There is no easy general method for this. As an example we'll see how it's done for the constraint $Ax = b$, where A is an m by d matrix.

$$\text{proj}_{Ax=b}(z) = \underset{\substack{x \\ Ax=b}}{\text{argmin}} \frac{1}{2} \|x - z\|^2$$

Since $\|x - z\|^2$ is a convex function, a KKT point will give us the global minimum.

$$L(x, \mu) = \frac{1}{2} \|x - z\|^2 + \mu^T (Ax - b)$$

By stationarity, we get

$$(\nabla_x L)(x, \mu) = (x - z) + A^T \mu = 0 \implies x = z - A^T \mu$$

By primal feasibility, we get

$$b = Ax = A(z - A^T \mu) \implies \mu = (AA^T)^{-1}(Az - b)$$
$$\implies x = (I - A^T (AA^T)^{-1} A)z - A^T (AA^T)^{-1} b$$

The above x is $\text{proj}_{Ax=b}(z)$.

We can plug the above equation into the algorithm to get a simpler expression:

$$x^{(t+1)} = x^{(t)} - \alpha_t (I - A^T (AA^T)^{-1} A) \nabla_f(x^{(t)})$$

2 Convergence Analysis

Theorem 1 (Proved eariler)**.** *Let $f \in C^1_L$. Then $\forall x, y \in \mathbb{R}^d$*

$$f(y) \leq f(x) + \nabla_f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2$$

Theorem 2 (Proved earlier)**.** *Let C be a convex set. Let $z \notin C$. Then $\forall x \in C$*

$$(\text{proj}_C(z) - z)^T(x - \text{proj}_C(z)) \geq 0$$

Let the objective function f in the projected gradient algorithm be C^1_L.

Theorem 3. *The projected gradient algorithm converges if we choose step size less than $\frac{2}{L}$ (but it may not converge to a local minimum).*

Proof.

$$\begin{aligned}
&x_{t+1} = \text{proj}_C(x_t - \alpha_t \nabla_f(x_t)) \\
&\implies (x_{t+1} - x_t + \alpha_t \nabla_f(x_t))^T(x_t - x_{t+1}) \geq 0 && \text{(by theorem 2)} \\
&\implies \nabla_f(x_t)^T(x_{t+1} - x_t) \leq -\frac{\|x_{t+1} - x_t\|^2}{\alpha_t} && (1)
\end{aligned}$$

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \nabla_f(x_t)^T(x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2 && \text{(by theorem 1)} \\
&\leq f(x_t) + \|x_{t+1} - x_t\|^2 \left(\frac{L}{2} - \frac{1}{\alpha_t}\right) && \text{(by equation (1))}
\end{aligned}$$

If we choose $\alpha_t < \frac{2}{L}$, then $f(x_{t+1}) < f(x_t)$. Assuming that f is lower-bounded, this means that the algorithm will converge. □

Theorem 4 (Proved earlier)**.** *Let f be C^1 and convex. Then*

$$\forall u, v \in \mathbb{R}^d, f(v) \geq f(u) + \nabla_f(u)^T(v - u)$$

Theorem 5. *When f is convex, x_{min} converges to a minimum if we choose step size less than $\frac{1}{L}$. Also, let x^* be a minimum of f, $E_T = \min_{0 \leq t \leq T}(f(x_t) - f(x^*))$ and $\alpha = \min_{t=0}^{T} \alpha_t$. Then*

$$E_T \leq \frac{\|x_0 - x^*\|^2}{2\alpha T}$$

Proof.

$$\begin{aligned}
&f(x^*) - f(x_t) \geq \nabla_f(x_t)^T(x^* - x_t) && \text{(by theorem 4)} \\
&\implies f(x_t) \leq f(x^*) + \nabla_f(x_t)^T(x_t - x^*)
\end{aligned}$$

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \nabla_f(x_t)^T(x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2 && \text{(by theorem 1)} \\
&\leq \big(f(x^*) + \nabla_f(x_t)^T(x_t - x^*)\big) + \nabla_f(x_t)^T(x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&\leq f(x^*) + \nabla_f(x_t)^T(x_{t+1} - x^*) + \frac{L}{2}\|x_{t+1} - x_t\|^2
\end{aligned}$$

$$
\begin{aligned}
& x_{t+1} = \operatorname{proj}_C(x_t - \alpha_t \nabla_f(x_t)) \\
& \implies (x_{t+1} - x_t + \alpha_t \nabla_f(x_t))^T (x^* - x_{t+1}) \geq 0 && \text{(by theorem 2)} \\
& \implies (x_{t+1} - x_t)^T (x_{t+1} - x^*) + \alpha_t \nabla_f(x_t)^T (x_{t+1} - x^*) \leq 0 \\
& \implies \nabla_f(x_t)^T (x_{t+1} - x^*) \leq -\frac{1}{\alpha_t} (x_{t+1} - x_t)^T (x_{t+1} - x^*)
\end{aligned}
$$

$$
\begin{aligned}
& f(x_{t+1}) - f(x^*) \\
& \leq \nabla_f(x_t)^T (x_{t+1} - x^*) + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
& \leq -\frac{1}{\alpha_t} (x_{t+1} - x_t)^T (x_{t+1} - x^*) + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
& = -\frac{1}{\alpha_t} (\delta_{t+1} - \delta_t)^T \delta_{t+1} + \frac{L}{2} \|\delta_{t+1} - \delta_t\|^2 && \text{(where } \delta_t = x_t - x^*\text{)} \\
& = -\frac{\|\delta_{t+1}\|^2}{\alpha_t} + \frac{L}{2} \|\delta_{t+1} - \delta_t\|^2 + \frac{\delta_t^T \delta_{t+1}}{\alpha_t} \\
& = -\frac{\|\delta_{t+1}\|^2}{\alpha_t} + \frac{L}{2} \|\delta_{t+1} - \delta_t\|^2 + \frac{\|\delta_{t+1}\|^2 + \|\delta_t\|^2 - \|\delta_{t+1} - \delta_t\|^2}{2\alpha_t} \\
& = \frac{\|\delta_t\|^2 - \|\delta_{t+1}\|^2}{2\alpha_t} + \frac{1}{2} \left(L - \frac{1}{\alpha_t} \right) \|\delta_{t+1} - \delta_t\|^2
\end{aligned}
$$

Let $E(x) = f(x) - f(x^*)$. If we always choose $\alpha_t < \frac{1}{L}$, then

$$
0 \leq E(x_{t+1}) < \frac{\|\delta_t\|^2 - \|\delta_{t+1}\|^2}{2\alpha_t} \implies \|\delta_t\| > \|\delta_{t+1}\|
$$

Let $E_T = \min_{0 \leq t \leq T} E(x_t)$ and $\alpha = \min_{t=0}^{T} \alpha_t$. Then

$$
\begin{aligned}
E_T &\leq \frac{1}{T} \sum_{t=0}^{T-1} E(x_{t+1}) \\
&< \frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\delta_t\|^2 - \|\delta_{t+1}\|^2}{2\alpha_t} \\
&\leq \frac{1}{2\alpha T} \sum_{t=0}^{T-1} (\|\delta_t\|^2 - \|\delta_{t+1}\|^2) \\
&= \frac{1}{2\alpha T} (\|\delta_0\|^2 - \|\delta_T\|^2) \\
&\leq \frac{\|\delta_0\|^2}{2\alpha T}
\end{aligned}
$$

When $T \to \infty$, $E_T \to 0$. Since $E_T = E(x_{\min})$, $x_{\min}$ converges to a minimum. □

CMO: Active Set Method

Eklavya Sharma

The active set method is a way of solving optimization problems with a convex quadratic objective and linear constraints. In this method, we guess the active set of constraints and then solve the resulting problem which has only equality constraints. We make multiple such guesses and each guess helps refine the next guess.

1 Equality constraints

We'll first figure out how to solve the optimization problem which has only equality constraints:

$$\min_x f(x) \text{ where } Ax = b$$

Here $f(x) = \frac{1}{2}x^TQx - h^Tx$, where Q is symmetric and positive definite. The i^{th} row of A is a_i^T. $x \in \mathbb{R}^d$ and $b \in \mathbb{R}^m$.

We'll find a KKT point for this problem. Since the objective is strictly convex and the constraints are linear, the KKT point gives us the unique global minimum of this problem.

$$L(x, \mu) = f(x) - \mu^T(Ax - b)$$

By stationarity, we get

$$\nabla_f(x) - A^T\mu = 0 \implies Qx - A^T\mu = h$$

By primal feasibility, we get

$$Ax = b$$

These 2 conditions can be written as a system of linear equations. Solving this system will give us the KKT point (x, λ).

$$\begin{bmatrix} Q & -A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} h \\ b \end{bmatrix}$$

This system of equations is guaranteed to have a solution, since (x, λ) is a KKT point iff it satisfies these conditions and the first-order necessary conditions for local minimum imply that a KKT point must exist (regularity is ensured because the constraints are linear).

2 Inequality constraints

We'll now try to solve this optimization problem:

$$\min_x f(x) \text{ where } Ax \geq b$$

Define argminset as

$$\operatorname*{argminset}_x g(x) = \{\widehat{x} : g(\widehat{x}) = \min_x g(x)\}$$

The subroutine eqsolve(Q, h, A, b) returns (x^*, μ^*) such that

$$\begin{bmatrix} Q & -A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \mu^* \end{bmatrix} = \begin{bmatrix} h \\ b \end{bmatrix}$$

This means that

$$x^* = \operatorname*{argmin}_{\substack{x \\ Ax=b}} \frac{1}{2} x^T Q x - h^T x$$

The subroutine blocking-constraints is defined as

$$\text{blocking-constraints}(A, x, u) = \operatorname*{argminset}_{\substack{i \\ a_i^T u < 0}} \frac{a_i^T x - b_i}{-a_i^T u}$$

Let A_B denote the matrix obtained by taking all rows of A whose indices lie in B.

Algorithm 1 `active-set-method`$(Q, h, A, b, x^{(0)})$: x_0 is the initial feasible point

1: $B^{(0)} = \{i : a_i^T x^{(0)} = b_i\}$
2: $t = 0$
3: **while** `true` **do**
4: **while** `true` **do**
5: $(u^{(t)}, \mu^{(t)}) = \text{eqsolve}(Q, h - Qx^{(t)}, A_{B^{(t)}}, 0)$
6: **if** $u^{(t)}$ `==` 0 **then**
7: `break`
8: **end if**
9: $I^{(t)} = \text{blocking-constraints}(A, x^{(t)}, u^{(t)})$
10: $\alpha^{(t)} = \left(1 \text{ if } I^{(t)} \texttt{ == } \{\} \text{ else } \min\left(1, \frac{a_i^T x^{(t)} - b_i}{-a_i^T u^{(t)}}\right)\right)$, where $i \in I^{(t)}$.
11: $x^{(t+1)} = x^{(t)} + \alpha^{(t)} u^{(t)}$
12: $B^{(t+1)} = B^{(t)} \cup (I^{(t)} \text{ if } \alpha^{(t)} < 1 \text{ else } \{\})$
13: $t = t + 1$
14: **end while**
15: Find $l^{(t)} \in B^{(t)}$ such that $\mu_{l^{(t)}}^{(t)} < 0$.
16: **if** $l^{(t)}$ `==` `null` **then**
17: $\lambda_i = \begin{cases} \mu_i^{(t)} & i \in B^{(t)} \\ 0 & i \notin B^{(t)} \end{cases}$
18: **return** $(x^{(t)}, \lambda)$
19: **else**
20: $B^{(t+1)} = B^{(t)} - \{l^{(t)}\}$
21: $x^{(t+1)} = x^{(t)}$
22: $t = t + 1$
23: **end if**
24: **end while**

Lemma 1 (Invariants)**.** *Let $\mathcal{A}(x) = \{i : a_i^T x = b_i\}$. The following are true for all t:*

- *$x^{(t)}$ is a feasible solution.*
- *$B^{(t)} \subseteq \mathcal{A}(x^{(t)})$.*

Proof by induction.
Base case: $x^{(0)}$ is given to be feasible. $B^{(0)} = \mathcal{A}(x^{(0)})$ by line 1. Therefore, the hypothesis holds for $t = 0$.

Inductive step: Assume the hypothesis holds at t.

Case 1: $u^{(t)} \neq 0$

If $u^{(t)} \neq 0$, then $x^{(t+1)} = x^{(t)} + \alpha^{(t)} u^{(t)}$. Let $J = \{i : a_i^T u^{(t)} < 0\}$. Then $\forall i \in J$,

$$\alpha^{(t)} \leq \frac{a_i^T x^{(t)} - b_i}{-a_i^T u^{(t)}} \implies b_i \leq a_i^T (x^{(t)} + \alpha^{(t)} u^{(t)}) = a_i^T x^{(t+1)}$$

When $i \notin J$, then $a_i^T u^{(t)} \geq 0$. Also, $a_i^T x^{(t)} \geq b_i$, by inductive hypothesis. Therefore,

$$a_i^T x^{(t+1)} = a_i^T x^{(t)} + \alpha^{(t)} a_i^T u^{(t)} \geq b_i$$

Therefore, $x^{(t+1)}$ is feasible.

$$i \in B^{(t)} \implies a_i^T x^{(t)} = b_i \qquad (B^{(t)} \subseteq \mathcal{A}(x^{(t)}) \text{ by inductive hypothesis})$$

$$(u^{(t)}, \mu^{(t)}) = \text{eqsolve}(Q, h - Qx^{(t)}, A_{B^{(t)}}, 0) \implies \forall i \in B^{(t)}, a_i^T u^{(t)} = 0$$

Therefore, $\forall i \in B^{(t)}$,

$$a_i^T x^{(t+1)} = a_i^T x^{(t)} + \alpha^{(t)} a_i^T u^{(t)} = b_i$$

Therefore, $B^{(t)} \subseteq \mathcal{A}(x^{(t+1)})$.

If $\alpha^{(t)} \geq 1$ or $I^{(t)} = \{\}$, then $B^{(t+1)} = B^{(t)} \subseteq \mathcal{A}(x^{(t+1)})$. Suppose $\alpha^{(t)} < 1$ and $I^{(t)} \neq \{\}$. Then $B^{(t+1)} = B^{(t)} \cup I^{(t)}$ and $\forall i \in I^{(t)}$,

$$\alpha^{(t)} = \frac{a_i^T x^{(t)} - b_i}{-a_i^T u^{(t)}} \implies b_i = a_i^T (x^{(t)} + \alpha^{(t)} u^{(t)}) = a_i^T x^{(t+1)}$$

Therefore, $I^{(t)} \subseteq \mathcal{A}(x^{(t+1)})$ which implies that $B^{(t+1)} \subseteq \mathcal{A}(x^{(t+1)})$.

Case 2: $u^{(t)} = 0$

By inductive hypothesis, $x^{(t)}$ is feasible. Therefore, $x^{(t+1)} = x^{(t)}$ is also feasible. By inductive hypothesis, $B^{(t)} \subseteq \mathcal{A}(x^{(t)})$. Therefore, $B^{(t+1)} \subset B^{(t)} \subseteq \mathcal{A}(x^{(t+1)})$.

Therefore, the inductive hypothesis holds at $t + 1$. By mathematical induction, the inductive hypothesis holds for all $t \geq 0$. □

Theorem 2 (Correctness)**.** *If* active-set-method *returns* $(x^{(t)}, \lambda)$*, then* $x^{(t)}$ *is a global minimum.*

Proof. We'll prove that $(x^{(t)}, \lambda)$ is a KKT point. This is a sufficient condition for minimum since this is a convex optimization problem.

The algorithm enters the conditional block on line 16 iff $u^{(t)} = 0$ and $\mu_i^{(t)} \geq 0$ for all $i \in B^{(t)}$.

$$\begin{aligned}
&(0, \mu) = \text{eqsolve}(Q, h - Qx^{(t)}, A_{B^{(t)}}, 0) \\
&\implies -A_{B^{(t)}}^T \mu = h - Qx^{(t)} \\
&\implies \nabla_f(x^{(t)}) = A_{B^{(t)}}^T \mu = \sum_{i \in B^{(t)}} \mu_i a_i = \sum_{i=1}^{m} \lambda_i a_i
\end{aligned}$$

This proves stationarity for $(x^{(t)}, \lambda)$.

Primal feasibility follows from lemma 1.

Dual feasibility follows from the fact that $\lambda_i = \mu_i^{(t)} \geq 0$ for all $i \in B^{(t)}$ and $\lambda_i = 0$ for all $i \notin B^{(t)}$.

When $i \notin B^{(t)}$, then $\lambda_i = 0$. When $i \in B^{(t)}$, then $a_i^T x^{(t)} = b_i$ (since $B^{(t)} \subseteq \mathcal{A}(x^{(t)})$ by lemma 1). This ensures complementary slackness.

Therefore, $(x^{(t)}, \lambda)$ is a KKT point. □

Lemma 3. $B^{(t)} = B^{(t+1)} \implies u^{(t+1)} = 0$

Proof.

$$
\begin{aligned}
&(I^{(t)} = \{\} \implies \alpha^{(t)} = 1) \\
&\implies (\alpha^{(t)} < 1 \implies I^{(t)} \neq \{\} \implies B^{(t+1)} \neq B^{(t)}) \\
&\implies (B^{(t+1)} = B^{(t)} \implies \alpha^{(t)} = 1) \\
&\implies (B^{(t+1)} = B^{(t)} \implies x^{(t+1)} = x^{(t)} + u^{(t)})
\end{aligned}
$$

By lemma 1, $x^{(t)} + u^{(t)}$ is feasible.

$$
\begin{aligned}
&(u^{(t)}, \mu^{(t)}) = \text{eqsolve}(Q, h - Qx^{(t)}, A_{B^{(t)}}, 0) \\
&\implies u^{(t)} = \underset{\substack{u \\ A_{B^{(t)}} u = 0}}{\text{argmin}} f(x^{(t)} + u) \\
&\implies x^{(t+1)} = x^{(t)} + u^{(t)} = \underset{\substack{x \\ \forall i \in B^{(t)}, a_i^T x = b_i}}{\text{argmin}} f(x) \\
&\implies x^{(t+1)} = \underset{\substack{x \\ \forall i \in B^{(t+1)}, a_i^T x = b_i}}{\text{argmin}} f(x) \qquad (B^{(t+1)} = B^{(t)})
\end{aligned}
$$

$$
\begin{aligned}
&(u^{(t+1)}, \mu^{(t+1)}) = \text{eqsolve}(Q, h - Qx^{(t+1)}, A_{B^{(t+1)}}, 0) \\
&\implies u^{(t+1)} = \underset{\substack{u \\ A_{B^{(t+1)}} u = 0}}{\text{argmin}} f(x^{(t+1)} + u) \\
&\implies x^{(t+1)} + u^{(t+1)} = \underset{\substack{x \\ \forall i \in B^{(t+1)}, a_i^T x = b_i}}{\text{argmin}} f(x)
\end{aligned}
$$

Therefore, $x^{(t+1)} = x^{(t+1)} + u^{(t+1)} \implies u^{(t+1)} = 0$. □

Therefore, if the algorithm gets stuck in the inner while loop, $B^{(t+1)} \neq B^{(t)}$. But in each iteration, we add a constraint and there are a finite number of constraints. Therefore, it's not possible to get stuck in the inner while loop.

Lemma 4. *If the algorithm does not terminate in the k^{th} step,*

$$u^{(k)} = 0 \implies (a_{l^{(k)}}^T u^{(k+1)} > 0 \wedge \nabla_f(x^{(k)})^T u^{(k+1)} < 0 \wedge f(x^{(k)} + u^{(k+1)}) < f(x^{(k)}))$$

Proof. Let $g^{(t)} = \nabla_f(x^{(t)})^T = Qx^{(t)} - h$. For any t,

$$
\begin{aligned}
&(u^{(t)}, \mu^{(t)}) = \text{eqsolve}(Q, h - Qx^{(t)}, A_{B^{(t)}}, 0) \\
&\implies \begin{bmatrix} Q & -A_{B^{(t)}}^T \\ A_{B^{(t)}} & 0 \end{bmatrix} \begin{bmatrix} u^{(t)} \\ \mu^{(t)} \end{bmatrix} = \begin{bmatrix} h - Qx^{(t)} \\ 0 \end{bmatrix} \\
&\implies Qu^{(t)} - A_{B^{(t)}}^T \mu^{(t)} = h - Qx^{(t)} = -g^{(t)} \wedge A_{B^{(t)}} u = 0 \\
&\implies \left(g^{(t)} + Qu^{(t)} = A_{B^{(t)}}^T \mu^{(t)} = \sum_{i \in B^{(t)}} \mu_i^{(t)} a_i \right) \wedge (\forall i \in B^{(t)}, a_i^T u = 0)
\end{aligned}
$$

For notational convenience, let $l = l^{(k)}$. Since $u^{(k)} = 0, x^{(k+1)} = x^{(k)} \implies g^{(k+1)} = g^{(k)}$ and $B^{(k+1)} = B^{(k)} - \{l\}$. Since the algorithm didn't terminate, $\mu_l^{(k)} < 0$.

$$g^{(k)} = g^{(k)} + Qu^{(k)} = \sum_{i \in B^{(k)}} \mu_i^{(k)} a_i = \mu_l^{(k)} a_l + \sum_{i \in B^{(k+1)}} \mu_i^{(k)} a_i$$

$$g^{(k+1)} + Qu^{(k+1)} = \sum_{i \in B^{(k+1)}} \mu_i^{(k+1)} a_i$$

On subtracting these 2 equations, we get

$$Qu^{(k+1)} = -\mu_l^{(k)} a_l + \sum_{i \in B^{(k+1)}} (\mu_i^{(k+1)} - \mu_i^{(k)}) a_i$$

$\forall i \in B^{(k+1)}, a_i^T \mu_i^{(k+1)} = 0$. Therefore,

$$\begin{aligned}
&(u^{(k+1)})^T Q u^{(k+1)} \\
&= -\mu_l^{(k)} a_l^T u^{(k+1)} + \sum_{i \in B^{(k+1)}} (\mu_i^{(k+1)} - \mu_i^{(k)}) a_i^T u^{(k+1)} \\
&= -\mu_l^{(k)} a_l^T u^{(k+1)} \\
&\implies a_l^T u^{(k+1)} = \frac{(u^{(k+1)})^T Q u^{(k+1)}}{-\mu_l^{(k)}} > 0 \qquad (Q \text{ is PD and } \mu_l^{(k)} < 0)
\end{aligned}$$

$$\begin{aligned}
&g^{(k)} + Qu^{(k+1)} = \sum_{i \in B^{(k+1)}} \mu_i^{(k+1)} a_i \\
&\implies {g^{(k)}}^T u^{(k+1)} + {u^{(k+1)}}^T Q u^{(k+1)} = \sum_{i \in B^{(k+1)}} \mu_i^{(k+1)} a_i u^{(k+1)} = 0 \\
&\implies {g^{(k)}}^T u^{(k+1)} = -{u^{(k+1)}}^T Q u^{(k+1)}
\end{aligned}$$

$$\begin{aligned}
&f(x^{(k)} + u^{(k+1)}) \\
&= f(x^{(k)}) + {g^{(k)}}^T u^{(k+1)} + \frac{1}{2} {u^{(k+1)}}^T Q u^{(k+1)} && \text{(Taylor series)} \\
&= f(x^{(k)}) - \frac{1}{2} {u^{(k+1)}}^T Q u^{(k+1)} \\
&< f(x^{(k)}) && (Q \text{ is PD})
\end{aligned}$$

□

Corollary 4.1. $u^{(k)} = 0 \implies u^{(k+1)} \neq 0$

Lemma 5. $\alpha^{(t)} > 0 \implies f(x^{(t+1)}) < f(x^{(t)})$

Proof. For $\alpha^{(t)}$ to exist, $u^{(t)}$ must be non-zero. Therefore, $x^{(t)} \neq x^{(t)} + u^{(t)}$.

$$u^{(t)} = \underset{\substack{u \\ A_{B^{(t)}} u = 0}}{\operatorname{argmin}} f(x^{(t)} + u) \qquad \implies x^{(t)} + u^{(t)} = \underset{\substack{x \\ (A_{B^{(t)}}) x = (b_{B^{(t)}})}}{\operatorname{argmin}} f(x)$$

Since $x^{(t)}$ satisfies $A_{B^{(t)}}x = b_{B^{(t)}}$, it is a feasible solution to the above optimization problem. However, $x^{(t)} + u^{(t)}$ is the optimial solution and $x^{(t)} \neq x^{(t)} + u^{(t)}$. Since f is a strictly convex function, it has a unique global minimum. Therefore, $f(x^{(t)} + u^{(t)}) < f(x^{(t)})$.

$$\begin{aligned}
f(x^{(t+1)}) &= f(x^{(t)} + \alpha^{(t)}u^{(t)}) \\
&= f((1-\alpha^{(t)})x^{(t)} + \alpha^{(t)}(x^{(t)} + u^{(t)})) \\
&\leq (1-\alpha^{(t)})f(x^{(t)}) + \alpha^{(t)}f(x^{(t)} + u^{(t)}) \\
&< (1-\alpha^{(t)})f(x^{(t)}) + \alpha^{(t)}f(x^{(t)}) \qquad (f(x^{(t)} + u^{(t)}) < f(x^{(t)}) \text{ and } \alpha^{(t)} > 0) \\
&= f(x^{(t)})
\end{aligned}$$

□

Lemma 6. *Let* $N^{(t)} = \{i : a_i^T u^{(t)} < 0\}$. *Then*

$$B^{(t)} \cap N^{(t)} = \{\} \qquad \mathcal{A}(x^{(t)}) \cap N^{(t)} = (\mathcal{A}(x^{(t)}) - B^{(t)}) \cap N^{(t)}$$

$$\alpha^{(t)} = 0 \iff \mathcal{A}(x^{(t)}) \cap N^{(t)} \neq \{\} \iff I^{(t)} = \mathcal{A}(x^{(t)}) \cap N^{(t)} \wedge I^{(t)} \neq \{\}$$

CMO: Duality

Eklavya Sharma

1 Duality

Consider the optimization problem P:

$$\min_{x \in \mathbb{R}^d} f(x) \text{ where } \forall i \in I, c_i(x) \geq 0 \wedge \forall j \in J, h_j(x) = 0$$

The corresponding Lagrangian is

$$L(x, \lambda, \mu) = f(x) - \lambda^T c(x) - \mu^T h(x)$$

Define g as

$$g(\lambda, \mu) = \min_{x \in \mathbb{R}^d} L(x, \lambda, \mu)$$

Let D be this optimization problem:

$$\max_{\lambda, \mu} g(\lambda, \mu) \text{ where } g(\lambda, \mu) \neq -\infty \wedge \lambda \geq 0$$

Then D is said to be the dual of P.

Theorem 1 (Weak duality theorem)**.** *Let x_0 be a feasible solution to P and (λ_0, μ_0) be feasible solution to D. Then*

$$g(\lambda_0, \mu_0) \leq L(x_0, \lambda_0, \mu_0) \leq f(x_0)$$

Proof.

$$\begin{aligned}
& g(\lambda_0, \mu_0) \\
& = \min_{x \in \mathbb{R}^d} L(x, \lambda_0, \mu_0) \\
& \leq L(x_0, \lambda_0, \mu_0) \\
& = f(x_0) - \lambda_0^T c(x_0) - \mu^T h(x_0) \\
& \leq f(x_0) && (\lambda_0 \geq 0 \wedge c(x_0) \geq 0 \wedge h(x_0) = 0 \text{ by feasibility})
\end{aligned}$$

□

Definition 1 (Duality gap)**.** *Let x^* be the optimal solution to P and (λ^*, μ^*) be the optimal solution to D. Then the duality gap is defined to be the quantity*

$$f(x^*) - g(\lambda^*, \mu^*)$$

Corollary 1.1. *Let x_0 be a feasible solution to P and (λ_0, μ_0) be a feasible solution to D. If $f(x_0) = g(\lambda_0, \mu_0)$, then the duality gap is 0 and x_0 and (λ_0, μ_0) are optimal solutions.*

Proof. Let x^* be the optimal solution to P and (λ^*, μ^*) be the optimal solution to D. Then

$$g(\lambda_0, \mu_0) \leq g(\lambda^*, \mu^*) \leq f(x^*) \leq f(x_0) = g(\lambda_0, \mu_0)$$

Therefore,

$$g(\lambda_0, \mu_0) = g(\lambda^*, \mu^*) = f(x^*) = f(x_0)$$

□

2 Wolfe Dual

We'll now focus our attention on convex optimization problems. In the optimization problem P:

- Let f be a convex function.
- Let $c_i(x) = -f_i(x)$, where f_i is a convex function.
- Let $h_j(x) = a_j^T x - b_j$, where $a_j \in \mathbb{R}^d$ and $b \in \mathbb{R}^{|I|}$. Let A be the matrix whose j^{th} column is a_j.

Let WD be the optimization problem

$$\max_{x,\lambda,\mu} L(x, \lambda, \mu) \text{ where } \lambda \geq 0 \wedge \nabla_x L(x, \lambda, \mu) = 0$$

This problem is called the Wolfe Dual of P.

Theorem 2 (Proved previously)**.** *Let f be C^1 and convex. Then*

$$\forall u, v \in \mathbb{R}^d, f(v) \geq f(u) + \nabla_f(u)^T (v - u)$$

Lemma 3 (Proved previously)**.** *Let (x^*, λ^*, μ^*) be a KKT point. Then $f(x^*) = L(x^*, \mu^*, \lambda^*)$.*

Theorem 4. *Let (x^*, λ^*, μ^*) be a KKT point of P. Then (x^*, λ^*, μ^*) is the optimal solution to WD.*

Proof. Let (x, λ, μ) be a feasible point of WD.

$$
\begin{aligned}
& L(x^*, \lambda^*, \mu^*) \\
& = f(x^*) && \text{(by lemma 3)} \\
& \geq L(x^*, \lambda, \mu) && \text{(by } \lambda \geq 0 \text{ and weak duality)} \\
& = f(x^*) + \sum_i \lambda_i f_i(x^*) + \sum_j \mu_j (a_j^T x^* - b_j) \\
& \geq (f(x) + \nabla f(x)^T (x^* - x)) \\
& \quad + \sum_i \lambda_i (f_i(x) + \nabla_{f_i}(x)^T (x^* - x)) \\
& \quad + \sum_j \mu_j (a_j^T (x^* - x) - (a_j^T x - b_j)) && \text{(by theorem 2)} \\
& = \left(f(x) + \sum_i \lambda_i f_i(x) + \sum_j \mu_j (a_j^T x - b_j) \right) \\
& \quad + (x^* - x)^T \left(\nabla f(x) + \sum_i \lambda_i \nabla_{f_i}(x) + \sum_j \mu_j a_j \right) \\
& = L(x, \lambda, \mu) + (x^* - x)^T (\nabla_x L(x, \lambda, \mu)) \\
& = L(x, \lambda, \mu) && \text{(feasibility of WD implies } \nabla_x L(x, \lambda, \mu) = 0)
\end{aligned}
$$

Therefore, (x^*, λ^*, μ^*) maximizes WD. □

Therefore, to find the KKT point of a problem, we can optimize its Wolfe Dual.

Example 1.

$$\min_x \frac{1}{2} \|x\|^2 \ \textit{where}\ A^T x \geq b$$

The Lagrangian for this problem is

$$L(x, \lambda) = \frac{1}{2} \|x\|^2 - \lambda^T (A^T x - b)$$

$$\nabla_x L(x, \lambda) = x - A\lambda$$

The Wolfe Dual is

$$\max_{x,\lambda} \frac{1}{2} \|x\|^2 - \lambda^T (A^T x - b) \ \textit{where}\ x - A\lambda = 0 \ \textit{and}\ \lambda \geq 0$$

We can simplify this by substituting $x = A\lambda$ and removing the constraint

$$\max_\lambda b^T \lambda - \frac{1}{2} \|A\lambda\|^2 \ \textit{where}\ \lambda \geq 0$$

Example 2.

$$\min_x c^T x \ \textit{where}\ x \geq 0 \wedge Ax \geq b$$

The Lagrangian for this problem is

$$L(x, \lambda, \pi) = c^T x - \lambda^T (Ax - b) - \pi^T x = (c - A^T \lambda - \pi)^T x + b^T \lambda$$

$$\nabla_x L(x, \lambda, \pi) = c - A^T \lambda - \pi$$

The Wolfe Dual is

$$\max_{x,\lambda,\pi} (c - A^T \lambda - \pi)^T x + b^T \lambda \text{ where } c - A^T \lambda - \pi = 0 \text{ and } \lambda \geq 0 \text{ and } \pi \geq 0$$

We can simplify this by substituting $\pi = c - A^T \lambda$ *and removing the constraint*

$$\max_{x,\lambda} b^T \lambda \text{ where } A^T \lambda \leq c \text{ and } \lambda \geq 0$$

This gives us the dual linear program for this problem.