STEEPEST DESCENT

David F. Gleich

August 21, 2023

Learning objectives
1. Appreciate how linear systems are closely related to minimizing quadratic functions
2. Witness a computation of the gradient for a multivariate function in matrix algebra
3. See a characterization of a quadratic minimizer as the solution of a linear system
4. Generalize the algorithm to the steepest descent algorithm for solving a linear system

1 LINEAR SYSTEMS AND QUADRATIC FUNCTION MINIMIZATION

We are studying quadratic function minimization because this turns out to a good way to understand how to solve $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ for symmetric positive definite matrices $\boldsymbol{A}$. A full understanding of this will involve some analysis of convex functions. This is all *straightforward* for this case (if not simple), but it is an instance of a far more general theory. Some of the notes will make references to more general results that could be proved but are not relevant for the linear system case.

1.1 MOTIVATION FROM THE SCALAR CASE

Recall that a scalar quadratic function can be written:

$$f(x) = ax^2 + bx + c.$$

These look like bowls or lines (when $a = 0$).

Consider the problem

$$\underset{x}{\text{minimize}}\; ax^2 + bx + c$$

The solution is undefined is $a < 0$ (or just ∞). Otherwise, $x = -b/(2a)$ is the point that achieves the minimum. This can be found by looking for a point where the derivative is 0:

$$f'(x) = 2ax + b = 0 \Rightarrow x = -b/(2a).$$

A multivariate quadratic *looks* very similar.

1.2 THE MULTIVARIATE QUADRATIC FOR $\boldsymbol{A}\mathbf{x} = \mathbf{b}$

For $\boldsymbol{A}\mathbf{x} = \mathbf{b}$, it turns out that for any positive definite matrix $\boldsymbol{A}$, that we can view it as the solution of an optimization problem

$$\underset{\mathbf{x}}{\text{minimize}} \quad \tfrac{1}{2}\mathbf{x}^T\boldsymbol{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}.$$

This is because if $\boldsymbol{A}$ is positive semi-definite, then this problem is convex with a unique global minimizer. A convex function is just one that always lies below any line connecting two points. Formally, this is $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \le \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$. A global minimizer is any point $\mathbf{x}^*$ where $f(\mathbf{x}^*) \le f(\mathbf{x})$ for any other point $\mathbf{x}$. Note that if $f(\mathbf{x})$ is convex and if we have two global minimizers, then any point on the line connecting them must be a minimizer by the property of convexity.

There is a stronger result to prove here too.

THEOREM 1 *Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\boldsymbol{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}$. Then $f(\mathbf{x})$ is convex if $\boldsymbol{A}$ is symmetric positive definite.*

Proof From the definition

$$\begin{aligned} f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= (\alpha\mathbf{x} + (1-\alpha)\mathbf{y})^T\boldsymbol{A}(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) - (\alpha\mathbf{x} + (1-\alpha)\mathbf{y})^T\mathbf{b} \\ &= \alpha(\alpha\mathbf{x}^T\boldsymbol{A}\mathbf{x} - \mathbf{x}^T\mathbf{b} + (1-\alpha)((1-\alpha)\mathbf{y}^T\boldsymbol{A}\mathbf{y} - \mathbf{y}^T\mathbf{y}) + 2\alpha(1-\alpha)\mathbf{x}^T\boldsymbol{A}\mathbf{y} \\ &= \alpha^2\mathbf{x}^T\boldsymbol{A}\mathbf{x} + (1-\alpha)^2\mathbf{y}^T\boldsymbol{A}\mathbf{y} + 2\alpha(1-\alpha)\mathbf{y}^T\boldsymbol{A}\mathbf{x} - \alpha\mathbf{x}^T\mathbf{b} - (1-\alpha)\mathbf{y}^T\mathbf{b} \end{aligned}$$

Our goal is to show that this is $\le \alpha\mathbf{x}^T\boldsymbol{A}\mathbf{x} + (1-\alpha)\mathbf{y}^T\boldsymbol{A}\mathbf{y} - \alpha\mathbf{x}^T\mathbf{b} - (1-\alpha)\mathbf{y}^T\mathbf{b}$, and so the idea is to show that

$$\alpha^2\mathbf{x}^T\boldsymbol{A}\mathbf{x} + (1-\alpha)^2\mathbf{y}^T\boldsymbol{A}\mathbf{y} + 2\alpha(1-\alpha)\mathbf{y}^T\boldsymbol{A}\mathbf{x} - \alpha\mathbf{x}^T\boldsymbol{A}\mathbf{x} - (1-\alpha)\mathbf{y}^T\boldsymbol{A}\mathbf{y} \le 0.$$

Note that we can simplify this to

$$(\alpha(\alpha-1))(\mathbf{x}^T\boldsymbol{A}\mathbf{x}+\mathbf{y}^T\boldsymbol{A}\mathbf{y}-2\mathbf{x}^T\boldsymbol{A}\mathbf{y})$$

where we have $(\alpha(\alpha-1)) \le 0$ and $(\mathbf{x}^T\boldsymbol{A}\mathbf{x}+\mathbf{y}^T\boldsymbol{A}\mathbf{y}-2\mathbf{x}^T\boldsymbol{A}\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T\boldsymbol{A}(\mathbf{x}-\mathbf{y}) \ge 0$. Hence, the entire expression is ≤ 0, and we are done! ∎

1.3 THE GRADIENT

Last time we proved that $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\boldsymbol{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}$ was a convex function.

Let's show that the gradient of $f(\mathbf{x})$ is really the vector $\boldsymbol{A}\mathbf{x}-\mathbf{b}$.

EXAMPLE 2 *Consider the function* $f(\mathbf{x}) = \frac{1}{2}\begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} 3 & -1 \\ -1 & 4 \end{bmatrix} - \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} 7 \\ -6 \end{bmatrix} = 3/2x^2 + 2y^2 - xy - 7x + 6y$. *Then the gradient is the vector*

$$\begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \end{bmatrix} = \begin{bmatrix} 3x-y-7 \\ 4y-x+6 \end{bmatrix} = \begin{bmatrix} 3 & -1 \\ -1 & 4 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} 7 \\ -6 \end{bmatrix}.$$

More generally,

$$f(x_1,\ldots,x_n) = 1/2\sum_{ij} A_{ij}x_ix_j - \sum_i x_ib_i$$

We like thining of this in terms of the following table:

$A_{11}x_1^2$	$A_{12}x_1x_2$	$\cdots$	$A_{1n}x_1x_n$
$A_{21}x_2x_1$	$A_{22}x_2^2$	$\cdots$	$\vdots$
$\vdots$	$\ddots$	$\ddots$	$\vdots$
$A_{n1}x_nx_1$	$\cdots$	$\cdots$	$A_{nn}x_n^2$

Now we have terms involving x_i in the ith row and ith column.

$$\partial f/\partial x_i = 1/2\sum_{j\neq i} A_{ij}x_i + A_{ii}x_i + 1/2\sum_{j\neq i} A_{ji}x_i - b_i = \text{ith row of } \boldsymbol{A}^T\mathbf{x} - b_i$$

1.4 THE MINIMIZER

The minimizer of a function is any point that is the lowest in some neighborhood. Formally, a point $\mathbf{x}^*$ is a local minimizer if $f(\mathbf{x}^*) \le f(\mathbf{x})$ for all $\mathbf{x}$ where $\|\mathbf{x}-\mathbf{x}^*\| \le \varepsilon$ for some positive value of ε. This just means that this is the lowest point in a neighborhood around the current point. The global minimizer $\mathbf{x}^*$ of a function is a point which is lower than everywhere else: $f(\mathbf{x}^*) \le f(\mathbf{x})$ for all $\mathbf{x}$.[1]

Convex functions are awesome because any local minimizer is a global minimizer!

This is easy to prove for continuous functions like the $f(\mathbf{x})$ that solves linear systems. Consider a point $\mathbf{x}$ and $\mathbf{y}$ where $\mathbf{x}$ is a local minizer and $\mathbf{y}$ is a global minimizer. Then along the line $\alpha\mathbf{x} + (1-\alpha)\mathbf{y}$ we must have that the function is bounded below by $\alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$. Because $\mathbf{x}$ isn't a global min, we know that $f(\mathbf{y}) < f(\mathbf{x})$. Hence, that we *must* reduce the value of the function for all positive α compared with $f(\mathbf{x})$. This means that $f(\mathbf{x})$ couldn't have been a local minimizer. Hence, any local minimizer is a global minimizer of a continuous convex function.

[1] For functions that aren't defined everywhere, this would be restricted to whereever the function is defined.

1.5 CHARACTERIZING THE MINIMIZER

Any point where the gradient is zero is a global minimizer for a continuous convex function.

This is true generally, but it's super easy to show for our function for linear systems.

THEOREM 3 [2] *Let* $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\boldsymbol{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}$ *where* $\boldsymbol{A}$ *is symmetric, positive definite. Then the vector of partial derivatives is* $\mathbf{g}(\mathbf{x}) = \boldsymbol{A}\mathbf{x}-\mathbf{b}$. *Let* $\mathbf{y}$ *be a point where* $\boldsymbol{A}\mathbf{x}-\mathbf{b} = 0$. *Then* $f(\mathbf{x}) \ge f(\mathbf{y})$.

[2] This theorem generalizes to any function with a positive definite Hessian, but that's for an optimization class.

Proof Let $\mathbf{x} = \mathbf{y} + \alpha\mathbf{z}$. Then:

$$f(\mathbf{x}) = \tfrac{1}{2}(\mathbf{y}+\alpha\mathbf{z})^T \boldsymbol{A}(\mathbf{y}+\alpha\mathbf{z}) - (\mathbf{y}+\alpha\mathbf{z})^T\mathbf{b} = \tfrac{1}{2}\mathbf{y}^T\boldsymbol{A}\mathbf{y} + \tfrac{1}{2}\alpha^2\mathbf{z}^T\boldsymbol{A}\mathbf{z} + \alpha\mathbf{z}^T\boldsymbol{A}\mathbf{y} - \mathbf{y}^T\mathbf{b} - \alpha\mathbf{z}^T\mathbf{b}.$$

Now, recall that $\boldsymbol{A}\mathbf{y} = \mathbf{b}$ because the gradient is zero. Then we have:

$$f(\mathbf{x}) = \tfrac{1}{2}\mathbf{y}^T\boldsymbol{A}\mathbf{y} + \tfrac{1}{2}\alpha^2\mathbf{z}^T\boldsymbol{A}\mathbf{z} + \alpha\mathbf{z}^T\mathbf{b} - \mathbf{y}^T\mathbf{b} - \alpha\mathbf{z}^T\mathbf{b} = f(\mathbf{y}) + \tfrac{1}{2}\alpha^2\mathbf{z}^T\boldsymbol{A}\mathbf{z} \geq f(\mathbf{y}). \qquad \blacksquare$$

1.6 FINDING THE MINIMIZER

If the gradient is not zero, then we can always reduce the function by moving a sufficiently along the negative gradient.

In general, this is just an application of Taylor's theorem for multivariate function, but we can again proof this easily for us, and get a cool result along the way!

Suppose $\mathbf{g}(\mathbf{x}) = \boldsymbol{A}\mathbf{x} - \mathbf{b} \neq 0$.[3] Then consider

[3] For the moment, we'll let $\mathbf{g} = \mathbf{g}(\mathbf{x}) = \boldsymbol{A}\mathbf{x} - \mathbf{b}$ for a fixed $\mathbf{x}$.

$$f(\mathbf{x}-\alpha\mathbf{g}) = \tfrac{1}{2}\mathbf{x}^T\boldsymbol{A}\mathbf{x}+\tfrac{1}{2}\alpha^2\mathbf{g}^T\boldsymbol{A}\mathbf{g}-\alpha\mathbf{g}^T\boldsymbol{A}\mathbf{x}-\mathbf{x}^T\mathbf{b}+\alpha\mathbf{g}^T\mathbf{b} = f(\mathbf{x})+\tfrac{1}{2}\alpha^2\mathbf{g}^T\boldsymbol{A}\mathbf{g}-\alpha\mathbf{g}^T\boldsymbol{A}\mathbf{x}+\alpha\mathbf{g}^T\mathbf{b} = f(\mathbf{x})+\alpha(\alpha/2\mathbf{g}^T\boldsymbol{A}\mathbf{g}+\alpha\mathbf{g}^T\mathbf{g}).$$

So if this result is going to be true, we need $(\alpha/2\mathbf{g}^T\boldsymbol{A}\mathbf{g} + \alpha\mathbf{g}^T\mathbf{g})$ for α small enough. Let

$$\rho = \begin{array}{ll} \text{maximize} & \frac{\mathbf{x}^T\boldsymbol{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \\ \text{subject to} & \mathbf{x} \neq 0 \end{array} \quad \text{then} \quad \rho \geq \frac{\mathbf{g}^T\boldsymbol{A}\mathbf{g}}{\mathbf{g}^T\mathbf{g}} \text{ for any vector } \mathbf{g}.$$

Hence, $\alpha/2\mathbf{g}^T\boldsymbol{A}\mathbf{g} \leq \rho\alpha/2\mathbf{g}^T\mathbf{g}$. Thus, if $\rho\alpha/2 \leq 1$ or $\alpha \leq 2/\rho$ we have

$$f(\mathbf{x}-\alpha\mathbf{g}) = f(\mathbf{x}) - \alpha\underbrace{(\alpha/2\mathbf{g}^T\boldsymbol{A}\mathbf{g} + \alpha\mathbf{g}^T\mathbf{g})}_{\geq 0} \leq f(\mathbf{x}).$$

Note this is exactly the same bound we got out of the Richardson method too!

2 THE STEEPEST DESCENT ALGORITHM FOR SOLVING LINEAR SYSTEMS

We now need to turn these insights into an algorithm for solving a linear system of equations. The idea in steepest descent is that we use the insight from the last section: we are trying to minimize $f(\mathbf{x})$ and we can make $f(\mathbf{x})$ smaller by taking a step along the gradient $g(\mathbf{x})$.

2.1 FROM RICHARDSON TO STEEPEST DESCENT

Steepest descent on $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\boldsymbol{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}$ is just a generalization of Richardson's iteration:

$$\begin{array}{rl} \text{Richardson} & \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha\underbrace{(\mathbf{b}-\boldsymbol{A}\mathbf{x}^{(k)})}_{\text{residual}} \\ \text{Steepest Descent} & \mathbf{x}^{(k)} = \mathbf{x}^{(k)} - \alpha\underbrace{\mathbf{g}(\mathbf{x})}_{\text{gradient}} = \mathbf{x}^{(k)} - \alpha(\boldsymbol{A}\mathbf{x}-\mathbf{b}). \end{array}$$

This means that if $0 < \alpha < 2/\rho$ then the steepest descent method will converge.

2.2 PICKING A BETTER VALUE OF α

The idea with the steepest descent method is that we can pick α at each step and use $f(\mathbf{x})$ to inform this choice. This method arose from a completely different place from Richardson's method for solving $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ (which was based on the Neumann series).

\begin{definition}[Steepest Descent Algorithm] Let $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ be a symmetric, positive definite linear system of equeations.

3 EXERCISES

1. (I'm not sure if this is true). Let $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ be a diagonally dominant M matrix, but where $\boldsymbol{A}$ is not symmetric. This means that $\boldsymbol{A}^{-1} \geq 0$. Suppose also that $\mathbf{b} \geq 0$. Develop an algorithm akin to steepest descent for this problem. Ideas include looking at functions like $f(x) = **e**^T **Ax**-