Conjugate Gradients

The Newton method we have encountered so far heavily relied on the need to solve a (full-rank) linear system. As we have seen, the direct solution of very large systems becomes impractical due to the rapidly increasing complexity of matrix inversion (or Cholesky factorization). However, numerical optimization algorithms can come to our aid, as the solution of a general (over-) complete $m \times n$ system, $m \geq n$

$$\mathbf{A}\mathbf{x} = \mathbf{y}$$

can be cast as the minimization of

$$\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} - 2\mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{y}^{\mathrm{T}}\mathbf{y}.$$

We leave as an exercise to show that if $\mathbf{A}$ is full rank, then $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is positive definite.

Exercise 1. *Let $\mathbf{A}$ be an $m \times n$ matrix with $m \geq n$ and $\mathrm{rank}(\mathbf{A}) = n$. Show that $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is an $n \times n$ symmetric positive-definite matrix.*

We will therefore dedicate our attention to minimizing a strictly convex quadratic function of the form

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x},$$

with and $n \times n$ $\mathbf{Q} \succ 0$. In terms of the previous problem, $\mathbf{Q} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\mathbf{b} = -2\mathbf{A}^{\mathrm{T}}\mathbf{y}$. In what follows, we will develop a power technique called *conjugate gradients* with complexity comparable to that of gradient descent (and, hence, scalable to large n's), but much faster convergence ratio.

1 Inner products

Before we start, we will need some preliminary notions in linear algebra. Recall that until now we have defined the inner (or "scalar") product in $\mathbb{R}^n$ as $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^{\mathrm{T}}\mathbf{y}$. This is the standard definition of inner product that induces the standard Euclidean (ℓ_2) norm. However, the notion of an inner product is more general.

Definition. *A function $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ is called an* inner product *on $\mathbb{R}^n$ if for every $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ and every scalars $\alpha, \beta \in \mathbb{R}$ it satisfies*

1. Commutativity*: $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ (a more general definition involving complex numbers requires a complex conjugate on the right-hand side);*

2. Distributivity *over vector addition (or* additivity*):* $\langle \mathbf{x}, \mathbf{y}+\mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$*;*

3. Homogeneity*:* $\langle \alpha\mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$*;*

4. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ *with* $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ *iff* $\mathbf{x} = 0$.

Sometimes, the additivity and the homogeneity axioms are combined into a single *bilinearity* property $\langle \mathbf{x}, \alpha\mathbf{y}+\mathbf{z} \rangle = \alpha\langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$ (again, with the complex conjugate of α in the complex case).

This axiomatic definition encompasses the standard Euclidean case; as before, a (general) inner product induces a (general) norm $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

Exercise 2. *Prove that* $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ *is a norm (recall the definition of a general norm).*

The inner product and the norm it induces obey the Cauchy-Schwarz inequality, $|\langle \mathbf{x}, \mathbf{x} \rangle| \leq \|\mathbf{x}\| \, \|\mathbf{y}\|$. Moreover, we can still think of the inner product as a measure of (the cosine of) the angle between two unit vectors, and say that $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* (in the sense of a certain inner product) iff $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ ($\mathbf{x}$ and $\mathbf{y}$ might not be orthogonal in the Euclidean sense in case of a general inner product).

Several commonly used inner products are listed below:

1. *Standard Euclidean inner product on* $\mathbb{R}^n$: $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^{\mathrm{T}}\mathbf{y}$ inducing the ℓ_2 norm $\|\mathbf{x}\|_2^2 = \mathbf{x}^{\mathrm{T}}\mathbf{x}$.

2. *Standard inner product on the space of functions on an interval* T:

$$\langle f, g \rangle = \int_T f(t) g^*(t) dt$$

(note the conjugate) inducing the so-called L_2 norm

$$\|f\|^2 = \int_T |f|^2(t) dt.$$

This can be thought of as a continuous version of the Euclidean inner product, and is ubiquitous in harmonic analysis.

3. The $\mathbf{Q}$-inner product $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{Q}} = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{y}$ for $\mathbf{Q} \succ 0$. This inner product induces the $\mathbf{Q}$-norm $\|\mathbf{x}\|_{\mathbf{Q}}^2 = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}$. In statistics, when $\mathbf{Q}$ is interpreted as the inverse covariance matrix, the metric induced by the $\mathbf{Q}$-norm is called the *Mahalanobis* distance, and can be thought of a normalized Euclidean distance taking the variances (and the covariances) of the vector elements into account.

Exercise 3. *Prove these are indeed inner products.*

For the rest of our discussion, the notion of the $\mathbf{Q}$-inner product and the $\mathbf{Q}$-norm it induces will be important.

2 Orthogonalization

Equipping the linear space with an inner product allows to define orthogonal bases. We will say that a collection $\{\mathbf{x}_i \in \mathbb{R}^n\}_{i=1}^m$ of linearly independent vectors is orthogonal if $\langle \mathbf{x}_i, \mathbf{x}_j \rangle \neq 0$ iff $i \neq j$. However, recall that orthogonality in the sense of one inner product does not imply orthogonality in the sense of another. It is therefore important, given a collection of vectors spanning a linear subspace of $\mathbb{R}^m$, to construct a new collection of orthogonal vectors spanning the same subspace. The latter collection is generally called an *orthogonal basis* with respect to a certain inner product (the basis is said *orthonormal* if the vectors are unit with respect to the norm induced by the chosen inner product).

A standard procedure for creating orthogonal bases is called the *Gram-Schmidt orthogonalization*. As the input, we are given a collection $\mathbf{x}_1, \dots, \mathbf{x}_m$ linearly independent vectors; as the output, we shall produce a new collection $\mathbf{y}_1, \dots, \mathbf{y}_m$ of orthogonal vectors (in the sense of some inner product that is assumed to be given). The procedure gradually constructs the vectors $\mathbf{y}_k$ from the input vectors $\mathbf{x}_k$ and the previously constructed $\mathbf{y}_k$'s. We start with assigning $\mathbf{y}_1 = \mathbf{x}_1$. Since $\mathbf{x}_2$ will generally not be orthogonal to $\mathbf{y}_1$, we cannot simply assign it to $\mathbf{y}_2$. Instead, we have to find a vector $\mathbf{y}_2$ in the linear space spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $\mathbf{y}_2$ is orthogonal to $\mathbf{y}_1$. This can be achieved by subtracting from $\mathbf{x}_2$ its projection onto $\mathbf{y}_1$.

Formally, we define the projection operator

$$\mathcal{P}_{\mathbf{u}}\mathbf{v} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle}\mathbf{u} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|^2}\mathbf{u}$$

projecting the vector $\mathbf{v}$ onto $\mathbf{u}$. An important property of the projection is that

$$\langle \mathbf{v} - \mathcal{P}_{\mathbf{u}}\mathbf{v}, \mathbf{u} \rangle = \left\langle \mathbf{v} - \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|^2}\mathbf{u}, \mathbf{u} \right\rangle = \langle \mathbf{v}, \mathbf{u} \rangle - \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|^2}\langle \mathbf{u}, \mathbf{u} \rangle = 0.$$

Due to this property,

$$\mathbf{y}_2 = \mathbf{x}_2 - \mathcal{P}_{\mathbf{y}_1}\mathbf{x_2} = \mathbf{x}_2 - \frac{\langle \mathbf{y}_1, \mathbf{x}_2 \rangle}{\|\mathbf{y}_1\|^2}\mathbf{y}_1$$

is orthogonal to $\mathbf{y}_1$.

The same procedure is repeated for subsequent vectors. In order to obtain the k-th output vector $\mathbf{y}_k$, we subtract from $\mathbf{x}_k$ its projection onto the subspace spanned by the previously constructed vectors $\mathbf{y}_1, \dots, \mathbf{y}_{k-1}$. The projection onto a linear subspace spanned by orthogonal vectors is simply given as the sum of the projections on each of the vectors individually,

$$\mathcal{P}_{\mathbf{u}_1 \perp \cdots \perp \mathbf{u}_k}\mathbf{v} = \sum_{i=1}^{k} \mathcal{P}_{\mathbf{u}_i}\mathbf{v}$$

(this is not true for non-orthogonal $\mathbf{v}_i$'s). Using this property, we can write

$$\mathbf{y}_k = \mathbf{x}_k - \mathcal{P}_{\mathbf{y}_1, \dots, \mathbf{y}_{k-1}}\mathbf{x_k} = \mathbf{x}_k - \sum_{i=1}^{k-1} \frac{\langle \mathbf{y}_i, \mathbf{x}_k \rangle}{\|\mathbf{y}_i\|^2}\mathbf{y}_i.$$

```
input  : set of linearly independent vectors $\{\mathbf{x}_i \in \mathbb{R}^n\}_{i=1}^m$
output: orthogonal vectors $\{\mathbf{y}_i \in \mathbb{R}^n\}_{i=1}^m$
Start with $\mathbf{y}_1 = \mathbf{x}_1$.
for $k = 2, \dots, m$ do
    $$\mathbf{y}_k = \mathbf{x}_k - \sum_{i=1}^{k-1} \frac{\langle \mathbf{y}_i, \mathbf{x}_k \rangle}{\|\mathbf{y}_i\|^2} \mathbf{y}_i.$$
end
```

Algorithm 1: Gram-Schmidt orthogonalization

3 Conjugate directions

Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be linearly independent vectors in $\mathbb{R}^n$. Substituting the particular selection of the $\mathbf{Q}$-inner product and the induced $\mathbf{Q}$-norm into the Gram-Schmidt procedure results in

$$\mathbf{d}_k = \mathbf{x}_k - \sum_{i=1}^{k-1} \frac{\mathbf{d}_i^{\mathrm{T}} \mathbf{Q} \mathbf{x}_k}{\mathbf{d}_i^{\mathrm{T}} \mathbf{Q} \mathbf{d}_i} \mathbf{d}_i,$$

where instead of $\mathbf{y}_i$ we used the notation $\mathbf{d}_i$. By construction, the $\mathbf{d}_i$'s are orthogonal. Such a family of vectors is called $\mathbf{Q}$-orthogonal or $\mathbf{Q}$-*conjugate directions*.

How is this related to minimization of quadratic functions? Recall that we are interested in minimizing

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^{\mathrm{T}} \mathbf{Q} \mathbf{x} + \mathbf{b}^{\mathrm{T}} \mathbf{x}.$$

Since the conjugate directions $\mathbf{d}_i$ form a basis for $\mathbb{R}^n$, we can reparametrize the problem in terms of a new system of coordinates, $\mathbf{x} = \mathbf{D}\boldsymbol{\xi} = \xi_1 \mathbf{d}_1 + \cdots + \xi_n \mathbf{d}_n$:

$$\begin{aligned}
\bar{f}(\boldsymbol{\xi}) &= \frac{1}{2} \boldsymbol{\xi}^{\mathrm{T}} \mathbf{D}^{\mathrm{T}} \mathbf{Q} \mathbf{D} \boldsymbol{\xi} + \mathbf{b}^{\mathrm{T}} \mathbf{D} \boldsymbol{\xi} \\
&= \frac{1}{2} \sum_{i,j=1}^{n} \xi_i \xi_j \mathbf{d}_i^{\mathrm{T}} \mathbf{Q} \mathbf{d}_j + \sum_{i=1}^{n} \xi_i \mathbf{b}^{\mathrm{T}} \mathbf{d}_i \\
&= \frac{1}{2} \sum_{i,j=1}^{n} \xi_i \xi_j \langle \mathbf{d}_i, \mathbf{d}_j \rangle_{\mathbf{Q}} + \sum_{i=1}^{n} \xi_i \mathbf{b}^{\mathrm{T}} \mathbf{d}_i \\
&= \sum_{i=1}^{n} \frac{\xi_i^2}{2} \|\mathbf{d}_i\|_{\mathbf{Q}}^2 + \xi_i \mathbf{b}^{\mathrm{T}} \mathbf{d}_i.
\end{aligned}$$

Observe that in the new system of coordinates, the function decomposes into the sum of n quadratic functions of the form

$$\varphi_i(\xi_i) = \frac{\xi_i^2}{2} \|\mathbf{d}_i\|_{\mathbf{Q}}^2 + \xi_i \mathbf{b}^{\mathrm{T}} \mathbf{d}_i.$$

In other words, $\bar{f}(\boldsymbol{\xi})$ is coordinate-wise separable. Hence,

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \min_{\boldsymbol{\xi}\in\mathbb{R}^n} \bar{f}(\boldsymbol{\xi}) = \min_{\boldsymbol{\xi}\in\mathbb{R}^n} \sum_{i=1}^{n} \varphi_i(\xi_i) = \sum_{i=1}^{n} \min_{\xi_i\in\mathbb{R}} \varphi_i(\xi_i).$$

Exercise 4. *Prove the above identity.*

The conclusion is far-reaching: if we are given a collection of n $\mathbf{Q}$-conjugate directions, the minimization of $f(\mathbf{x})$ splits into n one-dimensional minimizations of each of the φ_i's along the corresponding direction $\mathbf{d}_i$. This can be carried out by invoking line search n times. The resulting family of optimization algorithm are usually referred to as *conjugate direction methods.*

Let us start with some $\mathbf{x}_0$ and allow to optimize over the first k directions, $\mathbf{D}_k = (\mathbf{d}_1, \dots, \mathbf{d}_k)$, i.e., let $\boldsymbol{\xi}_k = (\xi_1, \dots, \xi_k)^{\mathrm{T}}$ be the optimization variables. This is akin to minimizing $f(\mathbf{x})$ over the affine subspace

$$G_k = \mathbf{x}_0 + \mathbf{D}_k\mathbb{R}^k = \{\mathbf{x}_0 + \mathbf{D}_k\boldsymbol{\xi}_k : \boldsymbol{\xi}_k \in \mathbb{R}^k\}.$$

Property (Expanding manifold)**.** *The sequence of the first k line searches produces*

$$\min_{\xi_1} \varphi_1(\xi_1) + \cdots + \min_{\xi_k} \varphi_1(\xi_k) = \min_{\mathbf{x}\in G_k} f(\mathbf{x}).$$

Exercise 5. *Give a formal proof to the expanding manifold property.*

Since $G_n = \mathbb{R}^n$, conjugate direction methods converge in n iterations.

4 Conjugate gradients

The conjugate direction family of optimization algorithms depends on the particular choice of the vectors $\mathbf{x}_k$ from which the $\mathbf{Q}$-conjugate directions $\mathbf{d}_k$ are built via the Gram-Schmidt procedure described above. A particular implementation of this idea, called *conjugate gradients*, uses the collection of gradients of the objective at the points visited on the optimization trajectory. The procedure is sequential: a new direction is created at every iteration, based on the previous directions and the current gradient.

For the quadratic objective we are considering, the gradient at point $\mathbf{x}_k$ is given by

$$\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k + \mathbf{b}.$$

Starting at $\mathbf{x}_0$, we set as in standard gradient descent

$$\mathbf{d}_0 = -\mathbf{g}_0 = -\mathbf{Q}\mathbf{x}_0 - \mathbf{b}.$$

The iterate at the k-th iteration is computed as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k,$$

where α_k is determined via line search. Note: conjugate directions methods (conjugate gradients in particular) require nearly exact minimization along each of the conjugate directions, and inexact linesearch may lead to divergence of the method.

The $k+1$-st direction is computed via Gram-Schmidt orthogonalization, using $-\mathbf{g}_k$ as the current vector:

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \sum_{i=0}^{k} \frac{\mathbf{d}_i^{\mathrm{T}} \mathbf{Q} \mathbf{g}_{k+1}}{\mathbf{d}_i^{\mathrm{T}} \mathbf{Q} \mathbf{d}_i} \mathbf{d}_i.$$

The latter expression can be greatly simplified. First, observe that from $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$, we have

$$\mathbf{d}_i = \frac{1}{\alpha_i}(\mathbf{x}_{i+1} - \mathbf{x}_i).$$

Hence,

$$\mathbf{Q}\mathbf{d}_i = \frac{1}{\alpha_i}\mathbf{Q}(\mathbf{x}_{i+1} - \mathbf{x}_i) = \frac{1}{\alpha_i}(\mathbf{g}_{i+1} - \mathbf{g}_i).$$

Second, due to the expanding manifold property, at iteration k, the function has been minimized over G_k, which has been spanned by $\mathbf{d}_0, \dots, \mathbf{g}_0$ or, equivalently, by $\mathbf{g}_0, \dots, \mathbf{g}_k$. Therefore, the projection of $\mathbf{g}_{k+1}$ on G_k is zero, or

$$\mathbf{g}_{k+1} \perp \mathbf{g}_k, \dots, \mathbf{g}_0,$$

with orthogonality interpreted in the standard Euclidean sense. Combining these two results, we have

$$\mathbf{d}_i^{\mathrm{T}} \mathbf{Q} \mathbf{g}_{k+1} = \frac{1}{\alpha_i} \mathbf{g}_{k+1}^{\mathrm{T}} (\mathbf{g}_{i+1} - \mathbf{g}_i),$$

vanishing for $i < k$. As the result, we can write

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k,$$

with

$$\beta_k = \frac{\mathbf{g}_{k+1}^{\mathrm{T}} (\mathbf{g}_{k+1} - \mathbf{g}_k)}{\mathbf{d}_k^{\mathrm{T}} (\mathbf{g}_{k+1} - \mathbf{g}_k)}.$$

The parameter β_k can be simplified even further. Observe that since $\mathbf{g}_{k+1} \perp G_k$, in particular, $\mathbf{g}_{k+1} \perp \mathbf{d}^k$ that belongs to G_k. Hence, the denominator can be simplified to

$$-\mathbf{d}_k^{\mathrm{T}} \mathbf{g}_k = -(-\mathbf{g}_k + \beta_{k-1} \mathbf{d}_{k-1})^{\mathrm{T}} \mathbf{g}_k.$$

However, by similar argument $\mathbf{g}_k \perp \mathbf{d}^{k-1}$, leading to the *Polack-Ribiere* method:

$$\beta_k = \frac{\mathbf{g}_{k+1}^{\mathrm{T}} (\mathbf{g}_{k+1} - \mathbf{g}_k)}{\|\mathbf{g}_k\|^2}.$$

Note that that the numerator can be further simplified by observing that $\mathbf{g}_{k+1} \perp \mathbf{g}_k$, leading to the *Fletcher-Reevs* variant:

$$\beta_k = \frac{\|\mathbf{g}_{k+1}\|^2}{\|\mathbf{g}_k\|^2}.$$

While for quadratic functions the two expressions are equivalent and the Fletcher-Reevs variant is preferable due to less computations, the methods differ for general functions, where the Polack-Ribiere formula is known to behave better.

```
input  : function $f$; initial point $\mathbf{x}_0$
output: (approximate) local minimizer $\mathbf{x}^*$ of $f$
Start with an initial guess $\mathbf{x}_0$, $\mathbf{g}_0 = \nabla f(\mathbf{x}_0)$, and set $\mathbf{d}_0 = -\mathbf{g}_0$.
for $k = 1, 2, \ldots,$ until convergence do
    Find descent direction $\mathbf{d}_k$
    Find step size $\alpha_k$
    Update $\mathbf{x}_k \leftarrow \mathbf{x}_{k-1} + \alpha_k \mathbf{d}_k$
end
Return $\mathbf{x}^* = \mathbf{x}^k$
```

Algorithm 2: Conjugate directions method for general functions

5 Convergence rate

6 Preconditioning