Reward Augmented Maximum Likelihood for Neural Structured Prediction

Mohammad Norouzi **Samy Bengio** **Zhifeng Chen** **Navdeep Jaitly**
Mike Schuster **Yonghui Wu** **Dale Schuurmans**

Google Brain

Abstract

A key problem in structured output prediction is direct optimization of the task reward function that matters for test evaluation. This paper presents a simple and computationally efficient approach to incorporate task reward into a maximum likelihood framework. We establish a connection between the log-likelihood and regularized expected reward objectives, showing that at a zero temperature, they are approximately equivalent in the vicinity of the optimal solution. We show that optimal regularized expected reward is achieved when the conditional distribution of the outputs given the inputs is proportional to their exponentiated (temperature adjusted) rewards. Based on this observation, we optimize conditional log-probability of edited outputs that are sampled proportionally to their scaled exponentiated reward. We apply this framework to optimize edit distance in the output label space. Experiments on speech recognition and machine translation for neural sequence to sequence models show notable improvements over a maximum likelihood baseline by using edit distance augmented maximum likelihood.

1 Introduction

Structured output prediction is ubiquitous in machine learning. Recent advances in natural language processing, machine translation, and speech recognition hinge on the development of better discriminative models for structured outputs and sequences. The foundations of learning structured output models were established by seminal work on graph transformer networks [21], conditional random fields (CRFs) [20], and structured large margin methods [40, 38], which demonstrate how generalization performance can be significantly improved, when one considers the joint effects of predictions across multiple output components during training. These models have evolved into their deep neural counterparts [35, 1] by the use of recurrent neural networks (RNN) with LSTM [16] and GRU [8] cells and attention mechanisms [3].

A key problem in structured output prediction has always been enabling direct optimization of the task reward (loss) that is used for test evaluation. For example, in machine translation one seeks better BLEU scores, and in speech recognition better word error rates. Not

surprisingly, almost all task reward metrics are not differentiable, and hence hard to optimize. Neural sequence models (*e.g.* [35, 3]) use a maximum likelihood (ML) framework to maximize the conditional probability of the ground-truth outputs given corresponding inputs. These models do not explicitly consider the task reward during training, hoping that conditional log-likelihood would serve as a good surrogate for the task reward. Such methods make no distinction between alternative incorrect outputs: log-probability is only measured on the ground-truth input-output pairs, and all alternative outputs are equally penalized, whether near or far from the ground-truth target. We believe that one can improve upon maximum likelihood sequence models, if the difference in the rewards of alternative outputs is taken into account.

Standard ML training, despite its limitations, enables training deep RNN models leading to revolutionary advances in machine translation [35, 3, 25] and speech recognition [7, 9, 10]. A key property of ML training for locally normalized RNN models is that the objective function factorizes into individual loss terms, which could be *efficiently* optimized using stochastic gradient descend (SGD). In particular, ML training does not require any form of inference or sampling from the model during training, which leads to computationally efficient and easy to implementations. By contrast, one may consider large margin and search-based structured prediction [11] formulations for training RNNs (*e.g.* [46]). Such methods incorporate some task reward approximation during training, but the behavior of the approximation is not well understood, especially for deep neural nets. Moreover, these methods require some form of inference at training time, which slows down training.

Alternatively, one can use reinforcement learning (RL) algorithms, such as policy gradient [44], to optimize expected task reward during training [30, 2]. Even though expected task reward seems like a natural objective, direct policy optimization faces significant challenges: unlike ML, the gradient for a mini-batch of training examples is extremely noisy and has a high variance; gradients need to be estimated via sampling from the model, which is a non-stationary distribution; the reward is often sparse in a high-dimensional output space, which makes it difficult to find any high value predictions, preventing learning from getting off the ground; and, finally, maximizing reward does not explicitly consider the supervised labels, which seems inefficient. In fact, all previous attempts at direct policy optimization for structured output prediction has started by bootstrapping from a previously trained ML solution [30, 2, 32] and they use several heuristics and tricks to make learning stable.

This paper presents a new approach to task reward optimization that combines the computational efficiency and simplicity of ML with the conceptual advantages of expected reward maximization. Our algorithm called *reward augmented maximum likelihood (RML)* simply adds a sampling step on top of the typical likelihood objective. Instead of optimizing conditional log-likelihood on training input-output pairs, given each training input, we first sample an output proportional to its exponentiated scaled reward. Then, we optimize log-likelihood on such auxiliary output samples given corresponding inputs. When the reward for an output is defined as its similarity to a ground-truth output, then the output sampling distribution is peaked at the ground-truth output, and its concentration is controlled by a temperature hyper-parameter.

Our theoretical analysis shows that the RML and regularized expected reward objectives optimize a KL divergence between the exponentiated reward and model distributions in opposite directions. Further, we show that at non-zero temperatures, the gap between the two criteria can be expressed by a difference of variances measured on interpolating distributions. This observation reveals how entropy regularized expected reward can be estimated by sampling from exponentiated scaled rewards, rather than sampling from the model distribution.

Remarkably, we find that the RML approach achieves significantly improved results over state of the art maximum likelihood RNNs. We show consistent improvement on both speech recognition (TIMIT dataset) and machine translation (WMT'14 dataset), where output sequences are sampled according to their edit distance to the ground-truth outputs. Surprisingly, we find that the best performance is achieved with output sampling distributions that put a lot of the weight away from the ground-truth outputs. In fact, in our experiments, the training algorithm rarely sees the original unperturbed outputs. Our results give further evidence that models trained with imperfect outputs and their reward values can improve upon models that are only exposed to a single ground-truth output per input [15, 24, 42].

2 Reward augmented maximum likelihood

Given a dataset of input-output pairs, $\mathcal{D} \equiv \{(\mathbf{x}^{(i)}, \mathbf{y}^{*(i)})\}_{i=1}^{N}$, structured output models learn a parametric score function $p_\theta(\mathbf{y} \mid \mathbf{x})$, which scores different output hypotheses, $\mathbf{y} \in \mathcal{Y}$. We assume that the set of possible output, $\mathcal{Y}$ is finite, *e.g.* English sentences up to a maximum length. In a probabilistic model, the score function is normalized, while in a large-margin model the score may not be normalized. In either case, once the score function is learned, given an input $\mathbf{x}$, the model predicts an output $\widehat{\mathbf{y}}$ achieving maximal score,

$$\widehat{\mathbf{y}}(\mathbf{x}) = \underset{\mathbf{y}}{\operatorname{argmax}}\, p_\theta(\mathbf{y} \mid \mathbf{x}) \,. \tag{1}$$

If this optimization is intractable, approximate inference (*e.g.* beam search) is used. We use a reward function $r(\mathbf{y}, \mathbf{y}^*)$ to evaluate different outputs against ground-truth outputs. Given a test dataset $\mathcal{D}'$, one computes $\sum_{(\mathbf{x},\mathbf{y}^*)\in\mathcal{D}'} r(\widehat{\mathbf{y}}(\mathbf{x}), \mathbf{y}^*)$ as a measure of empirical reward, and models with larger empirical reward are preferred. Ideally, one hopes to optimize empirical reward during training too.

However, since empirical reward is not amenable to numerical optimization, one often considers optimizing alternative differentiable objectives. Maximum likelihood (ML) framework tries to minimize negative log-likelihood of the parameters given the data,

$$\mathcal{L}_{\mathrm{ML}}(\boldsymbol{\theta};\mathcal{D}) = \sum_{(\mathbf{x},\mathbf{y}^*)\in\mathcal{D}} -\log p_\theta(\mathbf{y}^* \mid \mathbf{x}) \; . \tag{2}$$

Minimizing this objective, increases the conditional probability of the target outputs, $\log p_\theta(\mathbf{y}^* \mid \mathbf{x})$, while decreasing the conditional probability of alternative wrong outputs. According to this objective, all of the negative outputs are equally wrong, and none is preferred over the rest.

By contrast, reinforcement learning (RL) advocates optimizing expected reward (with a maximum entropy regularizer [45, 27]), which is formulated as minimization of the following objective,

$$\mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta};\tau,\mathcal{D}) = \sum_{(\mathbf{x},\mathbf{y}^*)\in\mathcal{D}} \Big\{ -\tau\mathbb{H}\left(p_\theta(\mathbf{y} \mid \mathbf{x})\right) - \sum_{\mathbf{y}\in\mathcal{Y}} p_\theta(\mathbf{y} \mid \mathbf{x})\, r(\mathbf{y},\mathbf{y}^*) \Big\}, \tag{3}$$

where $r(\mathbf{y},\mathbf{y}^*)$ denotes the reward function, *e.g.* negative edit distance or BLEU score, τ controls the degree of regularization, and $\mathbb{H}\left(p\right)$ is the entropy of a distribution p, *i.e.* $\mathbb{H}\left(p(\mathbf{y})\right) = -\sum_{\mathbf{y}\in\mathcal{Y}} p(\mathbf{y})\log p(\mathbf{y})$. It is well-known that optimizing $\mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta};\tau)$ using SGD is challenging because of the large variance of the gradients. Below we describe how ML and RL objectives are related, and propose a hybrid between the two that combines their benefits for supervised learning.

Let's define a distribution in the output space, termed *exponentiated payoff distribution*, that is central in linking ML and RL objectives:

$$q(\mathbf{y} \mid \mathbf{y}^*;\tau) = \frac{1}{Z(\mathbf{y}^*,\tau)} \exp\left\{r(\mathbf{y},\mathbf{y}^*)/\tau\right\} , \tag{4}$$

where $Z(\mathbf{y}^*,\tau) = \sum_{\mathbf{y}\in\mathcal{Y}} \exp\left\{r(\mathbf{y},\mathbf{y}^*)/\tau\right\}$. One can verify that the global minimum of $\mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta};\tau)$, *i.e.* optimal regularized expected reward, is achieved when the model distribution matches exactly with the exponentiated payoff distribution, *i.e.* $p_\theta(\mathbf{y} \mid \mathbf{x}) = q(\mathbf{y} \mid \mathbf{y}^*;\tau)$. To see this, we re-express the objective function in (3) in terms of a KL divergence between $p_\theta(\mathbf{y} \mid \mathbf{x})$ and $q(\mathbf{y} \mid \mathbf{y}^*;\tau)$,

$$\sum_{(\mathbf{x},\mathbf{y}^*)\in\mathcal{D}} D_{\mathrm{KL}}\left(p_\theta(\mathbf{y} \mid \mathbf{x}) \parallel q(\mathbf{y} \mid \mathbf{y}^*;\tau)\right) = \frac{1}{\tau}\mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta};\tau) + \mathrm{constant} \; , \tag{5}$$

where the constant on the RHS is $\sum_{(\mathbf{x},\mathbf{y}^*)\in\mathcal{D}} \log Z(\mathbf{y}^*,\tau)$. Thus, the minimum of $D_{\mathrm{KL}}\left(p_\theta \parallel q\right)$ and $\mathcal{L}_{\mathrm{RL}}$ is achieved when $p_\theta = q$. At $\tau = 0$, when there is no entropy regularization, the optimal p_θ is a delta distribution, $p_\theta(\mathbf{y} \mid \mathbf{x}) = \delta(\mathbf{y} \mid \mathbf{y}^*)$, where $\delta(\mathbf{y} \mid \mathbf{y}^*) = 1$ at $\mathbf{y} = \mathbf{y}^*$ and 0 at $\mathbf{y} \neq \mathbf{y}^*$. Note that $\delta(\mathbf{y} \mid \mathbf{y}^*)$ is equivalent to the exponentiated payoff distribution in the limit as $\tau \to 0$.

Going back to the log-likelihood objective, one can verify that (2) is equivalent to a KL divergence in the opposite direction between a delta distribution $\delta(\mathbf{y} \mid \mathbf{y}^*)$ and the model distribution $p_\theta(\mathbf{y} \mid \mathbf{x})$,

$$\sum_{(\mathbf{x},\mathbf{y}^*)\in\mathcal{D}} D_{\mathrm{KL}}\left(\delta(\mathbf{y} \mid \mathbf{y}^*) \parallel p_\theta(\mathbf{y} \mid \mathbf{x})\right) = \mathcal{L}_{\mathrm{ML}}(\boldsymbol{\theta}) \; . \tag{6}$$

There is no constant on the RHS, as the entropy of a delta distribution is zero, *i.e.* $\mathbb{H}\left(\delta(\mathbf{y} \mid \mathbf{y}^*)\right) = 0$.

We propose a method called *reward-augmented maximum likelihood (RML)* which generalizes ML by allowing a non-zero temperature parameter in the exponentiated payoff distribution, while still optimizing the KL divergence in the ML direction. The RML objective function takes the form,

$$\mathcal{L}_{\mathrm{RML}}(\boldsymbol{\theta};\tau,\mathcal{D}) = \sum_{(\mathbf{x},\mathbf{y}^*)\in\mathcal{D}} \Big\{ -\sum_{\mathbf{y}\in\mathcal{Y}} q(\mathbf{y} \mid \mathbf{y}^*;\tau)\log p_\theta(\mathbf{y} \mid \mathbf{x}) \Big\} , \tag{7}$$

which can be re-expressed in terms of a KL divergence as follows,

$$\sum_{(\mathbf{x},\mathbf{y}^*)\in\mathcal{D}} D_{\mathrm{KL}}\left(q(\mathbf{y} \mid \mathbf{y}^*;\tau) \parallel p_\theta(\mathbf{y} \mid \mathbf{x})\right) = \mathcal{L}_{\mathrm{RML}}(\boldsymbol{\theta};\tau) + \mathrm{constant} \; , \tag{8}$$

where the constant is $-\sum_{(\mathbf{x},\mathbf{y}^*)\in\mathcal{D}} \mathbb{H}\left(q(\mathbf{y} \mid \mathbf{y}^*,\tau)\right)$. Note that the temperature parameter, $\tau \geq 0$, serves as a hyper-parameter that controls the smoothness of the optimal distribution around correct targets by taking into account the reward function in the output space. The objective functions $\mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta};\tau)$ and $\mathcal{L}_{\mathrm{RML}}(\boldsymbol{\theta};\tau)$, have the same global optimum of p_θ, but they optimize a KL divergence in opposite directions. We characterize the difference between these two objectives below, showing that they are equivalent up to their first order Taylor approximations. For optimization convenience, we focus on minimizing $\mathcal{L}_{\mathrm{RML}}(\boldsymbol{\theta};\tau)$ to achieve a good solution for $\mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta};\tau)$.

2.1 Optimization

Optimizing the reward augmented maximum likelihood (RML) objective, $\mathcal{L}_{\mathrm{RML}}(\boldsymbol{\theta};\tau)$, is straightforward if one can draw unbiased samples from $q(\mathbf{y} \mid \mathbf{y}^*;\tau)$. We can express the gradient of $\mathcal{L}_{\mathrm{RML}}$ in terms of an expectation over samples from $q(\mathbf{y} \mid \mathbf{y}^*;\tau)$,

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{RML}}(\boldsymbol{\theta};\tau) \;=\; \mathbb{E}_{q(\mathbf{y}\mid\mathbf{y}^*;\tau)}\big[-\nabla_{\boldsymbol{\theta}}\log p_\theta(\mathbf{y}\mid\mathbf{x})\big]\,. \tag{9}$$

Thus, to estimate $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{RML}}(\boldsymbol{\theta};\tau)$ for a mini-batch of examples for SGD, one draws $\mathbf{y}$ samples given mini-batch $\mathbf{y}^*$'s and then optimizes log-likelihood on such samples by following the mean gradient. At a temperature $\tau = 0$, this reduces to always sampling $\mathbf{y}^*$, hence ML training with no sampling.

By contrast, the gradient of $\mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta};\tau)$, based on likelihood ratio methods, takes the form,

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta};\tau) \;=\; \mathbb{E}_{p_\theta(\mathbf{y}\mid\mathbf{x})}\big[-\nabla_{\boldsymbol{\theta}}\log p_\theta(\mathbf{y}\mid\mathbf{x})\cdot r(\mathbf{y},\mathbf{y}^*)\big]\,. \tag{10}$$

There are several critical differences between (9) and (10) that make SGD optimization of $\mathcal{L}_{\mathrm{RML}}(\boldsymbol{\theta};\tau)$ more desirable. First, in (9), one has to sample from a stationary distribution, the so called exponentiated payoff distribution, whereas in (10) one has to sample from the model distribution as it is evolving. Not only sampling from the model could slow down training, but also one needs to employ several tricks to get a better estimate of the gradient of $\mathcal{L}_{\mathrm{RL}}$ [30]. A body of literature in reinforcement learning focuses on reducing the variance of (10) by using smart techniques such as *actor-critique* methods [36, 12]. Further, the reward is often sparse in a high-dimensional output space, which makes finding any reasonable predictions challenging, when (10) is used to refine a randomly initialized model. Thus, smart model initialization is needed. By contrast, we initialize the models randomly and refine them using (9).

2.2 Sampling from the exponentiated payoff distribution

For computing the gradient of the model, using the RML approach, one needs to sample auxiliary outputs from the exponentiated payoff distribution, $q(\mathbf{y} \mid \mathbf{y}^*;\tau)$. This sampling is the price that we have to pay to learn with rewards. One should contrast this with loss-augmented inference in structured large margin methods, and sampling from the model in RL. We believe sampling outputs proportional to exponentiated rewards is more efficient and effective in many cases.

Experiments in this paper use reward values defined by either negative Hamming distance or negative edit distance. We sample from $q(\mathbf{y} \mid \mathbf{y}^*;\tau)$ by stratified sampling, where we first select a particular distance, and then sample an output with that distance value. Here we focus on edit distance sampling, as Hamming distance sampling is a simpler special case. Given a sentence $\mathbf{y}^*$ of length m, we count the number of sentences within an edit distance e, where $e \in \{0,\ldots,2m\}$. Then, we reweight the counts by $\exp\{-e/\tau\}$ and normalize. Let $c(e,m)$ denote the number of sentences at an edit distance e from a sentence of length m. First, note that a deletion can be thought as a substitution with a nil token. This works out nicely because given a vocabulary of length v, for each insertion we have v options, and for each substitution we have $v-1$ options, but including the nil token, there are v options for substitutions too. When $e = 1$, there are m possible substitutions and $m+1$ insertions. Hence, in total there are $(2m+1)v$ sentences at an edit distance of 1. Note, that exact computation of $c(e,m)$ is difficult if we consider all edge cases, for example when there are repetitive words in $\mathbf{y}^*$, but ignoring such edge cases we can come up with approximate counts that are reliable for sampling. When $e > 1$, we estimate $c(e,m)$ by

$$c(e,m) = \sum_{s=0}^{m} \binom{m}{s}\binom{m+e-2s}{e-s} v^e\,, \tag{11}$$

where s enumerates over the number of substitutions. Once s tokens are substituted, then those s positions lose their significance, and the insertions before and after such tokens could be merged. Hence, given s substitutions, there are really $m-s$ reference positions for $e-s$ possible insertions. Finally, one can sample according to BLEU score or other sequence metrics by importance sampling where the proposal distribution could be edit distance sampling above.

3 RML analysis

In the RML framework, we find the model parameters by minimizing the objective (7) instead of optimizing the RL objective, *i.e.* regularized expected reward in (3). The difference lies in minimizing $D_{\text{KL}}\left(q(\mathbf{y} \mid \mathbf{y}^*; \tau) \parallel p_\theta(\mathbf{y} \mid \mathbf{x})\right)$ instead of $D_{\text{KL}}\left(p_\theta(\mathbf{y} \mid \mathbf{x}) \parallel q(\mathbf{y} \mid \mathbf{y}^*; \tau)\right)$. For convenience, let's refer to $q(\mathbf{y} \mid \mathbf{y}^*; \tau)$ as q, and $p_\theta(\mathbf{y} \mid \mathbf{x})$ as p. Here, we characterize the difference between the two divergences, $D_{\text{KL}}\left(q \parallel p\right) - D_{\text{KL}}\left(p \parallel q\right)$, and use this analysis to motivate the RML approach.

We will initially consider the KL divergence in its more general form as a Bregman divergence, which will make some of the key properties clearer. A Bregman divergence is defined by a strictly convex, differentiable, closed potential function $F : \mathcal{F} \to \mathbb{R}$ [5]. Given F and two points $p, q \in \mathcal{F}$, the corresponding Bregman divergence $D_F : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^+$ is defined by

$$D_F\left(p \parallel q\right) = F\left(p\right) - F\left(q\right) - \left(p - q\right)^{\mathsf{T}} \nabla F\left(q\right) \ , \tag{12}$$

the difference between the strictly convex potential at p and its first order Taylor approximation expanded about q. Clearly this definition is not symmetric between p and q. By the strict convexity of F it follows that $D_F\left(p \parallel q\right) \geq 0$ with $D_F\left(p \parallel q\right) = 0$ if and only if $p = q$. To characterize the difference between opposite Bregman divergences, we provide a simple result that relates the two directions for an arbitrary Bregman divergence. Let H_F denote the Hessian of F.

Proposition 1. *For any twice differentiable strictly convex closed potential* F*, and* $p, q \in \text{int}(\mathcal{F})$*:*

$$D_F\left(q \parallel p\right) = D_F\left(p \parallel q\right) + \tfrac{1}{4}(q - p)^{\mathsf{T}} \big(H_F(b) - H_F(a)\big)(q - p) \tag{13}$$

for some $a = (1-\alpha)p + \alpha q$, $(0 \leq \alpha \leq \frac{1}{2})$, $b = (1-\beta)q + \beta p$, $(0 \leq \beta \leq \frac{1}{2})$*. (see supp. material)*

For probability vectors $p, q \in \Delta^{|\mathcal{Y}|}$ and a potential $F\left(p\right) = -\tau \mathbb{H}\left(p\right)$, $D_F\left(p \parallel q\right) = \tau D_{\text{KL}}\left(p \parallel q\right)$. Let $f^* : \mathbb{R}^{|\mathcal{Y}|} \to \Delta^{|\mathcal{Y}|}$ denote a normalized exponential operator that takes a real-valued logit vector and turns it into a probability vector. Let r and s denote real-valued logit vectors such that $q = f^*(r/\tau)$ and $p = f^*(s/\tau)$. Below, we characterize the gap between $D_{\text{KL}}\left(p(y) \parallel q(y)\right)$ and $D_{\text{KL}}\left(q(y) \parallel p(y)\right)$ in terms of the difference between $s(y)$ and $r(y)$.

Proposition 2. *The KL divergence between* p *and* q *in two directions can be expressed as,*

$$\begin{aligned} D_{\text{KL}}\left(p \parallel q\right) &= D_{\text{KL}}\left(q \parallel p\right) + \tfrac{1}{4\tau^2} \operatorname{Var}_{y \sim f^*(b/\tau)}\left[s(y) - r(y)\right] - \tfrac{1}{4\tau^2} \operatorname{Var}_{y \sim f^*(a/\tau)}\left[s(y) - r(y)\right] \\ &< D_{\text{KL}}\left(q \parallel p\right) + \|s - r\|_2^2, \end{aligned}$$

for some $a = (1-\alpha)r + \alpha s$, $(0 \leq \alpha \leq \frac{1}{2})$, $b = (1-\beta)s + \beta r$, $(0 \leq \beta \leq \frac{1}{2})$*. (see supp. material)*

Given Proposition 2, one can relate RL and RML objectives, $\mathcal{L}_{\text{RL}}(\boldsymbol{\theta}; \tau)$ (5) and $\mathcal{L}_{\text{RML}}(\boldsymbol{\theta}; \tau)$ (8), as,

$$\mathcal{L}_{\text{RL}} = \tau \mathcal{L}_{\text{RML}} + \tfrac{1}{4\tau} \sum_{(\mathbf{x}, \mathbf{y}^*) \in \mathcal{D}} \Big\{ \operatorname{Var}_{\mathbf{y} \sim f^*(b/\tau)}\left[s(\mathbf{y}) - r(\mathbf{y})\right] - \operatorname{Var}_{\mathbf{y} \sim f^*(a/\tau)}\left[s(\mathbf{y}) - r(\mathbf{y})\right] \Big\}, \tag{14}$$

where $s(\mathbf{y})$ denotes τ-scaled logits predicted by the model such that $p_\theta(\mathbf{y} \mid \mathbf{x}) = f^*(s(\mathbf{y})/\tau)$, and $r(\mathbf{y}) = r(\mathbf{y}, \mathbf{y}^*)$. The gap between regularized expected reward (5) and τ-scaled RML criterion (8) is simply a difference of two variances, whose magnitude decreases with increasing regularization. Proposition 2 also shows an opportunity for learning algorithms: if τ is chosen so that $q = f^*(r/\tau)$, then $f^*(a/\tau)$ and $f^*(b/\tau)$ have lower variance than p (which can always be achieved for sufficiently small τ provided p is not deterministic), then the expected regularized reward under p, and its gradient for training, can be exactly estimated, in principle, by including the extra variance terms and sampling from more focused distributions than p. Although we have not yet incorporated approximations to the additional variance terms into RML, this is an interesting research direction.

4 Related Work

The literature on structure output prediction is now quite vast, with three broad categories: (a) supervised learning approaches that ignore task reward and use only supervision; (b) reinforcement learning approaches that use only task reward and ignore supervision; and (c) hybrid approaches that attempt to exploit both supervision and task reward.

Work in category (a) includes classical conditional random field approaches and the more recent maximum likelihood training of RNNs. It also includes more recent approaches, such as [6, 19, 34], that ignore task reward, but attempt to perturb the training inputs and supervised training structures in a way that improves the robustness (and hopefully the generalization) of the resulting prediction model. Related are ideas for improving approximate maximum likelihood training for intractable models by passing the gradient calculation through an approximate inference procedure [13, 33]. Although these approaches offer improvements to standard maximum likelihood, we feel they are fundamentally limited by not incorporating task reward. The DAGGER method [31] also focuses on using supervision only, but can be extended by replacing the surrogate loss with a task loss; even then, this approach assumes that an expert is available to label every alternative sequence, which does not fit the current scenario.

By contrast, work in category (b) includes reinforcement learning approaches that only consider task reward and do not use any other supervision. Beyond the traditional reinforcement learning approaches, such as policy gradient [44, 37], and actor-critic [36], Q-learning [41], this category includes SEARN [11]. There is some relationship to the work presented here and work on relative entropy policy search [28], and policy optimization via expectation maximization [43] and KL-divergence [17, 39], however none of these bridge the gap between the two directions of the KL-divergence, nor do they consider any supervision data as we do here.

This paper clearly falls in category (c), where there is also a substantial body of related work that has considered how to both exploit supervision information and inform training by task reward. We have already mentioned large margin structured output training [38], which explicitly uses supervision but only considers an upper bound surrogate for task loss. Recently, this line of research has been improved to directly consider task reward [14], but is limited to a perceptron like training approach and continues to require reward augmented inference that cannot be efficiently achieved for general task rewards. A recent extension to gradient based training of deep RNN models has recently been achieved for structured prediction [4], but reward augmented inference remains required, and many heuristics were needed to apply the technique in practice. We have also already mentioned work that attempts to maximize task reward by bootstrapping from a maximum likelihood policy [30, 32], but such an approach only makes limited use of supervision. Some work in robotics has considered exploiting supervision as a means to provide indirect sampling guidance to improve policy search methods that maximize task reward [22, 23], but these approaches do not make use of maximum likelihood training. Most relevant is the work [18] which explicitly incorporates supervision in the policy evaluation phase of a policy iteration procedure that otherwise seeks to maximize task reward. Although interesting, this approach only considers a greedy policy form that does not lend itself to being represented as a deep RNN, and has not been applied to structured output prediction.

One advantage of the RML framework is its computational efficiency at training time. By contrast, RL and scheduled sampling [6] require sampling from the model, which can slow down the gradient computation by $2\times$. Structural SVM requires loss-augmented inference which is often more expensive than sampling from the model. Our framework only requires sampling from a fixed exponentated payoff distribution, which can be thought as a form of input pre-processing. This pre-processing can be parallelized by model training by having a specific thread handling loading the data and augmentation.

5 Experiments

We compare our approach, reward augmented maximum likelihood (*RML*), with standard maximum likelihood (*ML*) training on sequence prediction tasks using state-of-the-art attention-based recurrent neural networks [35, 3]. Our experiments demonstrate that the RML approach considerably outperforms ML baseline on both speech recognition and machine translation tasks.

5.1 Speech recognition

For experiments on speech recognition, we use the TIMIT dataset; a standard benchmark for clean phone recognition. This dataset comprises recordings from different speakers reading ten phonetically rich sentences covering major dialects of American English. We use the standard train / dev / test splits suggested by the Kaldi toolkit [29].

As the sequence prediction model, we use an attention-based encoder-decoder recurrent model of [7] with three 256-dimensional LSTM layers for encoding and one 256-dimensional LSTM layer for decoding. We do not modify the neural network architecture or its gradient computation in any way, but we only change the output targets fed into the network for gradient computation and SGD update.

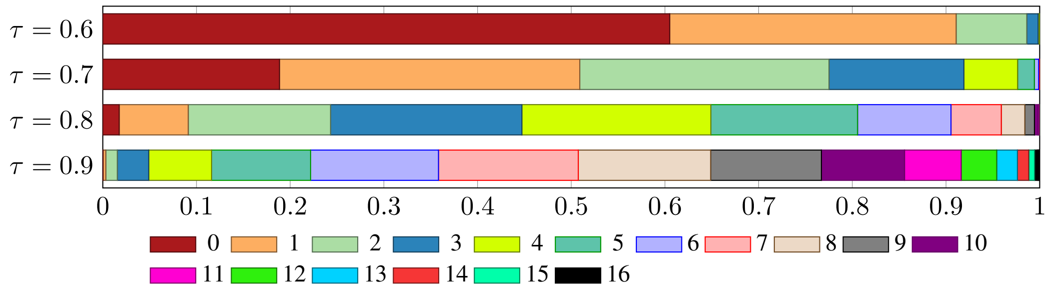

Figure 1: Fraction of different number of edits applied to a sequence of length 20 for different τ. At $\tau = 0.9$, augmentations with 5 to 9 edits are sampled with a probability > 0.1. [view in color]

Method	Dev set	Test set
ML baseline	20.87 $(-0.2, +0.3)$	22.18 $(-0.4, +0.2)$
RML, $\tau = 0.60$	19.92 $(-0.6, +0.3)$	21.65 $(-0.5, +0.4)$
RML, $\tau = 0.65$	19.64 $(-0.2, +0.5)$	21.28 $(-0.6, +0.4)$
RML, $\tau = 0.70$	18.97 $(-0.1, +0.1)$	21.28 $(-0.5, +0.4)$
RML, $\tau = 0.75$	18.44 $(-0.4, +0.4)$	20.15 $(-0.4, +0.4)$
RML, $\tau = 0.80$	18.27 $(-0.2, +0.1)$	19.97 $(-0.1, +0.2)$
RML, $\tau = 0.85$	18.10 $(-0.4, +0.3)$	19.97 $(-0.3, +0.2)$
RML, $\boldsymbol{\tau = 0.90}$	**18.00** $(-0.4, +0.3)$	**19.89** $(-0.4, +0.7)$
RML, $\tau = 0.95$	18.46 $(-0.1, +0.1)$	20.12 $(-0.2, +0.1)$
RML, $\tau = 1.00$	18.78 $(-0.6, +0.8)$	20.41 $(-0.2, +0.5)$

Table 1: Phone error rates (PER) for different methods on TIMIT dev and test sets. Average PER of 4 independent training runs is reported.

The input to the network is a standard sequence of 123-dimensional log-mel filter response statistics. Given each input, we generate new outputs around ground truth targets by sampling according to the exponentiated payoff distribution. We use negative edit distance as the measure of reward. Our output augmentation process allows insertions, deletions, and substitutions.

An important hyper-parameter in our framework is the temperature parameter, τ, controlling the degree of output augmentation. We investigate the impact of this hyper-parameter and report results for τ selected from a candidate set of $\tau \in \{0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0\}$. At a temperature of $\tau = 0$, outputs are not augmented at all, but as τ increases, more augmentation is generated. Figure 1 depicts the fraction of different numbers of edits applied to a sequence of length 20 for different values of τ. These edits typically include very small number of deletions, and roughly equal number of insertions and substitutions. For insertions and substitutions we uniformly sample elements from a vocabulary of 61 phones. According to Figure 1, at $\tau = 0.6$, more than 60% of the outputs remain intact, while at $\tau = 0.9$, almost all target outputs are being augmented with 5 to 9 edits being sampled with a probability larger than 0.1. We note that the augmentation becomes more severe as the outputs get longer.

The phone error rates (PER) on both dev and test sets for different values of τ and the ML baseline are reported in Table 1. Each model is trained and tested 4 times, using different random seeds. In Table 1, we report average PER across the runs, and in parenthesis the difference of average error to minimum and maximum error. We observe that a temperature of $\tau = 0.9$ provides the best results, outperforming the ML baseline by 2.9% PER on the dev set and 2.3% PER on the test set. The results consistently improve when the temperature increases from 0.6 to 0.9, and they get worse beyond $\tau = 0.9$. It is surprising to us that not only the model trains with such a large amount of augmentation at $\tau = 0.9$, but also it significantly improves upon the baseline. Finally, we note that previous work [9, 10] suggests several refinements to improve sequence to sequence models on TIMIT by adding noise to the weights and using more focused forward-moving attention mechanism. While these refinements are interesting and they could be combined with the RML framework, in this work, we do not implement such refinements, and focus specifically on a fair comparison between the ML baseline and the RML method.

Method	Average BLEU	Best BLEU
ML baseline	36.50	36.87
RML, $\tau = 0.75$	36.62	36.91
RML, $\tau = 0.80$	36.80	37.11
RML, $\boldsymbol{\tau = 0.85}$	$\mathbf{36.91}$	$\mathbf{37.23}$
RML, $\tau = 0.90$	36.69	37.07
RML, $\tau = 0.95$	36.57	36.94

Table 2: Tokenized BLEU score on WMT'14 English to French evaluated on newstest-2014 set. The RML approach with different τ considerably improves upon the maximum likelihood baseline.

5.2 Machine translation

We evaluate the effectiveness of the proposed approach on WMT'14 English to French machine translation benchmark. Translation quality is assessed using *tokenized BLEU* score, to be consistent with previous work on neural machine translation [35, 3, 26]. Models are trained on the full 36M sentence pairs from the training set of WMT'14, and evaluated on 3003 sentence pairs from newstest-2014 test set. To keep the sampling process efficient and simple on such a large corpus, we again augment output sentences based on edit distance, but we only allow substitution (no insertion or deletion). One may consider insertions and deletions or sampling according to exponentiated sentence BLEU scores, but we leave that to future work.

As the conditional sequence prediction model, we use an attention-based encoder-decoder recurrent neural network similar to [3], but we use multi-layer encoder and decoder networks comprising three layers of 1024 LSTM cells. As suggested by [3], for computing the softmax attention vectors, we build a feedforward neural network with 1024 hidden units, which is fed with the last encoder and the first decoder layers. Across all of our experiments we keep the network architecture and the hyper-parameters the same. We find that all of the models achieve their peak performance after about 4 epochs of training, once we anneal the learning rates. To reduce the noise in BLEU score evaluation, we report both peak BLEU score and BLEU score averaged among about 70 evaluations of the model while doing the fifth epoch of training.

Table 2 summarizes our experimental results on WMT'14. We note that our ML translation baseline is quite strong, if not the best among neural machine translation models [35, 3, 26], achieving very competitive performance for a single model. Even given such a strong baseline, the RML approach consistently improves the results. Our best model with a temperature $\tau = 0.85$ improves average BLEU by 0.4, and best BLEU by 0.35 points, which is a considerable improvement. Again we observe that as we increase the amount of augmentation from $\tau = 0.75$ to $\tau = 0.85$ the results consistently get better, and then they start to get worse with more augmentation.

Details. We train the models using asynchronous SGD with 12 replicas without momentum. We use mini-batches of size 128. We initially use a learning rate of 0.5, which we exponentially decay down to 0.05 after $800K$ training steps. We keep evaluating the models between 1.1 and 1.3 million steps and we report average and peak BLEU scores in Table 2. We use a vocabulary $200K$ words for the source language and $80K$ for the target language. We shard the $80K$-way softmax onto 8 GPUs for speedup. We only consider training sentences that are up to 80 tokens. We replace rare words with several UNK tokens based on their first and last characters. At inference time, we replace UNK tokens in the output sentences by copying source words according to largest attention dimensions. This rare word handling bares some similarity to [26].

6 Conclusion

We presented a learning algorithm for structured output prediction, which generalizes maximum likelihood training by enabling direct optimization of the task evaluation metric. Our method is computationally efficient and simple to implement, and it only requires augmenting the output targets used for training a maximum likelihood model. We present a method for sampling from output augmentations with increasing edit distance, and we show how using such augmented outputs for training improves maximum likelihood models by a considerable margin, on both machine translation and speech recognition tasks. We believe this framework is applicable to a wide range of probabilistic models with arbitrary reward functions. In future work, we intend to explore the applicability of this framework to other probabilistic models on tasks with other evaluations metrics.

References

[1] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins. Globally normalized transition-based neural networks. *arXiv:1603.06042*, 2016.

[2] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, and Y. Bengio. An actor-critic algorithm for sequence prediction. *arXiv:1607.07086*, 2016.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[4] D. Bahdanau, D. Serdyuk, P. Brakel, N. R. Ke, J. Chorowski, A. C. Courville, and Y. Bengio. Task loss estimation for sequence prediction. *arXiv:1511.06456*, 2015.

[5] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *JMLR*, 2005.

[6] S. Bengio, O. Vinyals, N. Jaitly, and N. M. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *NIPS*, 2015.

[7] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals. Listen, attend and spell. *ICASSP*, 2016.

[8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP*, 2014.

[9] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: first results. *arXiv:1412.1602*, 2014.

[10] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. *NIPS*, 2015.

[11] H. Daumé, III, J. Langford, and D. Marcu. Search-based structured prediction. *Mach. Learn. J.*, 2009.

[12] T. Degris, P. M. Pilarski, and R. S. Sutton. Model-free reinforcement learning with continuous action in practice. *ACC*, 2012.

[13] J. Domke. Generic methods for optimization-based modeling. *AISTATS*, 2012.

[14] T. Hazan, J. Keshet, and D. A. McAllester. Direct loss minimization for structured prediction. *NIPS*, 2010.

[15] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[17] H. J. Kappen, V. Gómez, and M. Opper. Optimal control as a graphical model inference problem. *Mach. Learn. J.*, 2012.

[18] B. Kim, A. M. Farahmand, J. Pineau, and D. Precup. Learning from limited demonstrations. *NIPS*, 2013.

[19] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. *ICML*, 2016.

[20] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, 2001.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient based learning applied to document recognition. *Proceedings of IEEE*, 1998.

[22] S. Levine and V. Koltun. Guided policy search. *ICML*, 2013.

[23] S. Levine and V. Koltun. Variational policy search via trajectory optimization. *NIPS*, 2013.

[24] D. Lopez-Paz, B. Schölkopf, L. Bottou, and V. Vapnik. Unifying distillation and privileged information. *ICLR*, 2016.

[25] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *EMNLP*, 2015.

[26] M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba. Addressing the rare word problem in neural machine translation. *ACL*, 2015.

[27] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *ICML*, 2016.

[28] J. Peters, K. Mülling, and Y. Altün. Relative entropy policy search. *AAAI*, 2010.

[29] D. Povey, A. Ghoshal, G. Boulianne, et al. The kaldi speech recognition toolkit. *ASRU*, 2011.

[30] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. *ICLR*, 2016.

[31] S. Ross, G. J. Gordon, and J. A. Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. *AISTATS*, 2010.

[32] D. Silver et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016.

[33] V. Stoyanov, A. Ropson, and J. Eisner. Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. *AISTATS*, 2011.

[34] E. Strubell, L. Vilnis, K. Silverstein, and A. McCallum. Learning dynamic feature selection for fast sequential prediction. *ACL*, 2015.

[35] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *NIPS*, 2014.

[36] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[37] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *NIPS*, 2000.

[38] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. *NIPS*, 2004.

[39] E. Todorov. Linearly-solvable markov decision problems. *NIPS*, 2006.

[40] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 2005.
[41] H. van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double q-learning. *arXiv:1509.06461*, 2015.
[42] V. Vapnik and R. Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *JMLR*, 2015.
[43] N. Vlassis, M. Toussaint, G. Kontes, and S. Piperidis. Learning model-free robot control by a Monte Carlo EM algorithm. *Autonomous Robots*, 2009.
[44] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn. J.*, 1992.
[45] R. J. Williams and J. Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 1991.
[46] S. Wiseman and A. M. Rush. Sequence-to-sequence learning as beam-search optimization. *arXiv:1606.02960*, 2016.

A Proofs

Proposition 1. *For any twice differentiable strictly convex closed potential F, and $p, q \in \operatorname{int}(\mathcal{F})$:*

$$D_F\left(q \,\|\, p\right) = D_F\left(p \,\|\, q\right) + \tfrac{1}{4}(q-p)^{\mathsf{T}}\big(H_F(b) - H_F(a)\big)(q-p) \tag{15}$$

for some $a = (1-\alpha)p + \alpha q$, $(0 \le \alpha \le \frac{1}{2})$, $b = (1-\beta)q + \beta p$, $(0 \le \beta \le \frac{1}{2})$.

Proof. Let $f(p)$ denote $\nabla F(p)$ and consider the midpoint $\frac{q+p}{2}$. One can express $F(\frac{q+p}{2})$ by two Taylor expansions around p and q. By Taylor's theorem there is an $a = (1-\alpha)p + \alpha q$ for $0 \le \alpha \le \frac{1}{2}$ and $b = \beta p + (1-\beta)q$ for $0 \le \beta \le \frac{1}{2}$ such that

$$\begin{aligned}
F(\tfrac{q+p}{2}) &= F(p) + (\tfrac{q+p}{2} - p)^\top f(p) + \tfrac{1}{2}(\tfrac{q+p}{2} - p)^\top H_F(a)(\tfrac{q+p}{2} - p) && (16)\\
&= F(q) + (\tfrac{q+p}{2} - q)^\top f(q) + \tfrac{1}{2}(\tfrac{q+p}{2} - q)^\top H_F(b)(\tfrac{q+p}{2} - q), && (17)\\
\text{hence,} \qquad 2F(\tfrac{q+p}{2}) &= 2F(p) + (q-p)^\top f(p) + \tfrac{1}{4}(q-p)^\top H_F(a)(q-p) && (18)\\
&= 2F(q) + (p-q)^\top f(q) + \tfrac{1}{4}(p-q)^\top H_F(b)(p-q). && (19)
\end{aligned}$$

Therefore,

$$\begin{aligned}
F(p) + F(q) - 2F(\tfrac{q+p}{2}) &= F(p) - F(q) - (p-q)^\top f(q) - \tfrac{1}{4}(p-q)^\top H_F(b)(p-q) && (20)\\
&= F(q) - F(p) - (q-p)^\top f(p) - \tfrac{1}{4}(q-p)^\top H_F(a)(q-p) && (21)\\
&= D_F\left(p \,\|\, q\right) - \tfrac{1}{4}(p-q)^\top H_F(b)(p-q) && (22)\\
&= D_F\left(q \,\|\, p\right) - \tfrac{1}{4}(q-p)^\top H_F(a)(q-p), && (23)
\end{aligned}$$

leading to the result. □

For the proof of Proposition 2, we first need to introduce a few definitions and background results. A Bregman divergence is defined from a strictly convex, differentiable, closed potential function $F : \mathcal{F} \to \mathbb{R}$, whose strictly convex conjugate $F^* : \mathcal{F}^* \to \mathbb{R}$ is given by $F^*(r) = \sup_{r \in \mathcal{F}} \langle r, q \rangle - F(q)$ [5]. Each of these potential functions have corresponding transfers, $f : \mathcal{F} \to \mathcal{F}^*$ and $f^* : \mathcal{F}^* \to \mathcal{F}$, given by the respective gradient maps $f = \nabla F$ and $f^* = \nabla F^*$. A key property is that $f^* = f^{-1}$ [5], which allows one to associate each object $q \in \mathcal{F}$ with its transferred image $r = f(q) \in \mathcal{F}^*$ and vice versa. The main property of Bregman divergences we exploit is that a divergence between any two domain objects can always be equivalently expressed as a divergence between their transferred images; that is, for any $p \in \mathcal{F}$ and $q \in \mathcal{F}$, one has [5]:

$$\begin{aligned}
D_F\left(p \,\|\, q\right) &= \quad F(p) - \langle p, r \rangle + F^*(r) = D_{F^*}\left(r \,\|\, s\right), && (24)\\
D_F\left(q \,\|\, p\right) &= F^*(s) - \langle s, q \rangle + F(q) \quad = D_{F^*}\left(s \,\|\, r\right), && (25)
\end{aligned}$$

where $s = f(p)$ and $r = f(q)$. These relations also hold if we instead chose $s \in \mathcal{F}^*$ and $r \in \mathcal{F}^*$ in the range space, and used $p = f^*(s)$ and $q = f^*(r)$. In general (24) and (25) are not equal.

Two special cases of the potential functions F and F^* are interesting as they give rise to KL divergences. These two cases include $F_\tau\left(p\right) = -\tau \mathbb{H}\left(p\right)$ and $F^*_\tau(s) = \tau \mathrm{l}se(s/\tau) = \tau \log \sum_y \exp\left(s(y)/\tau\right)$, where $\mathrm{l}se(\cdot)$ denotes the log-sum-exp operator. The respective gradient maps are $f_\tau(p) = \tau(\log(p) + \mathbf{1})$ and $f^*_\tau(s) = f^*(s/\tau) =$

$\frac{1}{\sum_y \exp(s(y)/\tau)} \exp(s/\tau)$, where f^*_τ denotes the normalized exponential operator for $\frac{1}{\tau}$-scaled logits. Below, we derive $D_{F^*_\tau}(r \parallel s)$ for such F^*_τ:

$$\begin{aligned}
D_{F^*_\tau}(s \parallel r) = &\ F^*_\tau(s) - F^*_\tau(r) - (s-r)^\mathsf{T} \nabla F^*_\tau(r) \\
= &\ \tau \mathrm{lse}(s/\tau) - \tau \mathrm{lse}(r/\tau) - (s-r)^\mathsf{T} f^*_\tau(r) \\
= &\ -\tau\big((s/\tau - \mathrm{lse}(s/\tau)) - (r/\tau - \mathrm{lse}(r/\tau)) \big)^\mathsf{T} f^*_\tau(r) \\
= &\ \tau f^*_\tau(r)^\mathsf{T} \big((r/\tau - \mathrm{lse}(r/\tau)) - (s/\tau - \mathrm{lse}(s/\tau)) \big) \\
= &\ \tau f^*_\tau(r)^\mathsf{T} \big(\log f^*_\tau(r) - \log f^*_\tau(s) \big) \\
= &\ \tau D_{\mathrm{KL}}\left(f^*_\tau(r) \parallel f^*_\tau(s) \right) \\
= &\ \tau D_{\mathrm{KL}}\left(q \parallel p \right)
\end{aligned} \tag{26}$$

Proposition 2. *The KL divergence between p and q in two directions can be expressed as,*

$$D_{\mathrm{KL}}(p \parallel q) = D_{\mathrm{KL}}(q \parallel p) + \tfrac{1}{4\tau^2} \mathrm{Var}_{y \sim f^*(b/\tau)}[s(y) - r(y)] - \tfrac{1}{4\tau^2} \mathrm{Var}_{y \sim f^*(a/\tau)}[s(y) - r(y)] \tag{27}$$

$$< D_{\mathrm{KL}}(q \parallel p) + \|s - r\|_2^2, \tag{28}$$

for some $a = (1-\alpha)r + \alpha s$, $(0 \le \alpha \le \frac{1}{2})$, $b = (1-\beta)s + \beta r$, $(0 \le \beta \le \frac{1}{2})$.

Proof. For a potential function $F^*_\tau(r) = \tau \mathrm{lse}(r/\tau)$, given (26), one can re-express Proposition 2 by multiplying both sides by τ as,

$$D_{F^*_\tau}(r \parallel s) = D_{F^*_\tau}(s \parallel r) + \tfrac{1}{4\tau} \mathrm{Var}_{y \sim f^*(b/\tau)}[s(y) - r(y)] - \tfrac{1}{4\tau} \mathrm{Var}_{y \sim f^*(a/\tau)}[s(y) - r(y)]\ . \tag{29}$$

Moreover, for the choice of $F^*_\tau(a) = \tau \mathrm{lse}(a/\tau)$, it is easy to verify that,

$$H_{F^*_\tau}(a) = \tfrac{1}{\tau}(\mathrm{Diag}(f^*_\tau(a)) - f^*_\tau(a) f^*_\tau(a)^\top)\ , \tag{30}$$

where $\mathrm{Diag}(\mathbf{v})$ returns a square matrix the main diagonal of which comprises a vector $\mathbf{v}$. Substituting this form for $H_{F^*_\tau}$ into the the quadratic terms in (15), one obtains,

$$(s-r)^\top H_{F^*_\tau}(a)(s-r) = \tfrac{1}{\tau}(s-r)^\top \big(\mathrm{Diag}(f^*_\tau(a)) - f^*_\tau(a) f^*_\tau(a)^\top\big)(s-r) \tag{31}$$

$$= \tfrac{1}{\tau}\big(E_{\mathbf{y} \sim f^*_\tau(a)}\left[(s(\mathbf{y}) - r(\mathbf{y}))^2\right] - E_{\mathbf{y} \sim f^*_\tau(a)}\left[s(\mathbf{y}) - r(\mathbf{y})\right]^2\big) \tag{32}$$

$$= \tfrac{1}{\tau} \mathrm{Var}_{\mathbf{y} \sim f^*_\tau(a)}\left[s(\mathbf{y}) - r(\mathbf{y})\right]\ . \tag{33}$$

Finally, we can substitude (33) and its variant for $H_{F^*_\tau}(b)$ into (15) to obtain (29) and equivalently (27).

Next, consider the inequality in (28). Let $\delta = s - r$ and note that

$$D_{F^*_\tau}(r \parallel s) - D_{F^*_\tau}(s \parallel r) = \tfrac{1}{4\tau}\delta^\top \big(H_{F^*_\tau}(b) - H_{F^*_\tau}(a)\big)\delta \tag{34}$$

$$= \tfrac{1}{4\tau}\delta^\top \mathrm{Diag}(f^*_\tau(b) - f^*_\tau(a))\delta + \tfrac{1}{4\tau}\big(\delta^\top f^*_\tau(a)\big)^2 - \tfrac{1}{4\tau}\big(\delta^\top f^*_\tau(b)\big)^2 \tag{35}$$

$$\le \tfrac{1}{4\tau}\|\delta\|_2^2 \|f^*_\tau(b) - f^*_\tau(a)\|_\infty + \tfrac{1}{4\tau}\|\delta\|_2^2 \|f^*_\tau(a)\|_2^2 \tag{36}$$

$$\le \tfrac{1}{2\tau}\|\delta\|_2^2 + \tfrac{1}{4\tau}\|\delta\|_2^2 \tag{37}$$

since $\|f^*_\tau(b) - f^*_\tau(a)\|_\infty \le 2$ and $\|f^*_\tau(a)\|_2^2 \le \|f^*_\tau(a)\|_1^2 \le 1$. □