5374 Fall 2022
Numerical Linear Algebra

Lecture 26

Gradient descent

... solving $Az = b$ (A is $n \times n$) via various algorithms:

but requires an upper triangular matrix

- Back-substitution takes $O(n^2)$ FLOPs

longer algorithm but handles all invertible matrices

- Gauss-Jordan takes $O(n^3)$ FLOPs
- QR decomposition $\longrightarrow$ $O(n^3)$ FLOPs
- (Today) k iterations of $\underline{\text{gradient descent}}$ takes $O(kn^2)$ FLOPs

requires a positive definite matrix

If A is positive definite and sparse (say each row has only $O(1)$ non-zero entries) then gradient descent takes $O(kn)$ FLOPs

Gradient descent

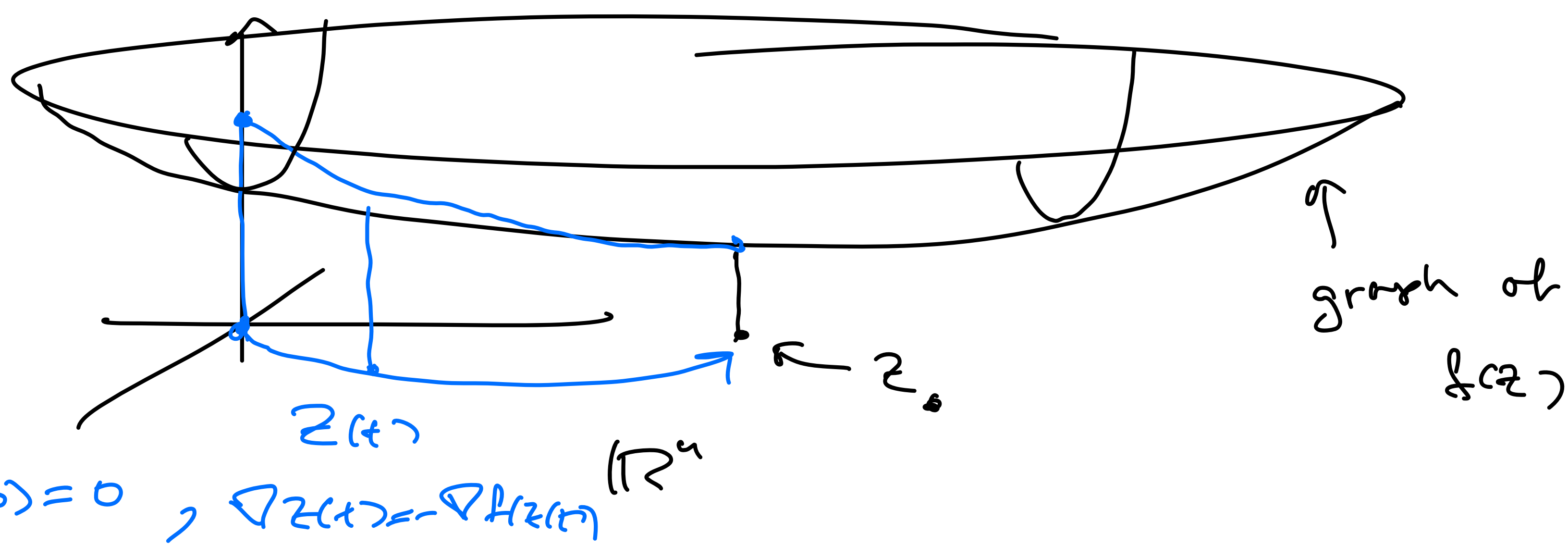
definite

If the matrix A is **positive**, the function

$$f(\boldsymbol{z}) = \frac{1}{2}(\mathrm{A}\boldsymbol{z}, \boldsymbol{z}) - (\boldsymbol{b}, \boldsymbol{z})$$

is a (strictly) convex quadratic polynomial of n variables.

The unique global minimum of $f(z)$ is the solution z_* to $Az_* = b$ ($z_* = A^{-1}b$)



Gradient descent

Some computations involving $f(\boldsymbol{z})$

Given $\boldsymbol{z}, \boldsymbol{v}$, and α, we considered the function

$$f(\boldsymbol{z}+\alpha\boldsymbol{v}) = \frac{1}{2}(\mathrm{A}(\boldsymbol{z}+\alpha\boldsymbol{v}), \boldsymbol{z}+\alpha\boldsymbol{v}) - (\boldsymbol{b}, \boldsymbol{z}+\alpha\boldsymbol{v})$$

$\frac{1}{2}(Az,z) - (b,z) = f(z)$

Let us expand the right hand side. Note that

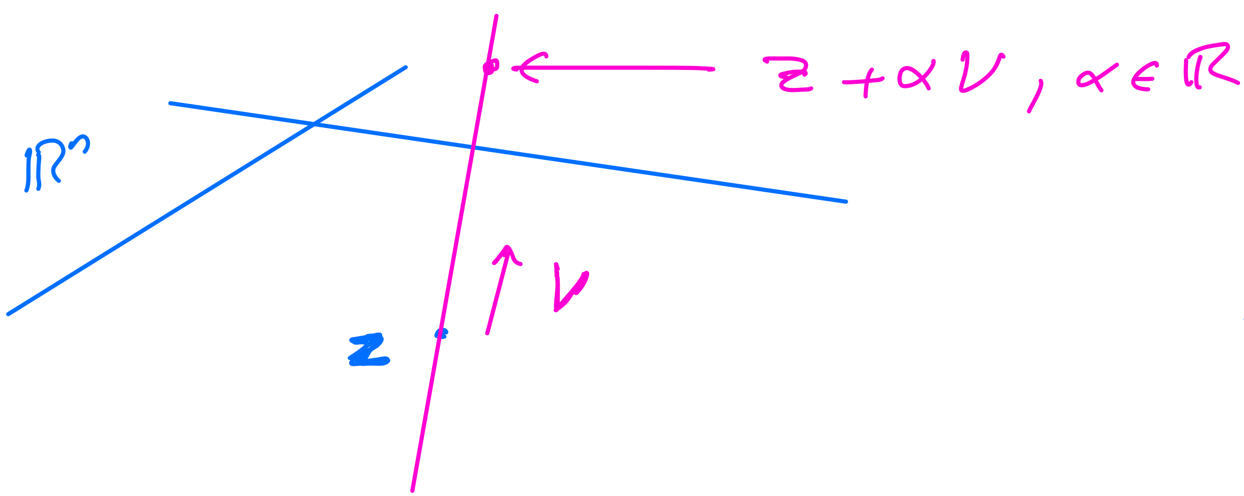
$$\begin{aligned} \frac{1}{2}(\mathrm{A}(\boldsymbol{z}+\alpha\boldsymbol{v}), \boldsymbol{z}+\alpha\boldsymbol{v}) &= \frac{1}{2}(\mathrm{A}(\boldsymbol{z}, \boldsymbol{z}) + \alpha(\mathrm{A}\boldsymbol{z}, \boldsymbol{v}) + \frac{1}{2}\alpha^2(\mathrm{A}\boldsymbol{v}, \boldsymbol{v})\big) \\ (\boldsymbol{b}, \boldsymbol{z}+\alpha\boldsymbol{v}) &= (\boldsymbol{b}, \boldsymbol{z}) + \alpha(\boldsymbol{b}, \boldsymbol{v}) \end{aligned}$$

$\frac{1}{2}\alpha^2(Av,v)$

$+\alpha(Az,v) - \alpha(b,v)$

$= \frac{1}{2}(Av,v)\alpha^2$

$+ (Az-b,v)\alpha$



Gradient descent

Some computations involving $f(\boldsymbol{z})$

Putting this together, we see that

$$f(\boldsymbol{z}+\alpha\boldsymbol{v}) = f(\boldsymbol{z}) + \alpha(\mathrm{A}\boldsymbol{z}-\boldsymbol{b},\boldsymbol{v}) + \tfrac{1}{2}\alpha^2(\mathrm{A}\boldsymbol{v},\boldsymbol{v})$$

Moreover,

$$\lim_{\alpha\to 0} \frac{f(z+\alpha v)-f(z)}{\alpha} = (Az-b, v), \ \forall\, v \Rightarrow Df(z) = Az-b$$

$$\lim_{\alpha\to 0} \frac{f(z+\alpha v)+f(z-\alpha v)-2f(z)}{\alpha^2} = (Av, v) \ \forall\, v$$

$$\Longrightarrow D^2 f(z) = A.$$

Gradient descent

Some computations involving $f(\boldsymbol{z})$

Remark: From the identity

$$f(\boldsymbol{z}+\alpha\boldsymbol{v}) = f(\boldsymbol{z}) + \alpha(\mathrm{A}\boldsymbol{z}-\boldsymbol{b},\boldsymbol{v}) + \tfrac{1}{2}\alpha^2(\mathrm{A}\boldsymbol{v},\boldsymbol{v})$$

$\frac{\partial^2 f}{\partial z_i \partial z_j}$

it follows immediately that

$$\nabla f(\boldsymbol{z}) = \mathrm{A}\boldsymbol{z}-\boldsymbol{b}, \quad D^2 f(\boldsymbol{z}) = \mathrm{A}$$

Gradient descent

Some computations involving $f(\boldsymbol{z})$

Remark:

(The following elementary observation will be key later)

If we write $\boldsymbol{x} = \boldsymbol{z} + \alpha \boldsymbol{v}$, then

$$f(\boldsymbol{x}) = f(\boldsymbol{z}) + (\mathrm{A}\boldsymbol{z} - \boldsymbol{b}, \boldsymbol{x} - \boldsymbol{z}) + \frac{1}{2}(\mathrm{A}(\boldsymbol{x} - \boldsymbol{z}), \boldsymbol{x} - \boldsymbol{z})$$

In particular, if $z = z_*$, the solution to $Az = b$,

$$f(x) = f(z_*) + \frac{1}{2}(A(x-z), x-z)$$

(since A is positive definite, this shows $f(x) \geq f(z_*)$ for all x, and $=$ if and only if $x = z_*$)

Note: If $\lambda_0 =$ smallest eigenvalue of A, then

$$f(x) - f(z_*) \geq \frac{\lambda_0}{2} \|x - z_*\|^2$$

Gradient descent

Line searches

How do we find the minimum of $f(\boldsymbol{z})$?

How do we find minimizers?

$(\equiv Az = b)$

Clearly, solving $\nabla f(\boldsymbol{z}) = 0$ is self-defeating in our context.

Let us settle for minimizing along **any given straight line**.

So we take $f(z + \alpha v)$ as before, and minimize in α, since this is a convex (1-d) parabola, the minimization is done by a basic algebraic operation.

Gradient descent

Line searches

(pick $v \in \mathbb{R}^n$, $v \neq 0$)

If current guess is $\boldsymbol{z}$, search for the minimizer on some line

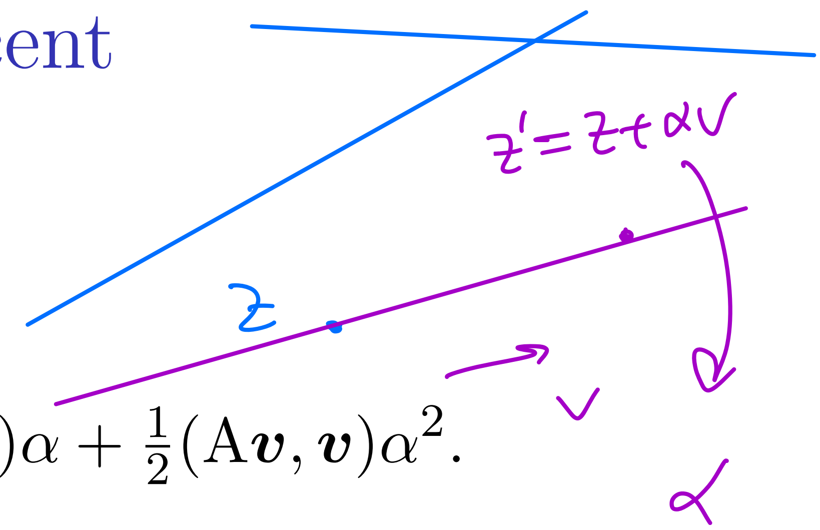
$$\boldsymbol{z} + \alpha \boldsymbol{v}, \;\; \alpha \in \mathbb{R}.$$

That is we minimize the function of one variable

$$f_{\boldsymbol{z},\boldsymbol{v}}(\alpha) := f(\boldsymbol{z} + \alpha \boldsymbol{v}).$$

Gradient descent

Line searches



Recall that

$$f_{\boldsymbol{z},\boldsymbol{v}}(\alpha) = f_{\boldsymbol{z},\boldsymbol{v}}(0) + (\mathrm{A}\boldsymbol{z} - \boldsymbol{b}, \boldsymbol{v})\alpha + \frac{1}{2}(\mathrm{A}\boldsymbol{v}, \boldsymbol{v})\alpha^2.$$

In particular,

$$\frac{d}{d\alpha} f_{\boldsymbol{z},\boldsymbol{v}}(\alpha) = (\mathrm{A}\boldsymbol{z} - \boldsymbol{b}, \boldsymbol{v}) + \alpha(\mathrm{A}\boldsymbol{v}, \boldsymbol{v}).$$

$\Rightarrow$ The minimum is achieved at α given by

$$\alpha = -\frac{(Az-b, v)}{(Av, v)}$$

(recall $(Av, v) \neq 0$ since $v \neq 0$ and A is positive definite)

Gradient descent

Line searches

Therefore the α that gives the minimum for $f_{\boldsymbol{z},\boldsymbol{v}}$ is

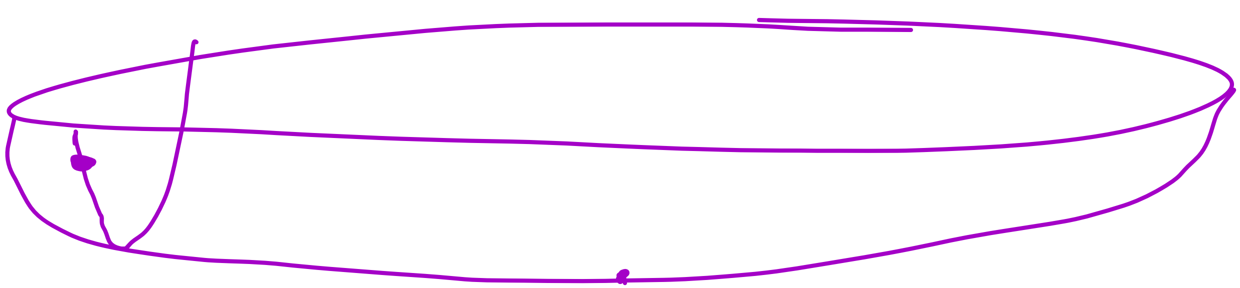
$$\alpha = -\frac{(\mathrm{A}\boldsymbol{z} - \boldsymbol{b}, \boldsymbol{v})}{(\mathrm{A}\boldsymbol{v}, \boldsymbol{v})}$$

This takes us from an initial guess $\boldsymbol{z}$ to

$$\boldsymbol{z}' = \boldsymbol{z} + \alpha\boldsymbol{v}, \ \alpha = -\frac{(\mathrm{A}\boldsymbol{z} - \boldsymbol{b}, \boldsymbol{v})}{(\mathrm{A}\boldsymbol{v}, \boldsymbol{v})}$$

This is called a line search, since we have searched (and found) the minimum value of the function along a given line. In this context, $\boldsymbol{v}$ is called “the search direction”.

Now, we could repeat the process, starting from $\boldsymbol{z}'$, and taking a **different** direction $\boldsymbol{v}'$ to look for yet a better candidate.



Gradient descent

Line searches

…so we arrive at the **Gradient Descent Method**.

Basically, Gradient Desccent consists in subsequent line searches, where at each stage we pick as search direction **the opposite to the gradient** of f at our current guess, that is

$$\boldsymbol{v} = -\nabla f(\boldsymbol{z}) = -(\mathrm{A}\boldsymbol{z} - \boldsymbol{b}) = \boldsymbol{b} - \mathrm{A}\boldsymbol{z}$$

Gradient descent

Line searches

In this case

$$\boldsymbol{v} = \boldsymbol{b} - \mathrm{A}\boldsymbol{z}, \quad \alpha = \frac{(\boldsymbol{v}, \boldsymbol{v})}{(\mathrm{A}\boldsymbol{v}, \boldsymbol{v})}.$$

$$\left(\alpha = -\frac{(Az-b, v)}{(Av, v)} \right.$$

$$= -\frac{(Az-b, b-Az)}{(Av, v)}$$

$$= \frac{(b-Az, b-Az)}{(Av, v)}$$

$$= \frac{(v, v)}{(Av, v)} \Big)$$

Gradient descent

The **Gradient Descent** algorithm

Initialization:

$$\boldsymbol{z}_0 \text{ (initial guess)} \quad \boldsymbol{v}_0 = \boldsymbol{b} - \mathrm{A}\boldsymbol{z}_0$$

At stage k:

$$\boldsymbol{v}_k = \boldsymbol{b} - \mathrm{A}\boldsymbol{z}_{k-1}$$

$$\alpha_k = \frac{(\boldsymbol{v}_k, \boldsymbol{v}_k)}{(\mathrm{A}\boldsymbol{v}_k, \boldsymbol{v}_k)}$$

$$\boldsymbol{z}_k = \boldsymbol{z}_{k-1} + \alpha_k \boldsymbol{v}_k$$

Stop the process at a pre-determined number of steps k_0.

Gradient descent

The **Gradient Descent** algorithm (fixed number of steps $= k_0$)

$$
\begin{aligned}
&\boldsymbol{z} = \boldsymbol{z}_0 \\
&\boldsymbol{v} = \boldsymbol{v}_0 \\
&\texttt{for } \ k = 1, \ldots, k_0\texttt{:} \\
&\qquad \boldsymbol{v} \leftarrow \boldsymbol{b} - \mathrm{A}\boldsymbol{z} \\
&\qquad \alpha \leftarrow (\boldsymbol{v}, \boldsymbol{v})/(\mathrm{A}\boldsymbol{v}, \boldsymbol{v}) \\
&\qquad \boldsymbol{z} \leftarrow \boldsymbol{z} + \alpha \boldsymbol{v} \\
&\texttt{return } \ \boldsymbol{z}
\end{aligned}
$$

Gradient descent

The **Gradient Descent** algorithm (fixed number of steps $= k_0$)

$$
\begin{aligned}
&\boldsymbol{z} = \boldsymbol{z}_0 \\
&\boldsymbol{v} = \boldsymbol{v}_0 \\
&\texttt{for}\ \ k = 1, \ldots, k_0\,\texttt{:} \\
&\qquad \boldsymbol{v} \leftarrow \boldsymbol{b} - \mathrm{A}\boldsymbol{z} \\
&\qquad \alpha \leftarrow (\boldsymbol{v}, \boldsymbol{v})/(\mathrm{A}\boldsymbol{v}, \boldsymbol{v}) \\
&\qquad \boldsymbol{z} \leftarrow \boldsymbol{z} + \alpha \boldsymbol{v} \\
&\texttt{return}\ \ \boldsymbol{z}
\end{aligned}
$$

Q: What is the number of FLOPs performed by this algorithm?

($O(k_0 n^2)$
$O(k_0 n)$ if A is sparse)

Gradient descent

The **Gradient Descent** algorithm (fixed number of steps $= k_0$)

$$
\begin{aligned}
&\boldsymbol{z} = \boldsymbol{z}_0\\
&\boldsymbol{v} = \boldsymbol{v}_0\\
&\texttt{for}\ \ k = 1, \ldots, k_0\texttt{:}\\
&\qquad \boldsymbol{v} \leftarrow \boldsymbol{b} - \mathrm{A}\boldsymbol{z}\\
&\qquad \alpha \leftarrow (\boldsymbol{v}, \boldsymbol{v})/(\mathrm{A}\boldsymbol{v}, \boldsymbol{v})\\
&\qquad \boldsymbol{z} \leftarrow \boldsymbol{z} + \alpha\boldsymbol{v}\\
&\texttt{return}\ \ \boldsymbol{z}
\end{aligned}
$$

Q: What is the # of FLOPs if A is sparse? $\left(\ O(k_0 n) \ \right)$

Gradient descent

Lemma

Let A be positive and $\boldsymbol{z}_*$ the solution to $A\boldsymbol{z}_* = \boldsymbol{b}$, and $\{\boldsymbol{z}_k\}_k$ a sequence generated by Gradient Descent, then

$$f(\boldsymbol{z}_k) - f(\boldsymbol{z}_*) \leq \left(1 - \frac{1}{\mathrm{cond(A)}}\right)(f(\boldsymbol{z}_{k-1}) - f(\boldsymbol{z}_*))$$

In particular

$$f(\boldsymbol{z}_k) - f(\boldsymbol{z}_*) \leq \left(1 - \frac{1}{\mathrm{cond(A)}}\right)^k (f(\boldsymbol{z}_0) - f(\boldsymbol{z}_*))$$