A Tutorial on Conjugate Gradient Descend

Richard Xu

February 13, 2021

1 Conjugate Gradient Descend Motivation

Imaging in coordinate descend, we have a 2-d function $f(x_1, x_2)$, where

$$\mathbf{x} = (x_1, x_2)^\top \tag{1}$$

suppose after optimizing along x_1-axis, led to $\mathbf{x}^{(1)}$ where $f(\mathbf{x}^{(1)})$ is minimized in its x_1 component:

$$\nabla_{x_1} f(\mathbf{x}^{(1)}) = 0 \tag{2}$$

next step is minimize along x_2-axis, and obtain $\mathbf{x}^{(2)}$ such that:

$$\nabla_{x_2} f(\mathbf{x}^{(2)}) = 0 \tag{3}$$

1.1 problem

the problem is that after optimizing in x_2 direction, it may "undo" effect of optimizing in x_1 direction previously

1.2 motivation

using previous example, one needs to move along a direction other than x_2-axis, s.t. $\nabla_{x_1} f(\mathbf{x}^{(2)})$ remains zero

in words, whilst minimizing a direction, function value along all previously optimized directions do not change, i.e., gradient at those previously-visited directions is zero! we need

to search for new non-axis directions:

1.3 where can it be used?

one common example is to minimize a **quadratic** problem:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \left(\mathbf{x}^\top \mathbf{Q} \mathbf{x} - \mathbf{b}^\top \mathbf{x} + c \right) \tag{4}$$

if matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is positive definite, then minimal value $\mathbf{x}^*$ is:

$$\mathbf{Q}\mathbf{x}^* = \mathbf{b} \tag{5}$$

1.3.1 other alternative to solve linear equation?

- general $\mathbf{Q}$: Gaussian elimination (LU factorization), but requires $\mathcal{O}(n^3)$
- $\mathbf{Q}$ is positive definite, but not Conjugate Gradient Descend
- **symmetric** positive definite (p.s.d) $\mathbf{Q}$, Cholesky decomposition

2 Q-conjugate

in general, $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ are said to be $\mathbf{Q}$-conjugate, such that:

$$\mathbf{q}_j^\top \mathbf{Q}\, \mathbf{q}_k = 0 \quad j \neq k \tag{6}$$

however, there is **no** requirement for $\mathbf{q}_k^\top \mathbf{Q}\, \mathbf{q}_k = 1$!

2.1 a special case

if $\mathbf{Q}$ is also symmetric, $\{\lambda_k, \mathbf{v}_k\}$ are eigen-(value,vector) pair:

$$\begin{aligned} \mathbf{Q}\mathbf{v}_k &= \lambda_k \mathbf{v}_k \\ \Longrightarrow\ \mathbf{v}_j^\top \mathbf{Q}\mathbf{v}_k &= \lambda_k \mathbf{v}_j^\top \mathbf{v}_k = 0 \quad j \neq k \\ \Longrightarrow\ \{\mathbf{q}_1, \ldots \mathbf{q}_n\} &= \{\mathbf{v}_1, \ldots \mathbf{v}_n\} \end{aligned} \tag{7}$$

$\{\mathbf{v}_1, \ldots \mathbf{v}_n\}$ can be thought as special case of $\mathbf{Q}$-conjugate vectors. These vectors are orthonormal without $\mathbf{Q}$

2.2 linear independence

let $\mathbf{Q}$ be **positive definite**, then all its $\mathbf{Q}$-conjugate vectors $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ are linearly independent

2.2.1 proof by contradiction

suppose an element $\mathbf{q}_k$ can be written in linear combination of $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\} \setminus \mathbf{q}_k$, i.e., linearly dependent:

$$\begin{aligned} \textbf{assume} \quad \mathbf{q}_k &= \alpha_1 \mathbf{q}_1 + \cdots + \alpha_{k-1}\mathbf{q}_{k-1} \\ \Longrightarrow\ \mathbf{q}_k^\top \mathbf{Q}\mathbf{q}_k &= \mathbf{q}_k^\top \mathbf{Q}\left(\alpha_1 \mathbf{q}_1 + \cdots + \alpha_{k-1}\mathbf{q}_{k-1}\right) \\ &= \mathbf{q}_k^\top \mathbf{Q}\, \alpha_1 \mathbf{q}_1 + \cdots + \mathbf{q}_k^\top \mathbf{Q}\, \alpha_{k-1}\mathbf{q}_{k-1} \\ &= 0 \end{aligned} \tag{8}$$

contradiction: by positive definiteness: $\mathbf{q}_k^\top \mathbf{Q}\mathbf{q}_k > 0 \quad \forall \mathbf{q}_k \neq 0$!
here we only prove linear independence, but they are in general **not** orthogonal!

2.3 compute α_k independently

let $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ be arbitary $\mathbf{Q}$-conjugate set, what is the corresonding $\{\alpha_1, \ldots, \alpha_n\}$?

write $\mathbf{x}^*$ as combination of linearly-independent basis:

$$\begin{aligned}
\mathbf{x}^* &= \alpha_1 \mathbf{q}_1 + \cdots + \alpha_n \mathbf{q}_n \\
\Longrightarrow \quad \mathbf{q}_k^\top \mathbf{Q} \mathbf{x}^* &= \mathbf{q}_k^\top \mathbf{Q} \left(\alpha_1 \mathbf{q}_1 + \cdots + \alpha_n \mathbf{q}_n\right) \qquad \times \text{ arbitrary } k^{\text{th}} \\
&= \alpha_k \mathbf{q}_k^\top \mathbf{Q} \mathbf{q}_k \\
\Longrightarrow \quad \alpha_k &= \frac{\mathbf{q}_k^\top \mathbf{Q} \mathbf{x}^*}{\mathbf{q}_k^\top \mathbf{Q} \mathbf{q}_k} = \frac{\mathbf{q}_k^\top \mathbf{b}}{\mathbf{q}_k^\top \mathbf{Q} \mathbf{q}_k}
\end{aligned} \tag{9}$$

change index from $\{1, \ldots, n\} \to \{0, \ldots, n-1\}$:

$$\begin{aligned}
\mathbf{x}^* &= \alpha_0 \mathbf{q}_0 + \cdots + \alpha_{n-1} \mathbf{q}_{n-1} \\
&= \sum_{k=0}^{n-1} \frac{\mathbf{q}_k^\top \mathbf{b}}{\mathbf{q}_k^\top \mathbf{Q}\, \mathbf{q}_k} \mathbf{q}_k \qquad \text{where } \alpha_k = \frac{\mathbf{q}_k^\top \mathbf{b}}{\mathbf{q}_k^\top \mathbf{Q}\, \mathbf{q}_k}
\end{aligned} \tag{10}$$

the above can be achieved in parallel where each $\mathbf{q}_k$ does **not** minimizing anything. It is **not** an algorithm, it simply decomposes $\mathbf{x}^*$

3 Putting into an algorithm

now, we want to find an iterative **algorithm**, with an initial point $\mathbf{x}_0$:

$$\begin{aligned}
\text{compute } (\alpha_0, \mathbf{q}_0) \qquad \mathbf{x}_1 &= \mathbf{x}_0 + \alpha_0 \mathbf{q}_0 \\
&\ldots \\
\text{compute } (\alpha_k, \mathbf{q}_k) \qquad \mathbf{x}_k &= \mathbf{x}_0 + \alpha_0 \mathbf{q}_0 + \cdots + \alpha_{k-1} \mathbf{q}_{k-1} \\
&\ldots \\
\text{compute } (\alpha_{n-1}, \mathbf{q}_{n-1}) \qquad \mathbf{x}^* &= \mathbf{x}_0 + \alpha_0 \mathbf{q}_0 + \cdots + \alpha_{n-1} \mathbf{q}_{n-1}
\end{aligned} \tag{11}$$

we need to two requirement for α_k and $\mathbf{q}_k$:

3.1 Two requirements

let gradient $\nabla f(\mathbf{x}_k)$ be oppsite direction of gradient descend:

$$\nabla f(\mathbf{x}_k) = \mathbf{Q} \mathbf{x}_k - \mathbf{b} \tag{12}$$

3.1.1 Requirement α_k

Given $\mathbf{q}_k$, α_k must be found such that, after computing:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{q}_k \tag{13}$$

At position $\mathbf{x}_{k+1}$, $f(\mathbf{x}_{k+1})$ mininizes in all directions of previous path vector $(\mathbf{x}_1 - \mathbf{x}_0, \ldots, \mathbf{x}_{k+1} - \mathbf{x}_k)$, i.e., gradient vector $\nabla f(\mathbf{x}_{k+1}) \in \mathbb{R}^n$ is $\perp$ to all previous path vectors:

$$\begin{aligned} \nabla f(\mathbf{x}_{k+1}) &\perp \text{sub-space span by path vectors } (\underbrace{\mathbf{x}_1 - \mathbf{x}_0}_{\alpha_0 \mathbf{q}_0}, \ldots, \underbrace{\mathbf{x}_{k+1} - \mathbf{x}_k}_{\alpha_k \mathbf{q}_k}) \\ &\perp \text{span}(\mathbf{q}_1, \ldots \mathbf{q}_k) \end{aligned} \tag{14}$$

3.1.2 Requirement $\mathbf{q}_{k+1}$

for next iteration, also find appropriate $\mathbf{q}_{k+1}$ such at, it satisfy all $\mathbf{Q}$-conjugate definition:

$$\mathbf{q}_{k+1}{}^{\top} \mathbf{Q}\, \mathbf{q}_i \quad \forall i \in 1, \ldots, k \tag{15}$$

4 Requirement α_k: $\nabla f(\mathbf{x}_{k+1}) \perp$ span$(\mathbf{q}_1, \ldots \mathbf{q}_k)$?

4.1 Why is it needed?

4.1.1 First iteration

starting from first step, given arbitrary point $\mathbf{x}_0$, and after picking a "sensible" $\mathbf{q}_0$", for example, $-\nabla f(\mathbf{x}_0)$:

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{x}_0 + \alpha_0 \mathbf{q}_0 \\ \implies \mathbf{x}_1 - \mathbf{x}_0 &= \alpha_0 \mathbf{q}_0 \end{aligned} \tag{16}$$

Obviously, we hope to find α_0 that makes location at $\mathbf{x}_1$ to minimize the **line** (direction) $\mathbf{x}_0 + \alpha_0 \mathbf{q}_0$, this is equivalently saying:

$$\begin{aligned} \nabla f(\mathbf{x}_1) &\perp (\mathbf{x}_0 + \alpha_0 \mathbf{q}_0) \\ &\perp \alpha_0 \mathbf{q}_0 \qquad \text{offset } \mathbf{x}_0 \text{ won't matter in } \perp \\ &\perp \text{span}(\mathbf{q}_0) \qquad \text{just a line} \end{aligned} \tag{17}$$

can be understood/visualized by moving $\mathbf{x}$ along the line:

$$\mathbf{x} = \mathbf{x}_0 + \alpha \mathbf{q}_0 \quad \text{for arbitary } \alpha \tag{18}$$

if gradient vector at $\mathbf{x}$, i.e., $\nabla_{\mathbf{x}} f(\mathbf{x})$ is **not** $\perp$ to the line, then there is some gradient component in the same direction of the line, until it gets to $\mathbf{x}_1$, where $\nabla_{\mathbf{x}} f(\mathbf{x}_1)$ has zero component in the line

4.1.2 Second iteration

$$\begin{aligned} \mathbf{x}_2 &= \mathbf{x}_1 + \alpha_1 \mathbf{q}_1 \\ &= \mathbf{x}_0 + \alpha_0 \mathbf{q}_0 + \alpha_1 \mathbf{q}_1 \end{aligned} \tag{19}$$

similarly, we want to find α_1, such that: $\mathbf{x}_2$ minimizes:

$$\nabla f(\mathbf{x}_2) \perp (\mathbf{x}_1 + \alpha_1 \mathbf{q}_1) \tag{20}$$

however, by doing so, $\mathbf{x}_2$ should also minimizes the plane span by:

$$\begin{aligned} \nabla f(\mathbf{x}_2) &\perp (\mathbf{x}_0 + \alpha_0 \mathbf{q}_0 + \alpha_1 \mathbf{q}_1) \\ &\perp (\alpha_0 \mathbf{q}_0 + \alpha_1 \mathbf{q}_1) \\ &\perp \text{span}(\mathbf{q}_0, \mathbf{q}_1) \end{aligned} \tag{21}$$

this is needed as one needs to ensure optimizing $\mathbf{x}_2$ should not "undo" the efforts by both $\mathbf{x}_1 - \mathbf{x}_0$ and $\mathbf{x}_2 - \mathbf{x}_1$.

4.1.3 k^{th} iteration

by finding appropriate α_k, we first can prove:

$$\nabla f(\mathbf{x}_{k+1}) \perp (\mathbf{x}_k + \alpha_k \mathbf{q}_k) \tag{22}$$

subsquently we can use induction to prove:

$$\nabla f(\mathbf{x}_{k+1}) \perp \text{span}(\underbrace{\mathbf{q}_0, \ldots, \mathbf{q}_k}_{k+1 \text{ terms}}) \tag{23}$$

i.e., $\mathbf{x}_{k+1}$ minimizes f over $\{\mathbf{x}_0 + \text{span}(\mathbf{q}_0, \ldots, \mathbf{q}_k)\}$. This is detailed in section(4.2.2)

4.1.4 What is α_k then?

$$\alpha_k = -\frac{\nabla f(\mathbf{x}_k)^\top \mathbf{q}_k}{\mathbf{q}_k^\top \mathbf{Q} \mathbf{q}_k} \tag{24}$$

we show why this choice of α_k leads to:

$$\begin{aligned} \nabla f(\mathbf{x}_{k+1}) \perp \mathbf{q}_k \quad &\text{or} \\ \mathbf{q}_k^\top \nabla f(\mathbf{x}_{k+1}) = 0 \end{aligned} \tag{25}$$

in section(4.2)

4.1.5 Last step

$$\mathbf{x}_n = \underset{\mathbf{x} \in \{\mathbf{x}_0 + \text{span}(\mathbf{q}_0, \ldots, \mathbf{q}_{n-1})\}}{\arg\min} \frac{1}{2} \mathbf{x}^\top \mathbf{Q} \mathbf{x} - b^\top \mathbf{x} \tag{26}$$

4.2 prove $\nabla f(\mathbf{x}_{k+1}) \perp \mathbf{span}(\mathbf{q}_0, \dots, \mathbf{q}_k)$

4.2.1 prove $\nabla f(\mathbf{x}_{k+1}) \perp \mathbf{q}_k$

First, we prove:

$$\alpha_k = -\frac{\nabla f(\mathbf{x}_k)^\top \mathbf{q}_k}{\mathbf{q}_k^\top \mathbf{Q} \mathbf{q}_k} \implies \nabla f(\mathbf{x}_{k+1}) \perp \mathbf{q}_k \tag{27}$$

write $\nabla f(\mathbf{x}_{k+1})$ in terms of $\nabla f(\mathbf{x}_k)$

by definition:

$$\begin{aligned}
\nabla f(\mathbf{x}_{k+1}) &= \mathbf{Q}\mathbf{x}_{k+1} - b \\
&= \mathbf{Q}(\mathbf{x}_k + \alpha_k \mathbf{q}_k) - b \\
&= (\mathbf{Q}\mathbf{x}_k - b) + \alpha_k \mathbf{Q}\mathbf{q}_k \\
&= \nabla f(\mathbf{x}_k) + \alpha_k \mathbf{Q}\mathbf{q}_k
\end{aligned} \tag{28}$$

substituting $\alpha_k = -\frac{\nabla f(\mathbf{x}_k)^\top \mathbf{q}_k}{\mathbf{q}_k^\top \mathbf{Q} \mathbf{q}_k}$ into $\nabla f(\mathbf{x}_{k+1}) = \nabla f(\mathbf{x}_k) + \alpha_k \mathbf{Q}\mathbf{q}_k$:

$$\begin{aligned}
\nabla f(\mathbf{x}_{k+1}) &= \nabla f(\mathbf{x}_k) + \alpha_k \mathbf{Q}\mathbf{q}_k \\
\implies \mathbf{q}_k^\top f(\mathbf{x}_{k+1}) &= \mathbf{q}_k^\top \left(\nabla f(\mathbf{x}_k) + \alpha_k \mathbf{Q}\mathbf{q}_k\right) \\
&= \mathbf{q}_k^\top \nabla f(\mathbf{x}_k) - \frac{\nabla f(\mathbf{x}_k)^\top \mathbf{q}_k}{\mathbf{q}_k^\top \mathbf{Q} \mathbf{q}_k} \mathbf{q}_k^\top \mathbf{Q}\mathbf{q}_k \\
&= \mathbf{q}_k^\top \nabla f(\mathbf{x}_k) - \mathbf{q}_k^\top \nabla f(\mathbf{x}_k) \\
&= 0 \\
\implies \nabla f(\mathbf{x}_{k+1}) &\perp \mathbf{q}_k
\end{aligned} \tag{29}$$

4.2.2 second we prove $\nabla f(\mathbf{x}_{k+1}) \perp \mathbf{span}(\mathbf{q}_0, \dots, \mathbf{q}_k)$ by induction

by induction, assume in the **previous** step:

$$\nabla f(\mathbf{x}_k) \perp \text{span}(\mathbf{q}_0, \dots, \mathbf{q}_{k-1}) \tag{30}$$

let $i < k$:

$$\begin{aligned}
\nabla f(\mathbf{x}_{k+1}) &= \nabla f(\mathbf{x}_k) + \alpha_k \mathbf{Q}\mathbf{q}_k \quad \text{using Eq.(28)} \\
\implies \mathbf{q}_i^\top \nabla f(\mathbf{x}_{k+1}) &= \mathbf{q}_i^\top \left(\nabla f(\mathbf{x}_k) + \alpha_k \mathbf{Q}\mathbf{q}_k\right) \\
&= \mathbf{q}_i^\top \nabla f(\mathbf{x}_k) + \alpha_k \underbrace{\mathbf{q}_i^\top \mathbf{Q}\mathbf{q}_k}_{=0} \quad \text{no need to subsitute } \alpha_k \\
&= \mathbf{q}_i^\top \nabla f(\mathbf{x}_k) \\
&= 0 \quad \text{by induction assumption } \nabla f(\mathbf{x}_k) \perp \text{span}(\mathbf{q}_0, \dots, \mathbf{q}_{k-1}) \\
\implies \nabla f(\mathbf{x}_{k+1}) &\perp \mathbf{q}_i \quad \text{for } i < k
\end{aligned} \tag{31}$$

5 Requirement $\mathbf{q}_{k+1}$: $\mathbf{q}_{k+1}{}^\top \mathbf{Q}\ \mathbf{q}_i \quad \forall i \in 1, \ldots, k$

one more thing missing, we know it works well for any arbitrary $\mathbf{Q}$-conjugate vectors $\{\mathbf{q}_0, \ldots, \mathbf{q}_n\}$.

looking at the first iteration: after letting $\mathbf{q}_0 = -\nabla f(\mathbf{x}_0)$:

$$\mathbf{q}_1 = -\nabla f(\mathbf{x}_1) + \beta_0 \mathbf{q}_0 \tag{32}$$

use definition of $\mathbf{Q}$-conjugacy:

$$\begin{aligned}
\mathbf{q}_1^\top \mathbf{Q} \mathbf{q}_0 &= 0 \\
\implies (-\nabla f(\mathbf{x}_1) + \beta_0 \mathbf{q}_0)^\top \mathbf{q}_0 &= 0 \\
-\nabla f(\mathbf{x}_1)^\top \mathbf{Q} \mathbf{q}_0 + \beta_0 \mathbf{q}_0^\top \mathbf{Q} \mathbf{q}_0 &= 0 \\
\beta_0 &= \frac{\nabla f(\mathbf{x}_1)^\top \mathbf{Q} \mathbf{q}_0}{\mathbf{q}_0{}^\top \mathbf{Q} \mathbf{q}_0}
\end{aligned} \tag{33}$$

it is easy to see that for k^{th} iterations:

$$\beta_k = \frac{\nabla f(\mathbf{x}_{k+1})^\top \mathbf{Q}\ \mathbf{q}_k}{\mathbf{q}_k{}^\top \mathbf{Q}\ \mathbf{q}_k} \tag{34}$$

6 Conjugate Gradient Algorithm

1. **initialize** $k = 0$, given $\mathbf{x}^0$:

$$\begin{aligned}
\mathbf{x}_0 &= \mathbf{x}^0 \\
\mathbf{q}_0 &= -\nabla_{\mathbf{x}} f(\mathbf{x}_0) = -\mathbf{Q}\mathbf{x}_0 + \mathbf{b}
\end{aligned} \tag{35}$$

2. **repeat** for k:

   (a) compute α_k:

$$\alpha_k = -\frac{\nabla f(\mathbf{x}_k)^\top \mathbf{q}_k}{\mathbf{q}_k^\top \mathbf{Q} \mathbf{q}_k} \tag{36}$$

   (b) update movement (main)

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha \mathbf{q}_k \tag{37}$$

   (c) update direction $\mathbf{q}_{k+1}$:

$$\mathbf{q}_{k+1} = -\nabla f(\mathbf{x}_{k+1}) + \beta_k \mathbf{q}_k \qquad \text{where} \quad \beta_k = \frac{\nabla f(\mathbf{x}_{k+1})^\top \mathbf{Q}\ \mathbf{q}_k}{\mathbf{q}_k{}^\top \mathbf{Q}\ \mathbf{q}_k} \tag{38}$$

   until at all n directions, or other criteria