Optimization for Data Science, FS22 (Bernd Gärtner and Niao He)
Graded Assignment 3

Student Name, Legi Number: Deepak Narayanan, 21-959-622

Residual Linear Neural Networks

Assignment 1

Need to show

$$f(A) = \mathbb{E}_{x,\xi}[\|y - \hat{y}\|]^2 = \|(\hat{R} - R)\Sigma^{1/2}\|_F^2 + \mathbb{E}_{\xi}[\|\xi\|^2]$$

We have,

$$\mathbb{E}_{x,\xi}[\|y - \hat{y}\|]^2 = \mathbb{E}_{x,\xi}[\|y\|^2 + \|\hat{y}^2\|^2 - 2\langle y, \hat{y}\rangle] = \mathbb{E}_{x,\xi}[\|y\|^2] + \mathbb{E}_{x,\xi}[\|\hat{y}^2\|^2] - 2\mathbb{E}_{x,\xi}[\langle y, \hat{y}\rangle]$$

by expanding the squares and linearity of expectation. We have,

$$y = Rx + \xi, \hat{y} = \hat{R}x$$

Using this, we have,

$$\mathbb{E}_{x,\xi}[\|y\|^2] = \mathbb{E}_{x,\xi}[y^T y] = \mathbb{E}_{x,\xi}[(Rx + \xi)^T(Rx + \xi)] = \mathbb{E}_{x,\xi}[(x^T R^T + \xi^T)(Rx + \xi)] = \mathbb{E}_{x,\xi}[x^T R^T Rx + x^T R^T \xi + \xi^T Rx + \xi^T \xi] = \mathbb{E}_{x,\xi}[x^T R^T Rx] + \mathbb{E}_{x,\xi}[x^T R^T \xi] + \mathbb{E}_{x,\xi}[\xi^T Rx] + \mathbb{E}_{x,\xi}[\xi^T \xi]$$

where we again used linearity of expectation and expanded the inner product. Since x, ξ are independent, we have,

$$\mathbb{E}_{x,\xi}[\xi^T Rx] = \mathbb{E}_{x,\xi}[x^T R^T \xi] = \mathbb{E}_x[x^T R^T]\mathbb{E}_{\xi}[\xi] = 0$$

where we use the fact that transpose of a scalar is the scalar itself and use the fact that $\mathbb{E}_{\xi}[\xi] = 0$. Therefore we have,

$$\mathbb{E}_{x,\xi}[\|y\|^2] = \mathbb{E}_{x,\xi}[x^T R^T Rx] + \mathbb{E}_{x,\xi}[\xi^T \xi] = \mathbb{E}_x[x^T R^T Rx] + \mathbb{E}_{\xi}[\xi^T \xi]$$

Now trace of a scalar is the scalar itself. Using linearity of expectation and cyclicity of trace (Equation 16 in Matrix Cookbook), we have,

$$Tr(\mathbb{E}_x[x^T R^T Rx]) = \mathbb{E}_x[Tr(x^T R^T Rx)] = \mathbb{E}_x[Tr(xx^T R^T R)] = Tr(\mathbb{E}_x[xx^T R^T R]) = Tr(\mathbb{E}_x[xx^T]R^T R) = Tr(\Sigma R^T R)$$

Therefore,

$$\mathbb{E}_{x,\xi}[\|y\|^2] = Tr(\Sigma R^T R) + \mathbb{E}_{\xi}[\xi^T \xi]$$

Now similarly, for $\hat{y}$, we have, using trace of scalar is the scalar itself and also cyclicity of trace,

$$\mathbb{E}_{x,\xi}[\|\hat{y}\|^2] = \mathbb{E}_{x,\xi}[x^T\hat{R}^T\hat{R}x] = \mathbb{E}_{x,\xi}[Tr(x^T\hat{R}^T\hat{R}x)] = \mathbb{E}_{x,\xi}[Tr(xx^T\hat{R}^T\hat{R})] = Tr(\mathbb{E}_{x,\xi}[xx^T]\hat{R}^T\hat{R}]) = Tr(\Sigma\hat{R}^T\hat{R})$$

Now for the inner product, we have

$$\mathbb{E}_{x,\xi}[\langle y, \hat{y}\rangle] = \mathbb{E}_{x,\xi}[(Rx+\xi)^T(\hat{R}x)] = \mathbb{E}_{x,\xi}[x^TR^T\hat{R}x + \xi^T\hat{R}x] = \mathbb{E}_x[x^TR^T\hat{R}x] + \mathbb{E}_{x,\xi}[\xi^T\hat{R}x] = \mathbb{E}_x[x^TR^T\hat{R}x]$$

where we used $\mathbb{E}_\xi[\xi] = 0$ and independence in the random variables. Again using cyclicity of trace and trace of scalar is scalar itself, we have

$$\mathbb{E}_x[x^TR^T\hat{R}x] = \mathbb{E}_x[Tr(x^TR^T\hat{R}x)] = \mathbb{E}_x[Tr(xx^TR^T\hat{R})] = Tr(\mathbb{E}_x[xx^T]R^T\hat{R})) = Tr(\Sigma R^T\hat{R})$$

Therefore we have,

$$f(A) = Tr(\Sigma R^TR) + \mathbb{E}_\xi[\xi^T\xi] + Tr(\Sigma\hat{R}^T\hat{R}) - 2Tr(\Sigma R^T\hat{R})$$

We then have,using square root of sigma and cyclicity of trace,

$$\begin{aligned} Tr(\Sigma R^TR) + Tr(\Sigma\hat{R}^T\hat{R}) - 2Tr(\Sigma R^T\hat{R}) = \\ Tr(\Sigma^{1/2}\Sigma^{1/2}R^TR) + Tr(\Sigma^{1/2}\Sigma^{1/2}\hat{R}^T\hat{R}) - 2Tr(\Sigma^{1/2}\Sigma^{1/2}R^T\hat{R}) = \\ Tr(\Sigma^{1/2}R^TR\Sigma^{1/2}) + Tr(\Sigma^{1/2}\hat{R}^T\hat{R}\Sigma^{1/2}) - 2Tr(\Sigma^{1/2}R^T\hat{R}\Sigma^{1/2}) \end{aligned}$$

Since Σ is symmetric, we have

$$\Sigma = \Sigma^T = (\Sigma^{1/2}\Sigma^{1/2})^T = \Sigma^{1/2^T}\Sigma^{1/2^T} = \Sigma^{1/2}\Sigma^{1/2}$$

By uniqueness $\Sigma^{1/2^T} = \Sigma^{1/2}$ Therefore we have,

$$\begin{aligned} Tr(\Sigma^{1/2}R^TR\Sigma^{1/2}) + Tr(\Sigma^{1/2}\hat{R}^T\hat{R}\Sigma^{1/2}) - 2Tr(\Sigma^{1/2}R^T\hat{R}\Sigma^{1/2}) = \\ Tr(\Sigma^{1/2}R^TR\Sigma^{1/2}) + Tr(\Sigma^{1/2}\hat{R}^T\hat{R}\Sigma^{1/2}) - Tr(\Sigma^{1/2}R^T\hat{R}\Sigma^{1/2}) - Tr(\Sigma^{1/2}\hat{R}^TR\Sigma^{1/2}) = \\ Tr(\Sigma^{1/2}R^T(R-\hat{R})\Sigma^{1/2}) + Tr(\Sigma^{1/2}\hat{R}^T(\hat{R}-R)\Sigma^{1/2})) = \\ Tr(\Sigma^{1/2}(R^T-\hat{R}^T)(R-\hat{R})\Sigma^{1/2}) = \\ Tr(((\hat{R}-R)\Sigma^{1/2})^T(\hat{R}-R)\Sigma^{1/2}) = \|(\hat{R}-R)\Sigma^{1/2}\|_F^2 \end{aligned}$$

Where we have used that $Tr(AB) = Tr((AB)^T) = Tr(B^TA^T)$ and that $\Sigma^{1/2^T} = \Sigma^{1/2}$ in the second equation above. Further we have $Tr(A^TA) = \|A\|_F^2$ Therefore, we finally have,

$$f(A) = \|(\hat{R}-R)\Sigma^{1/2}\|_F^2 + \mathbb{E}_\xi[\xi^T\xi] = \|(\hat{R}-R)\Sigma^{1/2}\|_F^2 + \mathbb{E}_\xi[\|\xi\|^2]$$

Assignment 2

From assignment 1, we have,

$$f(A) = Tr(\Sigma R^T R) + Tr(\Sigma \hat{R}^T \hat{R}) - 2Tr(\Sigma R^T \hat{R}) + \mathbb{E}_\xi[\|\xi\|^2]$$

We also have

$$\hat{R} = (I + A_l)(I + A_{l-1}) \dots (I + A_i)(I + A_{i-1}) \dots (I + A_2)(I + A_1)$$

We can group the terms containing $A_l, A_l - 1 \dots A_{i+1}$, and call it a matrix B, the terms containing $A_{i-1}, A_{i-2} \dots A_1$ and call it a matrix C. Therefore,

$$\hat{R} = B(I + A_i)C$$

Clearly the derivatives with respect to A_j are only concerned with the terms containing $\hat{R}$ in $f(A)$. Therefore, we are interested in the following:

$$\frac{\partial f}{\partial A_i} = \frac{\partial Tr(\Sigma \hat{R}^T \hat{R})}{\partial A_i} - 2\frac{\partial Tr(\Sigma R^T \hat{R})}{\partial A_i}$$

By cyclicity of trace, we have,

$$\frac{\partial f}{\partial A_i} = \frac{\partial Tr(\Sigma \hat{R}^T \hat{R})}{\partial A_i} - 2\frac{\partial Tr(R^T \hat{R} \Sigma)}{\partial A_i}$$

Now plugging in the value of $\hat{R}$, we have,

$$\frac{\partial f}{\partial A_i} = \frac{\partial Tr(\Sigma (B(I + A_i)C)^T B(I + A_i)C)}{\partial A_i} - 2\frac{\partial Tr(R^T B(I + A_i)C\Sigma)}{\partial A_i}$$

which is equal to,

$$\frac{\partial f}{\partial A_i} = \frac{\partial Tr(C^T (I + A_i^T) B^T \Sigma B(I + A_i)C)}{\partial A_i} - 2\frac{\partial Tr(R^T B(I + A_i)C\Sigma)}{\partial A_i}$$

Using cyclicity of trace, we have,

$$\frac{\partial f}{\partial A_i} = \frac{\partial Tr((I + A_i^T) B^T \Sigma B(I + A_i)CC^T)}{\partial A_i} - 2\frac{\partial Tr(R^T B(I + A_i)C\Sigma)}{\partial A_i}$$

Further

$$\frac{\partial (I + A_i)}{\partial A_i} = I$$

Recognizing the first term and second term are of the forms

$$\frac{\partial Tr(X^T Y X Z)}{\partial X} = YXZ + Y^T X Z^T, \frac{\partial Tr(YXZ)}{\partial X} = Y^T Z^T$$

and thus using equality 117 and 101 from matrix cookbook, we have,

$$\frac{\partial f}{\partial A_i} = B^T\Sigma B(I+A_i)CC^T + B^T\Sigma B(I+A_i)CC^T - 2B^TR\Sigma C^T = \\ 2(B^T\Sigma B(I+A_i)CC^T - B^TR\Sigma C^T) = 2B^T(\Sigma\hat{R}C^T - R\Sigma C^T) = \\ 2B^T\Sigma(\hat{R}-R)C^T$$

where we use that $B(I+A_i)C = \hat{R}$. Now plugging in the value of B and C from above, we have,

$$\frac{\partial f}{\partial A_i} = 2((I+A_l)(I+A_{l-1})\dots(I+A_{i+1}))^T\Sigma(\hat{R}-R)((I+A_{i-1})(I+A_{i-2})\dots(I+A_1))^T$$

which is equal to,

$$\frac{\partial f}{\partial A_i} = 2((I+A_{i+1}^T)(I+A_{i+2}^T)\dots(I+A_l^T))\Sigma(\hat{R}-R)((I+A_1^T)(I+A_2^T)\dots(I+A_{i-1}^T))$$

which is the required result.

Assignment 3

From Assignment 2, we have,

$$\frac{\partial f}{\partial A_i} = 2\left(\prod_{j=i+1}^{l}(I+A_j^T)\right)\Sigma(\hat{R}-R)\left(\prod_{k=1}^{i-1}(I+A_k^T)\right)$$

Claim 1:

$$||AB||_F^2 \geq \sigma_{\min}(A)^2\|B\|_F^2$$

where $\sigma_{\min}$ is the smallest singular value of A and A, B are square matrices in $\mathbb{R}^{d\times d}$. Proof: From Full SVD of A, we have, $A = \sum_{i=1}^{d} u_i\sigma_i v_i^T$. Further we have for the B_j being the j^{th} columns of B

$$||AB||_F^2 = \sum_{j=1}^{d}\|AB_j\|^2 = \sum_{j=1}^{d}\left\|\sum_{i=1}^{d}u_i\sigma_i v_i^TB_j\right\|^2$$

If A is not full rank with $r < d$, we can still pad the remaining $d-r$ columns of V in $A = U\Sigma V^T$, with the corresponding columns that make V full rank while also making the corresponding singular values 0 for the $d-r$ columns. Now expressing $B_j = \sum_{k=1}^{d}\alpha_k v_k$ as a vector in the column space of V, we have,

$$\sum_{j=1}^{d}\left\|\sum_{i=1}^{d}u_i\sigma_i v_i^TB_j\right\|^2 = \sum_{j=1}^{d}\left\|\sum_{i=1}^{d}u_i\sigma_i v_i^T\sum_{k=1}^{d}\alpha_k v_k\right\|^2 = \\ \sum_{j=1}^{d}\left\|\sum_{i=1}^{d}u_i\sigma_i\alpha_i\right\|^2 \geq \sigma_{\min}^2\sum_{j=1}^{d}\left\|\sum_{i=1}^{d}u_i\alpha_i\right\|^2 = \sigma_{\min}^2\sum_{j=1}^{d}\|B_j\|^2 = \sigma_{\min}^2\|B\|_F^2$$

where we use the fact that $v_i^T v_j = 1 \iff i = j, 0$ else , and $||\alpha_k u_k||^2 = \alpha_k^2$ and $\sum_{k=1}^{d} \alpha_k^2 = ||B_j||^2$ Now going to our original equation,

$$\frac{\partial f}{\partial A_i} = 2 \left(\prod_{j=i+1}^{l} (I + A_j^T) \right) \Sigma(\hat{R} - R) \left(\prod_{k=1}^{i-1} (I + A_k^T) \right)$$

Using the above fact and observing we have products of matrices and applying the result repeatedly, we have,

$$\left\| \frac{\partial f}{\partial A_i} \right\|_F^2 = 4 \left\| \left(\prod_{j=i+1}^{l} (I + A_j^T) \right) \Sigma(\hat{R} - R) \left(\prod_{k=1}^{i-1} (I + A_k^T) \right) \right\|_F^2 \geq$$
$$4\sigma_{\min}^{2(l-i)} (I + A_j^T) \left\| \Sigma(\hat{R} - R) \left(\prod_{k=1}^{i-1} (I + A_k^T) \right) \right\|_F^2$$

where $\sigma_{\min}(I + A_j^T)$ is minimum singular value over all $I + A_j^T \forall j \in [d]$. Now we know that f^* is the global minimum value of f, so therefore it is attained for some A^*. Therefore consider,

$$f^* = f(A^*) = \|(\hat{R}^* - R)\Sigma^{1/2}\|_F^2 + \mathbb{E}_\xi[\|\xi\|^2]$$

Also we have,

$$f(A) = \|(\hat{R} - R)\Sigma^{1/2}\|_F^2 + \mathbb{E}_\xi[\|\xi\|^2]$$

Consider $f(A) - f^*$. Using this, we have,

$$f(A) - f^* = \|(\hat{R} - R)\Sigma^{1/2}\|_F^2 - \|(\hat{R}^* - R)\Sigma^{1/2}\|_F^2 \iff$$
$$f(A) - f^* + \|(\hat{R}^* - R)\Sigma^{1/2}\|_F^2 = \|(\hat{R} - R)\Sigma^{1/2}\|_F^2 \implies$$
$$\|(\hat{R} - R)\Sigma^{1/2}\|_F^2 \geq f(A) - f^*$$
$$(\because \|(\hat{R}^* - R)\Sigma^{1/2}\|_F^2 \geq 0)$$

Now, we also have that $\|A^T\|_F^2 = \|A\|_F^2$. Now considering,

$$A = \Sigma(\hat{R} - R) \left(\prod_{k=1}^{i-1} (I + A_k^T) \right)$$

We have

$$A^T = \left(\prod_{k=i-1}^{1} (I + A_k) \right) (\Sigma(\hat{R} - R))^T$$

where the product decrements by step of 1. Therefore we have,

$$\left\|\Sigma(\hat{R}-R)\left(\prod_{k=1}^{i-1}(I+A_k^T)\right)\right\|_F^2 = \left\|\left(\prod_{k=i-1}^{1}(I+A_k)\right)(\Sigma(\hat{R}-R))^T\right\|_F^2$$

Again using the derived inequality we thus have,

$$\left\|\frac{\partial f}{\partial A_i}\right\|_F^2 \geq 4\sigma_{\min}^{2(l-i)}(I+A_j^T)\sigma_{\min}^{2(i-1)}(I+A_j)\left\|(\Sigma(\hat{R}-R))^T\right\|_F^2$$

where $\sigma_{\min}(I+A_j)$ is minimum singular value over all $I+A_j$ for $j \in [d]$. Singular values for matrices and their transposes are the same. Therefore, we can combine the two $\sigma_{\min}$ from above to get,

$$\left\|\frac{\partial f}{\partial A_i}\right\|_F^2 \geq 4\sigma_{\min}^{2(l-i)+2i-2}(I+A_j)\left\|(\Sigma(\hat{R}-R))^T\right\|_F^2$$

Using the relation that we derived above, and the fact norm of transpose is same as norm of the original matrix for frobenius norm, we have,

$$\left\|\frac{\partial f}{\partial A_i}\right\|_F^2 \geq 4\sigma_{\min}^{2l-2}(I+A_j)\left\|(\Sigma^{1/2}\Sigma^{1/2}(\hat{R}-R))^T\right\|_F^2 \geq 4\sigma_{\min}^{2l-2}(I+A_j)\sigma_{\min}(\Sigma^{1/2})\left\|\Sigma^{1/2}(\hat{R}-R))\right\|_F^2$$

From our above derivation, this quantity is, further,

$$\left\|\frac{\partial f}{\partial A_i}\right\|_F^2 \geq 4\sigma_{\min}^{2l-2}(I+A_j)\sigma_{\min}(\Sigma^{1/2})(f(A)-f^*)$$

Using SVD for $\Sigma^{1/2} = USV^T$, we have,

$$\Sigma = \Sigma^{1/2}\Sigma^{1/2} = \Sigma^{1/2^T}\Sigma^{1/2} = VSU^TUSV^T = VS^2V^T$$

We can see that the above quantity VS^2V^T is in fact the eigendecomposition of Σ and moreover, the squares of singular values of $\Sigma^{1/2}$ are the eigenvalues of Σ. This implies $\sigma_{\min}^2(\Sigma^{1/2}) = \lambda_{min}(\Sigma)$. Using this, we have,

$$\left\|\frac{\partial f}{\partial A_i}\right\|_F^2 \geq 4\sigma_{\min}^{2l-2}(I+A_j)\lambda_{min}(\Sigma)(f(A)-f^*)$$

We are given that $||A_j|| \leq \tau$. Further using the Courant Fisher Min Max Theorem (`https://en.wikipedia.org/wiki/Min-max_theorem`), we have,

$$\begin{aligned}\sigma_{\min}(I+A_j) &= \max_{S:dim(S)=d}\min_{x\in S,||x||=1}||(I+A_j)x|| = \max_{S:dim(S)=d}\min_{x\in S,||x||=1}||(I-(-A_j))x|| \\ &\geq \max_{S:dim(S)=d}\min_{x\in S,||x||=1}||x|| - ||-A_jx|| = \max_{S:dim(S)=d}\min_{x\in S,||x||=1}||x|| - ||A_jx|| = \\ &\max_{S:dim(S)=d}\min_{x\in S,||x||=1} 1 - ||A_jx||\end{aligned}$$

where we used triangle inequality ($||A-B|| \geq ||A||-||-B|| = ||A||-||B||$). By definition of spectral norm, $\frac{||A_j x||}{||x||} \leq ||A_j|| \leq \tau$. This implies, $\frac{||A_j x||}{||x||} \leq \tau \iff -\frac{||A_j x||}{||x||} \geq -\tau$. Using this, we have,

$$\sigma_{\min}(I + A_j) \geq \max_{S:dim(S)=d} \min_{x \in S, ||x||=1} 1 - ||A_j x|| \geq \max_{S:dim(S)=d} \min_{x \in S, ||x||=1} 1 - \tau = 1 - \tau$$

Using this in our inequality above, we thus have,

$$\left\| \frac{\partial f}{\partial A_i} \right\|_F^2 \geq 4\sigma_{\min}^{2l-2}(I + A_j)\lambda_{min}(\Sigma)(f(A) - f^*) \geq 4(1-\tau)^{2l-2}\lambda_{min}(\Sigma)(f(A) - f^*)$$

Now summing over all $i \in 1, 2, \ldots l$, we thus have,

$$\|\nabla f(A)\|_F^2 = \sum_{i=1}^{\ell} \left\| \frac{\partial f}{\partial A_i} \right\|_F^2 \geq 4\ell(1-\tau)^{2l-2}\lambda_{min}(\Sigma)(f(A) - f^*)$$

which is the required result.

Now if we have a critical point, we have,

$$\|\nabla f(A)\|_F^2 = 0 \geq 4\ell(1-\tau)^{2l-2}\lambda_{min}(\Sigma)(f(A) - f^*)$$

But $(1-\tau)^{2(\ell-1)}$ is a square therefore nonnegative, since $\lambda_{\min}(\Sigma)$ is nonnegative as Σ is a positive semidefinite matrix and all eigenvalues are non-negative, 4 and ℓ are nonnegative as well. This implies,

$$0 \geq f(A) - f^* \iff f(A) \leq f^*$$

However, f^* is the global minimum. This implies, $f^* \leq f(A)$ Combining these two, we must have, $f(A) = f^*$. This implies that every critical point is indeed a global minimum, which is the required result.

Frank Wolfe and Regularization Path

Assignment 4

We first consider the LMO step of the Frank Wolfe Algorithm. That is,

$$LMO(\mathbf{g}) = \mathbf{s} = \arg\min_{\mathbf{z}\in X} \mathbf{g}^T\mathbf{z} = \arg\max_{\mathbf{z}\in X} -\mathbf{g}^T\mathbf{z}$$

Suppose $X = \{x : \|x\|_1 \leq t\}$ This means

$$\mathbf{s} = \arg\max_{||\mathbf{z}||_1\leq t} -\mathbf{g}^T\mathbf{z} = \arg\max_{||\frac{\mathbf{z}}{t}||_1\leq 1} -\mathbf{g}^T\frac{\mathbf{z}}{t}.t = -t(\arg\max_{||\mathbf{y}||_1\leq 1} \mathbf{g}^T\mathbf{y})$$

where we can move the negative sign outside as the function is linear. By definition of dual norm, we have,

$$||\mathbf{g}||_\infty = \sup\{\mathbf{g}^T\mathbf{y} : ||\mathbf{y}||_1 \leq 1\}$$

As we can observe $-\mathbf{s}/t$ is indeed the vector that achieves the infinity norm of $\mathbf{g}$. In particular, when $\mathbf{g} = \nabla f(\mathbf{x})$, $\mathbf{s} = -t.\arg\max_{||\mathbf{y}||_1\leq 1} \nabla f(\mathbf{x})^T\mathbf{y}$. Let us call this vector $\mathbf{y}'$.Therefore $\mathbf{s} = -t\mathbf{y}'$. Further consider the following quantity,

$$g(\mathbf{x}) = \nabla f(\mathbf{x})^T(\mathbf{x} - \mathbf{s}) = \nabla f(\mathbf{x})^T(\mathbf{x} + t\mathbf{y}') = \nabla f(\mathbf{x})^T\mathbf{x} + t\nabla f(\mathbf{x})^T\mathbf{y}'$$

However by definition of dual norm, we have, $\nabla f(\mathbf{x})^T\mathbf{y}' = \|\nabla f(\mathbf{x})\|_\infty$. This is true because to obtain the value of the infinity norm, we can consider a vector that is basically ± 1 at exactly one element and zero everywhere else. Firstly this vector belongs to the L1 ball considered. Secondly, once we have access to the gradient, we can check which element and correspondingly which dimension of the gradient vector has the highest absolute value. Now we set this dimension in our vector $\mathbf{y}$ to be 1 if the highest absolute value is attained by a positive number, -1 if it is attained by a negative number. This way, we can always retrieve the infinity norm, therefore making the supremum attainable thus in fact giving us a maximum. Therefore we have,

$$g(\mathbf{x}) = \nabla f(\mathbf{x})^T\mathbf{x} + t\|\nabla f(\mathbf{x})\|_\infty$$

Now suppose $t \in [t_k, t_{k+1})$. Also let $\mathbf{x} = \tilde{\mathbf{x}}_t$. We then have,

$$g(\tilde{\mathbf{x}}_t) = \nabla f(\tilde{\mathbf{x}}_t)^T\tilde{\mathbf{x}}_t + t\|\nabla f(\tilde{\mathbf{x}}_t)\|_\infty$$

Since $t < t_{k+1} = t_k + \frac{\gamma}{2\|\nabla f(\tilde{\mathbf{x}}_{t_k})\|_\infty}$, we have, $\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t_k}$ as we set the value for this condition.

$$g(\tilde{\mathbf{x}}_t) = \nabla f(\tilde{\mathbf{x}}_t)^T\tilde{\mathbf{x}}_t + t\|\nabla f(\tilde{\mathbf{x}}_t)\|_\infty < f(\tilde{\mathbf{x}}_t)^T\tilde{\mathbf{x}}_t + t_{k+1}\|\nabla f(\tilde{\mathbf{x}}_t)\|_\infty =$$
$$f(\tilde{\mathbf{x}}_t)^T\tilde{\mathbf{x}}_t + \left(t_k + \frac{\gamma}{2\|\nabla f(\tilde{\mathbf{x}}_{t_k})\|_\infty}\right)\|\nabla f(\tilde{\mathbf{x}}_t)\|_\infty = \nabla f(\tilde{\mathbf{x}}_t)^T\tilde{\mathbf{x}}_t + t_k\|\nabla f(\tilde{\mathbf{x}}_t)\|_\infty + \frac{\gamma}{2}$$

Recognize that the first two terms are exactly the same as $g(\tilde{\mathbf{x}}_t)$ when $t = t_k$. We are given that this quantity is upper bounded by $\frac{\gamma}{2}$, since this is the duality gap!. Therefore we have,

$$g(\tilde{\mathbf{x}}_t) < \nabla f(\tilde{\mathbf{x}}_t)^T \tilde{\mathbf{x}}_t + t_k \|\nabla f(\tilde{\mathbf{x}}_t)\|_\infty + \frac{\gamma}{2} \leq \frac{\gamma}{2} + \frac{\gamma}{2} = \gamma$$

But we also know that since s is a minimizer, we have,

$$g(\mathbf{x}) = \nabla f(\mathbf{x})^T (\mathbf{x} - \mathbf{s}) \geq \nabla f(\mathbf{x})^T (\mathbf{x} - \mathbf{x}^*) \geq f(\mathbf{x}) - f(\mathbf{x}^*)$$

where the last inequality is from first order characterization of convexity. Therefore using this in the above derived inequality we have,

$$f(\tilde{\mathbf{x}}_t) - f(\mathbf{x}^*) \leq g(\tilde{\mathbf{x}}_t) \leq \gamma$$

Therefore for $t \in [t_k, t_{k+1})$, which is when $\mathbf{x}_t$ is set, we have,

$$f(\tilde{\mathbf{x}}_t) - f(\mathbf{x}^*) \leq \gamma$$

which is the required result.

Newton's Method with Backtracking

Assignment 5

We know the f is L-smooth. Therefore, we have

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$$

Suppose $y = \mathbf{x}_{t+1}$, $\mathbf{x} = \mathbf{x}_t$. We then have,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq \nabla f(\mathbf{x}_t)^T(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}||\mathbf{x}_{t+1} - \mathbf{x}_t||^2$$

We further know that,

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha\Delta\mathbf{x}_t; \Delta\mathbf{x}_t = -H_t^{-1}\nabla f(\mathbf{x}_t) \iff \nabla f(\mathbf{x}_t) = -H_t\Delta\mathbf{x}_t$$

Using these facts and plugging into the smoothness equation, we have,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -(H_t\Delta\mathbf{x}_t)^T(\alpha\Delta\mathbf{x}_t) + \frac{L}{2}||\alpha\Delta\mathbf{x}_t||^2 = -\alpha\Delta\mathbf{x}_t^T H_t\Delta\mathbf{x}_t + \frac{L}{2}\alpha^2||\Delta\mathbf{x}_t||^2$$

where we use the fact that hessian is symmetric. For our required result, if $\alpha \leq \min\{1, \frac{3\mu}{2L}\}$, we just require that the RHS above is upper bounded by $\alpha\frac{\Delta\mathbf{x}_t^T\nabla f(\mathbf{x}_t)}{4} = \frac{-\alpha\Delta\mathbf{x}_t^T H_t\Delta\mathbf{x}_t}{4}$. We require,

$$\begin{aligned}-\alpha\Delta\mathbf{x}_t^T H_t\Delta\mathbf{x}_t + \frac{L}{2}\alpha^2||\Delta\mathbf{x}_t||^2 \leq \frac{-\alpha\Delta\mathbf{x}_t^T H_t\Delta\mathbf{x}_t}{4} &\iff \frac{L}{2}\alpha^2||\Delta\mathbf{x}_t||^2 \leq \frac{3\alpha\Delta\mathbf{x}_t^T H_t\Delta\mathbf{x}_t}{4} \\ &\iff \alpha\left(\frac{3\Delta\mathbf{x}_t^T H_t\Delta\mathbf{x}_t}{2||\Delta\mathbf{x}_t||^2} - \alpha L\right) \geq 0\end{aligned}$$

The Hessian has a eigendecomposition: $H_t = V\Lambda V^T$. Now, consider the quantity, for $\mathbf{x} \neq 0$

$$\mathbf{x} \in \mathbb{R}^d, \frac{\mathbf{x}^T H_t\mathbf{x}}{||\mathbf{x}||^2} = \frac{\mathbf{x}^T V\Lambda V^T\mathbf{x}}{||\mathbf{x}||^2} = \frac{(V^T\mathbf{x})^T\Lambda(V^T\mathbf{x})}{||\mathbf{x}||^2} \geq \frac{\lambda_{\min}||V^T\mathbf{x}||^2}{||\mathbf{x}||^2} = \lambda_{\min}$$

Here we use the fact that since H_t is symmetric it's eigenvectors are orthogonal and therefore $V^T\mathbf{x}$ is a norm preserving transformation. From the result of Homework 10, exercise 3, we have any eigenvalue of the Hessian is at least μ, where μ is the strongly convex parameter. Combining the above two facts together, we therefore have,

$$\frac{\Delta\mathbf{x}_t^T H_t\Delta\mathbf{x}_t}{||\Delta\mathbf{x}_t||^2} \geq \lambda_{\min} \geq \mu \implies \frac{\Delta\mathbf{x}_t^T H_t\Delta\mathbf{x}_t}{||\Delta\mathbf{x}_t||^2} \geq \mu$$

Now consider the below quantity,

$$\alpha\left(\frac{3\Delta\mathbf{x}_t^T H_t\Delta\mathbf{x}_t}{2||\Delta\mathbf{x}_t||^2} - \alpha L\right)$$

For the given constraint on α, we just need to show this is non-negative. Using the result above,

$$\alpha\left(\frac{3\Delta\mathbf{x}_t^T H_t \Delta\mathbf{x}_t}{2||\Delta\mathbf{x}_t||^2} - \alpha L\right) \geq \alpha\left(\frac{3\mu}{2} - \alpha L\right) = \alpha L\left(\frac{3\mu}{2L} - \alpha\right)$$

Note that $\alpha \geq 0$, since we start at 1 and decrement by a factor $\gamma > 0$

$$\alpha \leq \min\left\{1, \frac{3\mu}{2L}\right\}$$

If $1 \leq \frac{3\mu}{2L} \implies \alpha \leq 1 \leq \frac{3\mu}{2L}$. If $1 > \frac{3\mu}{2L} \implies \alpha \leq \frac{3\mu}{2L} < 1$. This implies $\alpha \leq \frac{3\mu}{2L}$ in any case.

$$\because \alpha \leq \frac{3\mu}{2L} \implies \frac{3\mu}{2L} - \alpha \geq 0. \because \alpha, L \geq 0, \alpha L\left(\frac{3\mu}{2L} - \alpha\right) \geq 0.$$

Therefore for the given range of α, we indeed have,

$$\alpha\left(\frac{3\Delta\mathbf{x}_t^T H_t \Delta\mathbf{x}_t}{2||\Delta\mathbf{x}_t||^2} - \alpha L\right) \geq 0 \iff -\alpha\Delta\mathbf{x}_t^T H_t \Delta\mathbf{x}_t + \frac{L}{2}\alpha^2||\Delta\mathbf{x}_t||^2 \leq \frac{-\alpha\Delta\mathbf{x}_t^T H_t \Delta\mathbf{x}_t}{4}$$

$$\implies f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq \alpha\frac{\Delta\mathbf{x}_t^T \nabla f(\mathbf{x}_t)}{4}$$

which is the required result. Now for the second part,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq \alpha\frac{\Delta\mathbf{x}_t^T \nabla f(\mathbf{x}_t)}{4} = \frac{-\alpha\Delta\mathbf{x}_t^T H_t \Delta\mathbf{x}_t}{4} = \frac{-\alpha\Delta\mathbf{x}_t^T H_t \Delta\mathbf{x}_t||\Delta\mathbf{x}||^2}{4||\Delta\mathbf{x}||^2}$$

Now using the upper bound once again, we have,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq \frac{-\alpha\Delta\mathbf{x}_t^T H_t \Delta\mathbf{x}_t||\Delta\mathbf{x}||^2}{4||\Delta\mathbf{x}||^2} \leq -\frac{\alpha\mu||\Delta\mathbf{x}_t||^2}{4}$$

which is again the required result. Now considder the quantity,

$$\frac{\alpha\mu||\Delta\mathbf{x}_t||^2}{4} = \frac{\alpha\mu||H_t^{-1}\nabla f(\mathbf{x}_t)||^2}{4} = \frac{\alpha\mu||H_t^{-1}\nabla f(\mathbf{x}_t)||||H_t^{-1}\nabla f(\mathbf{x}_t)||}{4} \geq \frac{\alpha\mu||\nabla f(\mathbf{x}_t)||^2}{4\sigma_{\max}(H_t)^2}$$

For the inequality, consider the SVD of $H_t = USV^T$. Then $H_t^{-1} = VS^{-1}U^T$. Now the spectral norm of $H_t^{-1} = \max_{\mathbf{x}\in\mathbb{R}^d, \mathbf{x}\neq 0} \frac{||H_t^{-1}x||}{||x||}$.

$$\frac{||H_t^{-1}\mathbf{x}||}{||\mathbf{x}||} = \frac{||\sum_i v_i \frac{1}{\sigma_i} u_i^T \sum_j \alpha_j u_j||}{||\mathbf{x}||} = \frac{||\sum_i v_i \frac{1}{\sigma_i}\alpha_i||}{||\mathbf{x}||} \geq \frac{\frac{1}{\sigma_{\max}}||\sum_i v_i\alpha_i||}{||\mathbf{x}||} = \frac{1}{\sigma_{\max}(H_t)}$$

where we expressed $\mathbf{x}$ in the basis formed by U, and observe that since v_i's are orthonormal, they preserve the norm. Therefore in our case, instead of $\mathbf{x}$ we have $\nabla f(\mathbf{x}_t)$ and subsequently obtain the second inequality above. Therefore we have,

$$\frac{\alpha\mu||\Delta\mathbf{x}_t||^2}{4} \geq \frac{\alpha\mu||\nabla f(\mathbf{x}_t)||^2}{4\sigma_{\max}(H_t)^2}$$

Now we can invoke PL Inequality (since strong convexity $\implies$ PL Inequality) from Lecture Notes Definition 5.1 (or Lemma 5.2 in Handout 4d) on the gradient to further obtain that,

$$\frac{\alpha\mu||\nabla f(\mathbf{x}_t)||^2}{4\sigma_{\max}(H_t)^2} \geq \frac{\alpha\mu 2\mu(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{4\sigma_{\max}(H_t)^2} = \frac{\alpha\mu^2(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{2\sigma_{\max}(H_t)^2}$$

Assume that algorithm 2 runs for multiple iterations. And suppose it terminates at iteration k, at which iteration the condition for α is satisfied. That is,

$$\alpha_k = \gamma\alpha_{k-1} \leq \min\{1, 3\mu/2L\}$$

However,

$$\alpha_{k-1} > \min\{1, 3\mu/2L\} \iff \gamma\alpha_{k-1} = \alpha_k > \gamma\min\{1, 3\mu/2L\}$$

Therefore the α that we have at termination, α_k is greater than $\gamma\min\{1, 3\mu/2L\}$. Using this, we further have,

$$\frac{\alpha\mu||\nabla f(\mathbf{x}_t)||^2}{4\sigma_{\max}(H_t)^2} \geq \frac{\gamma\min\{1, 3\mu/2L\}\mu^2(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{2\sigma_{\max}(H_t)^2}$$

Now by combining the definitions of smoothness and strong convexity, we have the following:

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}||\mathbf{y} - \mathbf{x}||^2 \leq f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$$

This implies

$$\frac{\mu}{L} \leq 1$$

Now, if

$$\min\{1, 3\mu/2L\} = 1 \geq \mu/L$$

Now if

$$\min\{1, 3\mu/2L\} = 3\mu/2L;$$

Combining both cases, we have,

$$\min\{1, 3\mu/2L\} \geq K\frac{\mu}{L}$$

where $K > 0$. Using this and plugging this back, we have,

$$\frac{\gamma\min\{1, 3\mu/2L\}\mu^2(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{2\sigma_{\max}(H_t)^2} \geq \frac{\gamma K\mu\mu^2(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{2.L\sigma_{\max}(H_t)^2} = \frac{C\gamma\mu^3(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{L\sigma_{\max}(H_t)^2}$$

where $C = \frac{K}{2}$.
From Lemma 3.5 of Lecture Notes we know that, since f is $L-$smooth,

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\|$$

From theorem 2.9 in lecture notes, we have the spectral norm of the differentials of the gradients are bounded by L. However, the spectral norm is the highest singular value. Using this, we have,

$$||H_t|| = \sigma_{\max}(H_t) \leq L \implies \frac{1}{\sigma_{\max}(H_t)} \geq \frac{1}{L}$$

Using this, we have,

$$\frac{C\gamma\mu^3(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{L\sigma_{\max}(H_t)^2} \geq \frac{C\gamma\mu^3(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{LL^2} \geq \frac{C\gamma\mu^3(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{L^3}$$

Putting all the pieces together we have,

$$\frac{\alpha\mu||\Delta\mathbf{x}_t||^2}{4} \geq \frac{C\gamma\mu^3(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{L^3} \iff -\frac{C\gamma\mu^3(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{L^3} \geq -\frac{\alpha\mu||\Delta\mathbf{x}_t||^2}{4}$$

Therefore we have,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{\alpha\mu||\Delta\mathbf{x}_t||^2}{4} \leq -\frac{C\gamma\mu^3(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{L^3} \iff$$
$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq f(\mathbf{x}_t) - f(\mathbf{x}^*) - \frac{C\gamma\mu^3(f(\mathbf{x}_t) - f(\mathbf{x}^*))}{L^3} = \left(1 - C\frac{\gamma\mu^3}{L^3}\right)(f(\mathbf{x}_t) - f(\mathbf{x}^*))$$

Therefore we have,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq \left(1 - C\frac{\gamma\mu^3}{L^3}\right)(f(\mathbf{x}_t) - f(\mathbf{x}^*))$$

Now inductively repeating the inequality $t+1$ times, we have,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq \left(1 - C\frac{\gamma\mu^3}{L^3}\right)^{t+1}(f(\mathbf{x}_0) - f(\mathbf{x}^*))$$

Setting t = t-1, we obtain the required result, which is,

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \left(1 - \frac{C\gamma\mu^3}{L^3}\right)^{t}(f(\mathbf{x}_0) - f(\mathbf{x}^*))$$