A Appendix

A.1 RAML Background

The key idea behind RAML is to observe that the entropy-regularized policy gradient RL objective L_{RL} can be written as (up to constant and scaling):

$$L_{\mathrm{RL}} = \sum\nolimits_{(X,Y^\star)\in D} \mathrm{KL}(Q_\omega(Y|X) \parallel p^\star_\beta(Y|Y^\star)) \quad (18)$$

where $p^\star_\beta(Y|Y^\star)$ is the *exponentiated pay-off distribution* defined as:

$$p^\star_\beta(Y|Y^\star) = \exp\{r(Y,Y^\star)/\beta\}/Z(Y^\star,\beta) \quad (19)$$

$r(Y,Y^\star)$ is a reward function that measures some similarity of Y with respect to the ground truth $Y^\star$ (e.g. negative edit-distance). Whereas in RAML Norouzi et al. (2016), one optimizes the KL in the reverse direction:

$$L_{\mathrm{RAML}} = \sum\nolimits_{(X,Y^\star)\in D} \mathrm{KL}(p^\star_\beta(Y|Y^\star) \parallel Q_\omega(Y|X)) \quad (20)$$

It was shown that these two losses have the same global extremum and when away from it their gap is bounded under some conditions Norouzi et al. (2016). Compare the RAML gradient with the policy gradient:

$$\nabla L_{\mathrm{RAML}} = -E_{p^\star_\beta(Y|Y^\star)}\{\nabla \log Q_\omega(Y|X)\} \quad (21)$$

$$\nabla L_{\mathrm{RL}} = -E_{Q_\omega(Y|X)}\{r(Y,Y^\star)\nabla \log Q_\omega(Y|X)\} \quad (22)$$

RAML gradient samples from a stationary distribution, while policy gradient samples from the changing Q_ω distribution. Furthermore, samples from $p^\star_\beta(Y|Y^\star)$ has higher chance of landing in configurations of high reward by definition, while samples $Q_\omega(Y|X)$ relies on random exploration to discover sequences with high reward. For these reasons, RAML has much lower variance than RL.