MODERN OPTIMIZATION TECHNIQUES - GROUP 01

SECOND TAKE HOME EXAM

SUBMITTED BY: MUHAMMAD INAAM ASHRAF

MATRIKEL NR: 307524

Q2A: Coordinate Descent vs others

a. In Coordinate Descent, we take one coordinate i.e. feature at a time and move ~~toward~~ along this coordinate for a specific iteration. Then, in the next iteration, we chose another coordinate and move along its direction and so on. We can either randomly select coordinate for each iteration or we can loop through all coordinates sequentially to complete the cycle. Please note that we use the partial gradient with respect to the ~~gr~~ coordinate at hand for the update, as we keep all other coordinates constant.

In Gradient Descent, we use the complete derivative ~~and iterate~~ including all coordinates and move along the direction ~~of~~ towards the minimum. We use all samples for every iteration.

02)

M. Inaam Ashraf (307524)

In Stochastic Gradient Descent, we randomly select one sample or a mini-batch for each iteration. Rest is the same as GD.

Differences.

CD	GD	SGD
Choose one coordinate at a time	Use all coordinates every time.	Use all coordinates every time.
No step size is required	Use all samples every time	Use one sample or a mini-batch.
	Step size is required	Step size is required.

b)

Advantages	Disadvantages
No step size is required.	GD and SGD are better suited when we don't have many features in the data.
It converges faster as it switches coordinates at every iteration.	GD and SGD are better when it is not possible to compute partial derivative for each coordinate for CD.
Advantageous to use when we have many features	

M. Inaam Ashraf (307524)

2B):
$$\mathcal{L} = \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & -1 \end{bmatrix}, \quad Y = \begin{bmatrix} 7 \\ 9 \\ 5 \end{bmatrix}$$

a) We can write $\|Y - X\beta\|_2^2 \propto \beta^T X^T X\beta - 2y^T X\beta$

So,
$$\mathcal{L} = \frac{1}{2}\left(\beta^T X^T X\beta - 2y^T X\beta\right) + \lambda\sum_{m=1}^{M}|\beta_m|$$

$$= \frac{1}{2}\Big(x_m^T x_m \beta_m^2 + 2\beta_{-m}^T X_{-m}^T x_m\beta_m + \beta_{-m}^T X_{-m}^T X_{-m}\beta_{-m}$$
$$-2y^T x_m\beta_m - 2y^T X_{-m}\beta_{-m}\Big) + \lambda\sum_{m=1}^{M}|\beta_m|$$

$$\mathcal{L} = \frac{1}{2}\left(x_m^T x_m\beta_m^2 - 2\left(y - X_{-m}\beta_{-m}\right)^T x_m\beta_m\right) + \lambda\sum_{m=1}^{M}|\beta_m|$$

Taking derivative and putting it equal to 0.

$$\frac{d\mathcal{L}}{d\beta_m} = \frac{1}{2}\left(2x_m^T x_m\beta_m - 2\left(y - X_{-m}\beta_{-m}\right)^T x_m\right) + \frac{\partial}{\partial\beta_m}\lambda\sum_{m=1}^{M}|\beta_m|$$

Now,
$$\frac{\partial}{\partial\beta_m}\lambda\sum_{m=1}^{M}|\beta_m| = \begin{cases} -\lambda & \beta_m < 0 \\ \lambda & \beta_m > 0 \\ [-\lambda, \lambda] & \beta_m = 0 \end{cases}$$

$\rightarrow$ subdifferential.

M. Inaam Ashraf (307524)

so we can write.

$$\frac{\partial \mathcal{L}}{\partial \beta_m} \overset{!}{=} 0 \rightsquigarrow \beta_m \overset{\text{naive}}{=} \begin{cases} \dfrac{(y - X_{-m}\beta_{-m})^T . x_m - \lambda}{x_m^T x_m}, & \beta_m > 0 \\ \dfrac{(y - X_{-m}\beta_{-m})^T . x_m + \lambda}{x^T x_m}, & \beta_m < 0 \end{cases}$$

using soft thresholding.

$$\beta_m = \text{soft}\left(\frac{(y - X_{-m}\beta_{-m})^T . x_m}{x_m^T x_m}, \frac{\lambda}{x_m^T x_m} \right)$$

b) & c) Epoch 1: $\beta^0 = (2, 1)^T$, $\lambda = 0.1$

Iteration 1:

$$\beta_0^1 = \text{soft}\left(\frac{\left(\begin{bmatrix} 7 \\ 9 \\ 5 \end{bmatrix} - \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix} [1] \right)^T . \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}{[1 \quad 1 \quad 1] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}, \frac{0.1}{[1 \; 1 \; 1] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}} \right)$$

$$\beta_0^1 = \text{soft}\left(\frac{[5 \quad 6 \quad 6] . \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}{3}, \frac{0.1}{3} \right) = \text{soft}(5.67, 0.033)$$

$$\beta_0^1 = 5.63$$

$$\mathcal{L}_{RMSE} = \frac{(Y - X\beta)^T . (Y - X\beta)}{2} + \lambda \sum_{m \neq 0} |\beta|$$

M. Inaam Ashraf (307524)

$$\mathcal{L}_{RMSE} = Y - X\beta = \begin{bmatrix} 7 \\ 9 \\ 5 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 5.63 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 7 \\ 9 \\ 5 \end{bmatrix} - \begin{bmatrix} 7.63 \\ 8.63 \\ 4.63 \end{bmatrix} = \begin{bmatrix} -0.63 \\ 0.37 \\ 0.37 \end{bmatrix}$$

$$\mathcal{L}_{RMSE} = \frac{1}{2} \begin{bmatrix} -0.63 & 0.37 & 0.37 \end{bmatrix} \begin{bmatrix} -0.63 \\ 0.37 \\ 0.37 \end{bmatrix} + 0.1 (5.63 + 1)$$

$$= 0.335 + 0.663 = 0.998.$$

Iteration 2:

$$\beta_1^{1} = \text{soft}\left(\frac{\left(\begin{bmatrix} 7 \\ 9 \\ 5 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \{5.63\} \right)^T \cdot \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}}{\begin{bmatrix} 2 & 3 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}}, \frac{0.1}{\begin{bmatrix} 2 & 3 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}} \right)$$

$$= \text{soft}\left(\frac{13.47}{14}, \frac{0.1}{14} \right) = 0.955$$

$$Y - X\beta = \begin{bmatrix} 7 \\ 9 \\ 5 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 5.63 \\ 0.955 \end{bmatrix} = \begin{bmatrix} -0.54 \\ 0.5 \\ 0.32 \end{bmatrix}$$

$$\mathcal{L}_{RMSE} = \frac{1}{2} \begin{bmatrix} -0.54 & 0.5 & 0.32 \end{bmatrix} \begin{bmatrix} -0.54 \\ 0.5 \\ 0.32 \end{bmatrix} + 0.1 (5.63 + 0.955)$$

$$\mathcal{L}_{RMSE} = 0.98$$

M. Inaam Ashraf (307524)

Epoch 2: $\beta^1 = (5.63, 0.955)^T$

Iteration 3:

$$\beta_0^2 = \text{soft}\left(\frac{\left(\begin{bmatrix}7\\9\\5\end{bmatrix} - \begin{bmatrix}2\\3\\-1\end{bmatrix}[0.955]\right)^T \cdot \begin{bmatrix}1\\1\\1\end{bmatrix}}{3}, \frac{0.1}{3}\right)$$

$$= \text{Soft}\left(\frac{17.18}{3}, \frac{0.1}{3}\right) = 5.69.$$

$$Y - X\beta = \begin{bmatrix}7\\9\\5\end{bmatrix} - \begin{bmatrix}1 & 2\\1 & 3\\1 & -1\end{bmatrix}\begin{bmatrix}5.69\\0.955\end{bmatrix} = \begin{bmatrix}-0.6\\0.44\\0.26\end{bmatrix}$$

$$\ell_{RMSE} = \frac{1}{2}\begin{bmatrix}-0.6 & 0.44 & 0.26\end{bmatrix}\begin{bmatrix}-0.6\\0.44\\0.26\end{bmatrix} + 0.1(5.69 + 0.955)$$

$$= 0.978$$

Iteration 4

$$\beta_1^2 = \text{Soft}\left(\frac{\left(\begin{bmatrix}7\\9\\5\end{bmatrix} - \begin{bmatrix}1\\1\\1\end{bmatrix}[5.69]\right)^T \cdot \begin{bmatrix}2\\3\\-1\end{bmatrix}}{14}, \frac{0.1}{14}\right)$$

$$= \text{soft}\left(\frac{13.22}{14}, \frac{0.1}{14}\right) = 0.937$$

$$Y - X\beta = \begin{bmatrix}7\\9\\5\end{bmatrix} - \begin{bmatrix}1 & 2\\1 & 3\\1 & -1\end{bmatrix}\begin{bmatrix}5.69\\0.937\end{bmatrix} = \begin{bmatrix}-0.57\\0.49\\0.24\end{bmatrix}$$

$$\ell_{RMSE} = \frac{1}{2}\begin{bmatrix}-0.57 & 0.49 & 0.24\end{bmatrix}\begin{bmatrix}-0.57\\0.49\\0.24\end{bmatrix} + 0.1(5.69 + 0.937)$$

$$= 0.976.$$

M. Inaam Ashraf (307524) 07

7.1 -0.5

Loss is decreasing as

$0.998 > 0.98 > 0.978 > 0.976$

But decrease is getting slower which can mean that we are converging.

2C: a) If we converged at β^*, this means that for $\forall\, i \in R$, β^* minimized the function while keeping other variables constant.

From the lecture slides.

$$\exists\, s \in \partial(\lambda \|\beta\|_1), \quad \frac{\partial}{\partial \beta^*}\left(\frac{1}{2}\|Y - X\beta\|_2^2\right) + s = 0$$

where s is the subdifferential given by

$$s_i = \begin{cases} -\lambda & \beta^* < 0 \\ \lambda & \beta^* > 0 \\ [-\lambda, \lambda] & \beta^* = 0 \end{cases}$$

This mean that

$$0 \in \frac{\partial}{\partial \beta^*}\left(\frac{1}{2}\|Y - X\beta\|_2^2\right) + \partial(\lambda \|\beta\|_1)$$

meaning that β^* has minimized our objective function.

M. Inaam Ashraf (307524)

2C (b) $g(x) = |x_1 x_2| + 0.1(x_1+x_2)$

$$g(x) = |x_1|.|x_2| + 0.1(x_1+x_2)$$

$$= \begin{cases} -x_1|x_2| + 0.1(x_1+x_2), & x_1 < 0 \\ x_1|x_2| + 0.1(x_1+x_2), & x_1 \geq 0 \end{cases}$$

I will use a starting value of $x^0 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Now,

$$\frac{\partial g(x_0)}{\partial x_1} = \begin{cases} -|x_2| + 0.1, & x_1 < 0 \\ [-|x_2|+0.1, \ |x_2|+0.1], & x_1 = 0 \\ |x_2| + 0.1, & x_1 > 0 \end{cases}$$

$$= \left.\begin{cases} 0.1, & x_1 < 0 \\ [0.1, 0.1], & x_1 = 0 \\ 0.1 & x_2 > 0 \end{cases}\right\} \text{we are stuck.}$$

The same will happen if we try $\frac{\partial g(x)}{\partial x_2}$

Therefore, we get stuck after one step. ~~Hence,~~

M. Inaam Ashraf (307524)

09.

Q1. 1A: $f(x) = \ln(x)$

Computing gradient. $\nabla f(x) = \frac{\partial \ln(x)}{\partial x} = \frac{1}{x}$

Computing Hessian: $\nabla^2 f(x) = \frac{\partial}{\partial x}\left(\frac{1}{x}\right) = \frac{\partial}{\partial x}\left(x^{-1}\right)$

$$= -x^{-2}.(1) = -\frac{1}{x^2}$$

Computing Newton Update Step.

$$\Delta_{x,y} = -\nabla^2 f(x)^{-1}.\nabla f(x) \quad , \quad \nabla^2 f(x)^{-1} = -x^2$$

$$= -(-x^2)\left(\frac{1}{x}\right) = x.$$

9.1

Hence,

$$x^{t+1} = x^t + \mu^t (x^t)$$

−1.5

As can be seen, ~~it will never converge, as we are increasing the value, this is a maximization problem and we~~ it will converge slowly. even if it did. e.g. for $x^t = 0.1$ & $\mu^t = 0.01$

$$x^{t+1} = 0.1 + 0.01(0.1) = 0.101$$

whereas for Gradient ~~Ascent~~ Descent

$$x^{t+1} = x^t - \mu^t\left(\frac{1}{x^t}\right) = 0.1 - 0.01\left(\frac{1}{0.1}\right)$$

$$= 0.0$$

Hence, we will converge faster with Gradient Descent compared to Newton Method.

M. Inaam Ashraf (307524)

Q.2B. Adding Bias $X = \begin{bmatrix} 1 & 5 & 2 \\ 1 & 3 & -1 \\ 1 & 4 & -1 \end{bmatrix}$, $Y = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$

a. Cross entropy loss is given by.

$$\mathcal{L} = -\left(Y \log(\hat{Y}) + (1-Y)\log(1-\hat{Y})\right)$$

$$= -\left(Y \log(\sigma(X\beta)) + (1-Y)\log(1-\sigma(X\beta))\right)$$

wher $\sigma()$ is the sigmoid function. and its

derivative is

$$\nabla_\beta(\sigma(X\beta)) = \nabla_\beta\left(\frac{1}{1+e^{-X\beta}}\right)$$

$$\nabla_\beta(\sigma(X\beta)) = -\frac{\nabla_\beta(1+e^{-X\beta})}{(1+e^{-X\beta})^2} = \frac{e^{-X\beta}}{(1+e^{-X\beta})^2}$$

$$= \frac{1-1+e^{-X\beta}}{(1+e^{-X\beta})^2} = \frac{1}{1+e^{-X\beta}}\left(1-\frac{1}{1+e^{-X\beta}}\right)$$

$$\nabla_\beta(\sigma(X\beta)) = \sigma(X\beta)(1-\sigma(X\beta))$$

Computing First Derivative..

$$\nabla_\beta \mathcal{L} = -\left(\frac{Y}{\sigma(X\beta)}\nabla_\beta(\sigma(X\beta))\nabla_\beta(X\beta) + \frac{1-Y}{1-\sigma(X\beta)}\nabla_\beta(1-\sigma(X\beta)).\ \nabla_\beta(1-X\beta)\right)$$

$$\nabla_\beta \mathcal{L} = -\left(\frac{Y}{\sigma(X\beta)}\sigma(X\beta)(1-\sigma(X\beta))X + \frac{1-Y}{1-\sigma(X\beta)}(-\sigma(X\beta))(1-\sigma(X\beta))X\right)$$

$$\nabla_\beta \mathcal{L} = -\left(Y(1-\sigma(X\beta))X + (1-Y)(-\sigma(X\beta))X\right)$$

$$= -\left(YX + Y\sigma(X\beta)X - \sigma(X\beta)X + Y\sigma(X\beta)X\right)$$

$$= -(YX - \sigma(X\beta)X)$$

$$= - (Y-\hat{Y})X \text{ or } -X^T(Y-\hat{Y})$$

$$=$$

Computing Hessian

$$\nabla^2_\beta \mathcal{L} = \nabla_\beta\left(-(YX - \sigma(X\beta)X)\right)$$

$$= 0 - X^T.\sigma(X\beta)(1-\sigma(X\beta))X$$

$$\nabla^2_\beta \mathcal{L} = X^T.\hat{Y}(1-\hat{Y})X$$

Newton Step.

$$\beta^{t+1} = \beta^t - \mu^t \nabla^2\mathcal{L}^{-1}\nabla\mathcal{L}$$

$$= \beta^t + \mu^t (X^T.\hat{Y}(1-\hat{Y}).X)^{-1} X^T(Y-\hat{Y})$$

Q1B. b&c $\quad \beta^0 = (1,1,1)^T, \quad \mu = 0.1$

Iteration 1:

$$\hat{Y} = \sigma(X\beta^0) = \sigma\left(\begin{bmatrix} 1 & 5 & 2 \\ 1 & 3 & -1 \\ 1 & 4 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\right) = \sigma\left(\begin{bmatrix} 8 \\ 3 \\ 4 \end{bmatrix}\right)$$

M. Inaam Ashraf (307524)

Since $\sigma(X\beta) = \dfrac{1}{1+e^{-X\beta}}$

$$\hat{y} = \sigma(X\beta) = \begin{bmatrix} 0.999 \\ 0.95 \\ 0.98 \end{bmatrix}$$

Next,

$$W = \operatorname{diag}(\hat{y}(1-\hat{y})) = \begin{bmatrix} 0.0003 & 0 & 0 \\ 0 & 0.045 & 0 \\ 0 & 0 & 0.0177 \end{bmatrix}$$

$$X^T \hat{y}(1-\hat{y}) \cdot X = X^T W X$$

$$= \begin{bmatrix} 0.063 & 0.208 & -0.062 \\ 0.208 & 0.697 & -0.203 \\ -0.062 & -0.203 & 0.064 \end{bmatrix}$$

$$(X^T W X)^{-1} = \begin{bmatrix} 1508.27 & -303.5 & 501.8 \\ -303.5 & 78.7 & -45.1 \\ 501.8 & -45.1 & 359.06 \end{bmatrix}$$

$$\beta^1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + 0.1 \begin{bmatrix} 1508.27 & -303.5 & 501.8 \\ -303.5 & 78.7 & -45.1 \\ 501.8 & -45.1 & 359.06 \end{bmatrix} X^T(y-\hat{y})$$

$$= \text{''} + \text{''} \quad \text{''} \begin{bmatrix} -0.93 \\ -3.78 \\ 0.935 \end{bmatrix}$$

$$\beta^1 = \begin{bmatrix} 21.87 \\ -4.66 \\ 4.77 \end{bmatrix}$$

M. Inaam Ashraf (307524)

Q1B:

$$\mathcal{L} = -\left(Y^T \cdot \log(\hat{Y}) + (1-Y)^T \cdot \log(1-\hat{Y})\right) \quad \hat{Y} = X\beta^1$$

$$= -\left(\begin{bmatrix} -0.0003 \\ -0.0438 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 2.75 \end{bmatrix}\right)$$

$$\mathcal{L}^1 = -2.706$$

Iteration 2

$$\hat{Y} = \sigma(X\beta^1) = \sigma\left(\begin{bmatrix} 1 & 5 & 2 \\ 1 & 3 & -1 \\ 1 & 4 & -1 \end{bmatrix}\begin{bmatrix} 21.87 \\ -4.66 \\ 4.77 \end{bmatrix}\right)$$

$$\hat{Y} = \begin{bmatrix} 0.9997 \\ 0.957 \\ 0.1737 \end{bmatrix}$$

$$W = \mathrm{diag}(\hat{Y}(1-\hat{Y})) = \begin{bmatrix} 0.0003 & 0 & 0 \\ 0 & 0.041 & 0 \\ 0 & 0 & 0.143 \end{bmatrix}$$

$$X^T W X = \begin{bmatrix} 1 & 1 & 1 \\ 5 & 3 & 4 \\ 2 & -1 & -1 \end{bmatrix}\begin{bmatrix} 0.0003 & 0 & 0 \\ 0 & 0.041 & 0 \\ 0 & 0 & 0.143 \end{bmatrix}\begin{bmatrix} 1 & 5 & 2 \\ 1 & 3 & -1 \\ 1 & 4 & -1 \end{bmatrix}$$

$$X^T W X = \begin{bmatrix} 0.185 & 0.7 & -0.184 \\ 0.7 & 2.67 & -0.7 \\ -0.184 & -0.7 & 0.186 \end{bmatrix}$$

$$(X^T W X)^{-1} = \begin{bmatrix} 917.27 & -131.08 & 418.5 \\ -131.08 & 31.3 & -12.76 \\ 418.5 & -12.76 & 372.09 \end{bmatrix}$$

M. Inaam Ashraf (307524)

$$X^T \cdot (Y - \hat{Y}) = \begin{bmatrix} -0.13 \\ -0.564 \\ 0.131 \end{bmatrix}$$

$$\beta^2 = \beta^1 + \mu \, (X^T W X)^{-1} \, X^T \cdot (Y - \hat{Y})$$

$$= \begin{bmatrix} 21.87 \\ -4.66 \\ 4.77 \end{bmatrix} + 0.1 \begin{pmatrix} 917.27 & -131.08 & 418.5 \\ -131.08 & 31.32 & -12.76 \\ 418.5 & -12.76 & 372.09 \end{pmatrix} \begin{bmatrix} -0.13 \\ -0.564 \\ 0.131 \end{bmatrix}$$

$$\beta^2 = \begin{bmatrix} 22.8 \\ -4.89 \\ 4.92 \end{bmatrix}$$

$$\hat{Y} = X\beta^{(2)} = \begin{bmatrix} 0.9997 \\ 0.9611 \\ 0.157 \end{bmatrix}$$

$$\ell^2 = -\left(Y^T . \log(\hat{Y}) + (1 - Y)^T . \log(1 - \hat{Y}) \right)$$

$$= -\left(\begin{bmatrix} -0.00027 \\ -0.0395 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 2.85 \end{bmatrix} \right)$$

$$\ell^2 = -2.81$$

Loss is decreasing on we are getting closer to the prediction.

$$\hat{Y} = \begin{bmatrix} 0.987 \\ 0.957 \\ 0.1737 \end{bmatrix} \Rightarrow \begin{bmatrix} 0.9997 \\ 0.9611 \\ 0.157 \end{bmatrix}$$ is closer to classes. $$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = Y$$

M.Inaam Ashraf (307524). 15

Q1C. a) Quasi- Newton methods approximate the Hessian computation of the Newton Method in order to reduce the computational ~~co~~ cost. Here, a low-rank update is used. There criteria for the Hessian approximation is that these have some properties of the Hessian. The most important criteria is that the approximation fulfills the secant condition.

$$H(y-x) = \nabla f(y) - \nabla f(x)$$

approximately,

$$\nabla^2 f(x)\,(y-x) \approx \nabla f(y) - \nabla f(x) \text{ for } y \approx x$$

We can use the symmetric low rank update when there exists exactly one low-rank update such that

15.1 −0,5

i) Approximation (H) fulfills the secant condition

ii) H is symmetric

iii) and it is a rank-one update.

M. Inaam Ashraf (307524)

Q1C. b). The BFGS update is given by

$$H^{next} = H - \frac{Hs(Hs)^T}{s^T Hs} + \frac{gg^T}{g^T s}$$

It fulfills the secant condition, yields symmetric H and yields positive definite H, if $g^T s > 0$

BFGS requires N^2 storage to ~~materialize~~ & compute the inverse of the approximation of the Hessian. Materialization of ~~A~~ inverse A of the approximation of Hessian can be reduced by using recursive approach. But, still this approach require 2KN storage and is only an improvement if $K \ll N$, where K is the number of recursions.

In Limited Memory BFGS (L-BFGS), we can forget the older vectors (g, s) needed for computation of A. We only store and compute M≪N most recent ones of those vectors, which reduces the computation cost.

Index der Kommentare

7.1 incorrect loss

9.1 both GD and newton diverges here

15.1 symmetric rank one update is used when a positive definite hessian is not needed