MATH 565: Lecture 16 (03/05/2026)

Today: * L-BFGS
* subgradient method

L-BFGS Details

Newton update: $\bar{w}_{t+1} = \bar{w}_t - \alpha H_t^{-1} \nabla J(\bar{w}_t)$

BFGS: $\bar{w}_{t+1} = \bar{w}_t + G_t \bar{d}_t$ for descent direction $\bar{d}_t$

$d_t = -\nabla J(\bar{w}_t)$ by default

$$G_{t+1} = V_t^T G_t V_t + \rho_t \bar{q}_t \bar{q}_t^T \quad \text{——— (u)}$$

where $V_t = I - \rho_t \bar{v}_t \bar{q}_t^T$, $V_t^T = I - \rho_t \bar{q}_t \bar{v}_t^T$, and $\rho_t = \frac{1}{\bar{q}_t^T \bar{v}_t}$.

Applying (u) repeatedly for $t, t-1, t-2, \ldots$ we get

$$G_t = \underbrace{(V_{t-1}^T \cdots V_{t-m}^T)}_{\mathcal{V}_L} G_0 \underbrace{(V_{t-m} \cdots V_{t-1})}_{\mathcal{V}_R} + (\text{rank 2 correction terms})$$

In more detail,

$$G_t = \mathcal{V}_L G_0 \mathcal{V}_R + \sum_{j=t-m}^{t-1} \rho_j \left(\prod_{i=t-1}^{j+1} V_i^T \right) \bar{q}_j \bar{q}_j^T \left(\prod_{i=j+1}^{t-1} V_i \right)$$

In L-BFGS, we compute $G_t \bar{d}$ using only vector operations.

Let's apply G_t to $\bar{d}$ (descent direction) to compute $G_t \bar{d}$:

$$G_t \bar{d} = \mathcal{V}_L G_0 \mathcal{V}_R \bar{d} + \sum_{j=t-m}^{t-1} \rho_j \left(\prod_{i=t-1}^{j+1} V_i^T \right) \bar{q}_j \underbrace{\left(\bar{q}_j^T \prod_{i=j+1}^{t-1} V_i \bar{d} \right)}_{\text{scalar}}$$

$$G_t \bar{d} = \mathcal{V}_L G_0 \mathcal{V}_R \bar{d} + \sum_{j=t-m}^{t-1} \rho_j \left(\prod_{i=t-1}^{j+1} V_i^T \right) \bar{q}_j \left(\bar{q}_j^T \prod_{i=j+1}^{t-1} V_i \bar{d} \right)$$

Each $V_i \bar{d} = (I - \rho_i \bar{v}_i \bar{q}_i^T) \bar{d}$ } — rank-1 update

$= \bar{d} - \rho_i \bar{v}_i (\bar{q}_i^T \bar{d})$ ← scalar

We can do several, $O(m)$ to be precise, such updates, each taking only $O(d)$ time.

We first present the details of the calculations for $m=2$ case.

Details for $m=2$ case

We have $\{(\bar{q}_0, \bar{v}_0), (\bar{q}_1, \bar{v}_1)\}$, and the whole expression is

$$G_2 = V_1^T V_0^T G_0 V_0 V_1 + V_1^T \rho_0 \bar{q}_0 \bar{q}_0^T V_1 + \rho_1 \bar{q}_1 \bar{q}_1^T$$

Let's do a backward pass, G_0 initialization, and a forward pass to compute $G_2 \bar{d}$ for $\bar{d} = \nabla J(\bar{w}_2)$.

Backward pass

Step 1 $i=1$

* $\alpha_1 = \rho_1 \bar{q}_1^T \bar{d}$
* $\bar{d} \leftarrow \bar{d} - \alpha_1 \bar{v}_1$

} $\equiv V_1 \bar{d}$ $\quad (I - \rho_1 \bar{v}_1 \bar{q}_1^T) \bar{d}$

Step 2 $i=0$

* $\alpha_0 = \rho_0 \bar{q}_0^T \bar{d}$
* $\bar{d} \leftarrow \bar{d} - \alpha_0 \bar{v}_0$

} $= V_0 (V_1 \bar{d})$

G_0 initialization

$G_0 = \gamma_0 I$ for $\gamma_0 = \frac{\bar{q}_0^T \bar{v}_0}{\bar{v}_0^T \bar{v}_0}$ $\left(\text{in general, } \gamma_t = \frac{\bar{q}_{t-1}^T \bar{v}_{t-1}}{\bar{v}_{t-1}^T \bar{v}_{t-1}} \right)$

Set $\bar{z} = \gamma_0 \bar{d} \equiv G_0 (V_0 V_1 \bar{d})$

Forward Pass We need to take care of all rank-2 correction terms!

Step 1 $i=0$

* $\beta = \rho_0 \bar{v}_0^T \bar{z}$
* $\bar{z} \leftarrow \bar{z} + \bar{q}_0 (\alpha_0 - \beta)$

$\Big\} \equiv (V_0^T G_0 V_0 + \rho_0 \bar{q}_0 \bar{q}_0^T) x$ (part of $\bar{d}$)

Step 2 $i=1$

* $\beta = \rho_1 \bar{v}_1^T \bar{z}$
* $\bar{z} \leftarrow \bar{z} + \bar{q}_1 (\alpha_1 - \beta)$

Return $-\bar{z} = -G_2 \bar{d}$ (as the step for L-BFGS)

L-BFGS Algorithm

$\bar{d} = \nabla J(\bar{w}_t)$, $\{ (\bar{q}_{t-1}, \bar{v}_{t-1}), \dots (\bar{q}_{t-m}, \bar{v}_{t-m}) \}$.

↳ stored vectors from m previous iterations

1. Backward Pass

   for $i = t-1$ to $t-m$
   * $\alpha_i = \rho_i \bar{q}_i^T \bar{d}$
   * $\bar{d} \leftarrow \bar{d} - \alpha_i \bar{v}_i$

2. Initial scaling $G_0 = \gamma_t I$, for $\gamma_t = \frac{\bar{q}_{t-1}^T \bar{v}_{t-1}}{\bar{v}_{t-1}^T \bar{v}_{t-1}}$, $\bar{z} = \gamma_d \bar{d}$.

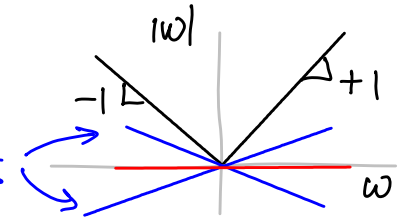
3. Forward Pass

   for $i = t-m$ to $t-1$
   * $\beta = \rho_i \bar{v}_i^T \bar{z}$
   * $\bar{z} \leftarrow \bar{z} + \bar{q}_i (\alpha_i - \beta)$

4. Return $-\bar{z}$

Complexity of BFGS is $O(d^2)$ while that of L-BFGS is $O(md)$, and usually $m \ll d$.

back to subgradients...



candidate subgradients (includes slope-0 line)

Recall the example of $J(w) = |w|$, which is not differentiable at $w=0$. The slope for $w>0$ is $+1$, while it is -1 for $w<0$. Any line with a slope in $[-1,1]$ can be a candidate tangent to $J(w)=|w|$ at $w=0$, including a slope of 0.

We now define subgradients for convex J.

Def A subgradient at $\bar{w}_0$ of a convex loss function $J(\bar{w})$ is a vector $\bar{v} \in \mathbb{R}^d$ such that

$$J(\bar{w}) \geq J(\bar{w}_0) + \bar{v}(\bar{w} - \bar{w}_0) \quad \forall\, \bar{w} \in \mathbb{R}^d.$$

We get the following property from the definition directly:

Prop If $\bar{v}_1, \bar{v}_2$ are subgradients of $J(\bar{w})$ at $\bar{w}_0$, then $\lambda \bar{v}_1 + (1-\lambda)\bar{v}_2$ is also a subgradient of $J(\bar{w})$ for $\lambda \in (0,1)$.

The presence of $\bar{v} = \bar{0}$ as a subgradient is equivalent to the optimality condition for convex loss function J, as we get $J(\bar{w}) \geq J(\bar{w}_0)$ from the definition condition.

Properties of subgradients

* If $J(\bar{w})$ is differentiable at $\bar{w}_0$, then $\nabla J(\bar{w}_0)$ is its unique subgradient at $\bar{w}_0$.

* If $J(\bar{w})$ is convex, then $\bar{0} \in \{\text{subgradients of } J(\bar{w}) \text{ at } \bar{w}_0\}$
  $\Rightarrow$ $\bar{w}_0$ is optimal.

* Let $\bar{v}_i$ be a subgradient of $J_i(\bar{w})$ at $\bar{w}_0$ for $i=1,2$.
  $\Rightarrow$ $\bar{v}_1 + \bar{v}_2$ is a subgradient of $(J_1 + J_2)(\bar{w})$.
  ↳ This result is quite useful in ML – many ML problems have additively separable loss functions.

We had used subgradients implicitly for the hinge loss SVM.

From Lecture 8:

$$\nabla J_{\text{H-SVM}} = -y_i \bar{x}_i \, \delta(1 - y_i(\bar{w}^T \bar{x}_i) > 0) + \lambda \bar{w} \qquad \text{(Hinge-SVM)}$$

↳ indicator function

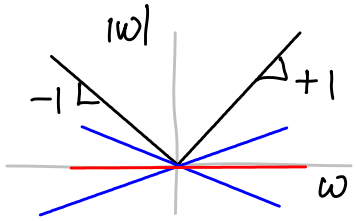
We were indirectly setting the subgradient to zero when the indicator function gives zero. Recall that the hinge loss function is not smooth.

We will describe the details of subgradients for regression with L_1-regularizer.

Application: Regression with L_1-regularization

$$J = \frac{1}{2}\|D\bar{w} - \bar{y}\|^2 + \lambda \|\bar{w}\|_1$$

$$= \frac{1}{2}\|D\bar{w} - \bar{y}\|^2 + \lambda \sum_{j=1}^{d} |w_j|$$



→ gradient descent

GD Update

$$\bar{w} \leftarrow \bar{w} - \alpha D^T(D\bar{w} - \bar{y}) - \alpha\lambda \bar{s}$$

→ subgradient

where $s_j = \begin{cases} +1, & w_j > 0 \\ -1, & w_j < 0 \\ \text{sample from } [-1,1], & w_j = 0 \end{cases}$.

This random choice could make J worse. So, we keep track of $\bar{w}_{best}$ based on the lowest $J(\bar{w})$ attained so far. If $\bar{w}_t$ is such that

$$J(\bar{w}_t) < J(\bar{w}_{best}), \text{ then}$$

$$\bar{w}_{best} \leftarrow \bar{w}_t.$$

Return $\bar{w}_{best}$ at the end.

→ Stop based on convergence criteria (as usual)

→ return $\bar{w}_{best}$ and not $\bar{w}_s$ when we stop after s iterations.

We could also choose $s_j = 0$ when $w_j = 0$.