# Maximum Entropy-Regularized Multi-Goal Reinforcement Learning (Appendix)

**Rui Zhao** [1 2] **Xudong Sun** [1] **Volker Tresp** [1 2]

## A. Proof of Theorem 1

**Theorem 1.** *The surrogate $\eta^{\mathcal{L}}(\boldsymbol{\theta})$ is a lower bound of the objective function $\eta^{\mathcal{H}}(\boldsymbol{\theta})$, i.e., $\eta^{\mathcal{L}}(\boldsymbol{\theta}) < \eta^{\mathcal{H}}(\boldsymbol{\theta})$, where*

$$\eta^{\mathcal{H}}(\boldsymbol{\theta}) = \mathcal{H}_p^w(\boldsymbol{\mathcal{T}}^g) = \mathbb{E}_p\left[\log \frac{1}{p(\boldsymbol{\tau}^g)} \sum_{t=1}^{T} r(S_t, G^e) \mid \boldsymbol{\theta}\right] \tag{1}$$

$$\eta^{\mathcal{L}}(\boldsymbol{\theta}) = Z \cdot \mathbb{E}_q\left[\sum_{t=1}^{T} r(S_t, G^e) \mid \boldsymbol{\theta}\right] \tag{2}$$

$$q(\boldsymbol{\tau}^g) = \frac{1}{Z} p(\boldsymbol{\tau}^g)\left(1 - p(\boldsymbol{\tau}^g)\right) \tag{3}$$

*$Z$ is the normalization factor for $q(\boldsymbol{\tau}^g)$. $\mathcal{H}_p^w(\boldsymbol{\mathcal{T}}^g)$ is the weighted entropy (Guiaşu, 1971; Kelbert et al., 2017), where the weight is the accumulated reward $\sum_{t=1}^{T} r(S_t, G^e)$ in our case.*

*Proof.*

$$\eta^{\mathcal{L}}(\boldsymbol{\theta}) = Z \cdot \mathbb{E}_q\left[\sum_{t=1}^{T} r(S_t, G^e) \mid \boldsymbol{\theta}\right] \tag{4}$$

$$= \sum_{\boldsymbol{\tau}^g} Z \cdot q(\boldsymbol{\tau}^g) \sum_{t=1}^{T} r(s_t, g^e) \tag{5}$$

$$= \sum_{\boldsymbol{\tau}^g} \frac{Z}{Z} p(\boldsymbol{\tau}^g)(1 - p(\boldsymbol{\tau}^g)) \sum_{t=1}^{T} r(s_t, g^e) \tag{6}$$

$$< \sum_{\boldsymbol{\tau}^g} -p(\boldsymbol{\tau}^g) \log p(\boldsymbol{\tau}^g) \sum_{t=1}^{T} r(s_t, g^e) \tag{7}$$

$$= \mathbb{E}_p\left[\log \frac{1}{p(\boldsymbol{\tau}^g)} \sum_{t=1}^{T} r(S_t, G^e) \mid \boldsymbol{\theta}\right] \tag{8}$$

$$= \mathcal{H}_p^w(\boldsymbol{\mathcal{T}}^g) \tag{9}$$

$$= \eta^{\mathcal{H}}(\boldsymbol{\theta}) \tag{10}$$

In the inequality, we use the property $\log x < x - 1$. □

[1]Faculty of Mathematics, Informatics and Statistics, Ludwig Maximilian University of Munich, Munich, Bavaria, Germany [2]Siemens AG, Munich, Bavaria, Germany. Correspondence to: Rui Zhao <zhaorui.in.germany@gmail.com>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

## B. Proof of Theorem 2

**Theorem 2.** *Let the probability density function of goals in the replay buffer be*

$$p(\boldsymbol{\tau}^g), \textit{where } p(\boldsymbol{\tau}_i^g) \in (0,1) \textit{ and } \sum_{i=1}^{N} p(\boldsymbol{\tau}_i^g) = 1. \tag{11}$$

*Let the proposal probability density function be defined as*

$$q(\boldsymbol{\tau}_i^g) = \frac{1}{Z} p(\boldsymbol{\tau}_i^g) \left(1 - p(\boldsymbol{\tau}_i^g)\right), \textit{ where } \sum_{i=1}^{N} q(\boldsymbol{\tau}_i^g) = 1. \tag{12}$$

*Then, the proposal goal distribution has an equal or higher entropy*

$$\mathcal{H}_q(\boldsymbol{\mathcal{T}}^g) - \mathcal{H}_p(\boldsymbol{\mathcal{T}}^g) \geq 0. \tag{13}$$

*Proof.* For clarity, we define the notations in this proof as $p_i = p(\boldsymbol{\tau}_i^g)$ and $q_i = q(\boldsymbol{\tau}_i^g)$.

Note that the definition of Entropy is

$$\mathcal{H}_p = \sum_i -p_i \log(p_i), \tag{14}$$

where the $i$th summand is $p_i \log(p_i)$, which is a concave function. Since the goal distribution has a finite support $I$, we have the real-valued vector $(p_1, \ldots, p_N)$ and $(\frac{1}{Z}q_1, \ldots, \frac{1}{Z}q_N)$.

We use Karamata's inequality (Kadelburg et al., 2005), which states that if the vector $(p_1, \ldots, p_N)$ majorizes $(\frac{1}{Z}q_1, \ldots, \frac{1}{Z}q_N)$ then the summation of the concave transformation of the first vector is smaller than the concave transformation of the second vector.

In our case, the concave transformation is the weighted information at the $i$th position -$p_i \log(p_i)$, where the weight is the probability $p_i$ (entropy is the expectation of information). Therefore, the proof of the theorem is also a proof of the majorizing property of $p$ over $q$ (Petrov).

We denote the proposal goal distribution as

$$q_i = f(p_i) = \frac{1}{Z} p_i (1 - p_i). \tag{15}$$

Note that in our case, the partition function $Z$ is a constant.

Majorizing has three requirements (Marshall et al., 1979).

The first requirement is that both vectors must sum up to one. This requirement is already met because

$$\sum_i p_i = \sum_i q_i = 1. \tag{16}$$

The second requirement is that monotonicity exits. Without loss of generality, we assume the probabilities are sorted:

$$p_1 \geq p_2 \geq \ldots \geq p_N \tag{17}$$

Thus, if $i > j$ then

$$f(p_i) - f(p_j) = \frac{1}{Z} p_i (1 - p_i) - \frac{1}{Z} p_j (1 - p_j) \tag{18}$$

$$= \frac{1}{Z} [(p_i - p_j) - (p_i + p_j)(p_i - p_j)] \tag{19}$$

$$= \frac{1}{Z} (p_i - p_j)(1 - p_i - p_j) \tag{20}$$

$$\geq 0. \tag{21}$$

which means that if the original goal probabilities are sorted, the transformed goal probabilities are also sorted,

$$f(p_1) \geq f(p_2) \geq \ldots \geq f(p_N). \tag{22}$$

The third requirement is that for an arbitrary cutoff index $k$, there is

$$p_1 + \ldots p_k < q_1 + \ldots + q_k. \tag{23}$$

To prove this, we have

$$\begin{aligned} p_1 + \ldots + p_k &= \frac{p_1 + \ldots + p_k}{1} && (24)\\ &= \frac{p_1 + \ldots + p_k}{p_1 + \ldots + p_N} && (25)\\ &\geq f(p_1) + ... + f(p_k) && (26)\\ &= \frac{1}{Z}[p_1(1-p_1) + ... + p_k(1-p_k)] && (27)\\ &= \frac{1}{Z}[p_1 + \ldots + p_k - (p_1^2 + \ldots + p_k^2)] && (28) \end{aligned}$$

Note that, we multiply $Z * 1$ to each side of

$$Z = p_1(1-p_1) + \ldots + p_N(1-p_N). \tag{29}$$

Then we have

$$(p_1 + \ldots + p_k)Z * 1 \geq p_1 + \ldots + p_k - (p_1^2 + \ldots + p_k^2) * 1. \tag{30}$$

Now, we substitute the expression of $Z$ and then have

$$(p_1 + \ldots + p_k)[p_1(1-p_1) + \ldots + p_N(1-p_N)] \geq [p_1 + \ldots + p_k - (p_1^2 + \ldots + p_k^2)] * 1. \tag{31}$$

We express 1 as a series of terms $\sum_i p_i$, we have

$$(p_1 + \ldots + p_k)[p_1(1-p_1) + \ldots + p_N(1-p_N)] \geq [p_1 + \ldots + p_k - (p_1^2 + \ldots + p_k^2)] * [(p_1 + \ldots + p_k) + (p_{k+1} + \ldots p_N)]. \tag{32}$$

We use the distributive law to the right side and have

$$\begin{aligned} &(p_1 + \ldots + p_k)[p_1(1-p_1) + \ldots + p_N(1-p_N)]\\ \geq &[p_1 + \ldots + p_k] * [(p_1 + \ldots + p_k) + (p_{k+1} + \ldots p_N)] - [(p_1^2 + \ldots + p_k^2)] * [(p_1 + \ldots + p_k) + (p_{k+1} + \ldots p_N)]. \end{aligned} \tag{33}$$

We move the first term on the right side to the left and use the distributive law then have

$$(p_1 + \ldots + p_k)[-1 * (p_1^2 + \ldots + p_N^2))] \geq -[(p_1^2 + \ldots + p_k^2)] * [(p_1 + \ldots + p_k) + (p_{k+1} + \ldots p_N)]. \tag{34}$$

We use the distributive law again on the right side and move the first term to the left and use the distributive law then have

$$(p_1 + \ldots + p_k)[-1 * (p_{k+1}^2 + \ldots + p_N^2))] \geq -[(p_1^2 + \ldots + p_k^2)] * [(p_{k+1} + \ldots p_N)]. \tag{35}$$

We remove the minus sign then have

$$(p_1 + \ldots + p_k)[(p_{k+1}^2 + \ldots + p_N^2))] \leq [(p_1^2 + \ldots + p_k^2)] * [(p_{k+1} + \ldots p_N)]. \tag{36}$$

To prove the inequality above, it suffices to show that the inequality holds true for each associated term of the multiplication on each side of the inequality.

Suppose that

$$i \leq k < j \tag{37}$$

then we have

$$p_i > p_j. \tag{38}$$

As mentioned above, the probabilities are sorted in descending order. We have

$$p_i p_j^2 - p_i^2 p_j = p_i p_j (p_j - p_i) < 0 \tag{39}$$

then

$$p_i p_j^2 < p_i^2 p_j. \tag{40}$$

Therefore, we have proved that the inequality holds true for an arbitrary associated term, which also applies when they are added up. □

## C. Insights

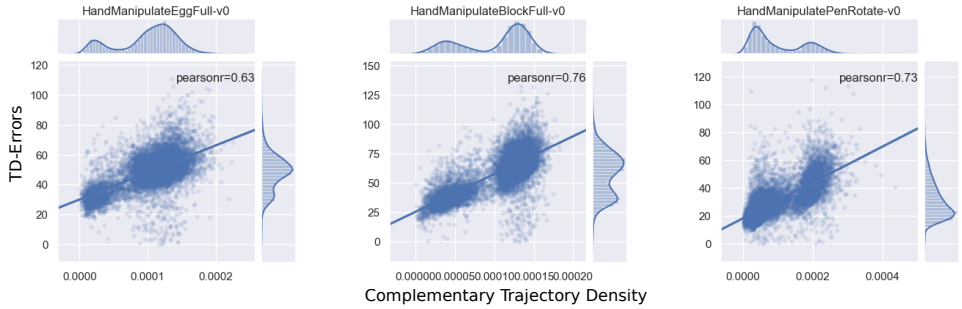


*Figure 1.* Pearson correlation between the complementary density $\bar{p}(\boldsymbol{\tau}^g)$ and TD-errors in the middle of training

To further understand why maximum entropy in goal space facilitates learning, we look into the TD-errors during training. We investigate the correlation between the complementary predictive density $\bar{p}(\boldsymbol{\tau}^g \mid \boldsymbol{\phi})$ and the TD-errors of the trajectory. The Pearson correlation coefficients, i.e., Pearson's r (Benesty et al., 2009), between the density $\bar{p}(\boldsymbol{\tau}^g \mid \boldsymbol{\phi})$ and the TD-errors of the trajectory are 0.63, 0.76, and 0.73, for the hand manipulation of egg, block, and pen tasks, respectively. The plot of the Pearson correlation is shown in Figure 1. The value of Pearson's r is between 1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. We can see that the complementary predictive density is correlated with the TD-errors of the trajectory with an average Pearson's r of 0.7. This proves that the agent learns faster from a more diverse goal distribution. Under-represented goals often have higher TD-errors, and thus are relatively more valuable to learn from. Therefore, it is helpful to maximize the goal entropy and prioritize the under-represented goals during training.

## References

Benesty, J., Chen, J., Huang, Y., and Cohen, I. Pearson correlation coefficient. In *Noise reduction in speech processing*, pp. 1–4. Springer, 2009.

Guiaşu, S. Weighted entropy. *Reports on Mathematical Physics*, 2(3):165–179, 1971.

Kadelburg, Z., Dukic, D., Lukic, M., and Matic, I. Inequalities of karamata, schur and muirhead, and some applications. *The Teaching of Mathematics*, 8(1):31–45, 2005.

Kelbert, M., Stuhl, I., and Suhov, Y. Weighted entropy: basic inequalities. *Modern Stochastics: Theory and Applications*, 4 (3):233–252, 2017. doi: 10.15559/17-VMSTA85. URL www.i-journals.org/vmsta.

Marshall, A. W., Olkin, I., and Arnold, B. C. *Inequalities: theory of majorization and its applications*, volume 143. Springer, 1979.

Petrov, F. Shannon entropy of $p(x)(1 - p(x))$ is no less than entropy of $p(x)$. MathOverflow. URL https://mathoverflow.net/q/320726. URL:https://mathoverflow.net/q/320726 (version: 2019-01-12).