Supplement to Sha & Pereira's paper

Definitions:

$$
\begin{aligned}
L_\lambda &\equiv \sum_k \ln P_\lambda(\mathbf{y}_k|\mathbf{x}_k) \\
P_\lambda(\mathbf{y}|\mathbf{x}) &\equiv \frac{P_\lambda(\mathbf{y},\mathbf{x})}{Z_\lambda(\mathbf{x})} \\
Z_\lambda(\mathbf{x}_k) &\equiv \sum_{\mathbf{y}} P_\lambda(\mathbf{y}_k,\mathbf{x}_k) \\
P_\lambda(\mathbf{y}_k,\mathbf{x}_k) &\equiv \exp(\boldsymbol{\lambda}\cdot\mathbf{F}(\mathbf{y}_k,\mathbf{x}_k)) = \exp(\sum_i \lambda^i \cdot F^i(\mathbf{y}_k,\mathbf{x}_k))
\end{aligned}
$$

Now let's differentiate L_λ wrt λ^i:

$$
\begin{aligned}
\frac{\partial}{\partial\lambda^i} L_\lambda &= \frac{\partial}{\partial\lambda^i}\sum_k \ln P_\lambda(\mathbf{y}_k|\mathbf{x}_k) && (1) \\
&= \frac{\partial}{\partial\lambda^i}\left(\sum_k \ln P_\lambda(\mathbf{y}_k,\mathbf{x}_k) - \sum_k \ln Z_\lambda(\mathbf{x}_k)\right) && (2) \\
&= \left(\frac{\partial}{\partial\lambda^i}(\sum_k \ln P_\lambda(\mathbf{y}_k,\mathbf{x}_k))\right) - \left(\frac{\partial}{\partial\lambda^i}(\sum_k \ln Z_\lambda(\mathbf{x}_k))\right) && (3)
\end{aligned}
$$

Starting with the rightmost sum of Eq.3, we will use $\frac{d}{dx}(\ln x) = \frac{1}{x}$ and the chain rule in Eq.5, the definition of Z_λ in Eq.6, and the definition of P_λ in Eq.7. Now use $\frac{d}{dx}(\exp x) = \exp(x)$ and the chain rule:

$$
\begin{aligned}
\frac{\partial}{\partial\lambda^i}\sum_k \ln Z_\lambda(\mathbf{x}_k) &= \sum_k \frac{\partial}{\partial\lambda^i} \ln Z_\lambda(\mathbf{x}_k) && (4) \\
&= \sum_k \frac{1}{Z_\lambda(\mathbf{x}_k)} \frac{\partial}{\partial\lambda^i} Z_\lambda(\mathbf{x}_k) && (5) \\
&= \sum_k \frac{1}{Z_\lambda(\mathbf{x}_k)} \frac{\partial}{\partial\lambda^i} \sum_{\mathbf{y}} P_\lambda(\mathbf{y},\mathbf{x}_k) && (6) \\
&= \sum_k \frac{1}{Z_\lambda(\mathbf{x}_k)} \frac{\partial}{\partial\lambda^i} \sum_{\mathbf{y}} \exp(\boldsymbol{\lambda}\cdot\mathbf{F}(\mathbf{y},\mathbf{x}_k)) && (7)
\end{aligned}
$$

Now use $\frac{d}{dx}(\exp(x)) = \exp(x)$ and the chain rule to continue the differentiation. Along the way we simplify in Eq.9 by multiplying the normalizer $\frac{1}{Z\lambda(x)}$ through, which gives us the expression for $P_\lambda(\mathbf{y}|\mathbf{x_k})$, which we can plug in.

$$
\begin{aligned}
\frac{\partial}{\partial \lambda^i} \sum_k \ln Z_\lambda(\mathbf{x}_k) &= \sum_k \frac{1}{Z_\lambda(\mathbf{x}_k)} \frac{\partial}{\partial \lambda^i} \sum_{\mathbf{y}} \exp(\boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{y}, \mathbf{x}_k)) && (8) \\
&= \sum_k \frac{1}{Z_\lambda(\mathbf{x}_k)} \sum_{\mathbf{y}} \exp(\boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{y}, \mathbf{x}_k)) \frac{\partial}{\partial \lambda^i} (\boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{y}, \mathbf{x}_k)) \\
&= \sum_k \sum_{\mathbf{y}} \frac{\exp(\boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{y}, \mathbf{x}_k))}{Z_\lambda(\mathbf{x}_k)} \frac{\partial}{\partial \lambda^i} (\boldsymbol{\lambda} \cdot \mathbf{F}(\mathbf{y}, \mathbf{x}_k)) && (9) \\
&= \sum_k \sum_{\mathbf{y}} P_\lambda(\mathbf{y}|\mathbf{x_k}) \cdot \frac{\partial}{\partial \lambda^i} (\sum_i \lambda^i \cdot F^i(\mathbf{y}, \mathbf{x}_k)) && (10) \\
&= \sum_k \sum_{\mathbf{y}} P_\lambda(\mathbf{y}|\mathbf{x_k}) \cdot F^i(\mathbf{y}, \mathbf{x}_k) && (11)
\end{aligned}
$$

This is something we can describe in words: it is the expected value of $F^i(\mathbf{y}, \mathbf{x}_k)$ under the distribution of $\mathbf{y}$'s induced by picking the $\mathbf{x}_k$'s in the sample uniformly, and then generating the $\mathbf{y}$'s using $\boldsymbol{\lambda}$, the current parameters. (Later we'll get to how to compute this!)

Going back to the leftmost sum of Eq.3 - this is the easy one - we see that this boils down to just the expected value of $F^i(\mathbf{y}, \mathbf{x}_k)$ in the sample:

$$
\begin{aligned}
\frac{\partial}{\partial \lambda^i} (\sum_k \ln P_\lambda(\mathbf{y}_k, \mathbf{x}_k)) &= \sum_k \frac{\partial}{\partial \lambda^i} (\sum_i \lambda^i \cdot F^i(\mathbf{y}_k, \mathbf{x}_k)) \\
&= \sum_k F^i(\mathbf{y}_k, \mathbf{x}_k))
\end{aligned}
$$