part 1

1.1.1)

write $f(x) = \ldots \underbrace{\sigma(W_3\sigma(W_2\sigma(W_1 x + b_1) + b_2) + b_3 \ldots}_{n \text{ layers}}$

$W_i = 1 \Rightarrow f(x) = \underbrace{\ldots\sigma(\sigma(\sigma(x + b_1) + b_2) + b_3)\ldots}_{n \text{ layers}}$.

Let $y_1 = x + b_1$

$y_2 = \sigma(y_1) + b_2$

$y_3 = \sigma(y_2) + b_3$

$\vdots$

$f(x) = y_n = \sigma(y_{n-1}) + b_n$

Then $\frac{df(x)}{dx} = \sigma'(y_{n-1})\, y'_{n-1}$

$= \sigma'(y_{n-1})\,\sigma'(y_{n-2})\, y'_{n-2}$

$\vdots$

$= \sigma'(y_{n-1})\,\sigma'(y_{n-2}) \cdots \sigma'(y_1)\, y_1'$

$= \sigma'(y_{n-1})\,\sigma'(y_{n-2}) \cdots \sigma'(y_1)\,(1)$

$= \prod_{i=1}^{n-1} \sigma'(y_i)$

$= \prod_{i=1}^{n-1} \left(\sigma(y_i) - \sigma^2(y_i)\right)$ since $\sigma'(x) = \sigma(x)(1-\sigma(x)) = \sigma(x) - \sigma^2(x)$

since for any $x \in \mathbb{R}$, $\frac{d(\sigma(x) - \sigma^2(x))}{d\sigma(x)} = 1 - 2\sigma(x) \overset{\text{set}}{=} 0$

$\Rightarrow \sigma(x) = \frac{1}{2}$ maximizes $\sigma'(x)$ at $\sigma'(x) = \frac{1}{4}$

and since $\sigma(x) \in (0,1) \Rightarrow 0 \le \sigma(x) - \sigma^2(x) \le \frac{1}{4}$

Then $0 \le \sigma(y_i) - \sigma^2(y_i) \le \frac{1}{4}$

$\Rightarrow 0 \le \prod_{i=1}^{n-1} \left(\sigma(y_i) - \sigma^2(y_i)\right) \le \left(\frac{1}{4}\right)^{n-1}$

Note that if there are n hidden layers plus an output layer, i.e. $f(x) = y_{n+1} = \sigma(y_n) + b_{n+1}$, then we can replace all $n-1$ with n to get

$0 \le \prod_{i=1}^{n} \left(\sigma(y_i) - \sigma^2(y_i)\right) \le \left(\frac{1}{4}\right)^{n}$

as wanted.

In either case, the gradients will vanish as n gets large.

1.2.1)

$$x_n = \dots \tanh(W \tanh(W \tanh(x_1))) \dots$$

$$x_2 = \tanh(W x_1)$$
$$x_3 = \tanh(W x_2)$$
$$\vdots$$
$$x_n = \tanh(W x_{n-1})$$

$$\frac{\partial x_n}{\partial x_1} = \prod_{t=1}^{n-1} \frac{\partial x_{t+1}}{\partial x_t}$$

$$= \prod_{t=1}^{n-1} \frac{\partial \tanh(W x_t)}{\partial W x_t} W$$

note that the Jacobian of tanh is $\begin{bmatrix} \tanh'(x_1) & & 0 \\ & \tanh'(x_2) & \\ 0 & & \ddots \tanh'(x_k) \end{bmatrix}$

and $\tanh'(x) = \operatorname{sech}^2(x) \in (0,1]$, for $x \in \mathbb{R}$. So $0 \le \sigma_{max}\left[\frac{\partial \tanh(W x_t)}{\partial W x_t}\right] \le 1$

$$\Rightarrow \sigma_{max}\left(\frac{\partial x_n}{\partial x_1}\right) = \sigma_{max}\left[\prod_{t=1}^{n-1} \frac{\partial x_{t+1}}{\partial x_t}\right] \le \prod_{t=1}^{n-1} \sigma_{max}\left[\frac{\partial x_{t+1}}{\partial x_t}\right] \quad \text{by hint.}$$

$$\text{and } \sigma_{max}\left[\frac{\partial x_{t+1}}{\partial x_t}\right] = \sigma_{max}\left[\frac{\partial \tanh(W x_t)}{\partial W x_t} W\right]$$

$$\le \sigma_{max}\left(\frac{\partial \tanh(W x_t)}{\partial W x_t}\right) \sigma_{max}(W)$$

$$\le 1 \cdot \frac{1}{2} = \frac{1}{2}$$

$$\Rightarrow \sigma_{max}\left(\frac{\partial x_n}{\partial x_1}\right) = \sigma_{max}\left[\prod_{t=1}^{n-1} \frac{\partial x_{t+1}}{\partial x_t}\right] \le \left(\frac{1}{2}\right)^{n-1}$$

Also, since singular values are non negative,

$$\Rightarrow 0 \le \sigma_{max}\left[\prod_{t=1}^{n-1} \frac{\partial x_{t+1}}{\partial x_t}\right] \le \left(\frac{1}{2}\right)^{n-1}$$

$$\Rightarrow 0 \le \sigma_{max}\left(\frac{\partial x_n}{\partial x_1}\right) \le \left(\frac{1}{2}\right)^{n-1} \quad \text{as wanted.}$$

1.3.1)

sub in kernel function: $\alpha_i = \frac{\sum_{j=1}^{n} \phi(Q_i)^T \phi(k_j) V_j}{\sum_{j=1}^{n} \phi(Q_i)^T \phi(k_j)}$

$= \frac{\phi(Q_i)^T \sum_{j=1}^{n} \phi(k_j) V_j^T}{\phi(Q_i)^T \sum_{j=1}^{n} \phi(k_j)}$

we can compute $\sum_{j=1}^{n} \phi(k_j) V_j^T$ and $\sum_{j=1}^{n} \phi(k_j)$ once and reuse them for every query. So attention can be calculated in $O(n)$.

1.3.2)

Assume we have SVD of $P = U \Sigma W^T$, $U, W \in \mathbb{R}^{n \times k}$ are semi-unitary s.t. $U^T U = W^T W = I$.

Then $PV = U \Sigma W^T V$

step ① computing $U\Sigma$ is multiplying the k n-dimentional vectors by the k non-zero singular values, which is $O(nk)$, and results in a $n \times k$ matrix

step ② computing $W^T V$ is multiplying $k \times n$ matrix by $n \times d$ matrix, which is $O(knd)$ and results in a $k \times d$ matrix

step ③ computing $U \Sigma W^T V$ is multiplying the $n \times k$ matrix by the $k \times d$ matrix which is $O(nkd)$

$\Rightarrow$ so the entire 3 steps combined is $O(nkd)$ as wanted.

(33).

P is rank k $\Rightarrow$ we can write $P = U\Sigma W^T$, $U, W \in \mathbb{R}^{n \times k}$ by SVD.

where U, W are semi-unitary.

Then we can take $C = D = W^T \in \mathbb{R}^{k \times n}$.

$$\Rightarrow Q(Ck)^T DV = Q(W^T K)^T W^T V$$
$$= Q K^T W W^T V \quad \text{since } WW^T = I, \text{ } W \text{ is orthogonal.}$$
$$= Q K^T V$$
$$= PV \quad \text{since } P = QK^T$$

Complexity:

step ① compute Ck ($k \times n$ times $n \times d$), $O(knd)$. output $k \times d$ matrix.
step ② compute DV ($k \times n$ times $n \times d$), $O(knd)$. output $k \times d$ matrix
step ③ $Q(Ck)^T$ ($n \times d$ times $d \times k$), $O(ndk)$. output $n \times k$ matrix.
step ④ $Q(Ck)^T DV$ ($n \times k$ times $k \times d$), $O(nkd)$. output $n \times d$ matrix.

$\Rightarrow$ so entire steps ① $\rightarrow$ ④ takes $O(nkd)$.

Part 2

2.1) $g[\theta, \tilde{a}] = f(\tilde{a}) \frac{\partial}{\partial \theta} \log p(a=\tilde{a}|\theta)$, $\tilde{a} \sim p(a|\theta)$

notice $\tilde{a}$ is either 1 or 0

if $\tilde{a}=0$: $g[\theta, \tilde{a}] = f(\tilde{a}) \frac{\partial}{\partial \theta} \log p(a=\tilde{a}|\theta)$
$= 0 \cdot \frac{\partial}{\partial \theta} \log p(a=\tilde{a}|\theta) = 0$

if $\tilde{a}=1$: $g[\theta, \tilde{a}] = f(\tilde{a}) \frac{\partial}{\partial \theta} \log p(a=\tilde{a}|\theta)$
$= 1 \cdot \frac{\partial}{\partial \theta} \log p(a=\tilde{a}|\theta)$
$= \frac{\partial}{\partial \theta} \log p(a=1|\theta)$
$= \frac{\partial}{\partial \theta} \log \mu$
$= \frac{1}{\mu} \frac{\partial \mu}{\partial \theta}$
$= \frac{1}{\mu} x \sigma'(\sum_{d=1}^{D} \theta_d x_d)$
$= \frac{1}{\mu} x \mu(1-\mu)$ since $\sigma'(x) = \sigma(x)(1-\sigma(x))$
$= x(1-\mu)$

combining them $\Rightarrow$ $\boxed{g[\theta, \tilde{a}] = \tilde{a} x (1-\mu)}$

2.2) $\mathrm{Var}[g[\theta, \tilde{a}]_i] = \mathrm{var}[\tilde{a} x_i (1-\mu)]$
$= x_i^2 (1-\mu)^2 \mathrm{Var}[\tilde{a}]$
$= x_i^2 (1-\mu)^2 \mu(1-\mu)$
$= x_i^2 \mu (1-\mu)^3$

Part 3

3.1.1

$$x^T L x = x^T (D-A) x$$

$$= x^T D x - x^T A x$$

$$= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \deg(1) & & 0 \\ & \ddots & \\ 0 & & \deg(n) \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \sum_{u=1}^{n} x_u \deg(u) x_u - \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

since $A_{uv} = 1$ if there is an edge between vertex u, v. and $A_{uv} = 0$ otherwise,

so $\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{(u,v) \in E} 2 x_u x_v$ since $(u,v) \in E$ $\Rightarrow A_{uv}$ and A_{vu} are both 1

$\Rightarrow$ Therefore

$$x^T L x = \sum_{u=1}^{n} x_u^2 \deg(u) - \sum_{(u,v) \in E} 2 x_u x_v$$

$$= \sum_{u=1}^{n} \sum_{(u,v) \in E} x_u^2 - \sum_{(u,v) \in E} 2 x_u x_v$$

$$= \sum_{(u,v) \in E} (x_u^2 + x_v^2) - \sum_{(u,v) \in E} 2 x_u x_v$$

$$= \sum_{(u,v) \in E} (x_u^2 - 2 x_u x_v + x_v^2)$$

$$= \sum_{(u,v) \in E} (x(u) - x(v))^2 \quad \text{as wanted,}$$

3.1.2)

Since L is a symetric matrix (because both A and D are symmetric) and $x^T L x = \sum_{(u,v)\in E} (x(u)-x(v))^2 \geq 0$, L must be PSD.

(by 3.1.1)

next consider the vector $v = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

$$Lv = (D-A)v = Dv - Av$$

$$= \begin{bmatrix} \deg(1) & & 0 \\ & \ddots & \\ 0 & & \deg(n) \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} \deg(1) \\ \deg(2) \\ \vdots \\ \deg(n) \end{bmatrix} - \begin{bmatrix} \sum_{i=1}^{n} A_{1i} \\ \sum_{i=1}^{n} A_{2i} \\ \vdots \\ \sum_{i=1}^{n} A_{ni} \end{bmatrix}$$

$$= 0 \qquad \left(\text{since } \sum_{i=1}^{n} A_{ji} = \deg(j) \right)$$

so the smallest eigenvalue $\lambda_{min}(L) = 0$ with eigenvector $v = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

3.2.1)

$$\lambda_{max}(\tilde{L}) = \max_{\|x\|=1} x^T \tilde{L} x \quad , x \in \mathbb{R}^n .$$

so it is sufficient to show that $x^T \tilde{L} x \leq 2$ for any $x \in \mathbb{R}^n$ st. $\|x\| = 1 \Leftrightarrow x^T x = 1$

First, $x^T(I + \tilde{A})x$

$$= x^T x + x^T \tilde{A} x$$

$$= \sum_{i=1}^{n} x(i)^2 + \sum_{(i,j) \in E} \frac{2x(i)x(j)}{\sqrt{d(i)\,d(j)}}$$

$$= \sum_{(i,j) \in E} \left(\frac{x(i)}{\sqrt{d(i)}} + \frac{x(j)}{\sqrt{d(j)}} \right)^2 \quad \text{same reasoning as 3.1.1}$$

$$\geq 0$$

$$\Rightarrow x^T(I + \tilde{A})x \geq 0$$

$$\Rightarrow x^T I x + x^T \tilde{A} x \geq 0$$

$$\Rightarrow -x^T \tilde{A} x \leq x^T x$$

$$\Rightarrow x^T x - x^T \tilde{A} x \leq 2x^T x$$

also, $x^T \tilde{L} x = x^T(I - \tilde{A})x = x^T x - x^T \tilde{A} x$

$$\Rightarrow x^T \tilde{L} x \leq 2x^T x .$$

$$\Rightarrow x^T \tilde{L} x \leq 2 \quad \text{since } \|x\| = x^T x = 1$$

as wanted.

3.2.2).

By 3.2.1, $\lambda_{max}(\tilde{L}) \le 2$.

$x^T \tilde{L} x \le 2$

$\Rightarrow x^T (I - \tilde{A}) x \le 2$

$\Rightarrow x^T x - x^T \tilde{A} x \le 2$

$\Rightarrow 1 - x^T \tilde{A} x \le 2$ if we take $\|x\| = x^T x = 1$

$\Rightarrow 1 - 2 \le x^T \tilde{A} x$

$\Rightarrow -1 \le x^T \tilde{A} x$

so $-1 \le \lambda_i(\tilde{A})$. ①

Also, $x^T \tilde{L} x = x^T (I - \tilde{A}) x$

$= x^T x - x^T \tilde{A} x$

$= \sum_{i=1}^{n} x(i)^2 - \sum_{(i,j) \in E} \frac{2x(i)x(j)}{\sqrt{d(i)\,d(j)}}$

$= \sum_{(i,j) \in E} \left(\frac{x(i)}{\sqrt{d(i)}} - \frac{x(j)}{\sqrt{d(j)}} \right)^2$ same reasoning as 3.1.1

≥ 0.

$\Rightarrow x^T (I - \tilde{A}) x \ge 0$

$\Rightarrow x^T x - x^T \tilde{A} x \ge 0$

$\Rightarrow x^T x \ge x^T \tilde{A} x$

$\Rightarrow 1 \ge x^T \tilde{A} x$ by hint 3.2.1.

so $\lambda_i(\tilde{A}) \le 1$ ②

combining ① + ② $\Rightarrow -1 \le \lambda_i(\tilde{A}) \le 1$

as wanted.

3.3.1)

we can take the eigendecomposition of $\tilde{A} = Q \Lambda Q^T$ where Q is an orthonormal matrix since $\tilde{A}$ is a square symmetric matrix.

Then $\tilde{A}^K = (Q \Lambda Q^T)^K$

$= \underbrace{Q \Lambda Q^T Q \Lambda Q^T \dots Q \Lambda Q^T}_{K \text{ times}}$

$= Q \Lambda I \Lambda I \dots I \Lambda Q^T$ since $Q^T Q = I$.

$= Q \Lambda^K Q^T$

since Λ is diagonal, Λ^K is equal to element wise power across the diagonal entries, which is much more efficient.

3.3.2)

No. It is not always beneficial to use a deeper model since the eigenvalues of $\tilde{A}$ are all between -1 and 1, so taking large number of layers K would result in Λ^K approaching 0 and thus $\tilde{A}^K$ approaching 0 (vanishing gradient) as well as the problem of over smoothing, and hence $H^{(K)}$.

3.4.1)

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{a}^T [W\vec{h}_i \,\|\, W\vec{h}_j]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\text{LeakyReLU}\left(\vec{a}^T [W\vec{h}_i \,\|\, W\vec{h}_k]\right)\right)}$$

→ GAT allows for (implicitly) assigning different importances to nodes of a same neighborhood, enabling greater model capacity.

→ In GAT, analyzing learned attention weights may result in better model interpretability.