

CS 8803-MDM Lecture 25

Parameter Optimization IV: Linear and Online

Alexander Gray

agray@cc.gatech.edu

Georgia Institute of Technology

Today

1. Linear Algebraic Optimization
2. Online Convex Optimization

Computational Linear Algebra

There is one main computation studied in numerical linear algebra: Solving a system of linear equations $Ax = b$ where A is $N \times N$, or perhaps $Ax_j = b_j$ for many j .

It has some common special cases:

- Inverting a matrix A to obtain A^{-1} ($AA^{-1} = I$; think $x = A^{-1}b$).
- Solving a linear least-squares problem $\min_x \|Ax - b\|_2$.
- Solving an eigenvalue problem $Ax = \lambda x$.

All three of these come up commonly in statistics, e.g. respectively in evaluating a Gaussian density, linear regression, PCA.

Types of Matrices

There are several special cases of matrices for which there are often faster custom methods:

- banded (diagonal, tridiagonal, triangular, Hessenberg...)
- block
- symmetric ($A^T = A$)
- definite (positive, semi-)
- Vandermonde
- Toeplitz

Mostly, these allow different matrix decompositions, which lead to different types of solutions.

Types of Matrices

These lead to implicit storage, and matrix multiplication:

- sparse
- kernel

Examples:

- Covariance matrices: symmetric positive definite, sometimes diagonal.
- AR models: Toeplitz.
- Kernel PCA: kernel matrix, or sparse.

Linear Regression

Recall linear regression:

$$y_i = X_i\beta + \epsilon_i = \sum_j \beta_j X_{ij} + \epsilon_i. \quad (1)$$

We want the parameters β which minimize the squared error, or residual sum of squares

$$\text{RSS}(\beta) = \sum_{i=1}^N \left(y_i - \hat{f}(x_i) \right)^2 \quad (2)$$

$$= (y - X\beta)^T (y - X\beta) \quad (3)$$

$$= \|y - X\beta\|_2. \quad (4)$$

Least Squares Problem

This has the form

$$\min_x \|Ax - b\|_2. \quad (5)$$

In linear algebra this is the standard linear *least-squares problem*.

This is basically solving $Ax = b$ when the system is *overdetermined* (more equations than unknowns). It can be put in the form of $A'x = b'$ with square A' as

$$(A^T A)x = A^T b, \quad (6)$$

called the *normal equations* of the least-squares problem.

Optimization Form

Consider the linear system $Ax = b$. Assume for the moment that A is symmetric and positive definite (thus square). Let's consider the best solution to the quadratic form

$$\phi(x) = \frac{1}{2}x^T Ax - x^T b. \quad (7)$$

Then the minimizer x^* is exactly the solution to $Ax = b$.

Optimization Form

To see this, suppose x^* is exactly the solution to $Ax = b$. Then by completing the square,

$$\phi(x) = \frac{1}{2}x^T Ax - x^T Ax^* \quad (8)$$

$$= \frac{1}{2}x^T Ax - x^T Ax^* + \frac{1}{2}x^{*T} Ax^* - \frac{1}{2}x^{*T} Ax^* \quad (9)$$

$$= \frac{1}{2}(x - x^*)^T A(x - x^*) - \frac{1}{2}x^{*T} Ax^*. \quad (10)$$

The last term does not depend on x , so $\phi(x)$ is minimized when $\frac{1}{2}(x - x^*)^T A(x - x^*)$ is minimized. Since A is positive definite, we know that this term is positive unless $x - x^*$ is 0.

So $\phi(x)$ takes its minimum when and only when $x = x^*$.

Optimization Form

More simply, looking at the derivative we find that

$$\nabla\phi = Ax - b. \quad (11)$$

Clearly the only point at which the gradient is zero is the solution of $Ax = b$. So the problem $Ax = b$ can be written as the optimization problem $\phi(x) = \frac{1}{2}x^T Ax - x^T b$.

Steepest Descent

At the current guess x_k , the function ϕ decreases most rapidly in the direction of the negative gradient:

$-\nabla\phi(x_k) = b - Ax_k$. If the *residual*

$$r_k = b - Ax_k \quad (12)$$

of x_k is nonzero, there exists a positive α such that $\phi(x_k + \alpha_k r_k) < \phi(x_k)$.

The residual gives us our search direction, $p_k = r_k$. Now we need the step size α_k ; determining it is called *line search*.

Steepest Descent

Line search is difficult in general, but for our quadratic function we can actually do *exact* line search, in which we choose α_k so that

$$\phi(x_{k+1}) = \min_{\alpha} \phi(x_k + \alpha p_k), \quad (13)$$

in which case we set

$$\alpha_k = \frac{p_k^T r_k}{p_k^T A p_k} \quad (14)$$

to obtain steepest descent with exact line search.

Conjugate Gradient

It turns out that this isn't that efficient. Taking the steepest direction from x_k is not the same as taking the direction which goes to the bottom of a long trough with steep sides, say.

We will do a smoothing of the directions over the iterations, by using

$$p_k = r_{k-1} + \beta_k p_{k-1} \quad (15)$$

where

$$\beta_k = \frac{\langle r_{k+1}, r_{k+1} \rangle}{\langle r_k, r_k \rangle}. \quad (16)$$

This is called *conjugate gradient*. We won't go into its theory.

Conjugate Gradient

What's good about this:

- Inside each iteration, we need only perform a multiplication of the matrix A with a different vector each time. We do not need to make a new matrix the size of A . This is good if A is large.
- It turns out this hones in quickly on the top eigenvalues.

It is the method of choice for large, sparse matrices.

Conjugate Gradient

The same basic approach can be derived this way, from the viewpoint of optimization, or from a different viewpoint involving methods called *Lanczos methods*. These viewpoints can be unified using the framework of *Krylov subspaces*.

There is a version of this approach for the least-squares problem, called *LSQR*, and also called *CGNE*. There is a version of this approach for the eigenvalue problem, called *Arnoldi iteration*.

There are versions for cases which are not symmetric positive definite.

Decomposition-based Methods

The downside of the Krylov-type approaches, also called *iterative methods*, is that their numerical stability is not the best possible.

For full-rank matrices, the methods of choice are *Gaussian elimination* or orthogonalization approaches such as *QR factorization*.

For rank-deficient or unknown-rank matrices, the method of choice is *singular value decomposition*.

Online Optimization Problems

One data point at a time.

Sherman-Morrison Formula

Suppose you have already obtained, after a lot of work, a matrix inverse A^{-1} . Now you want to make a small change to A , for example change one element, or one row, or one column:

$$A \rightarrow A + u \otimes v. \quad (17)$$

($u \otimes v$ is a matrix whose $(i, j)^{th}$ element is the product of the i^{th} component of u and the j^{th} component of v .) If u is a unit vector e_i , this adds v to the i^{th} row; if v is a unit vector e_j , this adds u to the j^{th} column; if both are proportional to unit vectors then a term is only added to one element.

Sherman-Morrison Formula

The *Sherman-Morrison* formula gives the inverse after the change

$$(A + u \otimes v)^{-1} = A^{-1} - \frac{(A^{-1}u) \otimes (vA^{-1})}{vA^{-1}u + 1}. \quad (18)$$

The advantage of this is that it only requires two matrix multiplications and a dot product, which is $O(N^2)$ rather than $O(N^3)$ to redo the entire inverse.

This can be used for obtaining

$$(A + u \otimes v)x = b \quad (19)$$

in an online fashion.

Stochastic Gradient Descent

Consider the batch optimization problem

$$\arg \min_w \left(\frac{1}{N} \sum_{i=1}^N \phi(w^T x_i, y_i) + \frac{\lambda}{2} \|w\|_2^2 \right). \quad (20)$$

where ϕ is convex. *Stochastic gradient descent* updates the parameters one datum at a time via

$$\hat{w}_k = \hat{w}_{k-1} - \alpha_k (\lambda \hat{w}_{k-1} + \phi'(\hat{w}_{k-1}^T x_k, y_k) x_k) \quad (21)$$

where $\phi' = \frac{\partial}{\partial w} \phi$.

Stochastic Gradient Descent

It can be shown that as $k \rightarrow \infty$ and $\alpha_k \rightarrow 0$,

$$\mathbb{E}_{x,y} \phi(\hat{w}_k^T, x, y) + \frac{\lambda}{2} \|\hat{w}_k\|_2^2 \rightarrow e^* \quad (22)$$

where

$$e^* = \arg \min_w \mathbb{E}_{x,y} \phi(w^T, x, y) + \frac{\lambda}{2} \|w\|_2^2, \quad (23)$$

i.e. the online estimate converges to the true optimum value.