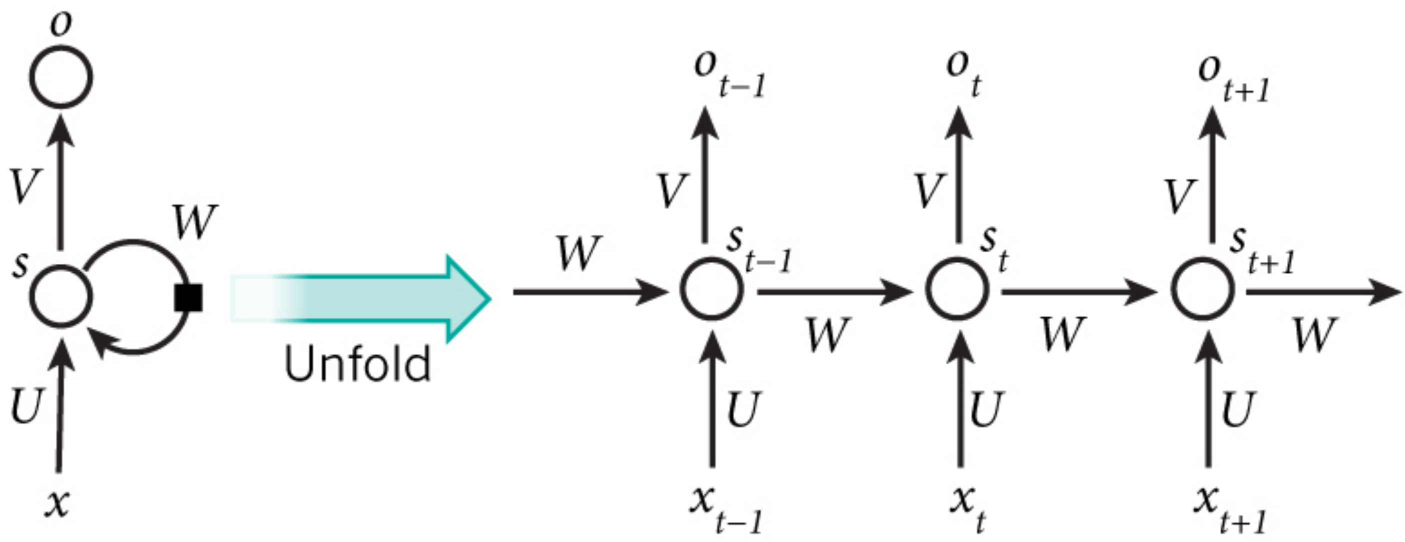
Summary of Last Month (2.22-3.20)

Liu, Chen. 20th, March

Basic Knowledge of RNN



1. Notation
    - Input data: $\mathbf{x} \in \mathbb{R}^N$
    - Hidden state: $\mathbf{s} \in \mathbb{R}^H$
    - Output data: $\mathbf{o} \in \mathbb{R}^K$
    - Input connection: $\mathbf{U} \in \mathbb{R}^{H \times N}$
    - Recurrent connection: $\mathbf{W} \in \mathbb{R}^{H \times H}$
    - Output connection: $\mathbf{V} \in \mathbb{R}^{H \times K}$

2. Computation:
    - $\mathbf{s}_t = \sigma(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{s}_{t-1})$
    - $\mathbf{o}_t = softmax(\mathbf{V}\mathbf{s}_t)$

3. Loss function:
    - Usually, we treat a squence $\{\mathbf{x}_i\}_{1 \le i \le M}$ as a training instance. Given the label of such sequence $\{\mathbf{y}_i\}_{1 \le i \le M}$, the loss function is defined below:

$$E = -\frac{1}{M}\sum_{t=1}^{M} \mathbf{y_t}^T \log \mathbf{o_t}$$

Loss Function of RNN

1. Obviously, we have that all the entries in $\mathbf{y}_t$ sum up to one. We can simplify the expression of loss function of RNN.

$$\begin{aligned} E &= -\frac{1}{M}\sum_{t=1}^{M}\sum_{i=1}^{K}\mathbf{y}_{t,i}\log\mathbf{o}_{t,i} \\ &= -\frac{1}{M}\sum_{t=1}^{M}\sum_{i=1}^{K}\mathbf{y}_{t,i}\log\frac{e^{\mathbf{V}_{i,:}\mathbf{s}_t}}{\sum_{j=1}^{K}e^{\mathbf{V}_{j,:}\mathbf{s}_t}} \\ &= \frac{1}{M}\sum_{t=1}^{M}(\log(\sum_{j=1}^{K}e^{\mathbf{V}_{j,:}\mathbf{s}_t}) - \mathbf{y}_t^T\mathbf{V}\mathbf{s}_t) \end{aligned}$$

2. We define $\theta = \{\mathbf{U}, \mathbf{W}, \mathbf{V}\}, f(\theta) = \sum_{t=1}^{M}\log(\sum_{j=1}^{K}e^{\mathbf{V_{j,:}s_t}})$ and $g(\theta) = \sum_{t=1}^{M}\mathbf{y}_t\mathbf{V}\mathbf{s}_t$, so

$$E = \frac{1}{M}(f(\theta) - g(\theta))$$

Loss Bounded with Respect to V

1. For $g(\theta)$ is linear with respect to $\mathbf{V}$, so we have:

$$g(\mathbf{V}^k + \Delta\mathbf{V}) = g(\mathbf{V}^k) + \langle\nabla_{\mathbf{V}}g(\mathbf{V}^k), \Delta\mathbf{V}\rangle$$

2. $f(\theta)$ is sum of several exponentiations. According to [Theorem 1 in "Stochastic Spetral Descent for Restricted Boltzmann Machine"](). Let $lse_{\mathbf{w}}(\mathbf{u}) = \log\sum_{j=1}^{K}w_j exp(u_j)$, we have:

$$f(\mathbf{V}) = lse_{\mathbf{1}}(\mathbf{V}\mathbf{s}_t)$$

Thus we can draw an upper bound with respect to $\mathbf{V}\mathbf{s}_t$:

$$\begin{aligned} f(\mathbf{U}^k, \mathbf{W}^k, \mathbf{V}^k + \Delta\mathbf{V}) &= lse_{\mathbf{1}}(\mathbf{V}^k\mathbf{s}_t + \Delta\mathbf{V}\mathbf{s}_t) \\ &\le lse_{\mathbf{1}}(\mathbf{V}^k\mathbf{s}_t) + \langle\nabla lse_{\mathbf{1}}(\mathbf{V}^k\mathbf{s}_t), \Delta\mathbf{V}\mathbf{s}_t\rangle + \frac{1}{2}\|\Delta\mathbf{V}\mathbf{s}_t\|_\infty^2 \\ &= f(\theta^k) + \langle\nabla_{\mathbf{V}\mathbf{s}_t}f(\theta^k), \Delta\mathbf{V}\mathbf{s}_t\rangle + \frac{1}{2}\|\Delta\mathbf{V}\mathbf{s}_t\|_\infty^2 \\ &\le f(\theta^k) + \langle\nabla_{\mathbf{V}}f(\theta^k), \Delta\mathbf{V}\rangle + \frac{1}{2}\|\Delta\mathbf{V}\|_{S^\infty}^2\|\mathbf{s}_t\|_2^2 \end{aligned}$$

3. $$E(\mathbf{U}^k, \mathbf{W}^k, \mathbf{V}^k + \Delta\mathbf{V}) \le E(\theta^k) + \langle\nabla_{vec(\mathbf{V})}E(\theta^k), vec(\Delta\mathbf{V})\rangle + \frac{1}{2}\|\Delta\mathbf{V}\|_{S^\infty}^2\|\mathbf{s}_t\|_2^2$$

Loss Bound with Respect to W

1. Let $\mathbf{\Phi} = \{\mathbf{W}, \mathbf{U}\}$, we can regard $\mathbf{s}_t$ as a function of Φ: $\mathbf{s}_t(\mathbf{\Phi})$. We have an upper bound of $E(\mathbf{\Phi} + \Delta\mathbf{\Phi})$: (Proof of this equation are attached in the handcraft appendix)

$$E(\mathbf{\Phi}+\Delta\mathbf{\Phi}) \leq E(\mathbf{\Phi}) + \langle \nabla_{\mathbf{\Phi}} E(\mathbf{\Phi}), \Delta\mathbf{\Phi}\rangle + \frac{1}{M}\sum_{t=1}^{M}(\frac{1}{2}\|\mathbf{V}\mathbf{s}_t(\mathbf{\Phi}+\Delta\mathbf{\Phi}) - \mathbf{V}\mathbf{s}_t(\mathbf{\Phi})\|_\infty^2 + 2\|\mathbf{V}\mathbf{s}_t(\mathbf{\Phi}+\Delta\mathbf{\Phi}) - \mathbf{V}\mathbf{s}_t(\mathbf{\Phi}) - \langle \mathbf{V}\nabla\mathbf{s}_t(\mathbf{\Phi}), \Delta\mathbf{\Phi}\rangle\|_\infty)$$

2. The upper bound of $\|\mathbf{V}\mathbf{s}_t(\mathbf{W}^k+\Delta\mathbf{W}) - \mathbf{V}\mathbf{s}_t(\mathbf{W}^k)\|_\infty^2$ to update $\mathbf{W}$.

    1. Last week, I made a mistake. $\frac{\partial[\mathbf{s}_t(\mathbf{W})]_p}{\partial W_{ij}} \neq 0$ even if $p \neq i$ when $t > 2$. So the linking equation is:

    $$\frac{[\partial \mathbf{V}\mathbf{s}_t(\mathbf{W})]_p}{\partial \mathbf{W}_{ij}} = \sum_{k=1}^{H} \mathbf{V}_{pk}\frac{\partial[\mathbf{s}_t(\mathbf{W})]_k}{\partial \mathbf{W}_{ij}}$$

    $$\frac{\partial[\mathbf{s}_t(\mathbf{W})]_p}{\partial \mathbf{W}_{ij}} = \sigma'(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{s}_{t-1})_p([\mathbf{s}_{t-1}]_j\delta_{ip} + \sum_{k=1}^{H}\frac{\partial[\mathbf{s}_{t-1}(\mathbf{W})]_k}{\partial \mathbf{W}_{ij}}\mathbf{W}_{pk})$$

    where $\delta_{ij} = 1$ iff $i = j$.

    2. Let P_t be a matrix of size $H \times H^2$ where $P_t(k, iH+j) = \frac{\partial[\mathbf{s}_t(\mathbf{W})]_k}{\mathbf{W}_{ij}}$, $\bar{\mathbf{S}}_t$ be a sparse matrix of the same size where $\bar{\mathbf{S}}_t(i,(i-1)H+j) = \mathbf{s}_{t,j}$ and $\mathbf{\Lambda}_t$ be a diagonal matrix whose diagonal is vector $\sigma'(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{s}_{t-1})$. We have:

    $$\frac{[\partial \mathbf{V}\mathbf{s}_t(\mathbf{W})]_p}{\partial \mathbf{W}} = Mtr(\mathbf{V}_{p,:}\mathbf{P}_t)$$

    $$\mathbf{P}_t = \mathbf{\Lambda}_t(\bar{\mathbf{S}}_{t-1} + \mathbf{W}\mathbf{P}_{t-1})$$

    $Mtr(\mathbf{s})$ is to turn a H^2-dim vector into a $H \times H$ matrix.

    3. Now, we can derive the upper bound of $\|\mathbf{V}\mathbf{s}_t(\mathbf{W}^k+\Delta\mathbf{W}) - \mathbf{V}\mathbf{s}_t(\mathbf{W}^k)\|_\infty^2$

    $$\begin{aligned}[\mathbf{V}\mathbf{s}_t(\mathbf{W}^k+\Delta\mathbf{W}) - \mathbf{V}\mathbf{s}_t(\mathbf{W}^k)]_p &= \int_0^1 tr[\frac{\partial[\mathbf{V}\mathbf{s}_t(\mathbf{W})]_p}{\partial \mathbf{W}}|_{\mathbf{W}=\mathbf{W}^k+t\Delta\mathbf{W}}\Delta\mathbf{W}]d_t \\ &= \int_0^1 tr[Mtr(\mathbf{V}_{p,:}\mathbf{\Lambda}_t(\bar{\mathbf{S}}_{t-1} + \mathbf{W}\mathbf{P}_{t-1}))|_{\mathbf{W}=\mathbf{W}^k+t\Delta\mathbf{W}}\Delta\mathbf{W}]d_t\end{aligned}$$

    Approximate this formula at $t = 0$ and note that every diagonal element of $\mathbf{\Lambda}_t$ is smaller than $\max_x \sigma'(x)$, so we have:

    $$\begin{aligned}[\mathbf{V}\mathbf{s}_t(\mathbf{W}^k+\Delta\mathbf{W}) - \mathbf{V}\mathbf{s}_t(\mathbf{W}^k)]_p &\leq max_x\sigma'(x)tr[Mtr(\mathbf{V}_{p,:}(\bar{\mathbf{S}}_{t-1} + \mathbf{W}\mathbf{P}_{t-1}))|_{\mathbf{W}=\mathbf{W}^k}\Delta\mathbf{W}^T] \\ &= \max_x \sigma'(x)\{tr[\mathbf{V}_{p,:}^T\mathbf{s}_{t-1}^T\Delta\mathbf{W}^T] + tr[Mtr(\mathbf{V}_{p,:}\mathbf{W}^k\mathbf{P}_{t-1})\Delta\mathbf{W}^T]\} \\ &= \max_x \sigma'(x)\{\mathbf{V}_{p,:}\Delta\mathbf{W}\mathbf{s}_{t-1} + tr[Mtr(\mathbf{V}_{p,:}\mathbf{W}^k\mathbf{P}_{t-1})\Delta\mathbf{W}]\} \\ &\leq \max_x \sigma'(x)\{\|\mathbf{V}_{p,:}\|_2\|\Delta\mathbf{W}\|_{S^\infty}\|\mathbf{s}_{t-1}\|_2 + \|Mtr(\mathbf{V}_{p,:}\mathbf{W}^k\mathbf{P}_{t-1})\|_{S^1}\|\Delta\mathbf{W}\|_{S^\infty}\} \\ &= \max_x \sigma'(x)\|\Delta\mathbf{W}\|_{S^\infty}\{\|\mathbf{V}_{p,:}\|_2\|\mathbf{s}_{t-1}\|_2 + \|Mtr(\mathbf{V}_{p,:}\mathbf{W}^k\mathbf{P}_{t-1})\|_{S^1}\}\end{aligned}$$

3. The upper bound of $\|\mathbf{V}\mathbf{s}_t(\mathbf{W}^k+\Delta\mathbf{W}) - \mathbf{V}\mathbf{s}_t(\mathbf{W}^k) - \langle \mathbf{V}\nabla_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k), \Delta\mathbf{W}\rangle\|_\infty$

    1. We analyze the p-th element of this vector.

$$
\begin{aligned}
&[\mathbf{V}\mathbf{s}_t(\mathbf{W}^k+\Delta\mathbf{W})-\mathbf{V}\mathbf{s}_t(\mathbf{W}^k)-\langle\mathbf{V}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k),\Delta\mathbf{W}\rangle]_p\\
&=\int_0^1 tr([\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k+t\Delta\mathbf{W})-\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k)]\Delta\mathbf{W})d_t\\
&\leq\|\Delta\mathbf{W}\|_{S^\infty}\|\int_0^1\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k+t\Delta\mathbf{W})-\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k)\|_{S^1}d_t
\end{aligned}
$$

2. Let's focus on the last integration part. We approximate it at point $t=0$:

$$
\begin{aligned}
&\int_0^1\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k+t\Delta\mathbf{W})-\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k)d_t\\
&\simeq\int_0^1\frac{d}{d_t}\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k+t\Delta\mathbf{W})|_{t=0}td_t\\
&=\frac{1}{2}\frac{d}{d_t}\mathbf{V}_{p,:}\triangledown_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k+t\Delta\mathbf{W})|_{t=0}
\end{aligned}
$$

3. We already have:

$$
\triangledown_{\mathbf{W}}\mathbf{V}_{p,:}\mathbf{s}_t(\mathbf{W})=Mtr[\mathbf{V}_{p,:}\mathbf{\Lambda}_t(\bar{\mathbf{S}}_{t-1}+\mathbf{W}\mathbf{P}_{t-1})]=\mathbf{\Lambda}_t\mathbf{V}_{p,:}^T\mathbf{s}_{t-1}^T+Mtr(\mathbf{V}_{p,:}\mathbf{\Lambda}_t\mathbf{W}\mathbf{P}_{t-1})
$$

so(λ_t is the diagonal elements of $\mathbf{\Lambda}_t$):We ignore the second-order derivative part, which vanish much faster then the first order part.

$$
\frac{d}{d_t}\triangledown_{\mathbf{W}}\mathbf{V}_{p,:}\mathbf{s}_t(\mathbf{W})=(\frac{d}{d_t}\lambda_t)\odot\mathbf{V}_{p,:}^T\mathbf{s}_{t-1}^T+\mathbf{\Lambda}_t\mathbf{V}_{p,:}^T(\frac{d}{d_t}\mathbf{s}_{t-1}^T)+Mtr(\mathbf{V}_{p,:}\frac{d}{d_t}\mathbf{\Lambda}_t\mathbf{W}\mathbf{P}_{t-1})+Mtr(\mathbf{V}_{p,:}\mathbf{\Lambda}_t\Delta\mathbf{W}\mathbf{P}_{t-1})
$$

1. Upper bound of $(\frac{d}{d_t}\lambda_t)\odot\mathbf{V}_{p,:}^T\mathbf{s}_{t-1}^T$

 - Let λ_t'' be a column vector of $\sigma''(\mathbf{U}\mathbf{x}_t+\mathbf{W}\mathbf{s}_t)$:

$$
\frac{d}{d_t}\lambda_t=\lambda_t''\odot[(\bar{\mathbf{S}}_{t-1}+\mathbf{W}\mathbf{P}_{t-1})vec(\Delta\mathbf{W})]=\lambda_t''\odot(\Delta\mathbf{W}\mathbf{s}_{t-1}+\mathbf{W}\mathbf{P}_{t-1}vec(\Delta\mathbf{W}))
$$

 - Note that $\mathbf{W}\mathbf{P}_{t-1}vec(\Delta\mathbf{W})$ is a column vector whose p-th element is $\mathbf{W}_{p,:}\mathbf{P}_{t-1}vec(\Delta\mathbf{W})=tr[Mtr(\mathbf{W}_{p,:}\mathbf{P}_{t-1})\Delta\mathbf{W}]\leq\|Mtr(\mathbf{W}_{p,:}\mathbf{P}_{t-1})\|_{S^1}\|\Delta\mathbf{W}\|_{S^\infty}$
 - As a result:

$$
\begin{aligned}
\|(\frac{d}{d_t}\lambda_t)\odot\mathbf{V}_{p,:}^T\mathbf{s}_{t-1}^T\|_{S^1}&\leq\|(\frac{d}{d_t}\lambda_t)\odot\mathbf{V}_{p,:}^T\|_{S^\infty}\|\mathbf{s}_{t-1}^T\|_{S^1}=\|(\frac{d}{d_t}\lambda_t)\odot\mathbf{V}_{p,:}^T\|_2\|\mathbf{s}_{t-1}^T\|_2\\
&\leq\|\lambda_t''\odot(\Delta\mathbf{W}\mathbf{s}_{t-1}+\mathbf{W}\mathbf{P}_{t-1}vec(\Delta\mathbf{W}))\|_2\|\mathbf{V}_{p,:}^T\|_\infty\|\mathbf{s}_{t-1}^T\|_2\\
&\leq\|\lambda_t''\|_\infty\|\mathbf{V}_{p,:}^T\|_\infty\|\mathbf{s}_{t-1}^T\|_2(\|\Delta\mathbf{W}\mathbf{s}_{t-1}\|_2+\|\mathbf{W}\mathbf{P}_{t-1}vec(\Delta\mathbf{W}))\|_2)\\
&\leq\|\lambda_t''\|_\infty\|\mathbf{V}_{p,:}^T\|_\infty\|\mathbf{s}_{t-1}^T\|_2(\|\Delta\mathbf{W}\|_{S^\infty}\|\mathbf{s}_{t-1}^T\|_2+\sqrt{H}\max_p(\|Mtr(\mathbf{W}_{p,:}\mathbf{P}_{t-1})\|_{S^1})\|\Delta\mathbf{W}\|_{S^\infty})\\
&=\|\lambda_t''\|_\infty\|\mathbf{V}_{p,:}^T\|_\infty\|\mathbf{s}_{t-1}^T\|_2\|\Delta\mathbf{W}\|_{S^\infty}(\|\mathbf{s}_{t-1}^T\|_2+\sqrt{H}\max_p(\|Mtr(\mathbf{W}_{p,:}\mathbf{P}_{t-1})\|_{S^1})
\end{aligned}
$$

 We have taken advantage of the conclusion: 1) $\|\mathbf{x}\odot\mathbf{y}\|_2\leq\|\mathbf{x}\|_\infty\|\mathbf{y}\|_2$. 2) $\|\mathbf{A}\mathbf{B}\|_{S^r}\leq\|\mathbf{A}\|_{S^q}\|\mathbf{B}\|_{S^p}$ if $r^{-1}=p^{-1}+q^{-1}$. 3) For vector(no matter raw or column) $\mathbf{x}$, we have $\|\mathbf{x}\|_{S^\infty}=\|\mathbf{x}\|_{S^1}=\|\mathbf{x}\|_2$.

2. Upper bound of $\mathbf{\Lambda}_t \mathbf{V}_{p,:}^T (\frac{d}{d_t}\mathbf{s}_{t-1}^T)$

- It is easy to find out that $\frac{d}{d_t}\mathbf{s}_{t-1}^T = \mathbf{P}_{t-1} vec(\Delta \mathbf{W})$ whose p-th elemnt is bounded by $\|Mtr(\mathbf{P}_{t-1,p})\|_{S^1} \|\Delta \mathbf{W}\|_{S^\infty}$
- Same as above:

$$
\begin{aligned}
\|\mathbf{\Lambda}_t \mathbf{V}_{p,:}^T (\frac{d}{d_t}\mathbf{s}_{t-1}^T)\|_{S^1} &\le \|\mathbf{\Lambda}_t \mathbf{V}_{p,:}^T\|_2 \|\mathbf{P}_{t-1} vec(\Delta \mathbf{W})\|_2 \\
\le \|\lambda_t\|_\infty \|\mathbf{V}_{p,:}^T\|_2 \|\Delta \mathbf{W}\|_{S^\infty} &\sqrt{H} \max_p (\|Mtr(\mathbf{P}_{t-1,p})\|_{S^1})
\end{aligned}
$$

3. Upper bound of $Mtr(\mathbf{V}_{p,:} \frac{d}{d_t} \mathbf{\Lambda}_t \mathbf{W} \mathbf{P}_{t-1})$

- We first estimate the upper bound of each element and then estimate the upper bound of Schatten-1 norm. Let $\mathbf{M} = Mtr(\mathbf{V}_{p,:} \frac{d}{d_t} \mathbf{\Lambda}_t \mathbf{W} \mathbf{P}_{t-1})$. For example, we can bound $\mathbf{M}_{ij}$ by:

$$
\begin{aligned}
\mathbf{M}_{ij} = (\frac{d}{d_t}\lambda_t^T) \odot \mathbf{V}_{p,:} \mathbf{W} \mathbf{P}_{t-1,:,iH+j} &\le \|(\frac{d}{d_t}\lambda_t^T) \odot \mathbf{V}_{p,:} \mathbf{W}\|_2 \|\mathbf{P}_{t-1,:.iH+j}\|_2 \\
&\le \|\frac{d}{d_t}\lambda_t^T\|_2 \|\mathbf{V}_{p,:}\|_\infty \|\mathbf{W}\|_{S^\infty} \|\mathbf{P}_{t-1,:.iH+j}\|_2 \\
\le \|\lambda_t''\|_\infty \|\Delta \mathbf{W}\|_{S^\infty} (\|\mathbf{s}_{t-1}^T\|_2 + \sqrt{H} \max_p (\|Mtr(\mathbf{W}_{p,:} \mathbf{P}_{t-1})\|_{S^1})&\|\mathbf{V}_{p,:}\|_\infty \|\mathbf{W}\|_{S^\infty} \|\mathbf{P}_{t-1,:.iH+j}\|_2
\end{aligned}
$$

- Now we estimate a upper bound of each element of $\mathbf{M}$. Note that Schatten-1 norm and element-wise-1 norm are equivalent, $\frac{1}{H}\|\mathbf{M}\| \le \|\mathbf{M}\|_{S^\infty} \le \|\mathbf{M}\|$, so we can approximately think that a matrix whose elements are all greater than other one has a greater Schatten-1 norm.

  $$\|\mathbf{M}\|_{S^1} \le \|\lambda_t''\|_\infty \|\Delta \mathbf{W}\|_{S^\infty} (\|\mathbf{s}_{t-1}^T\|_2 + \sqrt{H} \max_p (\|Mtr(\mathbf{W}_{p,:} \mathbf{P}_{t-1})\|_{S^1}) \|\mathbf{V}_{p,:}\|_\infty \|\mathbf{W}\|_{S^\infty} \|\bar{\mathbf{P}}_{t-1}\|_{S^1}$$

  where $\bar{\mathbf{P}}_{t-1}(i,j) = \|\mathbf{P}_{t-1}(:, iH+j)\|_2$.

4. Upper bound of $Mtr(\mathbf{V}_{p,:} \mathbf{\Lambda}_t \Delta \mathbf{W} \mathbf{P}_{t-1})$

- Very similar to above, first estimate the upper bound of each element.

$$\mathbf{M}_{ij} = \lambda_t^T \odot \mathbf{V}_{p,:} \Delta \mathbf{W} \mathbf{P}_{t-1,:,iH+j} \le \|\lambda_t\|_\infty \|\mathbf{V}_{p,:}\|_2 \|\Delta \mathbf{W}\|_{S^\infty} \|\mathbf{P}_{t-1,:,iH+j}\|_2$$

- So we can get the upper bound:

$$\|\mathbf{M}\|_{S^1} \le \|\lambda_t\|_\infty \|\mathbf{V}_{p,:}\|_2 \|\Delta \mathbf{W}\|_{S^\infty} \|\bar{\mathbf{P}}_{t-1}\|_{S^1}$$

Loss Bound with Respect to U

1. The upper bound of $\|\mathbf{V}\mathbf{s}_t(\mathbf{U}^k + \Delta \mathbf{U}) - \mathbf{V}\mathbf{s}_t(\mathbf{U}^k)\|_\infty^2$ to update $\mathbf{U}$.

   1. Linking equation:

$$\frac{\partial[\mathbf{V}\mathbf{s}_t(\mathbf{U})]_p}{\partial \mathbf{U}_{ij}} = \sum_{k=1}^{H} \mathbf{V}_{pk} \frac{\partial[\mathbf{s}_t(\mathbf{U})]_k}{\partial \mathbf{U}_{ij}}$$

$$\frac{\partial[\mathbf{s}_t(\mathbf{U})]_p}{\partial \mathbf{U}_{ij}} = \sigma'(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{s}_{t-1})_p[\mathbf{x}_{t,j}\delta_{ip} + \sum_{k=1}^{H} \mathbf{W}_{pk} \frac{\partial[\mathbf{s}_{t-1}(\mathbf{U})]_k}{\partial \mathbf{U}_{ij}}]$$

Similarly, we have the matrix form:

$$\frac{\partial[\mathbf{V}\mathbf{s}_t(\mathbf{U})]_p}{\partial \mathbf{U}} = Mtr(\mathbf{V}_{p,:}\mathbf{Q}_t)$$

$$\mathbf{Q}_t = \mathbf{\Lambda}_t(\bar{\mathbf{X}}_t + \mathbf{W}\mathbf{Q}_{t-1})$$

Similar to updating $\mathbf{W}$, $\mathbf{Q}_t$ is a $H \times HN$ matrix where $\mathbf{Q}_t(p, iN+j) = \frac{\partial[\mathbf{s}_t(\mathbf{U})]_p}{\partial \mathbf{U}_{ij}}$, $\bar{\mathbf{X}}_t$ is a sparse matrix of the same size where $\bar{\mathbf{X}}_t(i, iN+j) = \mathbf{x}_t(j)$ and $\mathbf{\Lambda}$ is a diagonal derivation matrix.

2. We can now get the upper bound:

$$\begin{aligned}
[\mathbf{V}\mathbf{s}_t(\mathbf{U}^k + \Delta\mathbf{U}) - \mathbf{V}\mathbf{s}_t(\mathbf{U}^k)]_p &= \int_0^1 tr[\frac{\partial[\mathbf{V}\mathbf{s}_t(\mathbf{U})]_p}{\partial \mathbf{U}}|_{\mathbf{U}=\mathbf{U}^k+t\Delta\mathbf{U}}\Delta\mathbf{U}]d_t \\
&= \int_0^1 tr[Mtr(\mathbf{V}_{p,:}\mathbf{\Lambda}_t(\bar{\mathbf{X}}_t + \mathbf{W}\mathbf{Q}_{t-1}))|_{\mathbf{U}=\mathbf{U}^k+t\Delta\mathbf{U}}\Delta\mathbf{U}]d_t
\end{aligned}$$

Approximate this equation at $t = 0$:

$$\begin{aligned}
[\mathbf{V}\mathbf{s}_t(\mathbf{U}^k + \Delta\mathbf{U}) - \mathbf{V}\mathbf{s}_t(\mathbf{U}^k)]_p &\le \max_x \sigma'(x)\{\mathbf{V}_{p,:}\Delta\mathbf{U}\mathbf{x}_t + tr[Mtr(\mathbf{V}_{p,:}\mathbf{W}^k\mathbf{Q}_{t-1})\Delta\mathbf{U}]\} \\
&\le \max_x \sigma'(x)\{\|\mathbf{V}_{p,:}\|_2\|\Delta\mathbf{U}\|_{S^\infty}\|\mathbf{x}_t\|_2 + \|Mtr(\mathbf{V}_{p,:}\mathbf{W}^k\mathbf{Q}_{t-1})\|_{S^1}\|\Delta\mathbf{U}\|_{S^\infty}\} \\
&= \max_x \sigma'(x)\|\Delta\mathbf{U}\|_{S^\infty}\{\|\mathbf{V}_{p,:}\|_2\|\mathbf{x}_t\|_2 + \|Mtr(\mathbf{V}_{p,:}\mathbf{W}^k\mathbf{Q}_{t-1})\|_{S^1}\}
\end{aligned}$$

2. The upper bound of $\|\mathbf{V}\mathbf{s}_t(\mathbf{U}^k + \Delta\mathbf{U}) - \mathbf{V}\mathbf{s}_t(\mathbf{U}^k) - \langle \mathbf{V}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k), \Delta\mathbf{U}\rangle\|_\infty$

1. Analyze the p-th element

$$\begin{aligned}
&[\mathbf{V}\mathbf{s}_t(\mathbf{U}^k + \Delta\mathbf{U}) - \mathbf{V}\mathbf{s}_t(\mathbf{U}^k) - \langle \mathbf{V}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k), \Delta\mathbf{U}\rangle]_p \\
&= \int_0^1 tr([\mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k + t\Delta\mathbf{U}) - \mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k)]\Delta\mathbf{U})d_t \\
&\le \|\Delta\mathbf{U}\|_{S^\infty}\|\int_0^1 \mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k + t\Delta\mathbf{U}) - \mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k)\|_{S^1}d_t
\end{aligned}$$

2. Similarly, focus on the last integration part:

$$\int_0^1 \mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k + t\Delta\mathbf{U}) - \mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k)$$
$$\simeq \int_0^1 \frac{d}{d_t}\mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k + t\Delta\mathbf{U})|_{t=0} t d_t$$
$$= \frac{1}{2}\frac{d}{d_t}\mathbf{V}_{p,:}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k + t\Delta\mathbf{U})|_{t=0}$$
$$= \frac{1}{2}\frac{d}{d_t}(\lambda_t \odot \mathbf{V}_{p,:}^T\mathbf{x}_t^T + Mtr(\lambda_t \odot \mathbf{V}_{p,:}\mathbf{W}\mathbf{Q}_{t-1}))|_{t=0}$$
$$\simeq \frac{1}{2}(\frac{d}{d_t}\lambda_t \odot \mathbf{V}_{p,:}^T\mathbf{x}_t^T + Mtr(\frac{d}{d_t}\lambda_t \odot \mathbf{V}_{p,:}\mathbf{W}\mathbf{Q}_{t-1}))$$

In this situation, we have ignored the second derivation part.

3. Upper bound of $\|\frac{d}{d_t}\lambda_t \odot \mathbf{V}_{p,:}^T\mathbf{x}_t^T\|_{S^1}$

   - Let λ_t'' be a column vector of $\sigma''(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{s}_{t-1})$:

$$\frac{d}{d_t}\lambda_t = \lambda_t'' \odot [(\bar{\mathbf{X}}_t + \mathbf{W}\mathbf{Q}_{t-1})vec(\Delta\mathbf{U})] = \lambda'' \odot (\Delta\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{Q}_{t-1}vec(\Delta\mathbf{U}))$$

$$\|(\frac{d}{d_t}\lambda_t) \odot \mathbf{V}_{p,:}^T\mathbf{x}_t^T\|_{S^1} \leq \|(\frac{d}{d_t}\lambda_t) \odot \mathbf{V}_{p,:}^T\|_{S^\infty}\|\mathbf{x}_t^T\|_{S^1} = \|(\frac{d}{d_t}\lambda_t) \odot \mathbf{V}_{p,:}^T\|_2\|\mathbf{x}_t^T\|_2$$
$$\leq \|\lambda_t'' \odot (\Delta\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{Q}_{t-1}vec(\Delta\mathbf{U}))\|_2\|\mathbf{V}_{p,:}^T\|_\infty\|\mathbf{x}_t^T\|_2$$
$$\leq \|\lambda_t''\|_\infty\|\mathbf{V}_{p,:}^T\|_\infty\|\mathbf{x}_t^T\|_2(\|\Delta\mathbf{U}\mathbf{x}_t\|_2 + \|\mathbf{W}\mathbf{Q}_{t-1}vec(\Delta\mathbf{U}))\|_2)$$
$$\leq \|\lambda_t''\|_\infty\|\mathbf{V}_{p,:}^T\|_\infty\|\mathbf{x}_t^T\|_2(\|\Delta\mathbf{U}\|_{S^\infty}\|\mathbf{x}_t^2\|_2 + \sqrt{H}\max_p(\|Mtr(\mathbf{W}_{p,:}\mathbf{Q}_{t-1})\|_{S^1})\|\Delta\mathbf{U}\|_{S^\infty})$$
$$= \|\lambda_t''\|_\infty\|\mathbf{V}_{p,:}^T\|_\infty\|\mathbf{x}_t^T\|_2\|\Delta\mathbf{U}\|_{S^\infty}(\|\mathbf{x}_t^2\|_2 + \sqrt{H}\max_p(\|Mtr(\mathbf{W}_{p,:}\mathbf{Q}_{t-1})\|_{S^1})$$

4. Upper bound of $Mtr(\frac{d}{d_t}\lambda_t \odot \mathbf{V}_{p,:}\mathbf{W}\mathbf{Q}_{t-1})$

   - Let $\mathbf{N} = Mtr(\frac{d}{d_t}\lambda_t \odot \mathbf{V}_{p,:}\mathbf{W}\mathbf{Q}_{t-1})$, we have:

$$\mathbf{N}_{ij} = \frac{d}{d_t}\lambda_t \odot \mathbf{V}_{p,:}\mathbf{W}\mathbf{Q}_{t-1,:,iN+j} \leq \|(\frac{d}{d_t}\lambda_t^T) \odot \mathbf{V}_{p,:}\mathbf{W}\|_2\|\mathbf{Q}_{t-1,:.iN+j}\|_2$$
$$\leq \|\frac{d}{d_t}\lambda_t^T\|_2\|\mathbf{V}_{p,:}\|_\infty\|\mathbf{W}\|_{S^\infty}\|\mathbf{Q}_{t-1,:.iN+j}\|_2$$
$$\leq \|\lambda_t''\|_\infty\|\Delta\mathbf{U}\|_{S^\infty}(\|\mathbf{x}_t\|_2 + \sqrt{H}\max_p(\|Mtr(\mathbf{W}_{p,:}\mathbf{Q}_{t-1})\|_{S^1})\|\mathbf{V}_{p,:}\|_\infty\|\mathbf{W}\|_{S^\infty}\|\mathbf{Q}_{t-1,:.iN+j}\|_2$$

Comments

- **Now, we have bounded the loss function based on Schatten-∞ norm. I am not sure it is completely right. However, it is obvious to find the following flaws.(Fonts in purple)**

  1. I have ignored the second derivative part while bounding $\|\mathbf{V}\mathbf{s}_t(\mathbf{W}^k + \Delta\mathbf{W}) - \mathbf{V}\mathbf{s}_t(\mathbf{W}^k) - \langle\mathbf{V}\nabla_{\mathbf{W}}\mathbf{s}_t(\mathbf{W}^k), \Delta\mathbf{W}\rangle\|_\infty$ as well as $\|\mathbf{V}\mathbf{s}_t(\mathbf{U}^k + \Delta\mathbf{U}) - \mathbf{V}\mathbf{s}_t(\mathbf{U}^k) - \langle\mathbf{V}\nabla_{\mathbf{U}}\mathbf{s}_t(\mathbf{U}^k), \Delta\mathbf{U}\rangle\|_\infty$, which is the same as what previous paper "Preconditioned Spectral Descent for Deep Learning" do. I have written a very simple script (input and

hidden units are both scalar) to stimulate the what RNN works. It is found that in fine-tuned model second-order derivation is indeed smaller than first-order if the activation is sigmoid, if the acitvation is relu or tanh, the second-order derivation is a bit bigger than the first one(they are of the same magnitute). However, it is only the comparison between $\frac{\partial \mathbf{Q}}{\partial \mathbf{U}}$ v.s $\mathbf{Q}$($\frac{\partial \mathbf{P}}{\partial \mathbf{U}}$ v.s. $\mathbf{P}$). **<u>At present, I can not prove</u>** $||Mtr(\mathbf{V}_{p,:}\mathbf{\Lambda}_t\mathbf{W}\frac{d\mathbf{P}_{t-1}}{dt})||_{S^1}$ **<u>is significantly smaller than (but it can't not much bigger than the first-order part) other parts of the derivative. The consequence will be catastrophic if it is not true, because computing second-order derivative is very expensive.</u>**

2. It seems that matrix operation is not enough to solve this optimization problem perfectly. As we see, $\frac{\partial \mathbf{s}_t}{\partial \mathbf{W}}$ (for $\mathbf{U}$ is similar) is the core element of linking equation which will be used repeatly for optimization. **<u>2D matrix is not powerfull enough to address this question, I think 3D tensor is better.</u>** However, I do not have enough mathematical tools to solve 3D operation, so I project the last two dimension into one, resulting $H \times H^2$ matrix. It is expedient and causes much problems, such as repeatly reshaping matrix. Some 'ugly' parts like $\sqrt{H}\ \max_p(||Mtr(\mathbf{W}_{p,:}\mathbf{P}_{t-1})||_{S^1})$ arises from this. **<u>I guess there exists a tighter and nicer upper bound if we use 3D tensor instead of 2D matrix.</u>**
3. I assume that if matrix $\mathbf{A}$'s every element is bigger than $\mathbf{B}$'s corresponding element, then $||\mathbf{A}||_{S^1} \leq ||\mathbf{B}||_{S^1}$. It is based on the fact that Schatten-1 norm and element-wise 1-norm are equivalent.($\frac{1}{\sqrt{mn}}||\mathbf{X}||_\infty \leq ||\mathbf{X}||_{S^1} \leq ||\mathbf{X}||_\infty$ for matrix $\mathbf{X}$ of size $n \times m$) This relation is strict if we use Schatten-2 norm, but S-2 norm is very expensive to compute.(It can be approximated by **super power** method) So, we approximate it by useing S-1 norm. (This problem is caused also by using 2D matrix instead of 3D tensor.)
4. $\mathbf{U}$ and $\mathbf{W}$ are bounded by Schatten-∞ norm, while $\mathbf{V}$ are bounded by element-wise ∞-norm. I will change it to Schatten-∞ norm later.

$$E = \frac{1}{M} \sum_{t=1}^{M} (\log \sum_{j=1}^{K} e^{V_{j,:} \cdot S_t}) - y_t^T V S_t .$$

$$f_{-t}(\theta) = \log \sum_{j=1}^{K} e^{V_{j,:} S_t} \qquad g_{-t}(\theta) = - y_t^T V S_t$$

Let $r = VS_t(\phi) - VS_t(\theta) - \langle \nabla_\theta VS_t(\theta), \phi - \theta \rangle$

Lemma 1 : $f_{-t}(\phi) \leq f_{-t}(\theta) + \langle \nabla_\theta f_{-t}(\theta), \phi - \theta \rangle + \frac{1}{2} \| VS_t(\phi) - VS_t(\theta) \|_\infty^2 + r$

We already know, for function $lse(x) = \log \sum_{j=1}^{J} e^{x_j}$, we have

$$lse(x_2) \leq lse(x_1) + \langle \nabla_{x_1} lse(x_1), x_2 - x_1 \rangle + \frac{1}{2} \| x_2 - x_1 \|_\infty^2$$

so. $f_{-t}(\phi) \leq f_{-t}(\theta) + \langle \nabla_{VS_t} f_{-t}(\theta), VS_t(\phi) - VS_t(\theta) \rangle + \frac{1}{2} \| VS_t(\phi) - VS_t(\theta) \|_\infty^2$

where $\langle \nabla_{VS_t} f_{-t}(\theta), VS_t(\phi) - VS_t(\theta) \rangle = \nabla_{VS_t} f_{-t}(\theta) [r + \langle \nabla_\theta VS_t(\theta), \phi - \theta \rangle]$

$= r \cdot \nabla_{VS_t(\theta)} f_{-t}(\theta) + \langle \nabla_{VS_t(\theta)} f_{-t}(\theta) \nabla_\theta VS_t(\theta), \phi - \theta \rangle$

$= r \cdot \nabla_{VS_t(\theta)} f_{-t}(\theta) + \langle \nabla_\theta f_{-t}(\theta), \phi - \theta \rangle$

It can be proved that $\nabla_{VS_t(\theta)} f_{-t}(\theta) \leq 1$ (softmax)

so $f_{-t}(\phi) \leq f_{-t}(\theta) + \langle \nabla_\theta f_{-t}(\theta), \phi - \theta \rangle + \frac{1}{2} \| VS_t(\phi) - VS_t(\theta) \|_\infty^2 + \underbrace{\max r}_{\| r \|_\infty}$

Lemma 2. $g_{-t}(\phi) \leq g_{-t}(\theta) + \langle \nabla_\theta g_{-t}(\theta), \phi - \theta \rangle + \max r$

$g_{-t}(\theta)$ is linear to $VS_t(\theta)$, so.

$g_{-t}(\phi) = g_{-t}(\theta) + \langle \nabla_{VS_t(\theta)} g_{-t}(\theta), VS_t(\phi) - VS_t(\theta) \rangle$

similarly, $\langle \nabla_{VS_t(\theta)} g_{-t}(\theta), VS_t(\phi) - VS_t(\theta) \rangle = \langle \nabla_\theta g_{-t}(\theta), \phi - \theta \rangle + r \nabla_{VS_t(\theta)} g(\theta)$

$\leq \langle \nabla_\theta g_{-t}(\theta), \phi - \theta \rangle + \max r$

so $g_{-t}(\phi) \leq g_{-t}(\theta) + \langle \nabla_\theta g_{-t}(\theta), \phi - \theta \rangle + \underbrace{\max r}_{\| r \|_\infty}$