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

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

(1)

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

$$P_{\text{plre}}(w_i|w_{i-1}) = P_{\text{alt}}^{D_0}(w_i|w_{i-1}) + \gamma_0(w_{i-1})\left(Z_{D_1}^{(\rho_1,\kappa_1)}(w_i|w_{i-1}) + \ldots + \gamma_{\eta-1}(w_{i-1})\left(Z_{D_\eta}^{(\rho_\eta,\kappa_\eta)}(w_i|w_{i-1}) + \gamma_{\eta}(w_{i-1})\left(Z_{D_{\eta+1}}^{(\rho_{\eta+1} = 0,\kappa_{\eta+1} = 1)}(w_i|w_{i-1})\right)\right)\right)$$

(2)

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

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

$$P_{\text{plre}}(w_i|w_{i-1}) = \sum_{j=0}^{\eta+1} \gamma_{0:j-1}(w_{i-1})Z_{D_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1})$$

where $\gamma_{0:j-1}(w_{i-1}) = \prod_{k=0}^{j-1} \gamma_k(w_{i-1})$ 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.

$$P_{\text{pwr}}(w_i|w_{i-1}) = \sum_{j=0}^{\eta} \gamma_{0:j}(w_{i-1})Y_{D_j}^{(\rho_j)}(w_i|w_{i-1})$$

(3)

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

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

(4)

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

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

$$\sum_{w_{i-1}} \hat{P}(w_{i-1})\gamma_{0:j-1}(w_{i-1})Y_{D_j}^{(\rho_j)}(w_i|w_{i-1}) = \sum_{w_{i-1}} \hat{P}(w_{i-1})\gamma_{0:j-1}(w_{i-1})Z_{D_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1})$$

(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:

\[
\prod_{h=0}^{j-1} \gamma_h(w_{i-1}) = \frac{d_s \prod_{i} c_{i,i-1}^{\rho_j} d_s \prod_{i} c_{i,i-1}^{\rho_{j-1}} \cdots d_s \prod_{i} c_{i,i-1}^{\rho_1}}{\sum_{i} c_{i,i-1}^{\rho_j} \sum_{i} c_{i,i-1}^{\rho_{j-1}} \cdots \sum_{i} c_{i,i-1}^{\rho_1}}.
\]

Using this form in Eq. 5 and simplifying yields:

\[
\sum_{w_{i-1}} \left(\sum_{i} c_{i,i-1}^{\rho_j}\right) Y_{D_j}^{(\rho_j)}(w_i | w_{i-1}) = \sum_{w_{i-1}} \left(\sum_{i} c_{i,i-1}^{\rho_j}\right) Z_{D_j}^{(\rho_j,\kappa_j)}(w_i | w_{i-1})
\]

which is equivalent to requiring that

\[
\sum_{w_{i-1}} Y_{D_j}^{(\rho_j)}(w_i, w_{i-1}) = \sum_{w_{i-1}} Z_{D_j}^{(\rho_j,\kappa_j)}(w_i, w_{i-1})
\]

which holds because rank minimization under \( gKL \) preserves row and column sums.

\[\square\]

2 Generalization to \( n > 2 \)

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

\[
\hat{P}(w) = \sum_{w_i^{-1}} P_{\text{plre}}(w_i | w_{i-n+1}^{i-1}) \hat{P}(w_{i-n+1}^{i-1})
\]

**Proof.** Recall that,

\[
P_{\text{plre}}(w_i | w_{i-n+1}^{i-1}) = P_{\text{alt}}^{D_0}(w_i | w_{i-n+1}^{i-1})
+ \gamma_0(w_{i-n+1}^{i-1}) \left( Z_{D_1}^{(\rho_{1},\kappa_1)}(w_i | w_{i-n+1}^{i-1}) + \cdots \right)
+ \gamma_{n-1}(w_{i-n+1}^{i-1}) \left( Z_{D_n}^{(\rho_{n},\kappa_n)}(w_i | w_{i-n+1}^{i-1}) \right)
+ \gamma_n(w_{i-n+1}^{i-1}) \left( P_{\text{plre}}(w_i | w_{i-n+2}^{i-1}) \right) \ldots
\]

Define,

\[
P_{\text{pwre}}(w_i | w_{i-n+1}^{i-1}) = P_{\text{alt}}^{D_0}(w_i | w_{i-n+1}^{i-1})
+ \gamma_0(w_{i-n+1}^{i-1}) \left( Y_{D_1}^{(\rho_{1},\kappa_1)}(w_i | w_{i-n+1}^{i-1}) + \cdots \right)
+ \gamma_{n-1}(w_{i-n+1}^{i-1}) \left( Y_{D_n}^{(\rho_{n},\kappa_n)}(w_i | w_{i-n+1}^{i-1}) \right)
+ \gamma_n(w_{i-n+1}^{i-1}) \left( P_{\text{pwre}}(w_i | w_{i-n+2}^{i-1}) \right) \ldots
\]

where with a little abuse of notation

\[
Y_{D_j}^{(\rho_j)}(w_i | w_{i-n+1}^{i-1}) = \frac{\delta(w_i, w_{i-n+1}^{i-1})^{\rho_j} - D_j(w_i, w_{i-n+1}^{i-1})}{\sum_{w_i} \delta(w_i, w_{i-n+1}^{i-1})^{\rho_j}}
\]
and

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

Furthermore, define

\[ P_{\text{terms}}^{\text{alt}}(w_i|w_{i-n'_{1-1}}^{i-1}) = P_{\text{alt}}^D(w_i|w_{i-n'_{1-1}}^{i-1}) + \gamma_0(w_{i-n'_{1-1}}^{i-1}) \left( Y(w_{i-n'_{1-1}}^{i-1}) + \ldots \right) + \gamma_{\eta-1}(w_{i-n'_{1-1}}^{i-1}) \left( Y(w_{i-n'_{1-1}}^{i-1}) \right) \ldots \]  

(12)

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

\[ P_{\text{terms}}^{\text{alt}}(w_i|w_{i-n'_{1-1}}^{i-1}) = P_{\text{alt}}^D(w_i|w_{i-n'_{1-1}}^{i-1}) \]  

(13)

for all \( n' \leq n \).

As a result, (for some choice of discount)

\[ P_{\text{terms}}(w_i|w_{i-n'_{1-1}}^{i-1}) = P_{\text{alt}}(w_i|w_{i-n'_{1-1}}^{i-1}) \]  

(14)

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

\[ \hat{P}(w) = \sum_{w_{i-1-n+1}} P_{\text{terms}}(w_i|w_{i-n'_{1-1}}^{i-1}) \hat{P}(w_{i-n'_{1-1}}^{i-1}) \]  

(15)

Thus, all we have to do is prove that

\[ \sum_{w_{i-1-n+1}} P_{\text{terms}}(w_i|w_{i-n'_{1-1}}^{i-1}) \hat{P}(w_{i-n'_{1-1}}^{i-1}) = \sum_{w_{i-1-n+1}} P_{\text{alt}}(w_i|w_{i-n'_{1-1}}^{i-1}) \hat{P}(w_{i-n'_{1-1}}^{i-1}) \]  

(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 \( P_{\text{terms}}(w_i|w_{i-n'_{1-1}}^{i-1}) \) as:

\[ P_{\text{terms}}(w_i|w_{i-n'_{1-2}}^{i-1}) = \sum_{j=0}^{\eta} \gamma_{0:j-1}(w_{i-2}^{i-1}) Y_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-2}^{i-1}) + \sum_{j=0}^{\eta+1} \gamma_{0: \eta}(w_{i-2}^{i-1}) \gamma_{0:j-1}(w_{i-1}) Y_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}) \]  

(17)

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

\[ \gamma_{0:j-1}(w_{i-2}^{i-1}) = \prod_{h=0}^{j-1} \gamma_h(w_{i-2}^{i-1}) = \frac{d_s \sum_i \tilde{c}^0_{i,i-1,i-2} d_s \sum_i \tilde{c}^0_{i,i-1,i-2} \ldots \sum_i \tilde{c}^0_{i,i-1,i-2}}{\sum_i \tilde{c}^0_{i,i-1,i-2} \sum_i \tilde{c}^0_{i,i-1,i-2} \ldots \sum_i \tilde{c}^0_{i,i-1,i-2}} \]  

(18)

Here \( \tilde{c}_{i,i-1,i-2} \) \( \tilde{c}_{i,i-1,i-2} \) is shorthand for \( \tilde{c}(w_i, w_{i-n'_{1-2}}^{i-1}) \).
Similarly, \( \gamma_{0,-1}(w_{i-1}) = 1 \) and
\[
\gamma_{0,j-1}(w_{i-1}) = \prod_{h=0}^{j-1} \gamma_h(w_{i-1}) = \frac{d_s \sum_i c_{i,j-1}^0 d_s \sum_i c_{i,j-1}^1 \ldots d_s \sum_i c_{i,j-1}^{p-1}}{\sum_i c_{i,j-1}^0 \sum_i c_{i,j-1}^1 \ldots \sum_i c_{i,j-1}^{p-1}} = (d_s)^j \frac{\sum_i c_{i,j-1}^p}{\sum_i c_{i,j-1}^0}
\] (19)

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

Analogously,
\[
P_{\text{ple}}(w_i|w_{i-2}^{i-1}) = \sum_{j=0}^n \gamma_{0,j-1}(w_{i-1}) Y_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-2}^{i-1}) \gamma_{0,j-1}(w_{i-1}) Z_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1})(20)
\]

Now consider any bigram term. We seek to show that:
\[
\sum_{w_{i-1}^{i-2}} \gamma_{0,j-1}(w_{i-1}^{i-2}) Y_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}^{i-2}) \gamma_{0,j-1}(w_{i-1}) Z_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}) = \sum_{w_{i-1}^{i-2}} \gamma_{0,j-1}(w_{i-1}^{i-2}) Y_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}^{i-2}) \gamma_{0,j-1}(w_{i-1}) Z_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1})(21)
\]

Plugging in the definition of \( \gamma_{0,j-1} \) and simplifying gives
\[
\sum_{w_{i-1}^{i-2}} (\sum_i c_{i,i-1,i-2}^{\rho_j}) Y_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}^{i-2}) = \sum_{w_{i-1}^{i-2}} (\sum_i c_{i,i-1,i-2}^{\rho_j}) Z_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}^{i-2})(22)
\]

which is equivalent to
\[
\sum_{w_{i-1}^{i-2}} Y_{D_j}^{(\rho_j, \kappa_j)}(w_i, w_{i-1}^{i-2}) = \sum_{w_{i-1}^{i-2}} Z_{D_j}^{(\rho_j, \kappa_j)}(w_i, w_{i-1}^{i-2})(23)
\]

which holds because of the definition of \( Z \) and the fact that rank minimization under \( gKL \) preserves row/column sums.

Now consider any bigram term. We seek to show that:
\[
\sum_{w_{i-1}^{i-2}} \gamma_{0,j-1}(w_{i-1}^{i-2}) Y_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}^{i-2}) \gamma_{0,j-1}(w_{i-1}) Z_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}) = \sum_{w_{i-1}^{i-2}} \gamma_{0,j-1}(w_{i-1}^{i-2}) Y_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}^{i-2}) \gamma_{0,j-1}(w_{i-1}) Z_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1})(24)
\]

Substituting definition of \( \gamma_{0,j-1}(w_{i-1}^{i-2}) \) gives
\[
\sum_{w_{i-1}^{i-2}} (d_s)^{j+1} \frac{\sum_i c_{i,i-1,i-2}^{\rho_{\eta+1}}}{\sum_i c_{i,i-1,i-2}^{0}} \gamma_{0,j-1}(w_{i-1}) Y_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}) \gamma_{0,j-1}(w_{i-1}) Z_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}) \gamma_{0,j-1}(w_{i-1}) Z_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1})(25)
\]

Simplifying and pushing in the sum over \( w_{i-2} \) gives,
\[
\sum_{w_{i-1}} (\sum_i c_{i,i-1,i-2}^{\rho_{\eta+1}}) Y_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1}) = \sum_{w_{i-1}} (\sum_i c_{i,i-1,i-2}^{\rho_{\eta+1}}) Z_{D_j}^{(\rho_j, \kappa_j)}(w_i|w_{i-1})(26)
\]
Note that since $\rho_{\eta+1} = 0, \sum_{i, i-2} \tilde{c}_{\rho_{\eta+1},i-1,i-2} = \sum_i \tilde{c}_{i,i-1}$ (by definition of $\tilde{c}$).

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

$$\sum_{w_{i-1}} (\sum_i \tilde{c}_{i,i-1}) \frac{(d_s)^j}{\sum_i \tilde{c}_{i,i-1}} Y_{D_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1}) = \sum_{w_{i-1}} (\sum_i \tilde{c}_{i,i-1}) \frac{(d_s)^j}{\sum_i \tilde{c}_{i,i-1}} Z_{D_j}^{(\rho_j,\kappa_j)}(w_i|w_{i-1}) \tag{27}$$

Simplifying gives,

$$\sum_{w_{i-1}} Y_{D_j}^{(\rho_j)}(w_i, w_{i-1}) = \sum_{w_{i-1}} Z_{D_j}^{(\rho_j,\kappa_j)}(w_i, w_{i-1}) \tag{28}$$

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

References