Generalized Entropy Regularization:

Or There's Nothing Special about Label Smoothing

Clara Meister, Elizabeth Salesky, and Ryan Cotterell

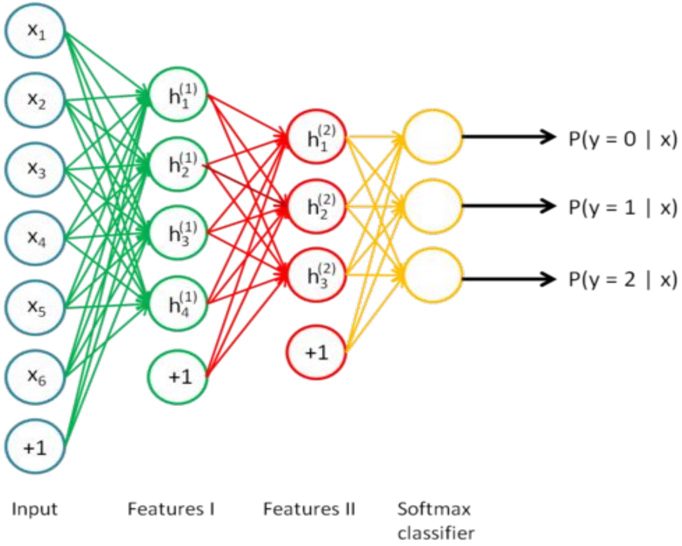
Organization:

- Regularizers for Probabilistic Models
- Interpretation as Entropy Regularizers
- A Single Framework
- Experimental Findings

Probabilistic Models

- General class of models that output a probability distribution, e.g., through the softmax over the final layer of a neural network.

*



- Used for many NLP tasks: machine translation, abstractive summarization, natural language inference, etc.

*https://rohanvarma.me/Neural-Net/

Regularization

What: Large neural networks need regularization during training to avoid overfitting!

Common forms of regularization:

- Dropout
- L2 weight normalization
- Label smoothing

Regularization

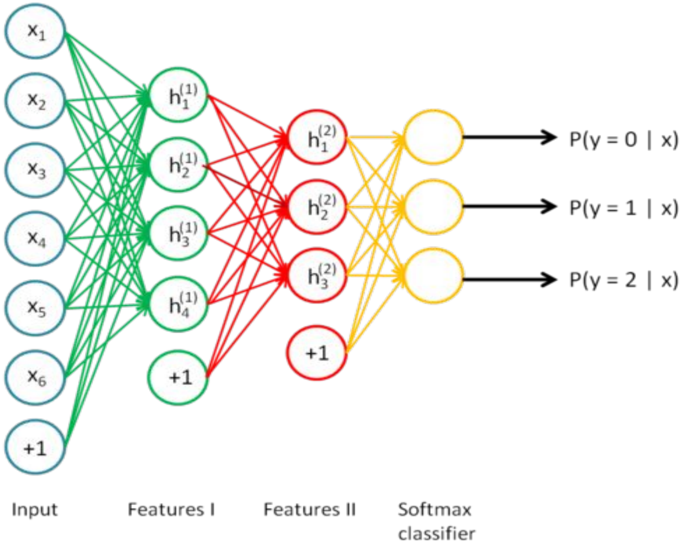
What: Large neural networks need regularization during training to avoid overfitting!

Common forms of regularization:

- Dropout
- L2 weight normalization
- **Label smoothing**

Regularization

Signs of overfitting: “peaky” (i.e. overconfident) output distributions

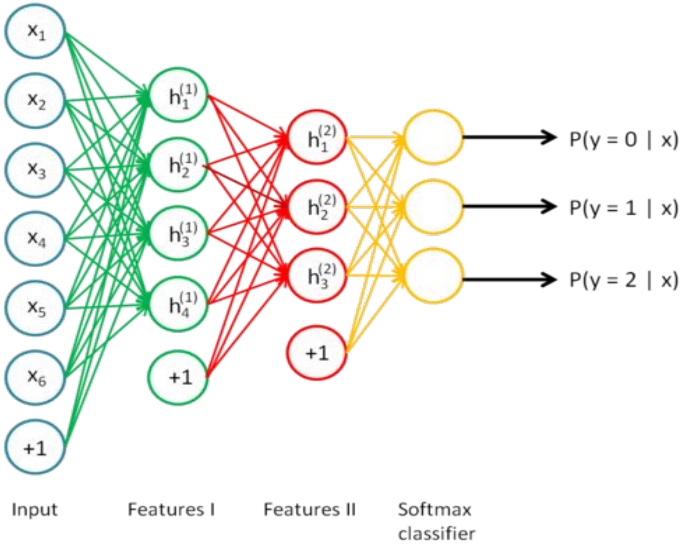


0.0001
0.9998
0.0001

← Model puts most of probability mass on ≅1 class

Regularization

Signs of overfitting: “peaky” (i.e. overconfident) output distributions



0.0001
0.9998
0.0001

Model puts most of probability mass on ≅1 class

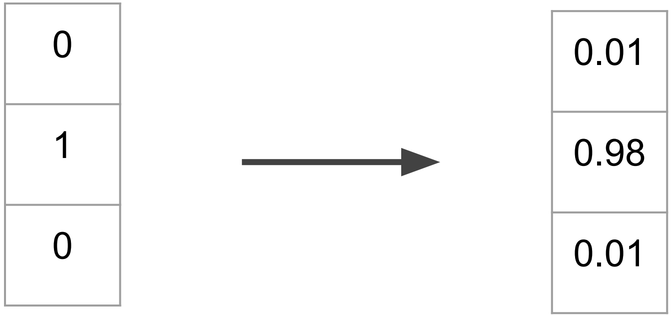
Low entropy distribution!!

Entropy Regularization

Label Smoothing (Szegedy et. al. 2016):

Entropy Regularization

Label Smoothing (Szegedy et. al. 2016):

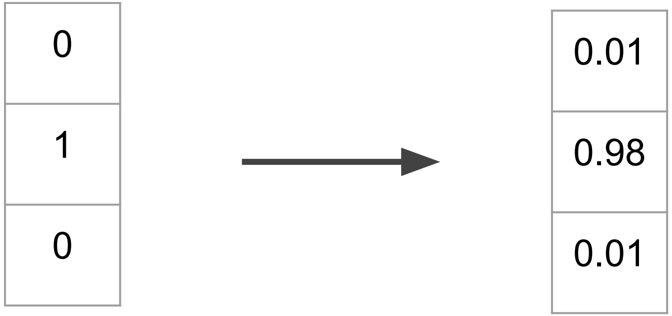


Original target distribution

Smoothed target distribution

Entropy Regularization

Label Smoothing (Szegedy et. al. 2016):



Original target distribution Smoothed target distribution

$$\mathrm{H}(\hat{p}, p_{\boldsymbol{\theta}}) \longrightarrow \mathrm{H}(\hat{p}', p_{\boldsymbol{\theta}})$$

Entropy Regularization

Label Smoothing (Szegedy et. al. 2016):

- Add-γ smoothing technique to ground truth (one-hot) labels. Cross entropy loss performed over augmented labels
- Interpretation as entropy regularization:

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{LS}} = (1-\gamma)\mathcal{L}(\boldsymbol{\theta}) + \gamma \mathrm{H}(u, p_{\boldsymbol{\theta}})$$

Entropy Regularization

Label Smoothing (Szegedy et. al. 2016):

- Add-γ smoothing technique to ground truth (one-hot) labels. Cross entropy loss performed over augmented labels
- Interpretation as entropy regularization:

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{LS}} = (1-\gamma)\mathcal{L}(\boldsymbol{\theta}) + \gamma \mathrm{H}(u, p_{\boldsymbol{\theta}})$$

Standard cross-entropy loss $\mathrm{H}(\hat{p}, p_{\boldsymbol{\theta}})$

Uniform distribution

Probabilistic model with parameters $\boldsymbol{\theta}$

Entropy Regularization

Label smoothing has become a default regularization method for many probabilistic modelling tasks!

Label Smoothing (Szegedy et. al. 2016):

- Add-γ smoothing technique to ground truth (one-hot) labels. Cross entropy loss performed over augmented labels
- Interpretation as entropy regularization:

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{LS}} = (1-\gamma)\mathcal{L}(\boldsymbol{\theta}) + \gamma \mathrm{H}(u, p_{\boldsymbol{\theta}})$$

Standard cross-entropy loss $\mathrm{H}(\hat{p}, p_{\boldsymbol{\theta}})$

Uniform distribution

Probabilistic model with parameters $\boldsymbol{\theta}$

Entropy Regularization

Label Smoothing (Szegedy et. al. 2016):

- Add-γ smoothing technique to ground truth (one-hot) labels. Cross entropy loss performed over augmented labels
- Interpretation as entropy regularization:

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{LS}} = (1-\gamma)\mathcal{L}(\boldsymbol{\theta}) + \gamma \mathrm{H}(u, p_{\boldsymbol{\theta}})$$

Confidence Penalty (Pereira et. al. 2017):

Entropy Regularization

Label Smoothing (Szegedy et. al. 2016):

- Add-γ smoothing technique to ground truth (one-hot) labels. Cross entropy loss performed over augmented labels
- Interpretation as entropy regularization:

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{LS}} = (1-\gamma)\mathcal{L}(\boldsymbol{\theta}) + \gamma \mathrm{H}(u, p_{\boldsymbol{\theta}})$$

Confidence Penalty (Pereira et. al. 2017):

- Add penalty to loss function for overconfident distributions $p_{\boldsymbol{\theta}}$
- Interpretation as entropy regularization:

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{CP}} = \mathcal{L}(\boldsymbol{\theta}) - \beta \mathrm{H}(p_{\boldsymbol{\theta}})$$

Entropy Regularization

Question: Are label smoothing and the confidence penalty the only forms of entropy regularization?

Entropy Regularization

Question: Are label smoothing and the confidence penalty the only forms of entropy regularization?

Answer: No! Otherwise, this would be a very boring paper.

Generalized Entropy Regularization

Framework for Entropy Regularization

Introducing: ***Generalized Entropy Regularization***

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

Explanation: in words

Introducing: ***Generalized Entropy Regularization***

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

Explanation: in words

Introducing: ***Generalized Entropy Regularization***

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

- J_α is a divergence measure between two distributions, e.g., u and $p_{\boldsymbol{\theta}}$.

Framework for Entropy Regularization

Explanation: in words

Introducing: ***Generalized Entropy Regularization***

$$\mathcal{L}(\boldsymbol{\theta})_{\text{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

- J_α is a divergence measure between two distributions, e.g., u and $p_{\boldsymbol{\theta}}$.
- Since u is the uniform (most entropic) distribution, adding a penalty for the divergence between u and $p_{\boldsymbol{\theta}}$ pushes $p_{\boldsymbol{\theta}}$ towards a higher entropy solution

Explanation: in math

Introducing: ***Generalized Entropy Regularization***

$$\mathcal{L}(\boldsymbol{\theta})_{\text{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

Explanation: in math

Introducing: ***Generalized Entropy Regularization***

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

$$D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}}) = \sum_{x,y \in \mathcal{C}} J_\alpha(u(\cdot) \,||\, p_{\boldsymbol{\theta}}(\cdot \mid x))$$

Fancy way of saying
“over the training corpus”

Framework for Entropy Regularization

Explanation: in math

Introducing: ***Generalized Entropy Regularization***

$$\mathcal{L}(\boldsymbol{\theta})_{\text{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

$$D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}}) = \sum_{x,y \in \mathcal{C}} J_\alpha(u(\cdot) \,||\, p_{\boldsymbol{\theta}}(\cdot \mid x))$$

$$J_\alpha(u \,||\, p_{\boldsymbol{\theta}}) := \frac{1}{\alpha(1-\alpha)} \Big((1-\alpha) G(u) + \alpha G(p_{\boldsymbol{\theta}}) - G((1-\alpha)u + \alpha p_{\boldsymbol{\theta}}) \Big)$$

Jensen-α divergence for generator function G

Framework for Entropy Regularization

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

For generator function $G(z) = -\mathrm{H}(z)$ (negative Shannon entropy) and $\boxed{\alpha \to 1}$:

$$\begin{aligned} J_\alpha(u \,||\, p_{\boldsymbol{\theta}}) &= \mathrm{KL}(u \,||\, p_{\boldsymbol{\theta}}) \\ &= \mathrm{H}(u, p_{\boldsymbol{\theta}}) + C \end{aligned}$$

Framework for Entropy Regularization

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

For generator function $G(z) = -\mathrm{H}(z)$ (negative Shannon entropy) and $\boxed{\alpha \to 1}$:

$$\begin{aligned} J_\alpha(u \,||\, p_{\boldsymbol{\theta}}) &= \mathrm{KL}(u \,||\, p_{\boldsymbol{\theta}}) \\ &= \mathrm{H}(u, p_{\boldsymbol{\theta}}) + C \end{aligned}$$

Equivalent to label smoothing!

Framework for Entropy Regularization

$$\mathcal{L}(\boldsymbol{\theta})_{\text{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

For generator function $G(z) = -\text{H}(z)$ (negative Shannon entropy) and $\alpha \to 0$:

$$\begin{aligned} J_\alpha(u \,||\, p_{\boldsymbol{\theta}}) &= \text{KL}(p_{\boldsymbol{\theta}} \,||\, u) \\ &= -\text{H}(p_{\boldsymbol{\theta}}) + C \end{aligned}$$

Framework for Entropy Regularization

$$\mathcal{L}(\boldsymbol{\theta})_{\text{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

For generator function $G(z) = -\mathrm{H}(z)$ (negative Shannon entropy) and $\alpha \to 0$:

$$\begin{aligned} J_\alpha(u \,||\, p_{\boldsymbol{\theta}}) &= \mathrm{KL}(p_{\boldsymbol{\theta}} \,||\, u) \\ &= -\mathrm{H}(p_{\boldsymbol{\theta}}) + C \end{aligned}$$

Equivalent to the confidence penalty!

Framework for Entropy Regularization

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

For generator function $G(z) = -\mathrm{H}(z)$ (negative Shannon entropy) and $\alpha \in (0,1)$:

$$J_\alpha(u \,||\, p_{\boldsymbol{\theta}}) = \frac{1-\alpha}{\alpha(1-\alpha)}\mathrm{KL}(u \,||\, (1-\alpha)u + \alpha p_{\boldsymbol{\theta}}) + \frac{\alpha}{\alpha(1-\alpha)}\mathrm{KL}(p_{\boldsymbol{\theta}} \,||\, (1-\alpha)u + \alpha p_{\boldsymbol{\theta}})$$

Framework for Entropy Regularization

$$\mathcal{L}(\boldsymbol{\theta})_{\mathrm{GER}} = \mathcal{L}(\boldsymbol{\theta}) + \beta D_{J_\alpha}(u \,||\, p_{\boldsymbol{\theta}})$$

For generator function $G(z) = -\mathrm{H}(z)$ (negative Shannon entropy) and $\alpha \in (0,1)$:

$$J_\alpha(u \,||\, p_{\boldsymbol{\theta}}) = \frac{1-\alpha}{\alpha(1-\alpha)}\mathrm{KL}(u \,||\, (1-\alpha)u + \alpha p_{\boldsymbol{\theta}}) + \frac{\alpha}{\alpha(1-\alpha)}\mathrm{KL}(p_{\boldsymbol{\theta}} \,||\, (1-\alpha)u + \alpha p_{\boldsymbol{\theta}})$$

Equivalent to ?????

Framework for Entropy Regularization

Question: Why do we need more forms of entropy regularization?

Framework for Entropy Regularization

Question: Why do we need more forms of entropy regularization?

- Label smoothing and the confidence penalty only cover 2 specific instances.

Framework for Entropy Regularization

Question: Why do we need more forms of entropy regularization?

- Label smoothing and the confidence penalty only cover 2 specific instances.
    - Label smoothing = minimize ***inclusive*** KL divergence between $\boldsymbol{u}$ and $\boldsymbol{p_\theta}$
    - Confidence penalty = minimize ***exclusive*** KL divergence between $\boldsymbol{u}$ and $\boldsymbol{p_\theta}$

Framework for Entropy Regularization

Question: Why do we need more forms of entropy regularization?

- Label smoothing and the confidence penalty only cover 2 specific instances.
  - Label smoothing = minimize ***inclusive*** KL divergence between u and p_θ
  - Confidence penalty = minimize ***exclusive*** KL divergence between u and p_θ
- There are a large number of other divergences we may choose to minimize.

Framework for Entropy Regularization

Question: Why do we need more forms of entropy regularization?

- Label smoothing and the confidence penalty only cover 2 specific instances.
  - Label smoothing = minimize ***inclusive*** KL divergence between u and p_θ
  - Confidence penalty = minimize ***exclusive*** KL divergence between u and p_θ
- There are a large number of other divergences we may choose to minimize.
- Specific divergence measures are more appropriate for certain tasks

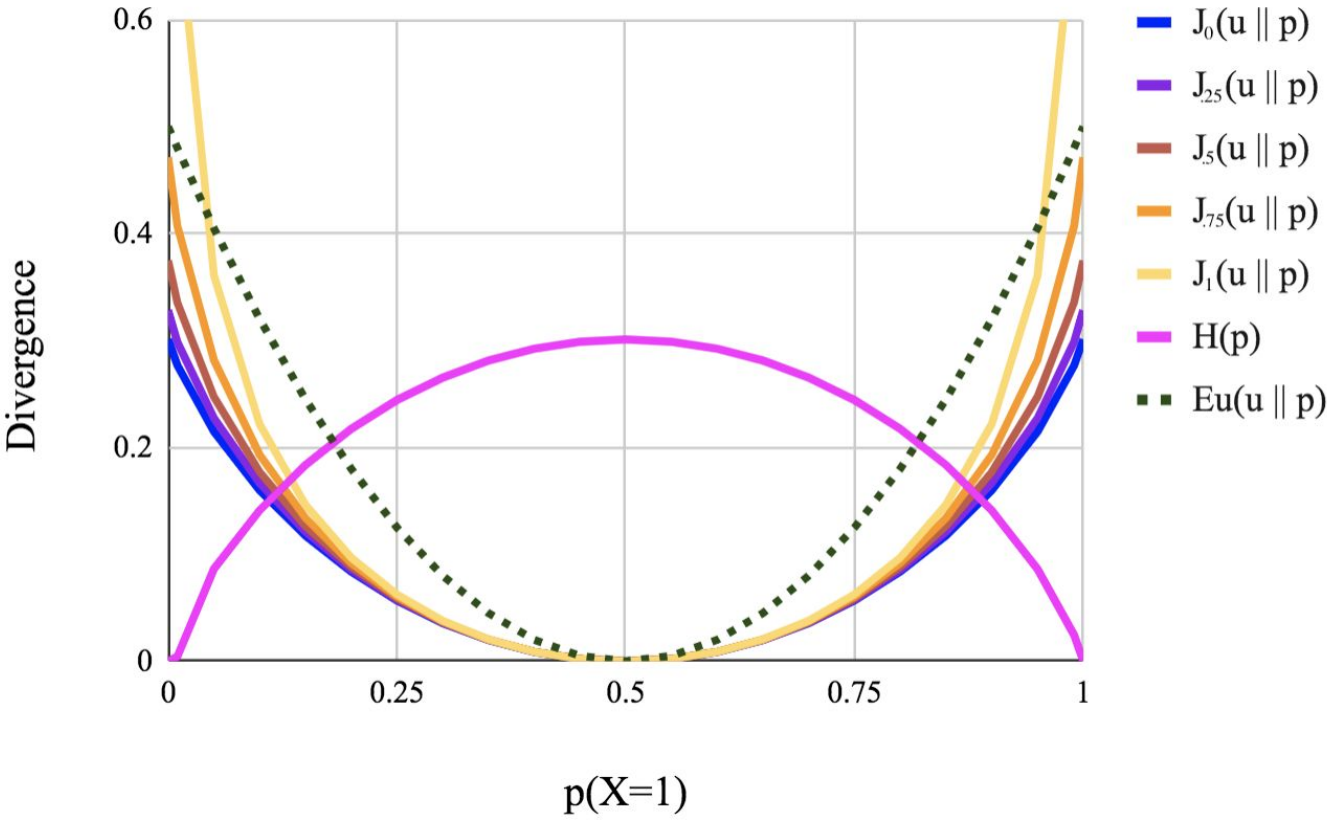
Framework for Entropy Regularization

Question: Why do we need more forms of entropy regularization?

- Label smoothing and the confidence penalty only cover 2 specific instances.
  - Label smoothing = minimize ***inclusive*** KL divergence between u and p_{θ}
  - Confidence penalty = minimize ***exclusive*** KL divergence between u and p_{θ}
- There are a large number of other divergences we may choose to minimize.
- **Specific divergence measures are more appropriate for certain tasks***

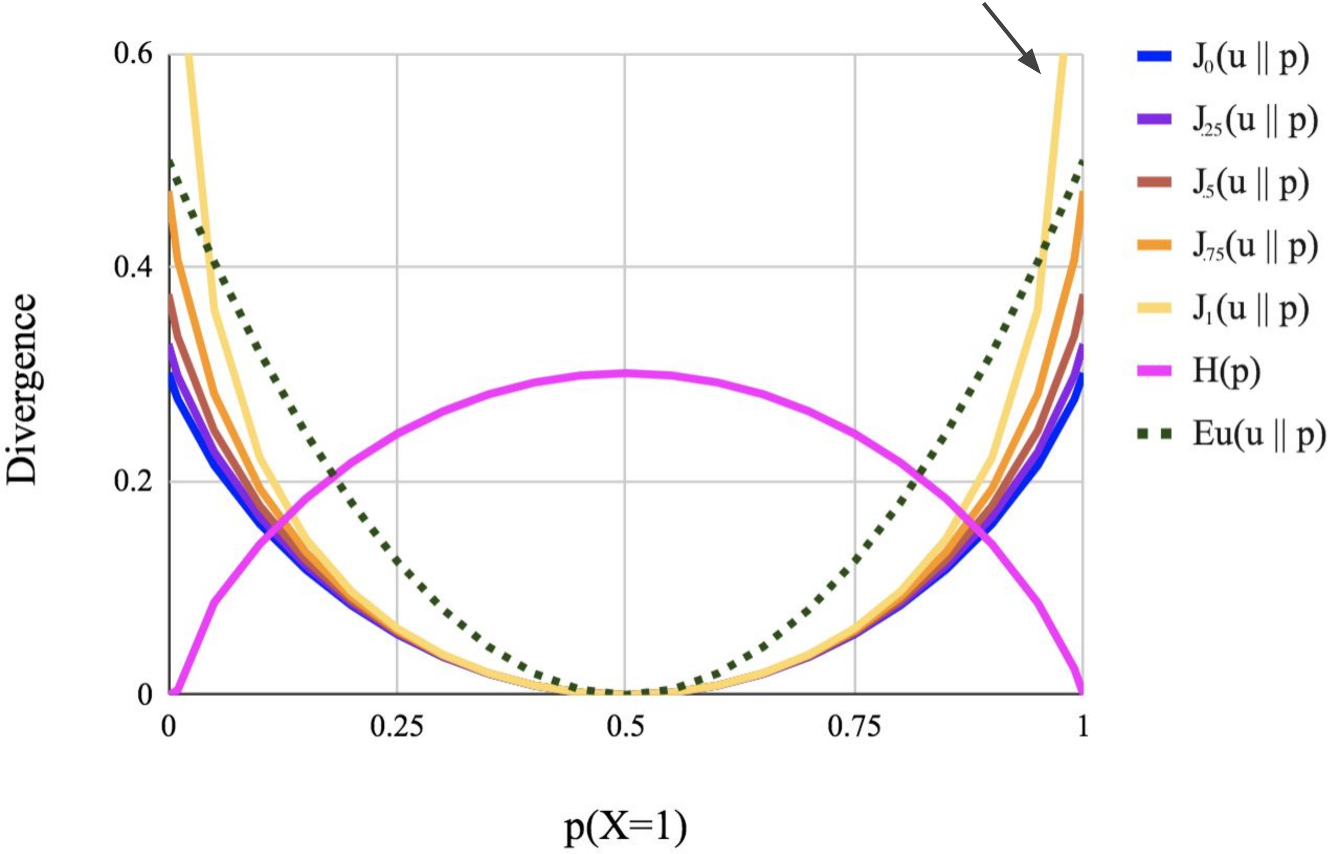
* See Minka's "Divergence Measures and Message Passing" for in-depth discussion

Framework for Entropy Regularization



Framework for Entropy Regularization

Label smoothing diverges for sparse solutions!



Experimental Findings

Experimental Findings

	WMT'14 De-En				IWSLT'14 De-En				MTTT Fr-En			
	α	β	$\hat{\mathrm{H}}$	BLEU	α	β	$\hat{\mathrm{H}}$	BLEU	α	β	$\hat{\mathrm{H}}$	BLEU
No Regularization	–	0	0.11	31.1	–	0	0.1	35.7	–	0	0.15	35.2
Label Smoothing D_{J_1} $(\gamma=0.1)$	1	0.11	0.23	31.3 +0.2	1	0.11	0.18	36.9 +1.2	1	0.11	0.18	36.5 +0.8
Label Smoothing D_{J_1}	1	0.35	0.38	31.7 +0.6	1	0.50	0.40	37.2 +1.5	1	0.693	0.47	37.5 +2.3
Confidence Penalty D_{J_0}	0	0.28	0.55	31.6 +0.5	0	0.76	0.81	37.5 +1.8	0	0.95	0.86	37.4 +2.2
GER D_{J_α}	0.7	0.65	0.47	32.0 +0.9	0.5	1.00	0.56	37.5 +1.8	0.85	0.52	0.37	37.6 +2.4

BLEU scores and normalized entropy of p_θ on test sets for WMT'14 De-En, WMT'14 De-En, and MTTT Fr-En. Results include baseline models with no (entropy) regularization and standard label smoothing with γ=1.

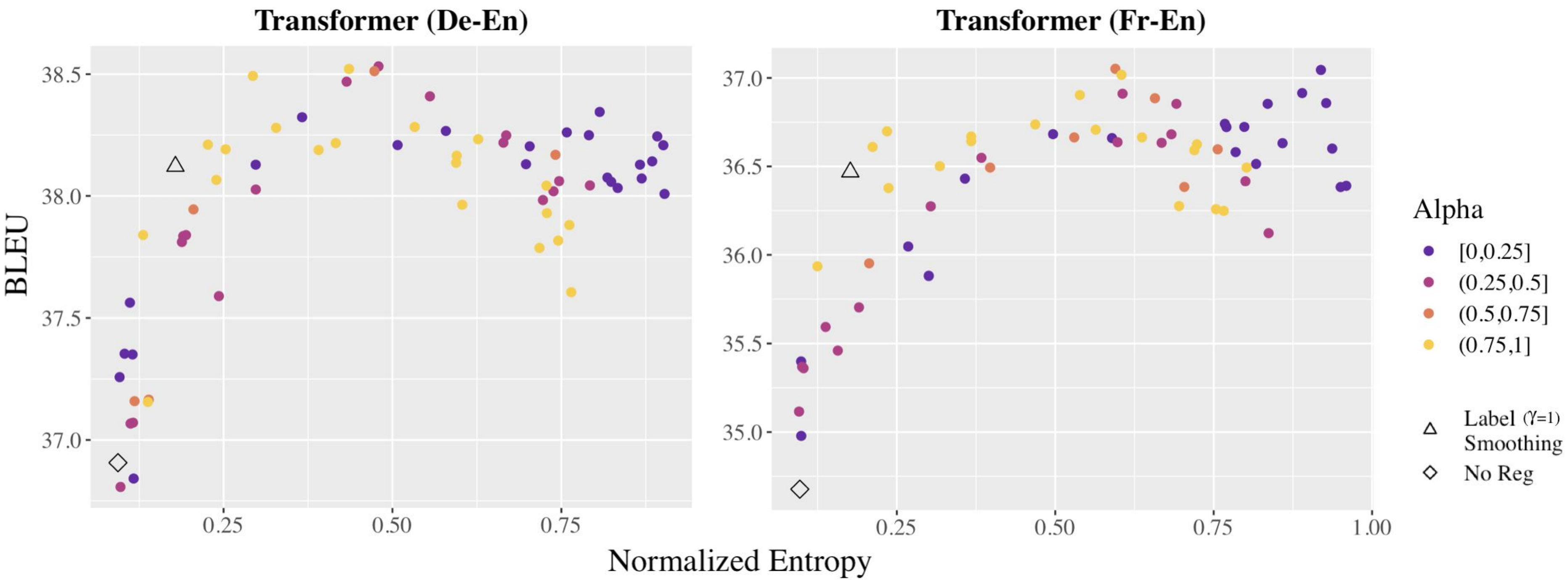
Experimental Findings

Empirically, we can do much better than standard label smoothing!

	WMT'14 De-En				IWSLT'14 De-En				MTTT Fr-En			
	α	β	$\hat{\text{H}}$	BLEU	α	β	$\hat{\text{H}}$	BLEU	α	β	$\hat{\text{H}}$	BLEU
No Regularization	–	0	0.11	31.1	–	0	0.1	35.7	–	0	0.15	35.2
Label Smoothing D_{J_1} $(\gamma=0.1)$	1	0.11	0.23	31.3 +0.2	1	0.11	0.18	36.9 +1.2	1	0.11	0.18	36.5 +0.8
Label Smoothing D_{J_1}	1	0.35	0.38	31.7 +0.6	1	0.50	0.40	37.2 +1.5	1	0.693	0.47	37.5 +2.3
Confidence Penalty D_{J_0}	0	0.28	0.55	31.6 +0.5	0	0.76	0.81	37.5 +1.8	0	0.95	0.86	37.4 +2.2
GER D_{J_α}	0.7	0.65	0.47	32.0 +0.9	0.5	1.00	0.56	37.5 +1.8	0.85	0.52	0.37	37.6 +2.4

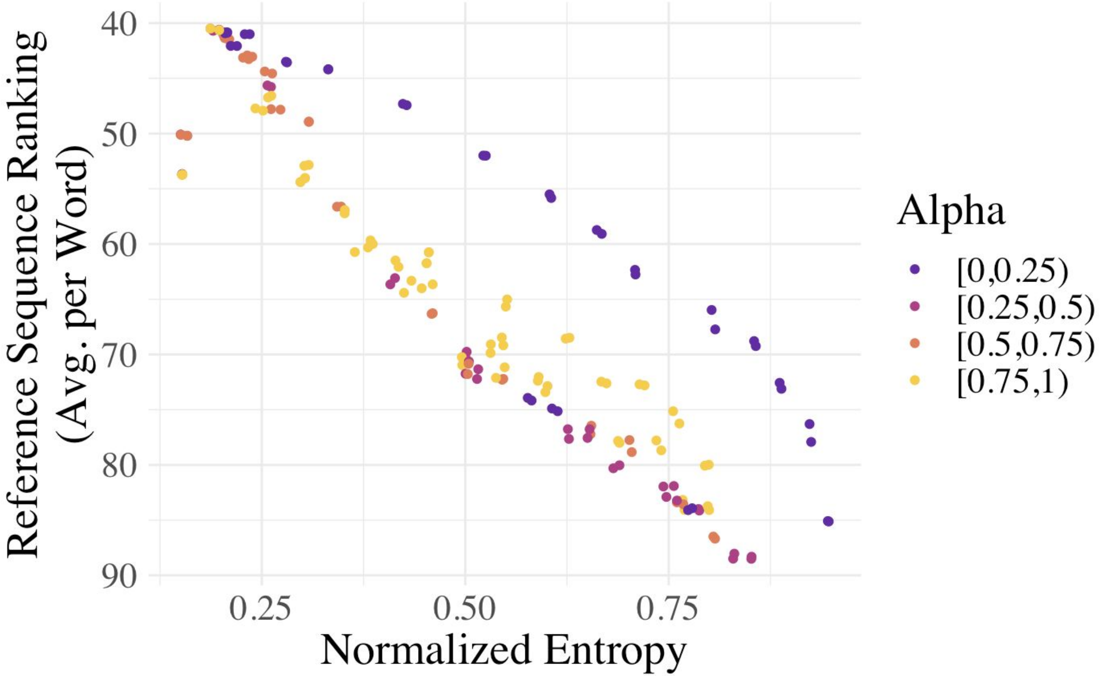
BLEU scores and normalized entropy of p_θ on test sets for WMT'14 De-En, WMT'14 De-En, and MTTT Fr-En. Results include baseline models with no (entropy) regularization and standard label smoothing with γ=1.

Experimental Findings



Experimental Findings

	Sparsity Threshold e^{-10}	Sparsity Threshold e^{-15}
Label Smoothing D_{J_1}	$38\% \pm 0.01\%$	$0.0\% \pm$ 5e-5%
Confidence Penalty D_{J_0}	$54\% \pm$ 5e-3%	$0.7\% \pm$ 4e-4%



Percentage of words with < ε probability mass at different values of ε. All models used in the calculation have entropy within the same 1%.

Summary + Conclusion

- Large probabilistic models need regularizers; various forms of entropy regularization have proven their merit in practice
- Many classes of entropy regularizers fit into our ***generalized entropy regularization*** framework.
- For the language generation tasks we consider, all regularizers can lead to good performance, suggesting we may generally desire a higher entropy solution $p_{\boldsymbol{\theta}}$.
- Some of these regularizers may be better suited for certain tasks due to the nature of the underlying divergence measure.

Thanks for watching

Title: Generalized Entropy Regularization: or There's Nothing Special about Label Smoothing
Authors: Clara Meister, Elizabeth Salezky, and Ryan Cotterell

Link to Paper