

# Supplementary Material

## 1 More on primal Bregman decoding; Comparison to temperature scaling

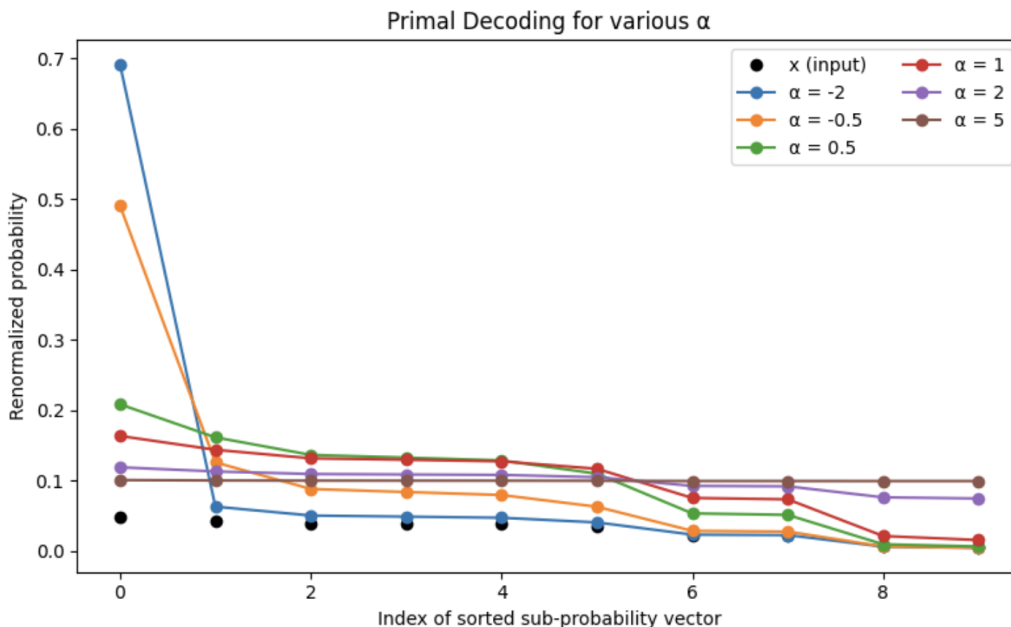


Figure 1: Primal  $\alpha$ -decodings of a fixed sub-probability vector for several representative  $\alpha$ .

Figure 1 illustrates the spectrum of primal Bregman  $\alpha$ -decoding of a fixed sub-probability vector  $x$  (denoted in the figure as  $x$  (input)). We can observe that, consistent with our derived theoretical limiting forms of  $\alpha$ -decoding (for  $\alpha \in \{-\infty, +\infty\}$ ), here for the most negative  $\alpha$  the missing mass is mostly redistributed onto the top token of the distribution, while for the most positive alpha, the missing mass is split between the tokens in a water-filling way. In fact, in this case, since all original sub-probability values are sufficiently small, the water-filling solution happens to “drown” all of the tokens, resulting in apparent convergence of the renormalized distribution to the uniform distribution.

One may notice the conceptual similarity with temperature scaling: The “very negative”  $\alpha$  case appears to be the most “greedy” one, and we could intuitively think of it as similar to temperature-0 scaling. (However, the similarity is indeed only purely intuitive in nature, as observe that for all  $\alpha$ , no matter how negative, our decodings always at least preserve the original probability mass on every token, and never expunge it down to 0 in favor of the top token unlike temperature scaling.) Conversely, the “very positive”  $\alpha$  case appears to send the original vector to a uniform distribution, similar to using high temperature scaling. (However, if our original vector’s largest sub-probability was large enough then the limiting vector as  $\alpha \rightarrow \infty$  would not be uniform, so the similarity is only conceptual.)

Thus, let us proceed to re-plot the same figure, with the best-fit ( $T$ -temperature scaling + standard renormalization) procedure compared to each  $\alpha$ -decoding result. In this case we found the best-fitting  $T$  for each  $\alpha$  by optimizing for the Euclidean distance between the temperature rescaling result and the  $\alpha$ -renormalized vector. The result is in Figure 2.

Observe that, as a sanity check, the case of  $\alpha = 1$  corresponds to temperature  $T = 1$ , as is should (because it recovers standard top- $k$  renormalization according to our theory). Further, observe that in this specific case, temperature scaling with increasingly high temperatures closely corresponds to the higher values of  $\alpha$ . The best-fitting temperature scaling for the lowest  $\alpha$  values indeed have low temperature as we anticipated, but don’t approximate the  $\alpha$ -renormalization as closely.

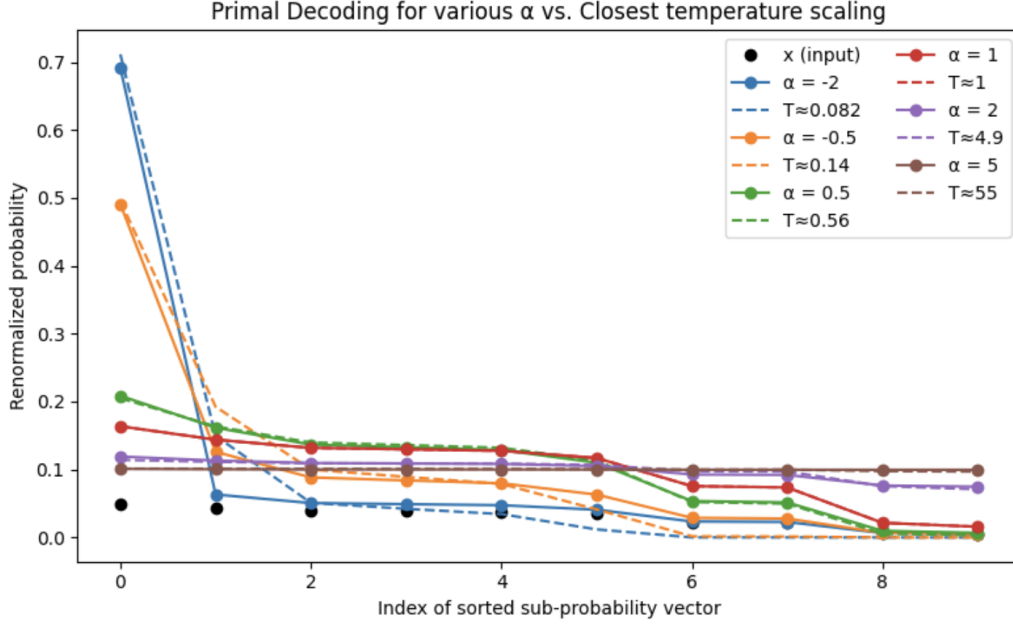


Figure 2: Primal  $\alpha$ -decodings of a fixed sub-probability vector vs. best-fitting temperature scaling.

## 29 2 The simultaneous effects of Bregman decoding and temperature scaling

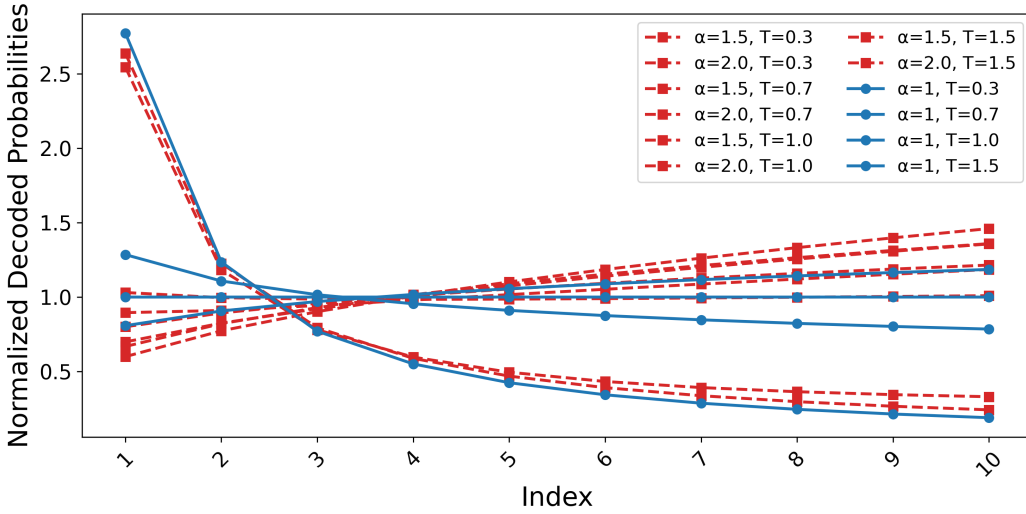


Figure 3: Comparison with changing the temperature.

30 Here, we provide a plot to help compare the simultaneous effects of Bregman decoding and tem-  
 31 perature scaling. We use the same simulation setting and plotting style as in our figure from the  
 32 introduction; except we only plot the nonzero probabilities (i.e., the top  $k = 10$  probabilities), and  
 33 we plot the *relative* sizes of the probabilities compared to the standard top- $k$  decoding. Further, we  
 34 use the same  $\alpha$  and temperature hyperparameters used in our experiments in Table 1 in the main  
 35 paper. The results are shown in Figure 3. Standard top- $k$  decoding corresponds to  $\alpha = 1$  and  $T = 1$ .  
 36 From the figure, it appears that the effect of  $\alpha > 1$  is to moderate/regularize the amount by which  
 37 the small probabilities are pushed to zero; which could potentially be one reason why  $\alpha$ -Bregman  
 38 decoding with  $\alpha > 1$  can perform better at high temperatures.