Ch 10 Conjugate Gradient Methods

Problem setting $\min_{x \in \mathbb{R}^n} f(x)$, $f(x) = \frac{1}{2}\|x\|_Q^2 - (b, x)$

$Q > 0$ i.e. Q is SPD.

Global minimizer x^* satisfies $Qx^* = b$, $x^* = Q^{-1}b$.

Gradient Methods $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$, $\alpha_k > 0$ step size

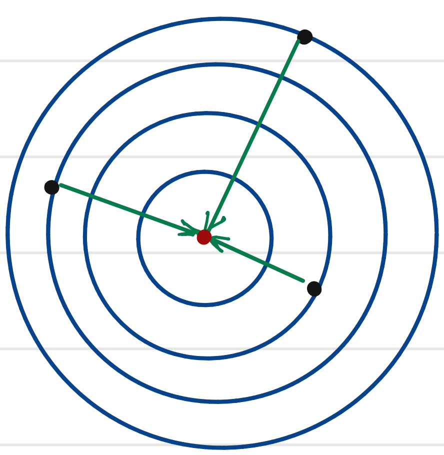
Steepest Gradient Descent $\alpha_k = \dfrac{\|g_k\|^2}{\|g_k\|_Q^2}$, $g_k = \nabla f(x_k)$.

$$\alpha_k = \underset{\alpha > 0}{\operatorname{argmin}}\, f(x_k - \alpha \nabla f(x_k)) = \underset{\alpha > 0}{\operatorname{argmin}}\, \|x_k - \alpha \nabla f(x_k) - x^*\|_Q^2$$

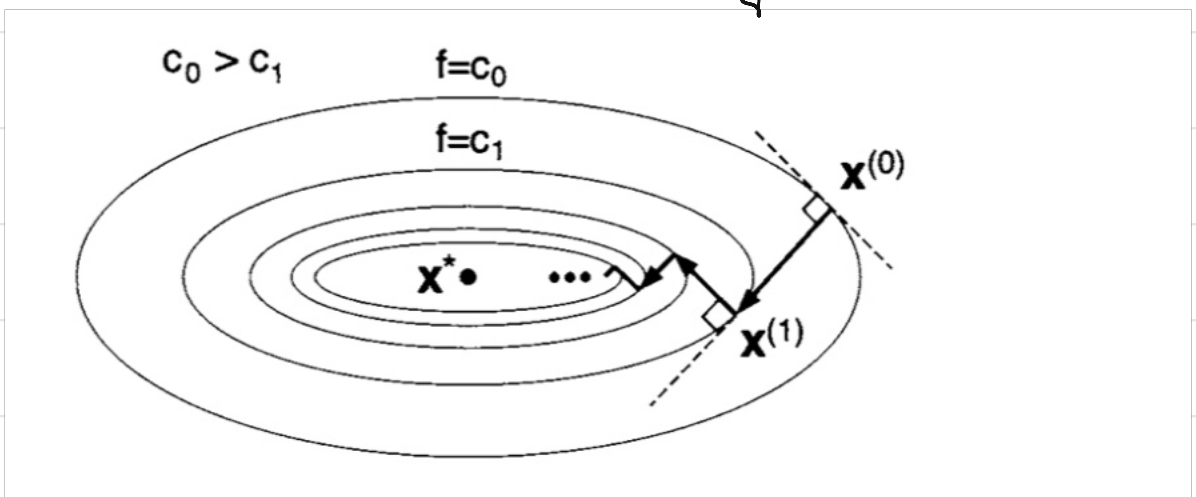
$$\boxed{\|x_{k+1} - x^*\|_Q \leqslant \frac{\kappa - 1}{\kappa + 1} \|x_k - x^*\|_Q ,}$$

where $\kappa = \dfrac{\lambda_{max}(Q)}{\lambda_{min}(Q)}$ is the condition number of Q.

Ideal case $\kappa = 1$.



When $\kappa \gg 1$, convergence is slow.



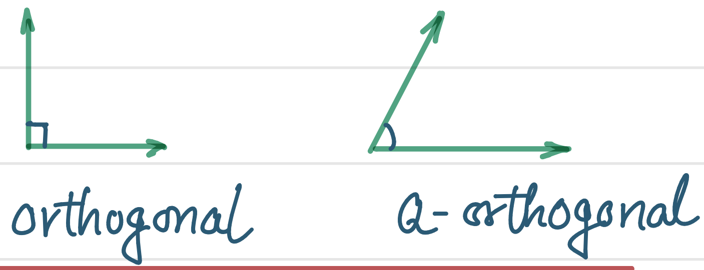
Steepest descent method in search for minimizer in a narrow valley.

Orthogonality

Classic l_2 inner product $(x, y) = y^T x = x^T y = (y, x)$

$x \perp y \iff (x, y) = 0$

$x \perp_Q y \iff (x, y)_Q = 0$

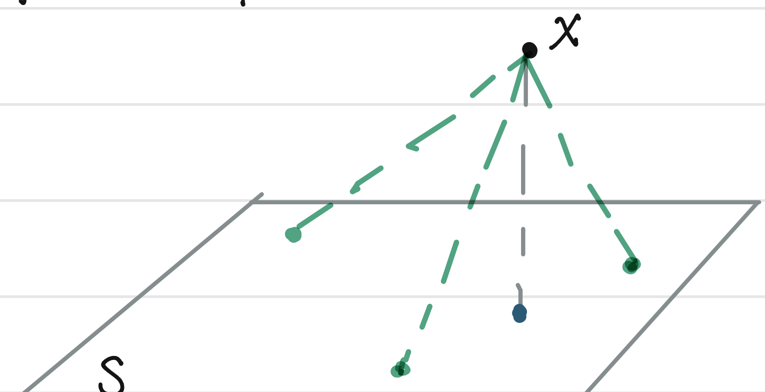


$$\boxed{(x, y)_Q := (Qx, y) = (x, Qy) = y^T Q x = x^T Q y.}$$

To speed up gradient methods, choose directions $\{d_i\}$ orthogonal in $(\cdot,\cdot)_Q$ inner product, or in short Q-orthogonal, or conjugate.
That is $(d_i, d_j)_Q = 0$, for $i \neq j$.

Projection, orthogonality, and shortest distance

A classical geometry problem: Given a plane S and a point $x \notin S$, find the point on S s.t. the distance to x is minimized.

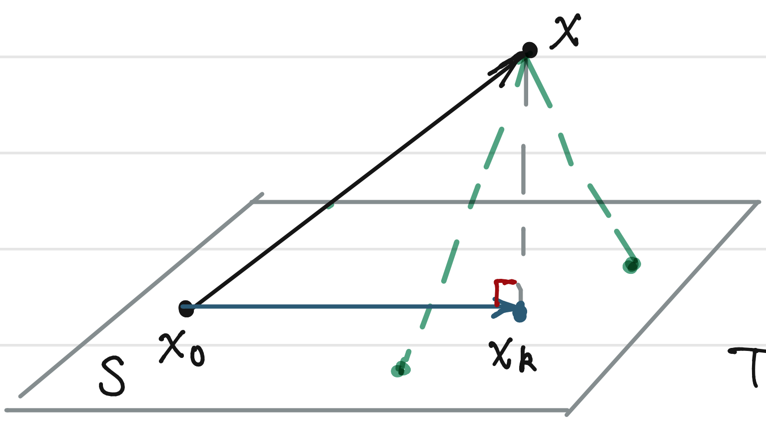


$$\min_{y \in S} f(y) := \frac{1}{2}\|x - y\|^2 \qquad (*)$$

$\nabla f(y) = y - x.$

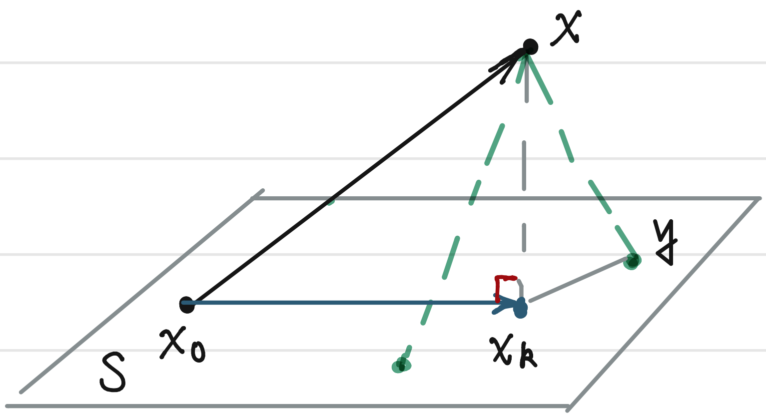
$\nabla f(y) = 0 \Rightarrow y = x.$ What's wrong?

This is a constrained minization problem.



Choose an arbitary point $x_0 \in S$

The solution to $(*)$ is given by the projection.



The vector $x_k - x_0 = \mathrm{Proj}_S(x - x_0)$

$$\boxed{\begin{array}{l}\text{Given a } u,\ \mathrm{Proj}_S u \in S \text{ satisfies} \\ (\mathrm{Proj}_S u, v) = (u, v) \quad \forall v \in S.\end{array}}$$

Lemma. Let $x_k - x_0 = \mathrm{Proj}_S(x - x_0)$. Then $x - x_k \perp S$, i.e.

$$\boxed{(x - x_k, v) = 0 \quad \forall v \in S} \qquad (1)$$

Consequently $\|x - x_k\| = \min_{y \in S} \|x - y\|$. $\qquad (2)$

Pf. By definition of Proj_S, $(x_k - x_0, v) = (x - x_0, v) \quad \forall v \in S.$

which is (1) by rearrangement.

$$\|x - x_k\|^2 = (x - x_k, x - x_k) = (x - x_k, x - y) + \cancel{(x - x_k, \overbrace{y - x_k}^{y - x_k \in S})}.$$

$$\leq \|x - x_k\| \, \|x - y\|$$

Cancel one $\|x - x_k\|$ to get $\|x - x_k\| \leq \|x - y\| \quad \forall y \in S.$

The equality holds when $y = x_k$. This completes the proof of (2). #.

Here we abuse notation S: treat it as a set of points, e.g. $x_k, y \in S$ and as a set of vectors, e.g. $v = y - x_k \in S$. The later one is indeed $T_{x_0}S$.

Question: What is the solution to $\min_{y \in S} \frac{1}{2} \|x - y\|_Q^2$?

Consider problem $\min\limits_{y \in S} \frac{1}{2}\|x-y\|_Q^2$.

Choose an arbitrary point $x_0 \in S$, let $x_k - x_0 = \mathrm{Proj}_S^Q(x-x_0)$, where for $u \in \mathbb{R}^n$, $\mathrm{Proj}_S^Q u \in S$ s.t. $\boxed{(\mathrm{Proj}_S^Q u, v)_Q = (u,v)_Q \quad \forall v \in S.}$

Then
$$\|x - x_k\|_Q = \min_{y \in S} \|x-y\|_Q .$$

Proof is almost identical. Leave as an exercise.

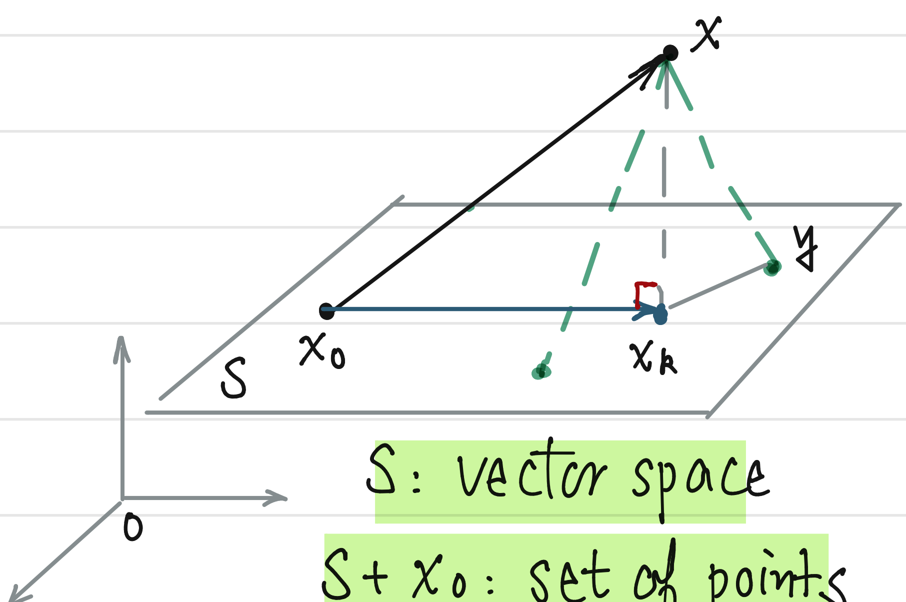
Quadratic Programming

$x = Q^{-1}b$ is the global min

$$\min_{y \in \mathbb{R}^n} f(y) = \frac{1}{2}\|y\|_Q^2 - (b,y) = \frac{1}{2}\|y - x\|_Q^2$$

Start from an arbitary x_0. Consider a k-dim subspace $S \subseteq \mathbb{R}^n$
Look for x_k s.t. $x_k - x_0 \in S$ and distance between x_k and x is minimized.

$$\min_{y - x_0 \in S} \|x-y\| \ ? \quad \text{or} \quad \min_{y - x_0 \in S} \|x-y\|_Q \ ?$$



S: vector space

S + x0: set of points

$\|\cdot\|_Q$ and $(\cdot,\cdot)_Q$ is the right choice

① Original problem
$$\min_y \|y-x\|_Q$$

② $\mathrm{Proj}_S^Q(x-x_0)$ is computable without knowing x.

$$(x-x_0, v)_Q = (Q(x-x_0), v) = (b - Qx_0, v) = -(\nabla f(x_0), v)$$

We don't know x but $Qx = b$ is known!

So $x_k - x_0 = \mathrm{Proj}_S^Q(x - x_0)$, $x_k = x_0 + \mathrm{Proj}_S^Q(x-x_0)$ is the solution.

Examples of subspaces.

① $\dim S = 1$. Now change notation $x_0 \to x_k$. Let $g_k = \nabla f(x_k)$.

$S = \mathrm{span}\{g_k\}$.

$$\min_{y - x_k \in S} \| x - y \|_Q$$

Optimal pt x_{k+1} is $x_{k+1} - x_k = \mathrm{Proj}_S^Q(x - x_k)$

$$(x_{k+1} - x_k, v)_Q = (x - x_k, v)_Q \quad \forall v \in S.$$

Choose $v = g_k$ basis of S. $\boxed{(x_{k+1} - x_k, g_k)_Q = (x - x_k, g_k)_Q}$ $(*)$

Now write $x_{k+1} - x_k = -\alpha_k g_k$ and notice $Q(x - x_k) = b - Qx_k = -g_k$

$(*)$ becomes $(\alpha_k g_k, g_k)_Q = (g_k, g_k)$ so $\alpha_k = \dfrac{\|g_k\|^2}{\|g_k\|_Q^2}$.

This is the steepest descent method and

$$\| x - x_{k+1} \|_Q = \min_{y - x_k \in S} \| x - y \|_Q = \min_{\alpha} \| x - (x_k - \alpha \nabla f(x_k)) \|_Q$$

② $\dim S = n$. Choose $x_0 \in \mathbb{R}^n$, what is $\displaystyle\min_{y - x_0 \in S} \| x - y \|_Q$?

As $x \in \mathbb{R}^n$, the answer is $y = x$! $S = \mathbb{R}^n$ non-constrained problem.

③ $S = \text{span}\{d_0, d_1, \cdots, d_k\}$. $(d_i, d_j)_Q = 0$, for $i \neq j$.

Choose arbitary x_0. $\min\limits_{y - x_0 \in S} \frac{1}{2} \|x - y\|_Q^2$

The best x_k can be found by $x_k - x_0 = \text{Proj}_S^Q (x - x_0)$

$$\boxed{(x_k - x_0, v)_Q = (x - x_0, v)_Q \quad \forall v \in S} \quad (*)$$

$x_k - x_0 \in S$ means $x_k - x_0 = \sum_{i=1}^{k} \alpha_i d_i$, with coefficients α_i to be determined.

Choose $v = d_j$ in $(*)$, we get

$$\left(\sum_{i=0}^{k} \alpha_i d_i, d_j\right)_Q = (x - x_0, d_j)_Q = (Q(x - x_0), d_j) = -(g_0, d_j)$$

$$\| \quad \alpha_j (d_j, d_j)_Q \quad \text{since } (d_i, d_j)_Q = 0 \text{ for } i \neq j.$$

So $\alpha_j = -\dfrac{(g_0, d_j)}{(d_j, d_j)_Q}$ for $j = 0, 1, \cdots, k$.

Recrusive formulae.

$$\boxed{x_{k+1} = x_k + \alpha_k d_k, \quad \alpha_k = -\frac{\|g_k\|^2}{\|g_k\|_Q^2}.}$$

When $k = n$. $S = \text{span}\{d_1, d_2, \cdots, d_n\}$. Then

$$\boxed{x - x_0 = \sum_{i=1}^{n} \alpha_i d_i, \quad \alpha_i = -\frac{(g_0, d_i)}{(d_i, d_i)_Q}, \quad \text{for } i = 1, \cdots, n.}$$

So we can find x by computing α_i using $g_0, d_i, i = 1, \cdots, n$.

x_0 is arbitary. But how to get Q-orthogonal basis $\{d_i\}$?

Question: How to find conjugate directions efficiently?

Assume we already have Q-orth directions $\{d_0, d_1, \cdots, d_k\}$

① add a new vector ② make it Q-orth

First two steps

To start with, chose $d_0 = -g(x_0)$. Now $S = \text{span}\{d_0\}$. Find x_1 by Q-projection: $x_1 - x_0 = \text{Proj}_S^Q(x - x_0)$.

Write $x_1 - x_0 = \alpha_0 d_0$. By definition $(x_1 - x_0, d_0)_Q = (x - x_0, d_0)_Q$

So $x_1 = x_0 + \alpha_0 d_0$, $\boxed{\alpha_0 = -\dfrac{(g_0, d_0)}{(d_0, d_0)_Q}.}$

$$\alpha_0 (d_0, d_0)_Q = (Qx - Qx_0, d_0) = -(g_0, d_0)$$

The new vector to be added is $-g_1 = -g(x_1) = b - Qx_1$.

Claim. $(g_1, d_0) = 0$.

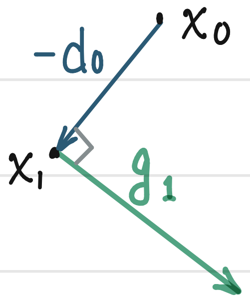
Pf. $\|x - x_1\|_Q^2 = \min_{y - x_0 \in S} \|x - y\|_Q^2 = \min_\alpha \phi(\alpha)$, $\phi(\alpha) = \|x - (x_0 + \alpha d_0)\|_Q^2$.

As $\alpha_0 = \text{argmin}\, \phi(\alpha)$, $\phi'(\alpha_0) = 0$. Compute $\phi'(\alpha) = 2(x - (x_0 + \alpha d_0), d_0)_Q$.

$\phi'(\alpha_0) = 0$ is equivalent to

$$(Q(x - (x_0 + \alpha_0 d_0)), d_0) = (b - Qx_1, d_0) = -(g_1, d_0) = 0. \qquad \#$$

In general, steepest descent along direction d_k, i.e. $x_{k+1} = x_k + \alpha_k d_k$ with $\alpha_k = -\dfrac{(g_k, d_k)}{(d_k, d_k)_Q}$ will imply $(g_{k+1}, d_k) = 0$.



The orthogonality is in $(\cdot,\cdot)$. But we need $(d_0, d_1)_Q = 0$.

Look for $d_1 \in \text{span}\{d_0, -g_1\}$. Write $d_1 = -g_1 + \beta d_0$

Use condition $(d_1, d_0)_Q = (-g_1 + \beta d_0, d_0)_Q = 0$ to get

$$\boxed{\beta = \frac{(g_1, d_0)_Q}{(d_0, d_0)_Q}.}$$

This is called Gram-Schmit algorithm

General steps.

Suppose we have Q-orthogonal vectors $\{d_0, d_1, \cdots, d_k\}$. Consider the subspace $S = \text{span}\{d_0, d_1, \cdots, d_k\}$ and compute x_{k+1} by Q-projection

$$x_{k+1} - x_0 = \text{Proj}_S^Q (x - x_0).$$

Write $x_{k+1} - x_0 = \sum_{i=0}^{k} \alpha_i d_i$. By definition, $(x_{k+1} - x_0, v)_Q = (x - x_0, v)_Q \quad \forall v \in Q \quad (*)$

Chose $v = d_i$ and use $(d_i, d_j)_Q = 0$ for $j \neq i$, we can compute

$$\left(\sum_{j=0}^{k} \alpha_j d_j, d_i\right)_Q = \alpha_i (d_i, d_i)_Q = (x - x_0, d_i)_Q = (Qx - Qx_0, d_i) = -(g_0, d_i)$$

So $\alpha_i = -\dfrac{(g_0, d_i)}{(d_i, d_i)_Q}$, for $i = 0, 1, \cdots, k$.

After we get x_{k+1}, compute $g_{k+1} = \nabla f(x_{k+1}) = Qx_{k+1} - b$.

Now we have $\{d_0, d_1, \cdots, d_k, -g_{k+1}\} \xRightarrow{} \{d_0, d_1, \cdots, d_k, d_{k+1}\}$

Gram-Schmit

Write $d_{k+1} = -g_{k+1} + \beta d_k + \gamma_{k-1} d_{k-1} + \cdots + \gamma_0 d_0$ and use $(d_{k+1}, d_i)_Q = 0$ to figure out coefficients for $i = 0, 1, \cdots, k$.

Magic fact: all $\gamma_i = 0$, i.e. $\boxed{d_{k+1} = -g_{k+1} + \beta d_k,}$ $\quad \beta = \dfrac{(g_{k+1}, d_k)_Q}{(d_k, d_k)_Q}$.

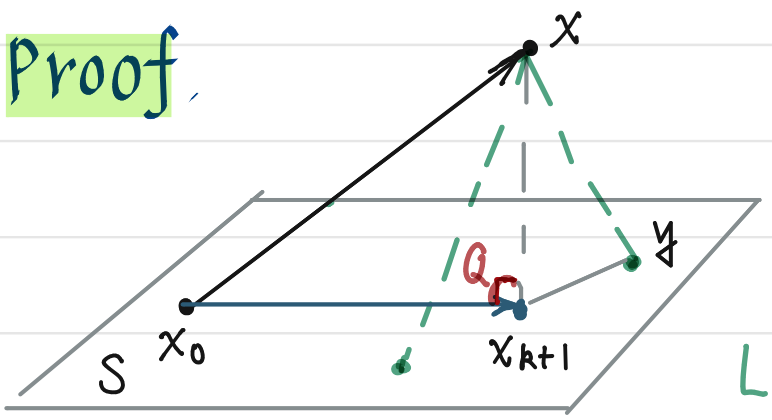
Explore the orthogonality of g_{k+1} from the Q-orthogonality $x - x_{k+1} \perp_Q S$.

Lemma. Let $x_{k+1} - x_0 = \mathrm{Proj}_S^Q (x - x_0)$ and $g_{k+1} = \nabla f(x_{k+1}) = Q x_{k+1} - b$.

Then ① $(g_{k+1}, d_i) = 0$ for $i = 0, 1, \cdots, k$.

② $(g_{k+1}, d_i)_Q = 0$ for $i = 0, 1, \cdots, k-1$.

Proof



① We have $x - x_{k+1} \perp_Q S$.

$(x - x_{k+1}, d_i)_Q = 0 \quad \forall\, i = 0, 1, \cdots, k$

LHS: $(Q(x - x_{k+1}), d_i) = -(g_{k+1}, d_i)$.

② $S = \mathrm{span}\{d_0, d_1, \cdots, d_k\} = \mathrm{span}\{g_0, g_1, \cdots, g_k\}$.

We can write $x_{i+1} = x_0 + \sum_{j=0}^{i} \alpha_j d_j = x_0 + \sum_{j=0}^{i-1} \alpha_j d_j + \alpha_i d_i = x_i + \alpha_i d_i$.

So $\alpha_i Q d_i = Q x_{i+1} - Q x_i = g_{i+1} - g_i \in S$ for $i = 0, 1, \cdots, k-1$.

As $Q d_i \in S$, from ①, $(g_{k+1}, d_i)_Q = (g_{k+1}, Q d_i) = 0$, for $i = 0, 1, \cdots, k-1$. #

Consequently, if $d_{k+1} = -g_{k+1} + \beta d_k + \sum_{i=0}^{k-1} \gamma_i d_i$, from $(d_{k+1}, d_i)_Q = 0$, we get $\gamma_i (d_i, d_i)_Q = (g_{k+1}, d_i)_Q = 0$ for $i = 0, 1, \cdots, k-1$.

Conjugate Gradient Methods

Problem setting.

$$\boxed{\min_{y\in\mathbb{R}^n} f(y) := \frac{1}{2}\|y\|_Q^2 - (b, y),} \qquad (1)$$

where Q is SPD, i.e. $Q = Q^T$, $Q > 0$.

Optimization problem (1) is equivalent to

$$\boxed{Qx = b} \qquad (2)$$

Key Steps ① Q-orth projection to $S = \text{span}\{d_0, d_1, \cdots, d_k\}$ $\quad \text{Proj}_S^Q$

$$x_{k+1} - x_0 = \text{Proj}_S^Q (x - x_0).$$

② Gram-Schmidt process to get d_{k+1}

$$d_{k+1} = -g_{k+1} + \beta d_k$$

③ Three-terms recrusive formula.

$$\begin{cases} x_{k+1} = x_k + \alpha_k d_k, \quad \alpha_k = \dfrac{(x - x_0, d_k)_Q}{(d_k, d_k)_Q} = -\dfrac{(g_0, d_k)}{(Qd_k, d_k)} \\ g_{k+1} = g_k + \alpha_k Q d_k \\ d_{k+1} = -g_{k+1} + \beta_k d_k, \quad \beta_k = \dfrac{(g_{k+1}, d_k)_Q}{(d_k, d_k)_Q} = \dfrac{(g_{k+1}, d_k)_Q}{(Qd_k, d_k)} \end{cases}$$

Remark. More efficient (less computational cost) formulae

$$\alpha_k = \frac{(g_k, g_k)}{(Qd_k, d_k)}, \quad \beta_k = \frac{(g_{k+1}, g_{k+1})}{(g_k, g_k)} \quad \text{in which } \|g_k\|^2 \text{ can be reused.}$$

Algorithm Due to the three-term formulae, there is no need to store all previous quantities.

```
function x = CG(Q,b,x,tol)

tol = tol*norm(b);
k = 1;
g = Q*x - b;
g0t = g0';
d = -g;
d2 = d'*(Q*d);
while sqrt(d2) >= tol && k<length(b)
        Qd = Q*d;
        d2 = d'*Qd;
        alpha = -g0t*d/d2;
        x = x + alpha*d;
        g = g + alpha*Qd;
        beta = g'*d/d2;
        d = -g + beta*d;
        k = k + 1;
end
```

Remark.

① The most time consuming part is matrix-vector product $Q*d$.

② The error measured by the relative error $\|d\|_Q < tol \|b\|$.

③ A maximum iteration step $n = \text{length}(b)$ is given to avoid infinite loops.