Supplementary Material for Language Modeling with Power Low Rank Ensembles

Ankur P. Parikh, Avneesh Saluja, Chris Dyer, Eric P. Xing
Carnegie Mellon University

[ ]

The primary purpose of the supplementary material is to provide a proof of Lemma 4. We also show that Lemma 4 extends to $n > 2$.

1 Proof of Lemma 4

Lemma 1. *Let $P_{plre}(w_i|w_{i-1})$ indicate the PLRE smoothed conditional probability and $\widehat{P}(w)$ indicate the MLE probability of w. Then,*

$$\widehat{P}(w) = \sum_{w_{i-1}} P_{plre}(w_i|w_{i-1})\widehat{P}(w_{i-1}) \tag{1}$$

Proof. Assume the following more general form where multiple low rank matrices can be used i.e.:

$$P_{\text{plre}}(w_i|w_{i-1}) = P^{\text{alt}}_{\boldsymbol{D}_0}(w_i|w_{i-1}) + \gamma_0(w_{i-1})\Bigg(\boldsymbol{Z}^{(\rho_1,\kappa_1)}_{\boldsymbol{D}_1}(w_i|w_{i-1}) + ...$$
$$+ \gamma_{\eta-1}(w_{i-1})\Bigg(\boldsymbol{Z}^{(\rho_\eta,\kappa_\eta)}_{\boldsymbol{D}_\eta}(w_i|w_{i-1}) + \gamma_\eta(w_{i-1})\Bigg(\boldsymbol{Z}^{(\rho_{\eta+1}=0,\kappa_{\eta+1}=1)}(w_i|w_{i-1})\Bigg)\Bigg)...\Bigg) \tag{2}$$

where we note that $\boldsymbol{Z}^{(\rho_{\eta+1}=0,\kappa_{\eta+1}=1)}(w_i|w_{i-1})$ is equivalent to $P^{\text{alt}}(w_i)$. It is assumed that $1 \geq \rho_0 \geq ....\rho_{\eta+1} = 0$.

First unroll the recursion and rewrite $P_{\text{plre}}(w_i|w_{i-1})$ as:

$$P_{\text{plre}}(w_i|w_{i-1}) = \sum_{j=0}^{\eta+1} \gamma_{0:j-1}(w_{i-1})\boldsymbol{Z}^{(\rho_j,\kappa_j)}_{\boldsymbol{D}_j}(w_i|w_{i-1})$$

where $\gamma_{0:j-1}(w_{i-1}) = \prod_{h=0}^{j-1}\gamma_h(w_{i-1})$ and $\gamma_{0:-1}(w_{i-1}) = 1$. Note that $P_{\text{pwr}}(w_i|w_{i-1})$ can be written in the same way.

$$P_{pwr}(w_i|w_{i-1}) = \sum_{j=0}^{\eta} \gamma_{0:j}(w_{i-1})\boldsymbol{Y}^{(\rho_j)}_{\boldsymbol{D}_j}(w_i|w_{i-1}) \tag{3}$$

Note that $P_{\text{pwr}}(w_i|w_{i-1})$ already satisfies the marginal constraint i.e.

$$\widehat{P}(w) = \sum_{w_{i-1}} P_{\text{pwr}}(w_i|w_{i-1})\widehat{P}(w_{i-1}) \tag{4}$$

because the discounts were chosen such that $P_{\text{pwr}}(w_i|w_{i-1}) = \widehat{P}(w_i|w_{i-1})$

Thus it suffices to show that for all $j = 0, ..., \eta + 1$:

$$\sum_{w_{i-1}} \widehat{P}(w_{i-1})\gamma_{0:j-1}(w_{i-1})\boldsymbol{Y}^{(\rho_j)}_{\boldsymbol{D}_j}(w_i|w_{i-1}) = \sum_{w_{i-1}} \widehat{P}(w_{i-1})\gamma_{0:j-1}(w_{i-1})\boldsymbol{Z}^{(\rho_j,\kappa_j)}_{\boldsymbol{D}_j}(w_i|w_{i-1}) \tag{5}$$

The statement above is trivially true when $j = 0$. For all other cases, note that due to the way we have set the discounts, $\gamma_{0:j-1}$ takes a special form:

$$\prod_{h=0}^{j-1} \gamma_h(w_{i-1}) = \frac{d_* \sum_i c_{i,i-1}^{\rho_1}}{\sum_i c_{i,i-1}^{\rho_0}} \frac{d_* \sum_i c_{i,i-1}^{\rho_2}}{\sum_i c_{i,i-1}^{\rho_1}} \cdots \frac{d_* \sum_i c_{i,i-1}^{\rho_j}}{\sum_i c_{i,i-1}^{\rho_{j-1}}}$$
$$= \frac{(d_*)^j \sum_i c_{i,i-1}^{\rho_j}}{\sum_i c_{i,i-1}} \tag{6}$$

Using this form in Eq. 5 and simplifying yields:

$$\sum_{w_{i-1}} \left(\sum_i c_{i,i-1}^{\rho_j} \right) \boldsymbol{Y}_{\boldsymbol{D}_j}^{(\rho_j)}(w_i|w_{i-1}) = \sum_{w_{i-1}} \left(\sum_i c_{i,i-1}^{\rho_j} \right) \boldsymbol{Z}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1})$$

which is equivalent to requiring that

$$\sum_{w_{i-1}} \boldsymbol{Y}_{\boldsymbol{D}_j}^{(\rho_j)}(w_i, w_{i-1}) = \sum_{w_{i-1}} \boldsymbol{Z}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i, w_{i-1}) \tag{7}$$

which holds because rank minimization under gKL preserves row and column sums. □

2 Generalization to $n > 2$

Theorem 1. *Let $P_{plre}(w_i|w_{i-n+1}^{i-1})$ indicate the PLRE smoothed conditional probability and $\widehat{P}(w)$ indicate the MLE probability of w. Then,*

$$\widehat{P}(w) = \sum_{w_{i-n+1}^{i-1}} P_{plre}(w_i|w_{i-n+1}^{i-1}) \widehat{P}(w_{i-n+1}^{i-1}) \tag{8}$$

Proof. Recall that,

$$P_{\text{plre}}(w_i|w_{i-n+1}^{i-1}) = P_{\boldsymbol{D}_0}^{\text{alt}}(w_i|w_{i-n+1}^{i-1})$$
$$+ \gamma_0(w_{i-n+1}^{i-1}) \Bigg(\boldsymbol{Z}_{\boldsymbol{D}_1}^{(\rho_1,\kappa_1)}(w_i|w_{i-n+1}^{i-1}) + \dots..$$
$$+ \gamma_{\eta-1}(w_{i-n+1}^{i-1}) \Bigg(\boldsymbol{Z}_{\boldsymbol{D}_\eta}^{(\rho_\eta,\kappa_\eta)}(w_i|w_{i-n+1}^{i-1})$$
$$+ \gamma_\eta(w_{i-n+1}^{i-1}) \Bigg(P_{\text{plre}}(w_i|w_{i-n+2}^{i-1}) \Bigg) \Bigg) \dots \Bigg) \tag{9}$$

Define,

$$P_{\text{pwr}}(w_i|w_{i-n+1}^{i-1}) = P_{\boldsymbol{D}_0}^{\text{alt}}(w_i|w_{i-n+1}^{i-1})$$
$$+ \gamma_0(w_{i-n+1}^{i-1}) \Bigg(\boldsymbol{Y}_{\boldsymbol{D}_1}^{(\rho_1,\kappa_1)}(w_i|w_{i-n+1}^{i-1}) + \dots..$$
$$+ \gamma_{\eta-1}(w_{i-n+1}^{i-1}) \Bigg(\boldsymbol{Y}_{\boldsymbol{D}_\eta}^{(\rho_\eta,\kappa_\eta)}(w_i|w_{i-n+1}^{i-1})$$
$$+ \gamma_\eta(w_{i-n+1}^{i-1}) \Bigg(P_{\text{pwr}}(w_i|w_{i-n+2}^{i-1}) \Bigg) \Bigg) \dots \Bigg) \tag{10}$$

where with a little abuse of notation

$$\boldsymbol{Y}_{\boldsymbol{D}_j}^{\rho_j}(w_i|w_{i-n'+1}^{i-1}) = \frac{\tilde{c}(w_i, w_{i-n'+1}^{i-1})^{\rho_j} - \boldsymbol{D}_j(w_i, w_{i-n'+1}^{i-1})}{\sum_{w_i} \tilde{c}(w_i, w_{i-n'+1}^{i-1})^{\rho_j}} \tag{11}$$

and

$$\tilde{c}(w_i, w_{i-n'+1}^{i-1}) = \begin{cases} c(w_i, w_{i-n'+1}^{i-1}), \text{ if } n' = n \\ \\ N_-(w_{i-n'+1}^{i}) \quad \text{if } n' < n \end{cases}$$

Furthermore, define

$$\begin{aligned} P_{\text{pwr}}^{\text{terms}}(w_i|w_{i-n'+1}^{i-1}) &= P_{\boldsymbol{D}_0}^{\text{alt}}(w_i|w_{i-n'+1}^{i-1}) \\ &+ \gamma_0(w_{i-n'+1}^{i-1})\Bigg(\boldsymbol{Y}_{\boldsymbol{D}_1}^{(\rho_1,\kappa_1)}(w_i|w_{i-n'+1}^{i-1}) + ... \\ &+ \gamma_{\eta-1}(w_{i-n'+1}^{i-1})\Bigg(\boldsymbol{Y}^{(\rho_\eta,\kappa_\eta)}(w_i|w_{i-n'+1}^{i-1})\Bigg)...\Bigg) \end{aligned} \tag{12}$$

Note that because of the way the discounts are computed in Algorithm 1,

$$P_{\text{pwr}}^{\text{terms}}(w_i|w_{i-n'+1}^{i-1}) = P^{\text{alt}}(w_i|w_{i-n'+1}^{i-1}) \tag{13}$$

for all $n' \leq n$.

As a result, (for some choice of discount)

$$P_{\text{pwr}}(w_i|w_{i-n+1}^{i-1}) = P_{\text{kn}}(w_i|w_{i-n+1}^{i-1}) \tag{14}$$

Since, we know that Kneser Ney satisfies the marginal constraint (Chen and Goodman, 1999) this implies that,

$$\widehat{P}(w) = \sum_{w_{i-n+1}^{i-1}} P_{\text{pwr}}(w_i|w_{i-n+1}^{i-1})\widehat{P}(w_{i-n+1}^{i-1}) \tag{15}$$

Thus, all we have to do is prove that

$$\sum_{w_{i-n+1}^{i-1}} P_{\text{pwr}}(w_i|w_{i-n+1}^{i-1})\widehat{P}(w_{i-n+1}^{i-1}) = \sum_{w_{i-n+1}^{i-1}} P_{\text{plre}}(w_i|w_{i-n+1}^{i-1})\widehat{P}(w_{i-n+1}^{i-1}) \tag{16}$$

Now, we follow the same argument as with $n = 2$ (i.e. unrolling the recursion and applying the fact that gKL preserves row/column sums).

For notational simplicity assume that $n = 3$. Then, we can write $P_{\text{pwr}}(w_i|w_{i-n+1}^{i-1})$ as:

$$P_{\text{pwr}}(w_i|w_{i-2}^{i-1}) = \sum_{j=0}^{\eta} \gamma_{0:j-1}(w_{i-2}^{i-1})\boldsymbol{Y}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-2}^{i-1}) + \sum_{j=0}^{\eta+1} \gamma_{0:\eta}(w_{i-2}^{i-1})\gamma_{0:j-1}(w_{i-1})\boldsymbol{Y}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1}) \tag{17}$$

where $\gamma_{0:-1}(w_{i-2}^{i-1}) = 1$ and

$$\begin{aligned} \gamma_{0:j-1}(w_{i-2}^{i-1}) = \prod_{h=0}^{j-1} \gamma_h(w_{i-2}^{i-1}) &= \frac{d_* \sum_i \tilde{c}_{i,i-1,i-2}^{\rho_1}}{\sum_i \tilde{c}_{i,i-1,i-2}^{\rho_0}} \frac{d_* \sum_i \tilde{c}_{i,i-1,i-2}^{\rho_2}}{\sum_i \tilde{c}_{i,i-1,i-2}^{\rho_1}} \cdots \frac{d_* \sum_i \tilde{c}_{i,i-1,i-2}^{\rho_j}}{\sum_i \tilde{c}_{i,i-1,i-2}^{\rho_{j-1}}} \\ &= \frac{(d_*)^j \sum_i \tilde{c}_{i,i-1,i-2}^{\rho_j}}{\sum_i \tilde{c}_{i,i-1,i-2}} \end{aligned} \tag{18}$$

Here $\tilde{c}_{i,i-1,i-2}$ is shorthand for $\tilde{c}(w_i, w_{i-2}^{i-1})$.

Similarly, $\gamma_{0:-1}(w_{i-1}) = 1$ and

$$\gamma_{0:j-1}(w_{i-1}) = \prod_{h=0}^{j-1} \gamma_h(w_{i-1}) = \frac{d_* \sum_i \tilde{c}_{i,i-1}^{\rho_1}}{\sum_i \tilde{c}_{i,i-1}^{\rho_0}} \frac{d_* \sum_i \tilde{c}_{i,i-1}^{\rho_2}}{\sum_i \tilde{c}_{i,i-1}^{\rho_1}} \dots \frac{d_* \sum_i \tilde{c}_{i,i-1}^{\rho_j}}{\sum_i \tilde{c}_{i,i-1}^{\rho_{j-1}}}$$
$$= \frac{(d_*)^j \sum_i \tilde{c}_{i,i-1}^{\rho_j}}{\sum_i \tilde{c}_{i,i-1}} \tag{19}$$

(Again, it is assumed that $1 \geq \rho_0 \geq \dots . \rho_{\eta+1} = 0$.)

Analogously,

$$P_{\text{plre}}(w_i|w_{i-2}^{i-1}) = \sum_{j=0}^{\eta} \gamma_{0:j-1}(w_{i-2}^{i-1}) \boldsymbol{Z}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-2}^{i-1}) + \sum_{j=0}^{\eta+1} \gamma_{0:\eta}(w_{i-2}^{i-1}) \gamma_{0:j-1}(w_{i-1}) \boldsymbol{Z}_{\boldsymbol{D}_j}^{(\rho_h,\kappa_j)}(w_i|w_{i-1}) \tag{20}$$

Now for any trigram term we prove that

$$\sum_{w_{i-2}^{i-1}} \gamma_{0:j-1}(w_{i-2}^{i-1}) \boldsymbol{Y}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-2}^{i-1}) \widehat{P}(w_{i-2}^{i-1}) = \sum_{w_{i-2}^{i-1}} \gamma_{0:j-1}(w_{i-2}^{i-1}) \boldsymbol{Z}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-2}^{i-1}) \widehat{P}(w_{i-2}^{i-1}) \tag{21}$$

Plugging in the definition of $\gamma_{0:j-1}$ and simplifying gives

$$\sum_{w_{i-2}^{i-1}} (\sum_i \tilde{c}_{i,i-1,i-2}^{\rho_j}) \boldsymbol{Y}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-2}^{i-1}) = \sum_{w_{i-2}^{i-1}} (\sum_i \tilde{c}_{i,i-1,i-2}^{\rho_j}) \boldsymbol{Z}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-2}^{i-1}) \tag{22}$$

which is equivalent to

$$\sum_{w_{i-2}^{i-1}} \boldsymbol{Y}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i, w_{i-2}^{i-1}) = \sum_{w_{i-2}^{i-1}} \boldsymbol{Z}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i, w_{i-2}^{i-1}) \tag{23}$$

which holds because of the definition of $\boldsymbol{Z}$ and the fact that rank minimization under gKL preserves row/column sums.

Now consider any bigram term. We seek to show that:

$$\sum_{w_{i-2}^{i-1}} \gamma_{0:\eta}(w_{i-2}^{i-1}) \gamma_{0:j-1}(w_{i-1}) \boldsymbol{Y}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1}) \widehat{P}(w_{i-2}^{i-1})$$
$$= \sum_{w_{i-2}^{i-1}} \gamma_{0:\eta}(w_{i-2}^{i-1}) \gamma_{0:j-1}(w_{i-1}) \boldsymbol{Z}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1}) \widehat{P}(w_{i-2}^{i-1}) \tag{24}$$

Substituting definition of $\gamma_{0:\eta}(w_{i-2}^{i-1})$ gives

$$\sum_{w_{i-2}^{i-1}} \frac{(d_*)^{\eta+1} \sum_i \tilde{c}_{i,i-1,i-2}^{\rho_{\eta+1}}}{\sum_i \tilde{c}_{i,i-1,i-2}} \gamma_{0:j-1}(w_{i-1}) \boldsymbol{Y}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1}) \widehat{P}(w_{i-2}^{i-1})$$
$$= \sum_{w_{i-2}^{i-1}} \frac{(d_*)^{\eta+1} \sum_i \tilde{c}_{i,i-1,i-2}^{\rho_{\eta+1}}}{\sum_i \tilde{c}_{i,i-1,i-2}} \gamma_{0:j-1}(w_{i-1}) \boldsymbol{Z}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1}) \widehat{P}(w_{i-2}^{i-1}) \tag{25}$$

Simplifying and pushing in the sum over w_{i-2} gives,

$$\sum_{w_{i-1}} (\sum_{i,i-2} \tilde{c}_{i,i-1,i-2}^{\rho_{\eta+1}}) \gamma_{0:j-1}(w_{i-1}) \boldsymbol{Y}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1}) = \sum_{w_{i-1}} (\sum_{i,i-2} \tilde{c}_{i,i-1,i-2}^{\rho_{\eta+1}}) \gamma_{0:j-1}(w_{i-1}) \boldsymbol{Z}_{\boldsymbol{D}_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1}) \tag{26}$$

Note that since $\rho_{\eta+1} = 0$, $\sum_{i,i-2} \tilde{c}^{\rho_{\eta+1}=0}_{i,i-1,i-2} = \sum_i \tilde{c}_{i,i-1}$ (by definition of $\tilde{c}$).

Using this fact and substituting definition of $\gamma_{0:j-1}(w_{i-1})$ gives

$$\sum_{w_{i-1}} (\sum_i \tilde{c}_{i,i-1}) \frac{(d_*)^j \sum_i \tilde{c}^{\rho_j}_{i,i-1}}{\sum_i \tilde{c}_{i,i-1}} \boldsymbol{Y}^{(\rho_j,\kappa_j)}_{\boldsymbol{D}_j}(w_i|w_{i-1}) = \sum_{w_{i-1}} (\sum_i \tilde{c}_{i,i-1}) \frac{(d_*)^j \sum_i \tilde{c}^{\rho_j}_{i,i-1}}{\sum_i \tilde{c}_{i,i-1}} \boldsymbol{Z}^{(\rho_j,\kappa_j)}_{\boldsymbol{D}_j}(w_i|w_{i-1}) \quad (27)$$

Simplifying gives,

$$\sum_{w_{i-1}} \boldsymbol{Y}^{(\rho_j)}_{\boldsymbol{D}_j}(w_i, w_{i-1}) = \sum_{w_{i-1}} \boldsymbol{Z}^{(\rho_j,\kappa_j)}_{\boldsymbol{D}_j}(w_i, w_{i-1}) \quad (28)$$

which holds because rank minimization under KL divergence preserves row and column sums.

□

References

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.