Steepest-Descent Method for minimizing a strongly convex function

Consider the strongly convex quadratic function

$$q(x) = \frac{1}{2}\langle Qx, x\rangle - \langle b, x\rangle + a$$

where Q is $n \times n$ symmetric positive definite matrix, $b \in \mathbb{R}^n$ and $a \in \mathbb{R}$. Let $r(x) = \nabla q(x) = Qx - b$.

We know that a strongly convex function has a unique minimizer. A sufficient condition for a point to be a global minimizer of q over $\mathbb{R}^n$ is that it is a critical point of q. Hence, the global minimizer x^* of q over $\mathbb{R}^n$ satisfies $r(x^*) = 0$, that is

$$x^* = Q^{-1}b.$$

By a theorem which we proved earlier we conclude that the limit point of the sequence $\{x_k\}_0^\infty$ generated by the steepest descent method is x^*.

The steepest descent $d_k = -\nabla q(x_k) = -r(x_k)$.

We denote $r(x_k)$ by r_k. Hence, $d_k = -r_k = -Qx_k + b$.

Since q is to be minimized over $\mathbb{R}^n$, we can apply exact minimization rule to calculate the step length α_k which minimizes $q(x_k - \alpha r_k)$ over $\mathbb{R}^n$. Now

$$\begin{aligned}
h(\alpha) &= q(x_k - \alpha r_k) \\
&= \frac{1}{2}\langle Q(x_k - \alpha r_k), x_k - \alpha r_k\rangle - \langle b, x_k - \alpha r_k\rangle + a \\
&= \frac{1}{2}\langle Qx_k, x_k\rangle - \alpha\langle Qx_k, r_k\rangle + \frac{\alpha^2}{2}\langle Qr_k, r_k\rangle \\
&\quad - \langle b, x_k\rangle + \alpha\langle b, r_k\rangle + a \qquad \left[\because \langle Qx_k, r_k\rangle = \langle Qr_k, x_k\rangle\right. \\
&= q(x_k) - \alpha\langle Qx_k - b, r_k\rangle + \frac{\alpha^2}{2}\langle Qr_k, r_k\rangle \\
&= q(x_k) - \alpha\|r_k\|^2 + \frac{\alpha^2}{2}\langle Qr_k, r_k\rangle
\end{aligned}$$

Exact minimizer α_k of h over $(0,\infty)$ is given by

$$0 = h'(\alpha_k) = -\|r_k\|^2 + \alpha_k \langle Q r_k, r_k \rangle$$

which implies $\alpha_k = \dfrac{\|r_k\|^2}{\langle Q r_k, r_k\rangle}$. Hence, the sequence generated by steepest-descent method is given by

$$\boxed{\begin{aligned} x_{k+1} &= x_k - \alpha_k r_k \\ \alpha_k &= \frac{\|r_k\|^2}{\langle Q r_k, r_k\rangle} \end{aligned}}$$

Try this scheme to find the next iterate starting from $(1,1)$ to minimize $q(x) = 2x_1^2 + x_2^2 - 3x_1 + 4$ over $\mathbb{R}^2$.

We next state the Kantorovich's inequality (without proof) to establish convergence rate for steepest-descent method for convex quadratic functions.

Kantorovich's Inequality If Q is a symmetric positive definite $n \times n$ matrix with eigenvalues $\{\lambda_i\}_1^n$ in the interval $[m, M]$ then

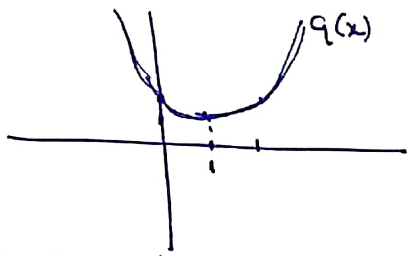
$$\frac{\langle Qx, x\rangle \langle Q^{-1}x, x\rangle}{\|x\|^4} \le \frac{(m+M)^2}{4mM}.$$

In the next theorem we establish convergence rate.

~~Theorem~~ Define the optimality gap

$$E(x) = q(x) - \min_{x \in \mathbb{R}^n} q(x).$$

For example if $q(x) = x^2 - 2x + 3$



$$\min_{x \in \mathbb{R}} q(x) = q(1) = 2$$

$$E(x) = x^2 - 2x + 1$$

We can see that the value of $E(x)$ decreases in each iteration of steepest-descent method. The question is at what rate?

Conditional number τ of a positive definite symmetric matrix Q is defined as

$$\tau = \frac{\lambda_{max}}{\lambda_{min}}$$

where λ_{max} and λ_{min} are maximum and minimum eigenvalues of Q.

For instance consider $Q = \begin{bmatrix} 3 & -2 \\ -2 & 4 \end{bmatrix}$ which is a positive definite matrix. Check the eigen values are $\frac{7 \pm \sqrt{17}}{2}$. Conditional number τ of $Q = \frac{7+\sqrt{17}}{7-\sqrt{17}}$.

In the next theorem we establish $E(x)$ decreases at a geometric rate.

Theorem In the steepest descent method for minimizing a strongly convex quadratic function $q(x)$, the optimality gap $E(x)$ decreases at a geometric rate

$$E(x_{k+1}) \leq \left(\frac{\tau-1}{\tau+1}\right)^2 E(x_k).$$

Proof We know $x_{k+1} = x_k - \alpha_k r_k$ where $\alpha_k = \frac{\|r_k\|^2}{\langle Qr_k, r_k\rangle}$. We can easily show that

$$\begin{aligned} E(x_{k+1}) &= E(x_k - \alpha_k r_k) \\ &= q(x_k - \alpha_k r_k) - \min_{x \in \mathbb{R}^n} q(x) \\ &= q(x_k) - \alpha_k \|r_k\|^2 + \frac{\alpha_k^2}{2} \langle Qr_k, r_k\rangle - \min_{x \in \mathbb{R}^n} q(x) \\ &= E(x_k) - \alpha_k \|r_k\|^2 + \frac{\alpha_k^2}{2} \langle Qr_k, r_k\rangle . \end{aligned}$$

As $\alpha_k = \frac{\|r_k\|^2}{\langle Qr_k, r_k\rangle}$ we have

$$E(x_{k+1}) = E(x_k) - \frac{\|r_k\|^4}{\langle Qr_k, r_k\rangle} + \frac{\|r_k\|^4}{2\langle Qr_k, r_k\rangle}$$

$$E(x_{k+1}) = E(x_k) - \frac{1}{2}\frac{\|r_k\|^4}{\langle Qr_k, r_k\rangle}$$

On dividing by $E(x_k)$ we have

$$\frac{E(x_{k+1})}{E(x_k)} = 1 - \frac{1}{2}\frac{\|r_k\|^4}{\langle Qr_k, r_k\rangle E(x_k)} . \qquad (1)$$

Let x^* be global minimizer of q over $\mathbb{R}^n$. Then $\nabla q(x^*) = 0$ and hence $\nabla E(x^*) = 0$ where $x^* = Q^{-1}b$. By Taylor's formula

$$E(x_k) = E(x^*) + \langle \nabla E(x^*), x_k - x^*\rangle + \frac{1}{2}\langle Q(x_k - x^*), x_k - x^*\rangle$$

as $\nabla^2 E(x^*) = Q$. As $E(x^*) = 0$, $\nabla E(x^*) = 0$ we have

$$E(x_k) = \frac{1}{2}\langle Q(x_k - x^*), x_k - x^*\rangle .$$

As $r_k = Qx_k - b = Qx_k - Qx^* = Q(x_k - x^*)$ we have

$$E(x_k) = \frac{1}{2}\langle r_k, Q^{-1}r_k\rangle = \frac{1}{2}\langle Q^{-1}r_k, r_k\rangle$$

Substituting in (1) we get

$$\frac{E(x_{k+1})}{E(x_k)} = 1 - \frac{\|r_k\|^4}{\langle Qr_k, r_k\rangle\langle Q^{-1}r_k, r_k\rangle}$$

Using Kantorovich's inequality in the interval $[\lambda_{min}, \lambda_{max}]$ we have

$$\frac{\langle Qr_k, r_k\rangle\langle Q^{-1}r_k, r_k\rangle}{\|r_k\|^4} \le \frac{[\lambda_{min} + \lambda_{max}]^2}{4\lambda_{min}\lambda_{max}}$$

$$= \frac{[1+\tau]^2}{4\tau}$$

$$\frac{4\tau}{(1+\tau)^2} \le \frac{\|r_k\|^4}{\langle Qr_k, r_k\rangle\langle Q^{-1}r_k, r_k\rangle}$$

$$\frac{-\|r_k\|^4}{\langle Qr_k, r_k\rangle\langle Q^{-1}r_k, r_k\rangle} \le \frac{-4\tau}{[\tau+1]^2}$$

$$\frac{E(x_{k+1})}{E(x_k)} \le 1 - \frac{4\tau}{(\tau+1)^2} = \left(\frac{\tau-1}{\tau+1}\right)^2 .$$

In the next corollary of the above theorem we show that the optimality gap $E(x)$ is halved in every $O(\tau)$ operations.

<u>Corollary</u> In the steepest-descent method for minimizing $q(x)$ the optimality gap $E(x)$ is halved in every $O(\tau)$ operations.

<u>Proof</u> Using the above theorem we have

$$\frac{E(x_m)}{E(x_0)} = \frac{E(x_m)}{E(x_{m-1})} \frac{E(x_{m-1})}{E(x_{m-2})} \cdots \frac{E(x_1)}{E(x_0)}$$

$$\leq \left(\frac{\tau-1}{\tau+1}\right)^{2m}$$

We want to see for what least value of m the value of $E(x_m) \leq \frac{1}{2} E(x_0)$. Let m be the smallest positive integer such that

$$\left(\frac{\tau-1}{\tau+1}\right)^{2m} \leq \frac{1}{2}.$$

$$\frac{\tau-1}{\tau+1} = 1 - \frac{2}{\tau+1}$$

If τ is large

$$-\ln 2 \simeq 2m \ln\left(1 - \frac{2}{\tau+1}\right)$$

$$\simeq \frac{-4m}{\tau+1} \simeq \frac{-4m}{\tau}.$$

Hence, $m = O(\tau)$.

$$\ln(1-x) = x + \frac{x^2}{2} + \text{---} \simeq x$$

if x is small.

<u>Constrained Optimization Problem</u>

We next recall the problem

$$\text{Minimize } f(x)$$
$$\text{subject to } x \in C$$

where ~~[illegible]~~ C is a closed convex subset of $\mathbb{R}^n$. Let f be differentiable on an open set containing C.

<u>Necessary Optimality</u> If $x^* \in C$ is a minimizer of f then

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \forall x \in C. \qquad (2)$$

We recall projection map $\Pi_C : \mathbb{R}^n \to C$ defined as

$$\Pi_C(x^*) = \{u \in C : \|x^* - u\| = \inf_{z \in C} \|x^* - z\|\}$$

It is known that $\Pi_C(x)$ is singleton for every $x^* \in \mathbb{R}^n$ as C is a closed convex set.

Projection inequality

$$\langle x^* - \Pi_C(x^*), z - \Pi_C(x^*) \rangle \leq 0 \ \forall z \in C$$

The angle between $x^* - \Pi_C(x^*)$ and $z - \Pi_C(x^*)$ is obtuse at the most right angle $\forall z \in C$

In the next lemma we give an equivalent condition to (2)

Lemma Let $C \subseteq \mathbb{R}^n$ be a closed convex set and f be differentiable on an open set containing C. Let $s > 0$. For $x^* \in C$ we have

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0 \ \forall x \in C \iff \Pi_C(x^* - s\nabla f(x^*)) = x^*.$$

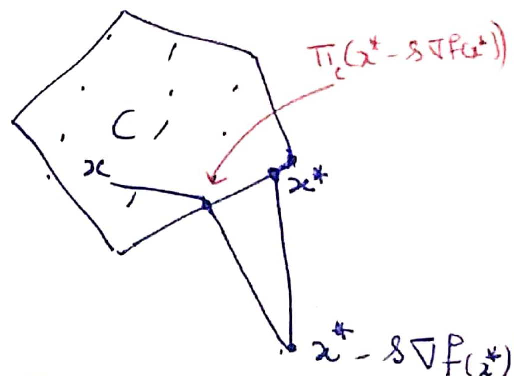
Proof By projection inequality

$$\langle x^* - s\nabla f(x^*) - \Pi_C(x^* - s\nabla f(x^*)), x - \Pi_C(x^* - s\nabla f(x^*)) \rangle \leq 0 \quad \forall x \in C$$

Hence $\Pi_C(x^* - s\nabla f(x^*)) = x^*$

$$\iff \langle x^* - s\nabla f(x^*) - x^*, x - x^* \rangle \leq 0 \ \forall x \in C$$

$$\iff \langle \nabla f(x^*), x - x^* \rangle \geq 0 \ \forall x \in C.$$



We now discuss the Gradient-Projection method which is modification of steepest descent method to deal with the problem

$$\text{Min } f(x)$$
$$\text{subject to } x \in C$$

where C is a closed convex set. The sequence $\{x_k\}$

generated by this method should be such that $x_k \in C$ for every k. Hence even though we initially move along the direction ~~$\nabla f(x_k)$~~ $-\nabla f(x_k)$ the new direction is obtained after taking projection onto C. The algorithm of the Gradient Projection method given below is self explanatory.

Step 0 Choose $x_0 \in C$, $s > 0$, $0 < \beta < 1$, $0 < \sigma < 1$.

Step k Given x_k compute

$$\bar{x}_k = \Pi_C (x_k - s \nabla f(x_k))$$

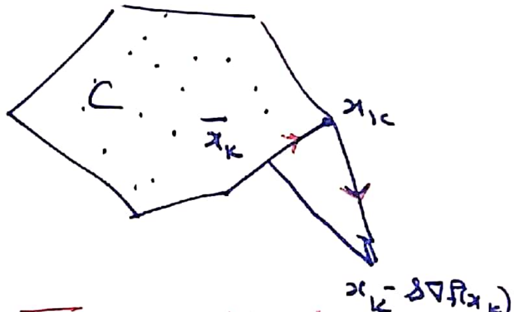
Perform an Armijo type line search by recursively testing the inequality

$$f(x_k) - f(x_k + \beta^m (\bar{x}_k - x_k)) \geq - \sigma \beta^m \langle \nabla f(x_k), \bar{x}_k - x_k \rangle$$

$$m = 0, 1, 2 \ldots$$

until it is satisfied at $m_k = m$. Set

$$x_{k+1} = x_k + \beta^{m_k} (\bar{x}_k - x_k)$$



The new direction is $x_k - \bar{x}_k$

In the next theorem we show that limit point of sequence generated by Gradient Projection Method satisfies the necessary optimality condition (2).

Theorem Let $C \subseteq \mathbb{R}^n$ be a closed convex set and f be a differentiable function defined on an open set containing C. Then ~~the~~ a limit point x^* of the sequence $\{x_k\}_0^\infty$ generated by gradient projection method with Armijo's step size rule satisfies the condition

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \forall x \in C.$$

Proof. From Armijo's rule we have

$$f(x_k) - f(x_{k+1}) = f(x_k) - f(x_k + \alpha_k d_k)$$
$$\geq -\sigma \alpha_k \langle \nabla f(x_k), d_k \rangle \qquad (3)$$

where $d_k = \bar{x}_k - x_k$ where $\bar{x}_k = \Pi_c(x_k - s\nabla f(x_k))$ and $\alpha_k = \beta^{m_k}$. Let $\nabla f(x_k) \neq 0$.

Claim d_k is a strict descent direction. Using projection inequality

$$\langle x_k - s\nabla f(x_k) - \Pi_c(x_k - s\nabla f(x_k)), x_k - \Pi_c(x_k - s\nabla f(x_k)) \rangle \leq 0$$

[by taking $x = x_k$]

~~Using~~ + ~~Substituting the value of d_k~~ As $d_k = \bar{x}_k - x_k$ we have

$$\langle -s\nabla f(x_k) - d_k, -d_k \rangle \leq 0$$

$$\Rightarrow \quad \|d_k\|^2 \leq -s\langle \nabla f(x_k), d_k \rangle \qquad (4)$$

As x_k is not a local minimizer $\langle \nabla f(x_k), x - x_k \rangle < 0$ for some $x \in C$. Using the lemma proved earlier

$$\Pi_c(x_k - s\nabla f(x_k)) \neq x_k.$$

Hence, $d_k \neq 0$ which implies $\|d_k\|^2 > 0$. Hence from (4) we have $\langle \nabla f(x_k), d_k \rangle < 0$.

Let $\{x_{k_\ell}\}$ be a subsequence of $\{x_k\}$ that converges to x^*. Since d_{k_ℓ} is a descent direction

$$\langle \nabla f(x_{k_\ell}), d_{k_\ell} \rangle < 0.$$

which implies

$$f(x_{k_\ell+1}) < f(x_{k_\ell}).$$

Also

$$f(x_{k_{\ell+1}}) \leq f(x_{k_\ell+1})$$

take care
$k_{\ell+1} \neq k_\ell + 1$

Hence

$$f(x_{k_{\ell+1}}) \leq f(x_{k_\ell+1}) \leq f(x_{k_\ell}).$$

As $f(x_{k_\ell}) \downarrow f(x^*)$ ~~implies~~ we have.

$$f(x_{k_\ell+1}) - f(x_{k_\ell}) \downarrow 0.$$

From (3) we have $\lim_{\ell \to \infty} \alpha_{k_\ell} \langle \nabla f(x_{k_\ell}), d_{k_\ell} \rangle = 0$ (5

As $x_{k_\ell} \to x^*$ it follows that $d_{k_\ell} \to d^* = \Pi_c(x^* - s\nabla f(x^*)) - x^*$.

Hence from (4) we have

$$0 \le \|d^*\|^2 \le -s\langle \nabla f(x^*), d^* \rangle \qquad (6)$$

<u>Claim</u> $\langle \nabla f(x^*), d^* \rangle = 0$.

If $\alpha_{k_l} \not\to 0$ then the claim follows from (5).

Otherwise by Armijo's unsuccessful step we have

$$f(x_{k_l}) - f(x_{k_l} + \frac{\alpha_{k_l}}{\beta} d_{k_l}) < -\sigma \frac{\alpha_{k_l}}{\beta} \langle \nabla f(x_{k_l}), d_{k_l} \rangle$$

Using mean value theorem there exists $z_{k_l} \in (x_{k_l}, x_{k_l} + \frac{\alpha_{k_l}}{\beta} d_{k_l})$

$$-\frac{\alpha_{k_l}}{\beta} \langle \nabla f(z_{k_l}), d_{k_l} \rangle < -\sigma \frac{\alpha_{k_l}}{\beta} \langle \nabla f(x_{k_l}), d_{k_l} \rangle$$

which implies

$$(1-\sigma)\langle \nabla f(x^*), d^* \rangle \ge 0. \qquad (7)$$

As $s > 0$ and $1-\sigma > 0$ the claim follows from (6) & (7). Hence from (6) we have $\|d^*\|^2 = 0 \Rightarrow d^* = 0$.

$$\Rightarrow \Pi_C(x^* - s\nabla f(x^*)) = x^*.$$

Hence by the previsiously proved lemma

$$\langle \nabla f(x^*), x - x^* \rangle \ge 0 \quad \forall\, x \in C.$$