A Supplemental Material

A.1 Scaling properties of power objectives

The scaling properties of both α and β-divergence, shown in Table 1, imply that if we do not enforce the scale of the model to be 1 by normalizing it, the model will be unable to learn. Indeed, we can rewrite $D_\alpha(p_\mathcal{D}||\exp(s_\theta))$ as:

$$\begin{aligned}
&D_\alpha(p_\mathcal{D}||\exp(s_\theta)) \\
&\quad = D_\alpha(p_\mathcal{D}||p_\theta \times Z_\theta) \\
&\quad = \frac{1}{\alpha(\alpha-1)} \sum_{(x,y)\in\mathcal{D}} (Z_\theta(x))^{1-\alpha} (p_\theta(y|x))^{1-\alpha}
\end{aligned}$$

That makes the objective possible to minimize by simply minimizing $Z_\theta(x) \forall x \in \mathcal{X}$ — which does not imply any learning from the data. It is also the case with $D_\beta(p_\mathcal{D}||\exp(s_\theta))$:

$$\begin{aligned}
&D_\beta(p_\mathcal{D}||\exp(s_\theta)) \\
&\quad = D_\beta(p_\mathcal{D}||p_\theta \times Z_\theta) \\
&\quad = \frac{1}{\beta(\beta-1)} \sum_{(x,y)\in\mathcal{D}} \Big[-\beta (Z_\theta(x))^{\beta-1} (p_\theta(y|x))^{\beta-1} \\
&\qquad\qquad + \quad (Z_\theta(x))^\beta \sum_{y'\in\mathcal{Y}} (p_\theta(y'|x))^\beta \Big] \\
&\quad = \frac{1}{\beta(\beta-1)} \sum_{(x,y)\in\mathcal{D}} (Z_\theta(x))^{\beta-1} \times \\
&\qquad \left[-\beta (p_\theta(y|x))^{\beta-1} + \quad Z_\theta(x) \sum_{y'\in\mathcal{Y}} (p_\theta(y'|x))^\beta \right]
\end{aligned}$$

However, it is easy to derive that it is not the case for the γ-divergence:

$$\begin{aligned}
&D_\gamma(p_\mathcal{D}||p_\theta) \\
&\quad = \sum_{(x,y)\in\mathcal{D}} \left[\log p_\theta(y|x) - \frac{1}{\gamma} \log \sum_{y'\in\mathcal{Y}} \big(p_\theta(y'|x)\big)^\gamma \right] \\
&\quad = \sum_{(x,y)\in\mathcal{D}} \Big[s_\theta(x,y) - \log Z_\theta(x) \\
&\qquad\qquad - \frac{1}{\gamma} \log \frac{1}{(Z_\theta(x))^\gamma} \sum_{y'\in\mathcal{Y}} \exp(\gamma s_\theta(x,y')) \Big] \\
&\quad = \sum_{(x,y)\in\mathcal{D}} \left[s_\theta(x,y) - \frac{1}{\gamma} \log \sum_{y'\in\mathcal{Y}} \exp(\gamma s_\theta(x,y')) \right] \\
&\quad = D_\gamma(p_\mathcal{D}||\exp(s_\theta))
\end{aligned}$$

A.2 NCE as a binary divergence

The divergence $D_{KL}(p^C_\mathcal{D}||p^C_\theta)$ can be written as:

$$\begin{aligned}
D_{KL}(p^C_\mathcal{D}||p^C_\theta) &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p^C_\mathcal{D}(x,y) \log \frac{p^C_\mathcal{D}(x,y)}{p^C_\theta(x,y)} \\
&= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{p_\mathcal{D}(y|x)}{p_\mathcal{D}(y|x)+kp_n(y)} \times \\
&\quad \left[\log \frac{p_\mathcal{D}(y|x)}{p_\mathcal{D}(y|x)+kp_n(y)} - \log \frac{p_\theta(y|x)}{p_\theta(y|x)+kp_n(y)} \right]
\end{aligned}$$

We can remove the first term, which is not dependent on θ, and will not intervene in the objective function. If we do the same with the divergence $D_{KL}(1-p^C_\mathcal{D}||1-p^C_\theta)$ and add them, we obtain the following:

$$\begin{aligned}
&-\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \Bigg(\frac{p_\mathcal{D}(y|x)}{p_\mathcal{D}(y|x)+kp_n(y)} \log \frac{p_\theta(y|x)}{p_\theta(y|x)+kp_n(y)} \\
&\qquad + \frac{kp_n(y)}{p_\mathcal{D}(y|x)+kp_n(y)} \log \frac{kp_n(y)}{p_\theta(y|x)+kp_n(y)} \Bigg)
\end{aligned}$$

With NCE, we consider that examples are coming from the mixture $\frac{1}{k+1}p_\mathcal{D} + \frac{k}{k+1}p_n$, instead of being uniformly spread, which transforms the objective into:

$$\begin{aligned}
&-\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \Bigg(p_\mathcal{D}(y|x) \log \frac{p_\theta(y|x)}{p_\theta(y|x)+kp_n(y)} \\
&\qquad + kp_n(y) \log \frac{kp_n(y)}{p_\theta(y|x)+kp_n(y)} \Bigg)
\end{aligned}$$

We can then rewrite it as a sum of expectations:

$$\begin{aligned}
&-\sum_{x\in\mathcal{X}} \Bigg(\mathbb{E}_{y\sim p_\mathcal{D}(.|x)} \left[\log \frac{p_\theta(y|x)}{p_\theta(y|x)+kp_n(y)} \right] \\
&\qquad + k\mathbb{E}_{\hat{y}\sim p_n} \left[\log \frac{kp_n(\hat{y})}{p_\theta(\hat{y}|x)+kp_n(\hat{y})} \right] \Bigg)
\end{aligned}$$

That becomes the NCE objective once we approximate the second expectation over k samples:

$$\begin{aligned}
&-\sum_{(x,y)\in\mathcal{D}} \Bigg[\log \frac{p_\theta(y|x)}{p_\theta(y|x)+kp_n(y)} \\
&\qquad + \sum_{i=1}^{k} \log \frac{kp_n(\hat{y}_i)}{p_\theta(\hat{y}_i|x)+kp_n(\hat{y}_i)} \Bigg]
\end{aligned}$$

We should note that minimizing the NCE objective is then equivalent to minimizing both a f-divergence and a Bregman divergence. For example, by making a variable change between p^C_θ and

p_θ, we can circle back to writing the NCE objective as a Bregman divergence $D_\phi(p_\mathcal{D}||p_\theta)$, with $\phi(x) = x \log x - (1+x)\log(1+x)$, as shown in Gutmann and Hirayama (2011).

A.3 Objective-specific 'perplexity'

See Figure 5. We can observe that the behavior of the objective-specific counterparts to perplexity closely mirrors it, even when the values are quite distant.

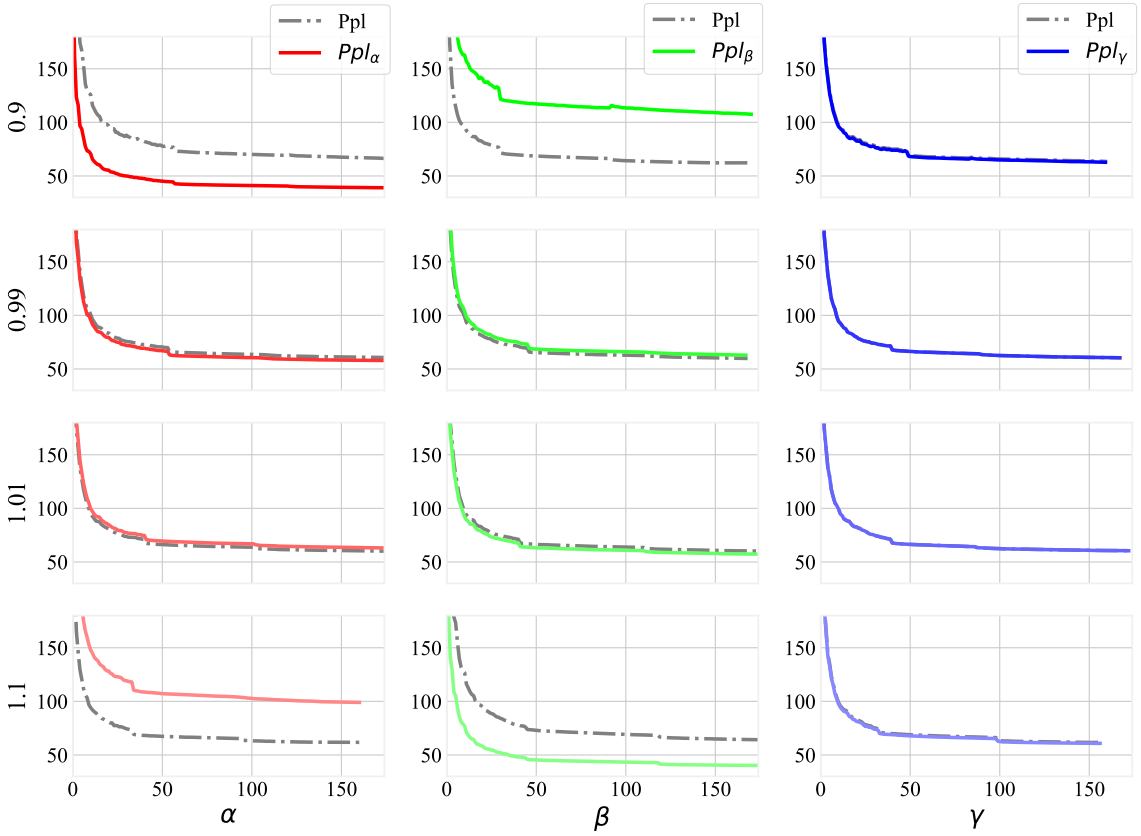

Figure 5: Validation results by epoch obtained on the PTB with the exact objectives , derived from MLE, during training. We give the validation perplexity (dotted gray) and the 'counterpart' to perplexity corresponding to the training objective (in color). Each color corresponds to a different objective, and we use different shades to indicate that changing the value of the power parameter makes the tracked values different.

A.4 Complete sampling-based objectives

See Table 6.

A.5 Detailed performance of sampling based-objectives

See Figures 6,7,8 and 9.

	Objective
Approximated Softmax	$-\sum_{(x,y)\in\mathcal{D}}\left[s_\theta(x,y)-\log p_n(y)-\log\sum_{i=1}^{k}\exp\left(s_\theta(x,\hat{y}_i)-\log p_n(\hat{y}_i)\right)\right]$
$\alpha\in\mathbb{R}\setminus\{0,1\}$	$\frac{1}{\alpha(\alpha-1)}\sum_{(x,y)\in\mathcal{D}}\left(\frac{\exp\left(s_\theta(x,y)-\log p_n(y)\right)}{\sum_{i=1}^{k}\exp\left(s_\theta(x,\hat{y}_i)-\log p_n(\hat{y}_i)\right)}\right)^{1-\alpha}$
$\beta\in\mathbb{R}\setminus\{0,1\}$	$\frac{1}{\beta(\beta-1)}\sum_{(x,y)\in\mathcal{D}}\left[\frac{\sum_{i=1}^{k}\exp\left(\beta(s_\theta(x,\hat{y}_i)-\log p_n(\hat{y}_i))\right)}{\left(\sum_{i=1}^{k}\exp\left(s_\theta(x,\hat{y}_i)-\log p_n(\hat{y}_i)\right)\right)^{\beta}}-\beta\frac{\exp\left((\beta-1)(s_\theta(x,y)-\log p_n(y))\right)}{\left(\sum_{i=1}^{k}\exp\left(s_\theta(x,\hat{y}_i)-\log p_n(\hat{y}_i)\right)\right)^{\beta-1}}\right]$
$\gamma\in\mathbb{R}\setminus\{0,1\}$	$-\sum_{(x,y)\in\mathcal{D}}\left[s_\theta(x,y)-\log p_n(y)-\frac{1}{\gamma}\log\sum_{i=1}^{k}\exp\left(\gamma(s_\theta(x,\hat{y}_i)-\log p_n(\hat{y}_i))\right)\right]$
Noise Contrastive Estimation	$-\sum_{(x,y)\in\mathcal{D}}\left[\log\frac{\exp\left(s_\theta(x,y)\right)}{\exp\left(s_\theta(x,y)\right)+kp_n(y)}+\sum_{i=1}^{k}\log\frac{kp_n(\hat{y}_i)}{\exp\left(s_\theta(x,\hat{y}_i)\right)+kp_n(\hat{y}_i)}\right]$
$\alpha\in\mathbb{R}\setminus\{0,1\}$	$\frac{1}{\alpha(\alpha-1)}\sum_{(x,y)\in\mathcal{D}}\left[\left(\frac{\exp\left(s_\theta(x,y)\right)}{\exp\left(s_\theta(x,y)\right)+kp_n(y)}\right)^{1-\alpha}+\sum_{i=1}^{k}\left(\frac{kp_n(\hat{y}_i)}{\exp\left(s_\theta(x,\hat{y}_i)\right)+kp_n(\hat{y}_i)}\right)^{1-\alpha}\right]$
$\beta\in\mathbb{R}\setminus\{0,1\}$	$\sum_{(x,y)\in\mathcal{D}}\left[-\frac{1}{\beta-1}\left(\left(\frac{\exp\left(s_\theta(x,y)\right)}{\exp\left(s_\theta(x,y)\right)+kp_n(y)}\right)^{\beta-1}+\sum_{i=1}^{k}\left(\frac{kp_n(\hat{y}_i)}{\exp\left(s_\theta(x,\hat{y}_i)\right)+kp_n(\hat{y}_i)}\right)^{\beta-1}\right)\right.$ $\left.+\ \frac{1}{\beta}\left(\frac{\exp\left(\beta s_\theta(x,y)\right)+(kp_n(y))^{\beta}}{\left(\exp\left(s_\theta(x,y)\right)+kp_n(y)\right)^{\beta}}+\sum_{i=1}^{k}\frac{\exp\left(\beta s_\theta(x,\hat{y}_i)\right)+(kp_n(\hat{y}_i))^{\beta}}{\left(\exp\left(s_\theta(x,\hat{y}_i)\right)+kp_n(\hat{y}_i)\right)^{\beta}}\right)\right]$
$\gamma\in\mathbb{R}\setminus\{0,1\}$	$\sum_{(x,y)\in\mathcal{D}}\left[-\log\frac{\exp\left(s_\theta(x,y)\right)}{\exp\left(s_\theta(x,y)\right)+kp_n(y)}-\frac{1}{\gamma-1}\log\sum_{i=1}^{k}\left(\frac{kp_n(\hat{y}_i)}{\exp\left(s_\theta(x,\hat{y}_i)\right)+kp_n(\hat{y}_i)}\right)^{\gamma-1}\right.$ $\left.+\ \frac{1}{\gamma}\log\left(\frac{\exp\left(\gamma s_\theta(x,y)\right)+(kp_n(y))^{\gamma}}{\left(\exp\left(s_\theta(x,y)\right)+kp_n(y)\right)^{\gamma}}+\sum_{i=1}^{k}\frac{\exp\left(\gamma s_\theta(x,\hat{y}_i)\right)+(kp_n(\hat{y}_i))^{\gamma}}{\left(\exp\left(s_\theta(x,\hat{y}_i)\right)+kp_n(\hat{y}_i)\right)^{\gamma}}\right)\right]$

Table 6: Complete objectives of power generalizations of the Approximated Softmax and Noise Contrastive Estimation objective functions, based on α, β, and γ divergences. All the samples $(\hat{y}_i)_{i=1}^{k}$ are drawn from the auxiliary distribution p_n.

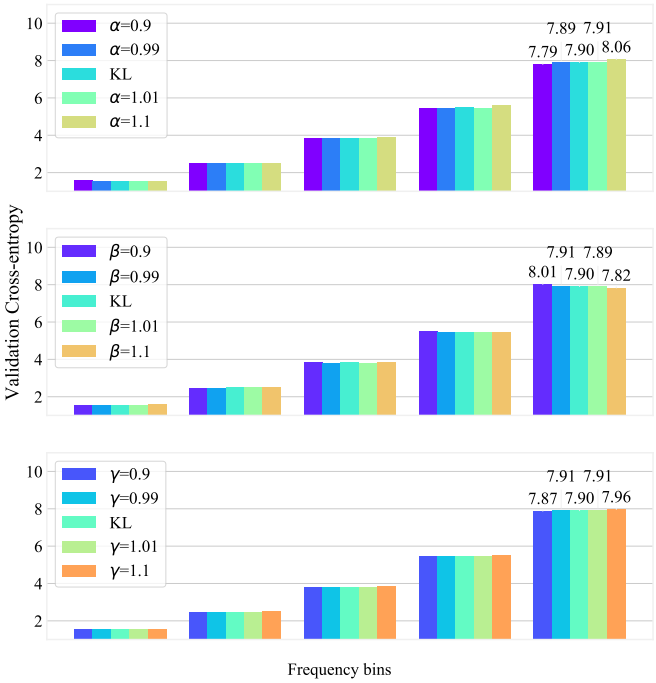

Figure 6: Validation cross-entropy values for the best epoch obtained for models trained with objectives derived from the AS objective with α-divergences (top), β-divergences (middle) and γ-divergences (bottom) on the PTB. Words are grouped into 5 buckets of equal size, following their frequencies. We display values for each bucket from the most frequent words (left) to less frequent ones (right).

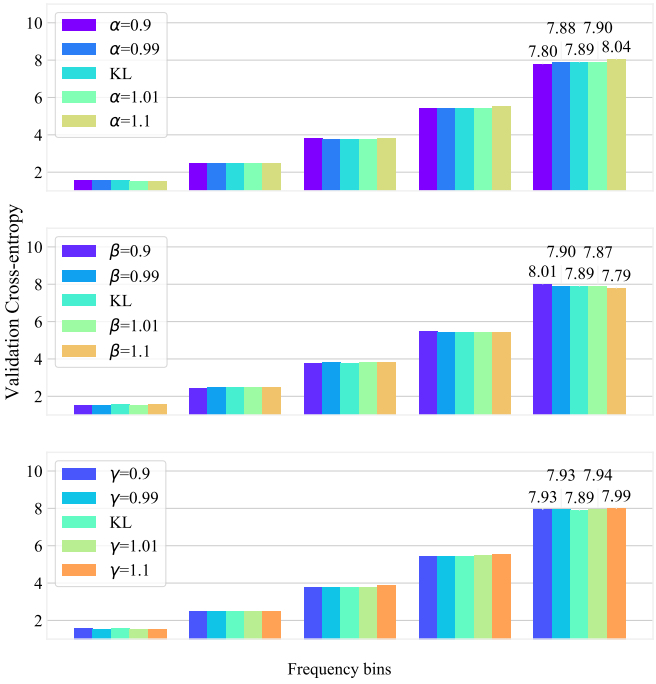

Figure 7: Validation cross-entropy values for the best epoch obtained for models trained with objectives derived from the NCE objective with α-divergences (top), β-divergences (middle) and γ-divergences (bottom) on the PTB. Words are grouped into 5 buckets of equal size, following their frequencies. We display values for each bucket from the most frequent words (left) to less frequent ones (right).

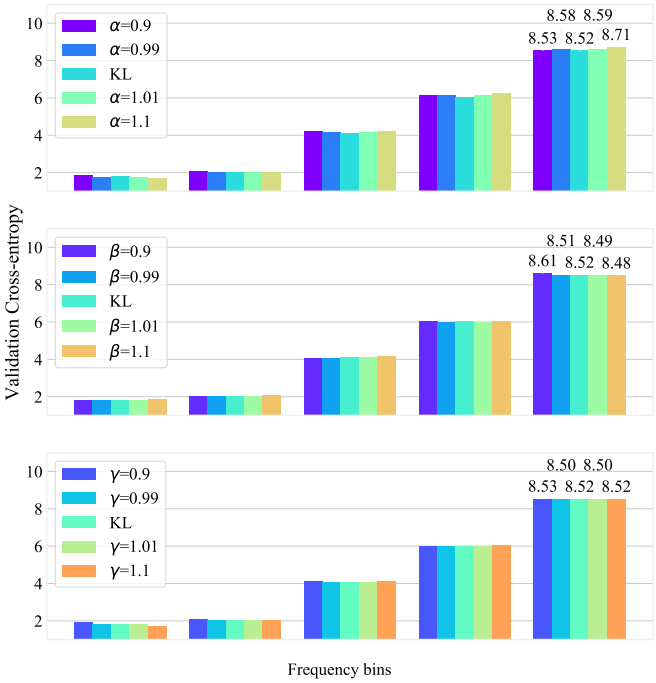

Figure 8: Validation cross-entropy values for the best epoch obtained for models trained with objectives derived from the AS objective with α-divergences (top), β-divergences (middle) and γ-divergences (bottom) on the WT2. Words are grouped into 5 buckets of equal size, following their frequencies. We display values for each bucket from the most frequent words (left) to less frequent ones (right).

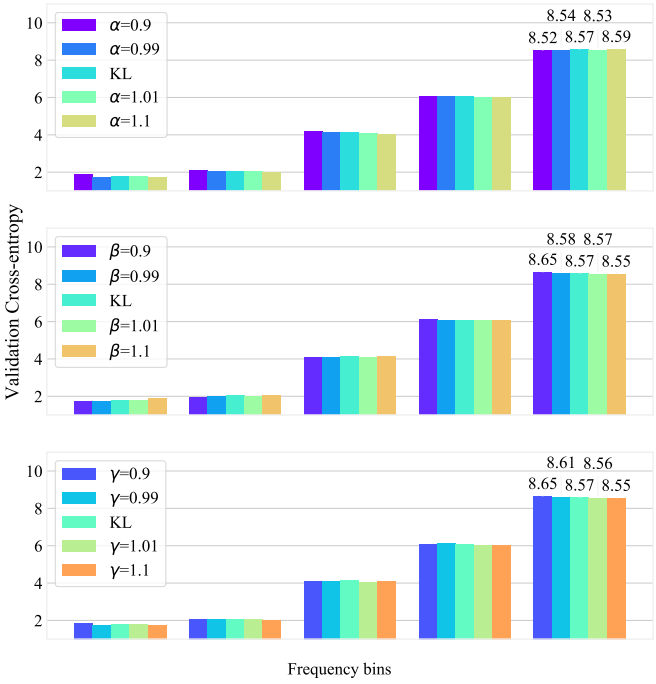

Figure 9: Validation cross-entropy values for the best epoch obtained for models trained with objectives derived from the NCE objective with α-divergences (top), β-divergences (middle) and γ-divergences (bottom) on the WT2. Words are grouped into 5 buckets of equal size, following their frequencies. We display values for each bucket from the most frequent words (left) to less frequent ones (right).