MATH 565: Lecture 14 (02/26/2026)

Today: * Conjugate gradient method

Computationally Efficient Variants of Newton Method

Recall: Newton update: $\bar{w} \leftarrow \bar{w} - \alpha H^{-1} \nabla$. But inverting large, e.g., $10^6 \times 10^6$, Hessian is impractical...

There is a class of methods called quasi-Newton methods, which approximate the Hessian. But we first introduce the conjugate gradient method.

B. Conjugate Gradient Method (CGM)

This is considered a Hessian-free optimization method. CGM expresses the Newton step as a sequence of d steps by choosing d directions that are orthogonal, i.e., conjugate directions.

The CGM takes d steps to get to the optimum of a quadratic J.

Idea: Any quadratic J can be represented as a sum of additively separable univariate functions using an appropriate basis's.

With $\bar{w} \to \bar{w}'$ (in a different basis), we write

$$J(\bar{w}') = \sum_{i=1}^{d} J_i(w_i') \quad \to \text{separable}$$

↳ optimize each dimension (direction) individually/separately.

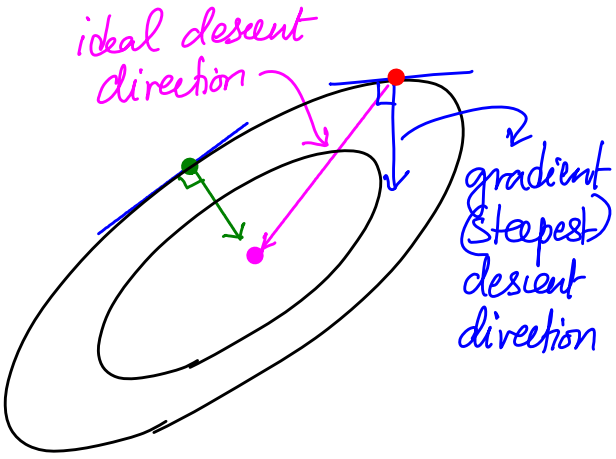
$\bar{w}'$ ← such a basis for quadratic $J(\bar{w})$.

We can always find such a basis for quadratic $J(\bar{w})$. Hence, we consider that case in detail first.

Let $J(\bar{w}) = \frac{1}{2}\bar{w}^T H \bar{w} - \bar{b}^T \bar{w} + \bar{c}$, where H is symmetric and PSD.

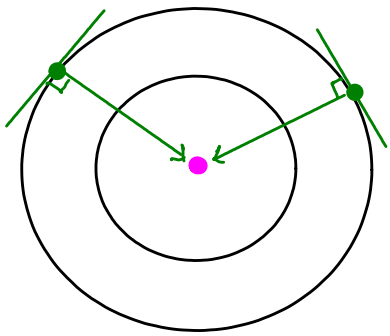
The level sets of J are ellipsoids.

The gradient (steepest) descent direction may not be ideal at all points, e.g., at •, this direction does not point toward the origin •, which the ideal descent direction does.



The gradient descent direction can be ideal at some points, e.g., •. But from many points, it will not be.

If the level sets are spheres (instead of ellipsoids), then the gradient descent direction is ideal at every point, i.e., it always points to the origin. Hence, we solve the problem in a single step from any starting point.



In the general case where the level sets are ellipsoids, we need to choose the descent directions more carefully.

With $J(\bar{w}) = \frac{1}{2}\bar{w}^T H \bar{w} - \bar{b}^T \bar{w} + \bar{c}$, we get

$$\nabla J(\bar{w}) = H\bar{w} - \bar{b}, \text{ and}$$

$$HJ(\bar{w}) = H.$$

Hence, the first order condition $\nabla J = \bar{0}$ amounts to solving the linear system $H\bar{w} = \bar{b}$.

We wrote $-\bar{b}^T\bar{w}$ (instead of $+\bar{b}^T\bar{w}$) to get this convenient result!

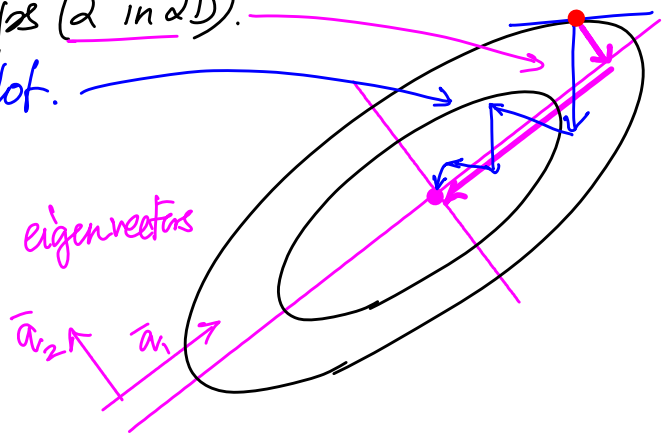
The default Newton update is $\bar{w} \leftarrow \bar{w} - H^{-1}\nabla J$.

Here, if $H^{-1} = I$ (which means $H = I$), then we solve the problem in one update (single step). This corresponds to the case where the level sets are spheres.

If $H \neq I$, we need to choose the directions carefully, and the eigenvectors of H represent these good directions!

Moving along the eigenvectors will get us to the origin in (at most) d steps (2 in 2D).

Gradient descent can oscillate a lot.



To find these good directions,

We rewrite $J(\cdot)$ in the basis using the eigenvectors of H.

This will be an additively separable sum of univariate quadratic loss functions (in the new basis). We can optimize each (new) dimension individually.

If $\bar{q}_j$ is an eigenvector of H, then

$$H\bar{q}_j = \lambda_j \bar{q}_j \qquad (\lambda_i: \text{eigenvalue} \equiv \bar{q}_j).$$

$\Rightarrow$ For any other eigenvector $\bar{q}_i$ of H, since $\bar{q}_i^T \bar{q}_j = 0$, we get

$$\bar{q}_i^T H \bar{q}_j = \lambda \bar{q}_i^T \bar{q}_j = 0, \quad \text{i.e.,}$$

$$\boxed{\bar{q}_i^T H \bar{q}_j = 0}.$$

This is called H-orthogonality in linear algebra, or mutual conjugacy in optimization.

But, eigenvectors of H are not the only vectors satisfying (H-orthogonality) mutual conjugacy ($\bar{q}_i^T H \bar{q}_j = 0$).

Let $Q = [\bar{q}_0 \; \bar{q}_1 \cdots \bar{q}_{d-1}]$ be a set of d H-orthogonal directions.

With $\bar{w} = Q\bar{w}'$, we get

$$J(\bar{w}) = J(Q\bar{w}')$$

$$= \frac{1}{2} \bar{w}'^T Q^T H Q \bar{w}' - \bar{b}^T Q \bar{w}' + \bar{c}$$

$$= \frac{1}{2} \bar{w}'^T \Delta_Q \bar{w}' - \bar{b}'^T \bar{w}' + \bar{c} = J(\bar{w}'), \text{ where}$$

Δ_Q is a diagonal matrix and $\bar{b}' = Q^T \bar{b}$.

$\Rightarrow$ Coordinate descent (using, e.g., line search) along each $\bar{q}_j$ direction ($\bar{w}_j$ direction) gets to the optimum of $J(\bar{w})$ in at most d steps.

Result If J is strictly convex, then J has linearly independent conjugate directions, i.e., such a Q exists.

Q. How to find conjugate directions?

Default approach : GSO(H) Gram-Schmidt Orthogonalization.

But GSO can be costly, as it runs in $O(d^2)$ time.

Can we get away with something that runs in $O(d)$ time?

Yes, as conjugacy to only the most recent direction is enough!

Let's go back to the original $\bar{w}$ (instead of $\bar{w}'$).

We define $\bar{q}_{t+1} = -\nabla J(\bar{w}_{t+1}) + \beta_t \bar{q}_t$ and choose β_t to ensure $\bar{q}_{t+1}^T H \bar{q}_t = 0$.

$$0 = \bar{q}_t^T H \left(\bar{q}_{t+1} = -\nabla J(\bar{w}_{t+1}) + \beta_t \bar{q}_t \right)$$

$$\Rightarrow \beta_t = \frac{\bar{q}_t^T H \nabla J(\bar{w}_{t+1})}{\bar{q}_t^T H \bar{q}_t} \qquad (*)$$

Conjugate Gradient Method (CGM)

Initialization. $\bar{w} = \bar{w}_0$; $\bar{q}_0 = -\nabla J(\bar{w}_0)$

general step t $(t = 0, \dots, T)$ → $T = d$, but can be set smaller

1. $\bar{w}_{t+1} \leftarrow \bar{w}_t + \alpha_t \bar{q}_t$
   where α_t is found using line search to minimize J.
2. $\bar{q}_{t+1} = -\nabla J(\bar{w}_{t+1}) + \beta_t \bar{q}_t$
   where β_t is computed using $(*)$

CGM is supposed to be Hessian-free, but (*) uses H??

In practice, we can do all computations involving H (for β_t's) using projections of H using finite differences.

For a given direction (vector) $\bar{v}$, we can write

$$H\bar{v} \approx \frac{\nabla J(\bar{w}+\delta\bar{v}) - \nabla J(\bar{w})}{\delta} \quad \text{for "small" } \delta.$$

How to choose δ?

If δ is chosen "too large", the approximation may be inaccurate.

If δ is chosen "too small", we may get numerical errors.

A common value: $\delta = \sqrt{\text{machine precision}}$

so, $\delta = 10^{-8}$ for double point floating point representation.