MATH 565: Lecture 15 (03/03/2026)

Today: * CGM for non-quadratic J
* quasi-Newton methods
* BFGS

We saw conjugate gradient method (CGM) for quadratic J.

What about more general J?

At iteration t, with $\bar{w}_t$ known,

- construct $F(\bar{w}_t)$ at $\bar{w}_t$, the second order Taylor expansion of J at $\bar{w}_t$, i.e., quadratic approximation of J at $\bar{w}_t$.

Then, two options

1. Solve $F(\bar{w})$ for a few iterations with $HF(\bar{w}_t)$ fixed (linear CGM); or
2. change $HF(\bar{w}_t)$ at every iteration that updates $\bar{w}_t$ (nonlinear CGM).

Advantages (✓) and Disadvantages (✗)

1. Linear CGM
   - ✓ linear CGM can be solved exactly ($\leq d$ steps)
   - ✓ if quadratic approximation is good, the overall # iterations is not too large.
   - ✗ Hessian may not be accurate for all substeps.
2. Nonlinear CGM
   - ✓ Hessian is accurate at each substep.
   - ✗ mutual conjugacy (H-orthogonality) deteriorates with iterations
     - could restart computation of conjugate directions every few steps.

Linear CGM may work better in practice for ML problems.

C. Quasi-Newton Methods and BFGS

Recall, Newton update: $\bar{w}_{t+1} \leftarrow \bar{w}_t - \alpha_t H_t^{-1} \nabla J_t$

It's too expensive to invert H_t at each step. Instead, in a quasi-Newton method, we do the update as

$$\bar{w}_{t+1} = \bar{w}_t - \alpha_t G_t \nabla J(\bar{w}_t), \text{ where}$$

α_t is the optimal learning rate identified using line search, which need not be exact as in CGM, as conjugacy is no longer required.

But, approximate conjugacy may be maintained by starting $G_0 = I$ (and/or by setting $G_t = I$ after every m, say, iterations).

How do we update G_t?

Using the quasi-Newton condition or **secant condition**:

$$\bar{w}_{t+1} - \bar{w}_t = G_{t+1}\left[\nabla J(\bar{w}_{t+1}) - \nabla J(\bar{w}_t)\right] \quad \text{——— (*)}$$

How did we get (*)? In 1D, the mean value theorem (MVT) gives

$$\frac{J'(w_{t+1}) - J'(w_t)}{w_{t+1} - w_t} \approx J''(\bar{w}_t)$$

In d-dimensions, we consider the first order Taylor expansion of $\nabla J(\bar{w}_{t+1})$ about $\bar{w}_t$:

$$\nabla J(\bar{w}_{t+1}) \approx \nabla J(\bar{w}_t) + HJ(\bar{w}_t)(\bar{w}_{t+1} - \bar{w}_t)$$

$$\Rightarrow (\bar{w}_{t+1} - \bar{w}_t) \approx \underbrace{[HJ(w_t)]^{-1}}_{G_{t+1}} \left(\nabla J(\bar{w}_{t+1}) - \nabla J(\bar{w}_t)\right).$$

$$\underbrace{\bar{w}_{t+1} - \bar{w}_t}_{\bar{q}_t} = G_{t+1}\underbrace{\left[\nabla J(\bar{w}_{t+1}) - \nabla J(\bar{w}_t)\right]}_{\bar{v}_t} \quad \text{——— } (*)$$

$\bar{q}_t = \bar{w}_{t+1} - \bar{w}_t$ and $\bar{v}_t = \nabla J(\bar{w}_{t+1}) - \nabla J(\bar{w}_t)$

$(*)$ can be written as $\bar{q}_t = G_{t+1}\bar{v}_t$,

which is a grossly underdetermined system ($\bar{q}_t, \bar{v}_t \in \mathbb{R}^d$, while $G_{t+1} \in \mathbb{R}^{d \times d}$: d equations in d^2 unknowns).

The <u>BFGS Method</u> (Broyden-Fletcher-Goldfarb-Shanno)

- finds closest (in the sense of Frobenius norm) G_{t+1} to G_t that is symmetric by solving:

$$\min \; \|G_{t+1} - G_t\|_F$$
$$\text{s.t.} \quad \bar{q}_t = G_{t+1}\bar{v}_t \quad (\text{or } (*))$$
$$G_{t+1}^T = G_{t+1}$$

This is a quadratic optimization problem with linear constraints, and can be solved in closed form:

$$G_{t+1} = \underbrace{\left(I - \rho_t \bar{q}_t \bar{v}_t^T\right) G_t \left(I - \rho_t \bar{v}_t \bar{q}_t^T\right)}_{V_t} + \rho_t \bar{q}_t \bar{q}_t^T \quad \text{——— } (u)$$

where $\rho_t = \frac{1}{\bar{q}_t^T \bar{v}_t}$

This update maintains PD (positive definiteness) of G_t!

Lemma 10 If $G_t \succ 0$ (PD) and $\bar{q}_t^T \bar{v}_t > 0$ (curvature condition), then G_{t+1} as given by (u) satisfies $G_{t+1} \succ 0$, i.e., it is PD.

Proof Let $V_t = I - \rho_t \bar{v}_t \bar{q}_t^T$. Then (u) becomes

$$G_{t+1} = \underbrace{V_t^T G_t V_t}_{\succeq 0} + \underbrace{\rho_t \bar{q}_t \bar{q}_t^T}_{\succeq 0}.$$

The first term is PSD, as $\forall \bar{v} \in \mathbb{R}^q \setminus \{\bar{0}\}$,

$$\bar{v}^T V_t^T G_t V_t \bar{v} = \bar{v}'^T G_t \bar{v}' \text{ for } \bar{v}' = V_t \bar{v}$$
$$\geq 0 \text{ as } G_t \succ 0 \text{ (and hence } \succeq 0).$$

The second term is also PSD, as $\rho_t > 0$ by assumption, and $\forall \bar{v} \in \mathbb{R}^d \setminus \{\bar{0}\}$,

$$\bar{v}^T \bar{q}_t \bar{q}_t^T \bar{v} = (\bar{v}^T \bar{q}_t)^2 \geq 0.$$

But how do we know G_{t+1} is PD (and not just PSD)?

Let $\bar{v}^T G_{t+1} \bar{v} = 0$

$\Rightarrow \bar{v}^T V_t G_t V_t \bar{v} = 0$ and $\bar{v}^T \bar{q}_t = 0$, i.e.,

$V_t \bar{v} = \bar{0}$ and $\bar{v}^T \bar{q}_t = 0$.

$$\Rightarrow \left(I - \rho_t \bar{v}_t \bar{q}_t^T\right)\bar{v} = \bar{v} - \rho_t \bar{v}_t \underbrace{(\bar{q}_t^T \bar{v})}_{=0} = 0$$

$\Rightarrow \bar{v} = \bar{0}$.

$\Rightarrow G_{t+1} \succ 0$!

□

But BFGS still needs to maintain and update G_t, which takes $O(d^2)$ time. Can we do something in $O(d)$?

L-BFGS: Limited memory BFGS

— store only the most recent $m \approx 30$ (say) $\bar{q}_t, \bar{v}_t$ vectors.
Set $G_{t-m+1} = I$ and apply (u) m times to derive G_{t+1} (using only vector-vector computations).
→ details in the next lecture...

Non-Differentiable Loss Functions

Examples 1. L_1-loss or L_1-regularizer → $J \sum_{i=1}^{d} |w_i|$
$|w_i| \rightarrow$ not differentiable at $w_i = 0$.

2. ranking loss functions. → rank the positive instances first — minimize sum of ranks of all instances.

D. Subgradient Method for Convex J

(more next time...)