Optimization Algorithms: Theory and Software Implementation

Prof. Thirumulanathan D

Department of Mathematics

Institute of IIT Kanpur

Lecture: 29

This lecture continues the discussion on quasi-Newton methods, focusing on proving two key properties of the DFP (Davidon-Fletcher-Powell) algorithm. We begin by recalling that the quasi-Newton condition is $\delta_k = B_{k+1}\gamma_k$, and we have studied two methods that satisfy it: the rank-one correction and the DFP method. The core difference between them lies in the formula for updating the matrix B_k. The process for choosing the descent direction and the step length remains identical.

From our numerical experiments, both algorithms minimized a five-dimensional quadratic function in exactly five iterations. For the non-quadratic function $f(x_1, x_2) = x_1^2e^{x_2} + x_2^2e^{x_1}$, their performance was also very similar. Starting from initial points like (1, 1) or (-0.5, -0.5), both converged to a solution. However, when starting from $(-\sqrt{2}, -\sqrt{2})$, both algorithms converged to the saddle point (-2, -2). This demonstrates the significant common ground between the two methods.

This lecture will prove two fundamental properties of the DFP method:

1. For a quadratic function, the algorithm converges to the minimizer in at most *n* steps, and upon completion, the matrix B_n is equal to the inverse of the Hessian, H^{-1}.

2. For general functions, when using an exact line search, the matrices B_{k+1} remain positive definite.

Note: $B_{k+1}= B^{k+1}$, $\gamma_k= \gamma^k$, $\delta_k= \delta^k$, $x_{n+1}= x^{n+1}$ (Notation)

We will prove the first property in this lecture. Recall the DFP update formula:

$B_{k+1} = B_k + (\delta_k\delta_k^T)/(\delta_k^T\gamma_k) - (B_k\gamma_k\gamma_k^TB_k)/(\gamma_k^TB_k\gamma_k)$

Theorem: Let $f(x) = (1/2)x^THx + b^Tx + c$ be a quadratic function where H is a positive definite matrix. If the DFP method with an exact line search is used to minimize f, then the method converges to the minimizer x* in at most n steps, i.e., $x_{n+1} = x^*$. Furthermore, upon convergence, $B_n = H^{-1}$.

Proof:

The proof establishes two properties by mathematical induction: the hereditary property and the conjugacy of the steps with respect to H.

Let $\delta_i = x_{i+1} - x_i$ and $\gamma_i = \nabla f(x_{i+1}) - \nabla f(x_i) = g_{i+1} - g_i$. For a quadratic function, the gradient is linear, so $\gamma_i = H\delta_i$.

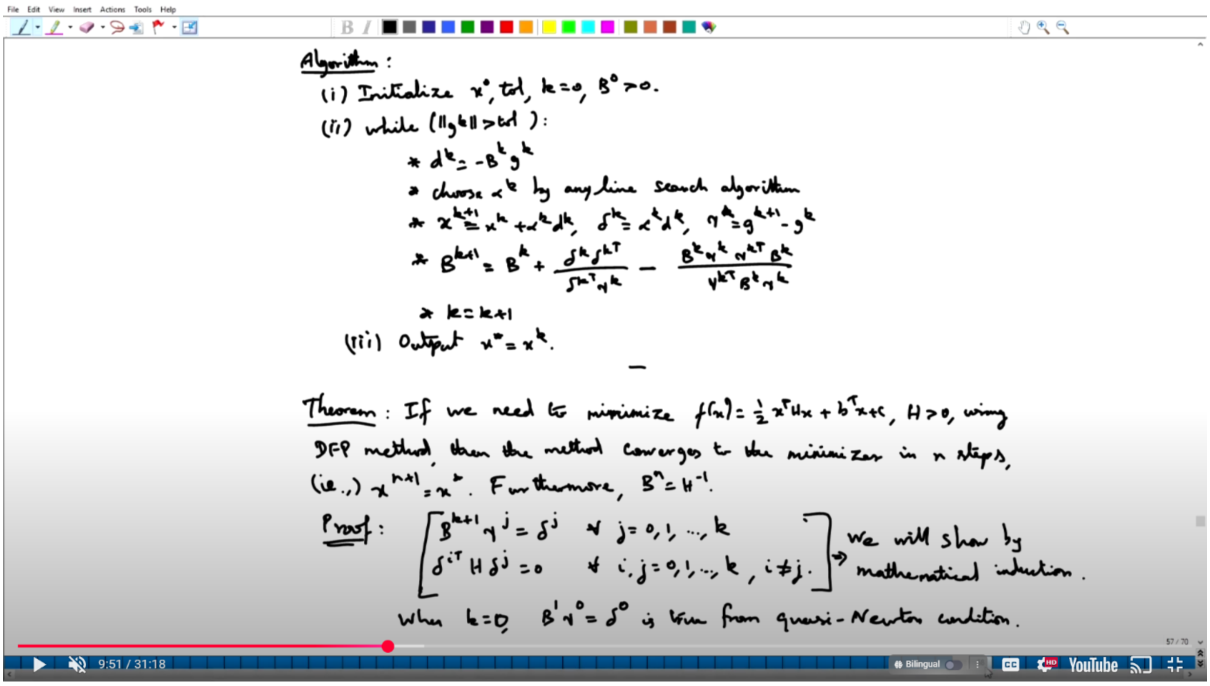
We will show that for all k, the following hold:

1. **Hereditary Property:** $B_{k+1}\gamma_j = \delta_j$ for all $j = 0, 1, ..., k$.
2. **Conjugacy (H-orthogonality):** $\delta_i^T H \delta_j = 0$ for all $i, j \in \{0, 1, ..., k\}$ where $i \neq j$.

Base Case (k=0):

The hereditary property, $B_1\gamma_0 = \delta_0$, is true by the quasi-Newton condition. The conjugacy condition is vacuously true as there are no distinct indices i and j. Thus, both properties hold for k=0.

(Refer Slide Time 9:51)



Inductive Step:

Assume both properties hold for all indices up to k-1. We will prove they hold for k.

Part 1: Proving $\delta_i^T H \delta_k = 0$ for all $i < k$ (Conjugacy for step k)

First, consider the gradient at iteration k. For a quadratic function, $g_k = Hx_k + b$. We can express g_k relative to an earlier iteration $i < k$:

$$g_k = g_{i+1} + H(\delta_{i+1} + \delta_{i+2} + ... + \delta_{k-1})$$

Now, consider the inner product $\delta_i^T g_k$:

$$\delta_i^T g_k = \delta_i^T g_{i+1} + \delta_i^T H \delta_{i+1} + \delta_i^T H \delta_{i+2} + ... + \delta_i^T H \delta_{k-1}$$

We now show that each term in this sum is zero.

* $\delta_i^T g_{i+1} = 0$: This results from the exact line search. The step size α_i is chosen to minimize $f(x_i + \alpha d_i)$. The optimality condition is $d_i^T \nabla f(x_i + \alpha_i d_i) = 0$. Since $\delta_i = \alpha_i d_i$ and $\nabla f(x_i + \alpha_i d_i) = g_{i+1}$, it follows that $\delta_i^T g_{i+1} = 0$.

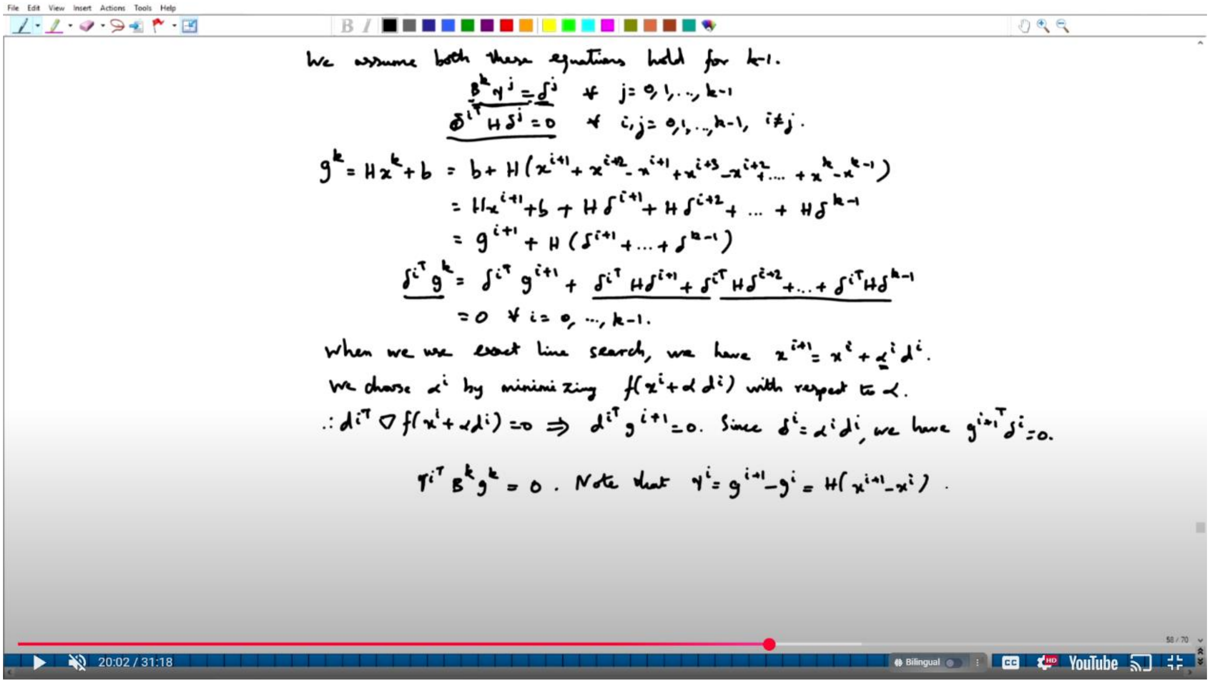
* $\delta_i^T H \delta_j = 0$ for $j = i+1$ to $k-1$: This is true by the conjugacy property of the induction hypothesis.

Therefore, $\delta_i^T g_k = 0$ for all $i < k$.

By the hereditary property (induction hypothesis), we have $\delta_i = B_k \gamma_i$. Substituting gives:

$\delta_i^T g_k = (B_k \gamma_i)^T g_k = \gamma_i^T B_k g_k = 0$

(Refer Slide Time 20:02)



Since $\gamma_i = H\delta_i$, we get:

$\delta_i^T H B_k g_k = 0$

Note that the search direction is $d_k = -B_k g_k$, and $\delta_k = \alpha_k d_k$. Therefore, $B_k g_k = -\delta_k/\alpha_k$. Substituting yields:

$\delta_i^T H (-\delta_k/\alpha_k) = 0 \Rightarrow \delta_i^T H \delta_k = 0$ for all $i < k$.

This establishes the conjugacy property for step k.

Part 2: Proving $B_{k+1}\gamma_j = \delta_j$ for all $j \leq k$ (Hereditary property for step k)

We now use the DFP update formula. For any $j \leq k$:

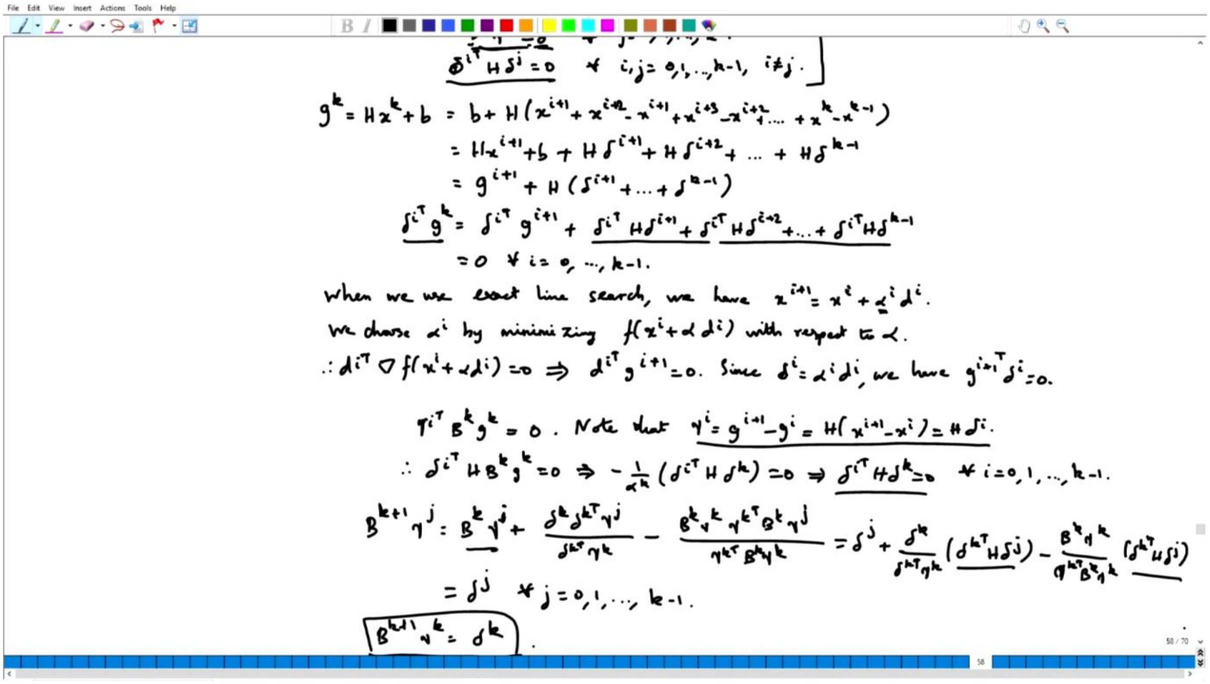
$B_{k+1}\gamma_j = B_k\gamma_j + (\delta_k \delta_k^T \gamma_j)/(\delta_k^T \gamma_k) - (B_k \gamma_k \gamma_k^T B_k \gamma_j)/(\gamma_k^T B_k \gamma_k)$

We analyze this equation case by case:

* For $j < k$:

* By the induction hypothesis (hereditary property for B_k), $B_k\gamma_j = \delta_j$.

(Refer Slide Time 25:40)



* The term $\delta_k^T\gamma_j = \delta_k^T(H\delta_j) = (\delta_k^T H\delta_j)$. From the conjugacy property we just proved for k, this is zero for $j < k$.

* The term $\gamma_k^T B_k\gamma_j = \gamma_k^T\delta_j = (H\delta_k)^T\delta_j = \delta_k^T H\delta_j$, which is also zero for $j < k$.

Therefore, for $j < k$, the second and third terms vanish, and we get $B_{k+1}\gamma_j = B_k\gamma_j = \delta_j$.

* For $j = k$:

The formula becomes $B_{k+1}\gamma_k = B_k\gamma_k + (\delta_k\delta_k^T\gamma_k)/(\delta_k^T\gamma_k) - (B_k\gamma_k\gamma_k^T B_k\gamma_k)/(\gamma_k^T B_k\gamma_k)$.

Simplifying:

* The second term is $(\delta_k\delta_k^T\gamma_k)/(\delta_k^T\gamma_k) = \delta_k$.

* The third term is $(B_k\gamma_k\gamma_k^T B_k\gamma_k)/(\gamma_k^T B_k\gamma_k) = B_k\gamma_k$.

Thus, $B_{k+1}\gamma_k = B_k\gamma_k + \delta_k - B_k\gamma_k = \delta_k$, which is the quasi-Newton condition.

Therefore, the hereditary property $B_{k+1}\gamma_j = \delta_j$ holds for all $j = 0, 1, \ldots, k$. This completes the inductive step.

Final Step: Showing $B_n = H^{-1}$ and $x_{n+1} = x^*$

By the hereditary property, after n steps we have:

$B_n [\gamma_0 \mid \gamma_1 \mid \ldots \mid \gamma_{n-1}] = [\delta_0 \mid \delta_1 \mid \ldots \mid \delta_{n-1}]$

Since $\gamma_i = H\delta_i$, this can be rewritten as:

$B_n H [\delta_0 \mid \delta_1 \mid \ldots \mid \delta_{n-1}] = [\delta_0 \mid \delta_1 \mid \ldots \mid \delta_{n-1}]$

We now show the vectors $\delta_0, \delta_1, \ldots, \delta_{n-1}$ are linearly independent. Suppose, for contradiction, that they are not. Then there exist coefficients $\beta_0, \beta_1, \ldots, \beta_{n-1}$, not all zero, such that $\sum\beta_i\delta_i = 0$. Multiplying this equation on the left by $\delta_j^T H$ for any j gives:

$$\delta_j^T H \left(\sum\beta_i\delta_i\right) = \sum\beta_i(\delta_j^T H\delta_i) = \beta_j(\delta_j^T H\delta_j) = 0$$

The last equality follows because $\delta_j^T H\delta_i = 0$ for $i \neq j$ by the conjugacy property.

Since H is positive definite, $\delta_j^T H\delta_j > 0$, which forces $\beta_j = 0$.

This argument holds for every j, so all β_i must be zero, contradicting our assumption. Therefore, the vectors δ_i are linearly independent.

Since the δ_i's form a basis for R^n, the matrix $[\delta_0 \mid \delta_1 \mid \ldots \mid \delta_{n-1}]$ is invertible.

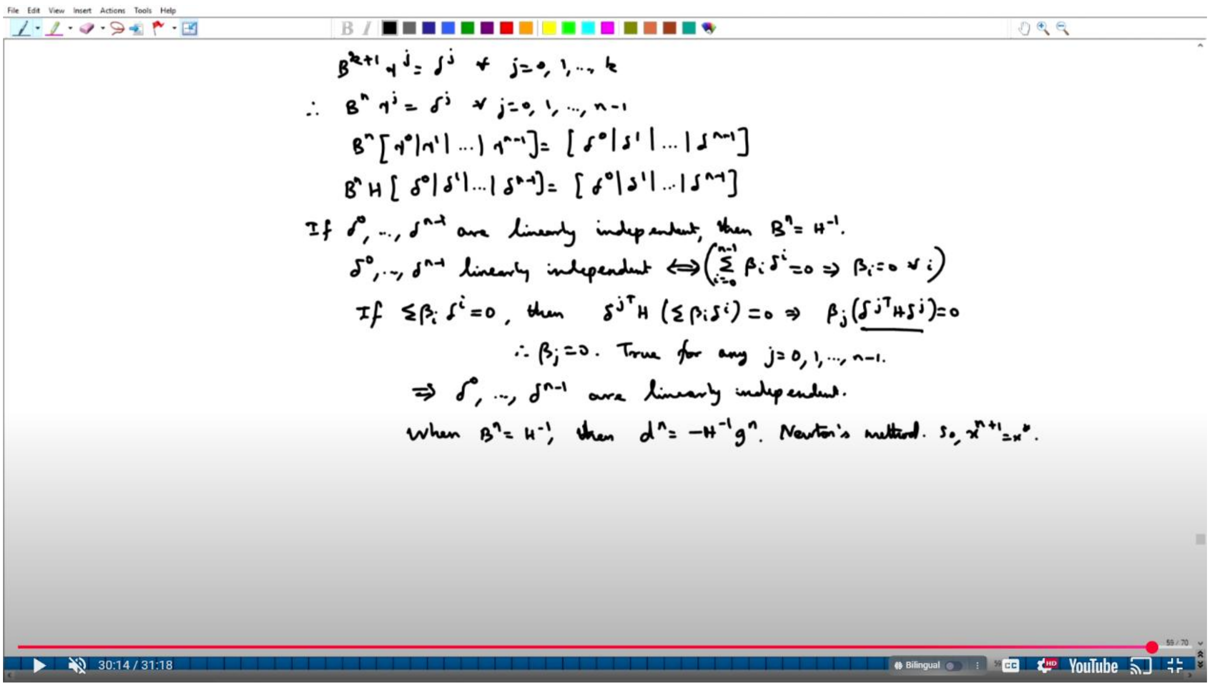
From the equation $B_n H [\delta] = [\delta]$, it follows that $B_n H = I$, and thus $B_n = H^{-1}$.

With $B_n = H^{-1}$, the search direction at step n is the Newton direction, $d_n = -B_n g_n = -H^{-1}g_n$.

For a quadratic function, the Newton step converges to the minimizer in one step.

Therefore, $x_{n+1} = x^*$.

(Refer Slide Time 30:30)



This proves that the DFP method converges in at most n steps for a quadratic function and exactly computes the inverse Hessian. This property is shared by the rank-one correction method as well.

The second property, regarding the preservation of positive definiteness in the general case, will be addressed in the next lecture. Thank you.