Stop Regressing: Training Value Functions via Classification for Scalable Deep RL

Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taïga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, Aviral Kumar, Rishabh Agarwal

Donghu Kim

tl;dr

- For value learning, we should probably choose cross-entropy loss over MSE loss.
  - Classification is often superior to regression (e.g., for large models)
- There's a better cross-entropy loss than C51: Histogram Loss.
  - Instead of modeling the distribution of returns, model the distribution of target value.

Chapter I

What is HL-Gauss?

Value Learning: MSE vs Cross Entropy

- MSE (Regression)

$$\mathrm{TD}_{\mathrm{MSE}}(\theta) = \mathbb{E}_{\mathcal{D}}\left[\left((\widehat{\mathcal{T}}Q)(S_t, A_t; \theta^-) - Q(S_t, A_t; \theta)\right)^2\right]$$

Target value: $(\widehat{\mathcal{T}}Q)(s, a; \theta^-) = R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a'; \theta^-) \mid S_t = s,\ A_t = a$

- Cross Entropy (Categorical distribution)

$$\mathrm{TD}_{\mathrm{CE}}(\theta) = \mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^{m} p_i(S_t, A_t; \theta^-) \log \hat{p}_i(S_t, A_t; \theta)\right]$$

Target distribution: $\sum_{i=1}^{m} p_i(S_t, A_t; \theta^-)\, z_i \approx (\widehat{\mathcal{T}}Q)(S_t, A_t; \theta^-)$

$$Q(s, a; \theta) = \mathbb{E}\left[\, Z(s, a; \theta)\,\right], \qquad Z(s, a; \theta) = \sum_{i=1}^{m} \hat{p}_i(s, a; \theta) \cdot \delta_{z_i}, \qquad \hat{p}_i(s, a; \theta) = \frac{\exp\left(l_i(s, a; \theta)\right)}{\sum_{j=1}^{m} \exp\left(l_j(s, a; \theta)\right)}$$

Previous Value Learning Methods

- Q-Learning (MSE)

$$\mathrm{TD}_{\mathrm{MSE}}(\theta) = \mathbb{E}_{\mathcal{D}}\left[\left((\widehat{\mathcal{T}}Q)(S_t, A_t; \theta^-) - Q(S_t, A_t; \theta)\right)^2\right]$$

Target: $(\widehat{\mathcal{T}}Q)(s, a; \theta^-) = R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a'; \theta^-) \mid S_t = s,\ A_t = a$

- Conservative Q-Learning (MSE)

$$\min_{\theta}\ \alpha\left(\mathbb{E}_{\mathcal{D}}\left[\log\left(\sum_{a'} \exp(Q(S_{t+1}, a'; \theta))\right)\right] - \mathbb{E}_{\mathcal{D}}\left[Q(S_t, A_t; \theta)\right]\right) + \mathrm{TD}_{\mathrm{MSE}}(\theta)$$

Previous Value Learning Methods

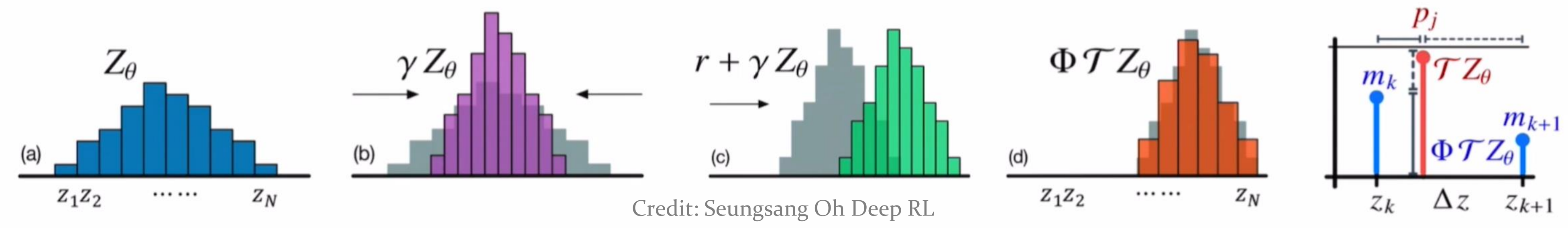
- C51 (Cross Entropy)
  - Model the distribution of returns, use that to compute the target distribution.

$$(\widehat{\mathcal{T}}Z)(s,a;\theta^-) \overset{D}{=} \sum_{i=1}^{m} \hat{p}_i(S_{t+1},A_{t+1};\theta^-)\cdot\delta_{R_{t+1}+\gamma z_i} \,\Big|\, S_t=s,\ A_t=a$$ (Bellman equation)

$$p_i(S_t,A_t;\theta^-) = \sum_{j=1}^{m} \hat{p}_j(S_{t+1},A_{t+1};\theta^-)\cdot\xi_j(R_{t+1}+\gamma z_i)$$

$$\xi_j(x) = \frac{x-z_j}{z_{j+1}-z_j}\mathbb{1}\{\lfloor x\rfloor = z_j\} + \frac{z_{j+1}-x}{z_{j+1}-z_j}\mathbb{1}\{\lceil x\rceil = z_j\}$$

(Matching support by force)

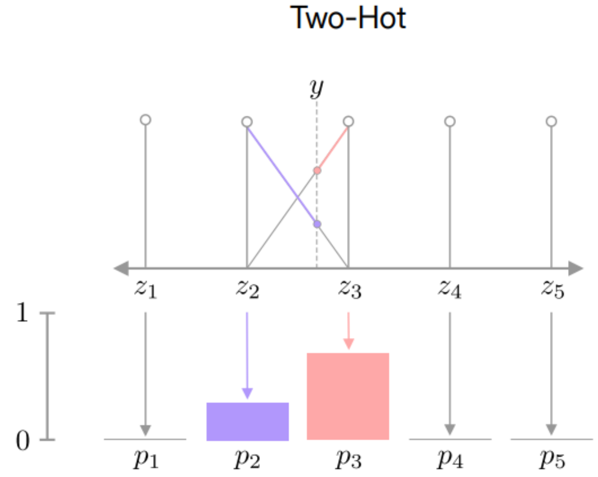


Credit: Seungsang Oh Deep RL

Why Not Just Directly Model the Target Distribution?

- Return to single-output (C51 has 51 outputs), but use it to model a categorical dist.
- 3 methods considered

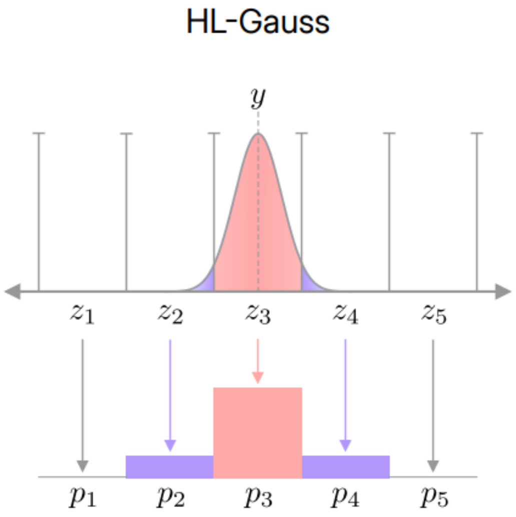
1. One hot distribution ('binning')
   - Bad: Loss of information.
2. Two hot distribution
   - Interpolation between two nearest z's.
   - Good: No loss of information.
   - Bad: Does not fully harness the ordinal structure of discrete regression.
     (Cannot tell the 'distance' between two adjacent classes)



Why Not Just Directly Model the Target Distribution?

3. Histrograms as categorical distribution (Histogram Loss Family) (Imani and White)

- Model any distribution based on the prediction (e.g., Gaussian)
- Slice the graph into multiple bins, measure each area to make a histogram.
  (Easily computed using CDF)
- Good: Better exploits the ordinal structure
  (Can actually tell that classes are equidistant)
- Good: Analogous to label smoothing.
  (Can control the degree by controlling the std. of Gaussian)



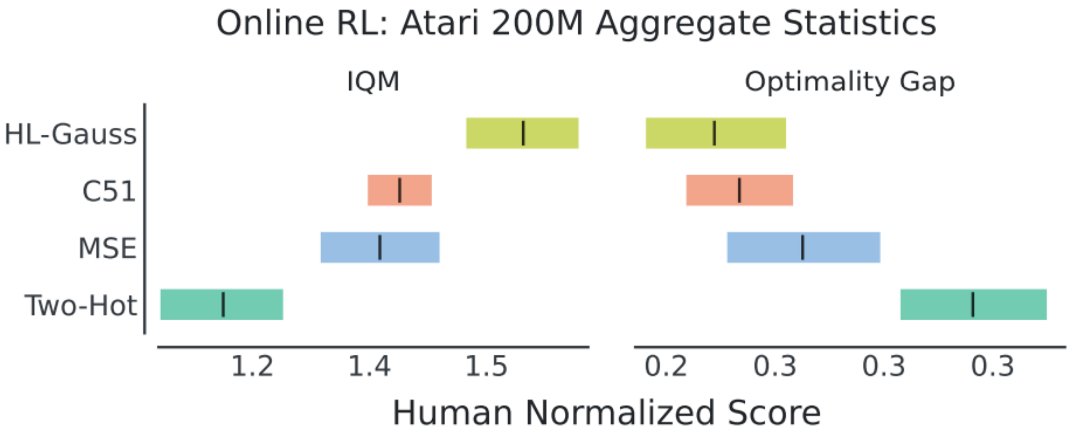
[1] Ehsan Imani and Martha White. Improving regression performance with distributional losses. ICML 2018.

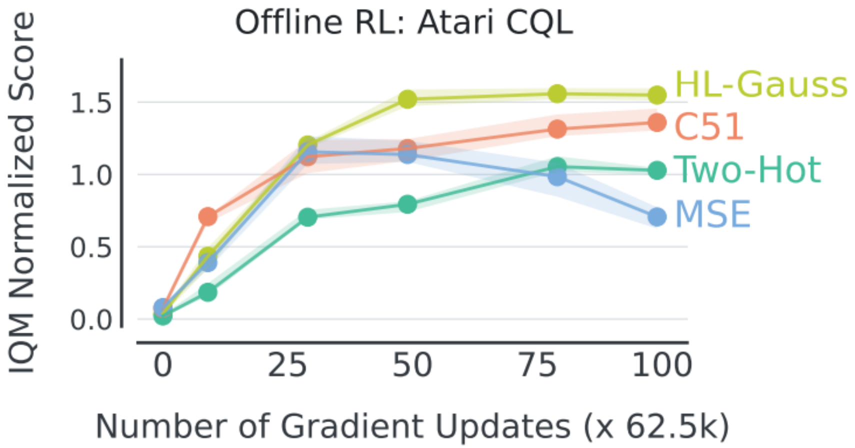
Chapter II

How Good is HL-Gauss?

Experiment1: Atari, Single-game

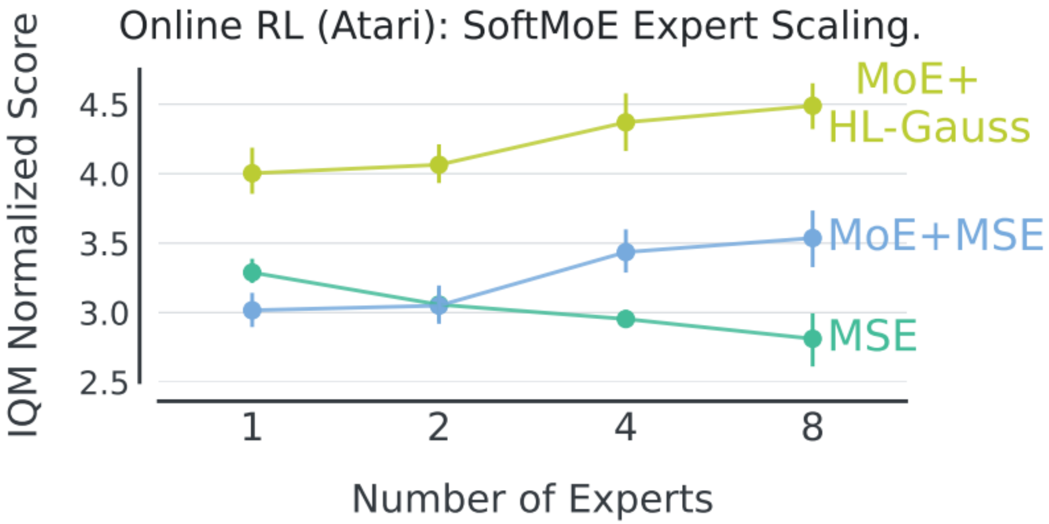
- Online – DQN 200M, 60 games
- Offline – CQL 6.25M, 17 games
  - MSE degrades; Cross-Entropy methods retain performance.





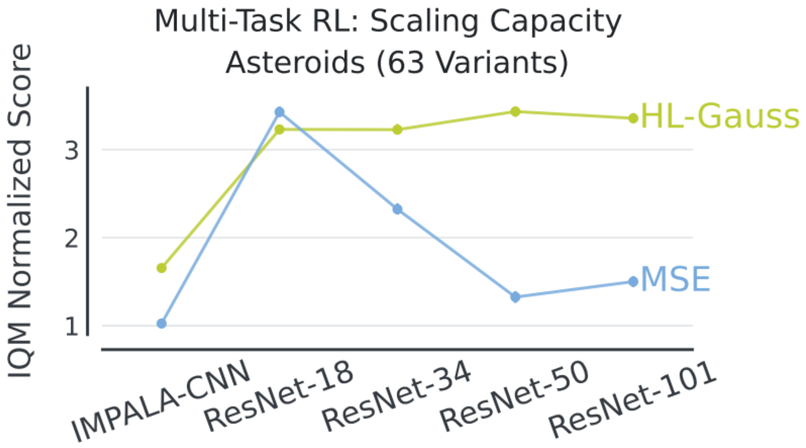
Experiment2: Atari, Scaling with SoftMoE

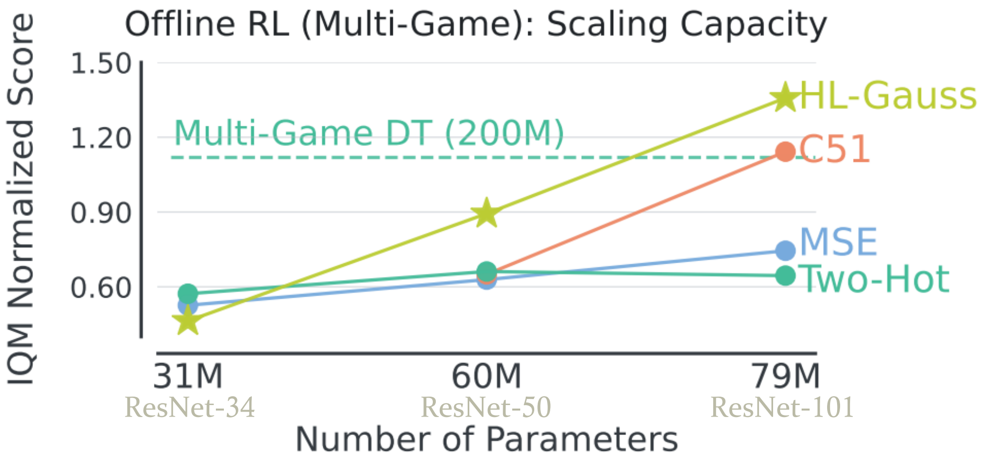
- Online – Impala, 20 games
- HL-Gauss is complementary to MoE



Experiment3: Atari, Scaling Generalist Policies

- Online – Impala, 63 gamemodes of Asteroids
- Offline – ScaledQL(ResNet), 40 games

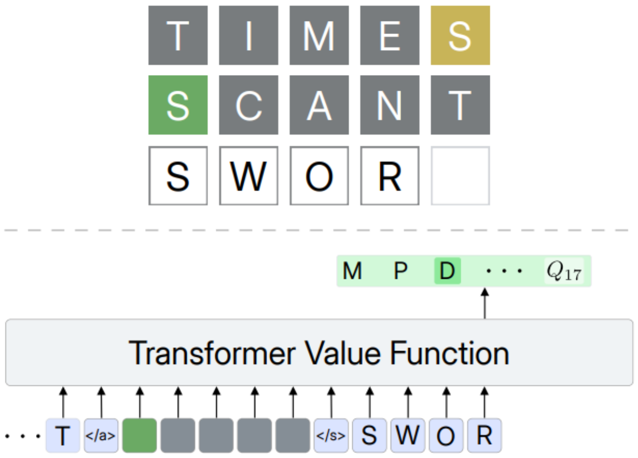


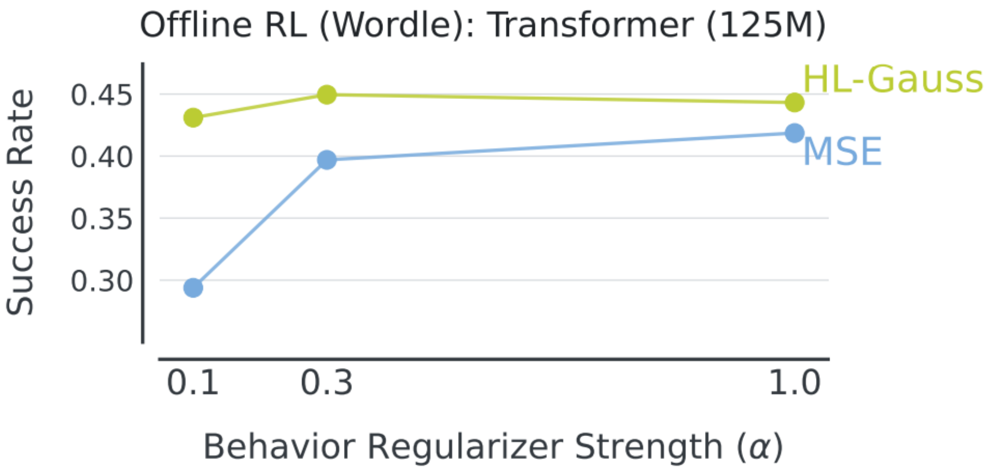


[1] Kumar et al. Offline Q-Learning on Diverse Multi-Task Data Both Scales and Generalizes. ICML 2023.

Experiment4: Wordle

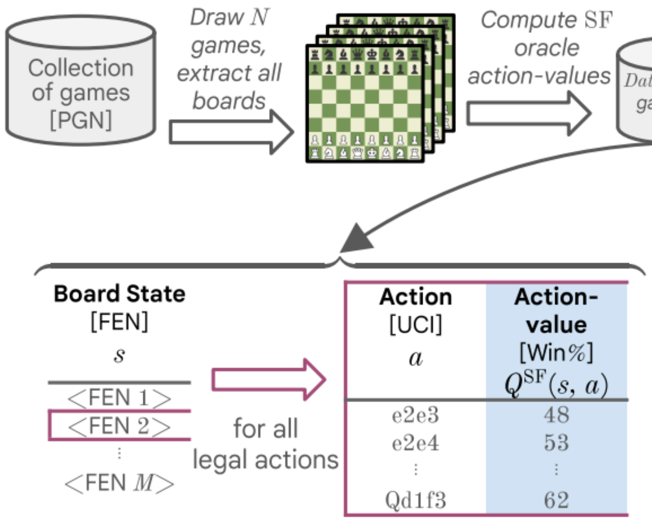
- Offline – Wordle dataset, 125M GPT-like transformer, DQN+CQL
- Cross-entropy (HL-Gauss) is more suitable for training transformers as well.

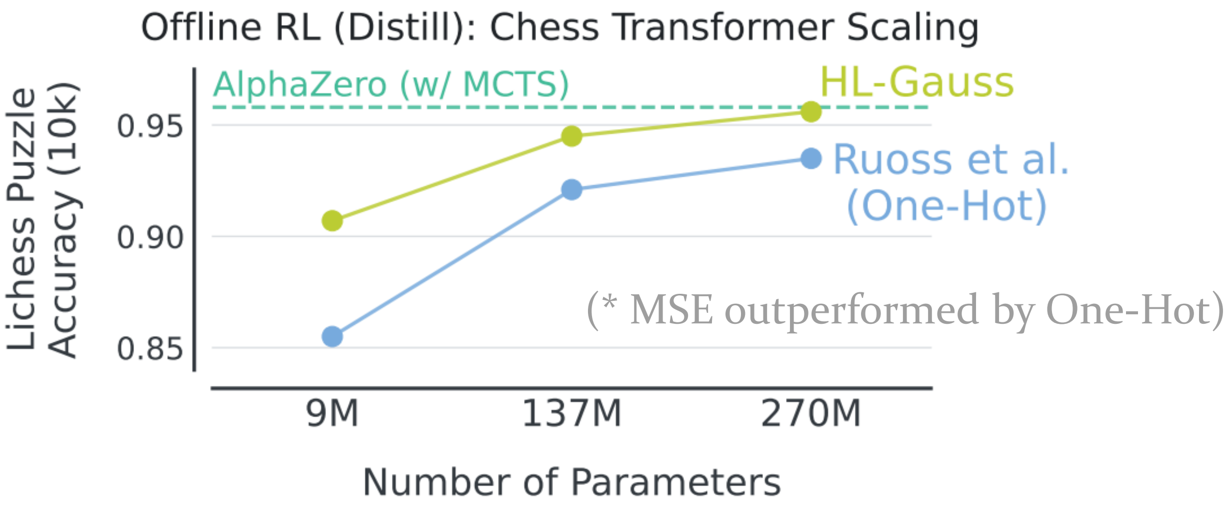




Experiment5: Chess without MCTS

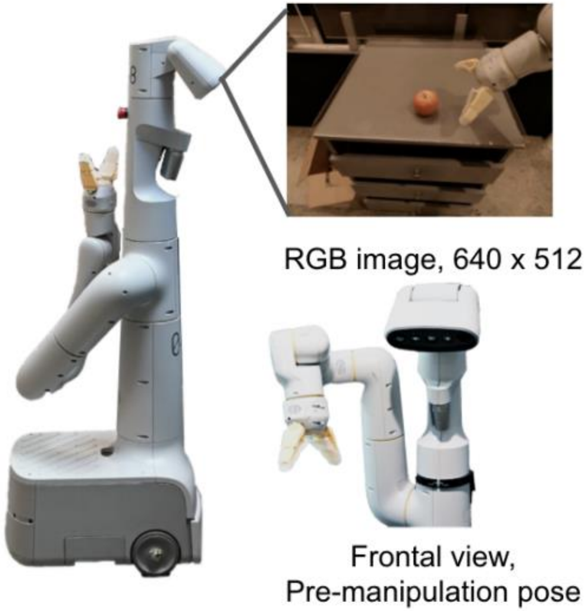
- Offline – Chess dataset, Transformer, action-value distillation from Stockfish16.
- Competitive to AlphaZero without any searching.

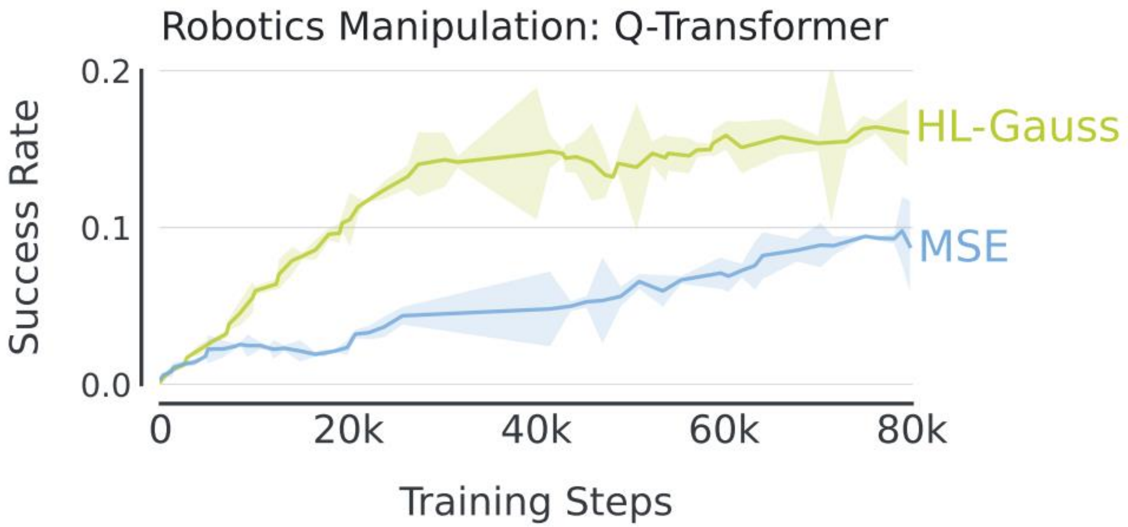




Experiment6: Manipulation Tasks

- Collected dataset, 60M Q-Transformer.



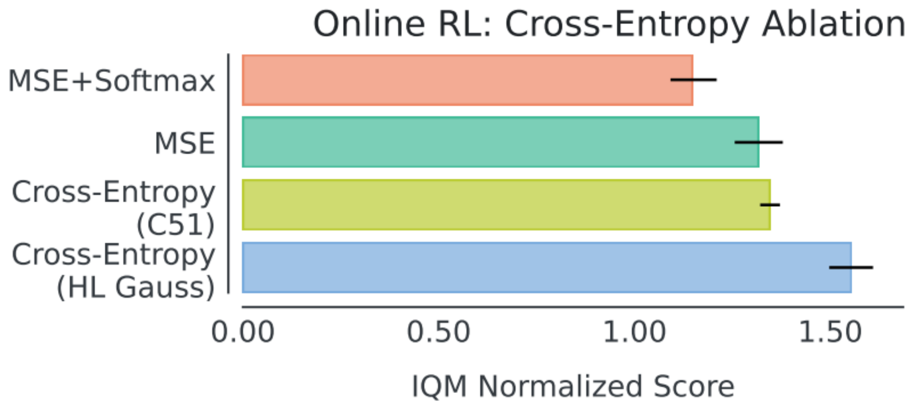


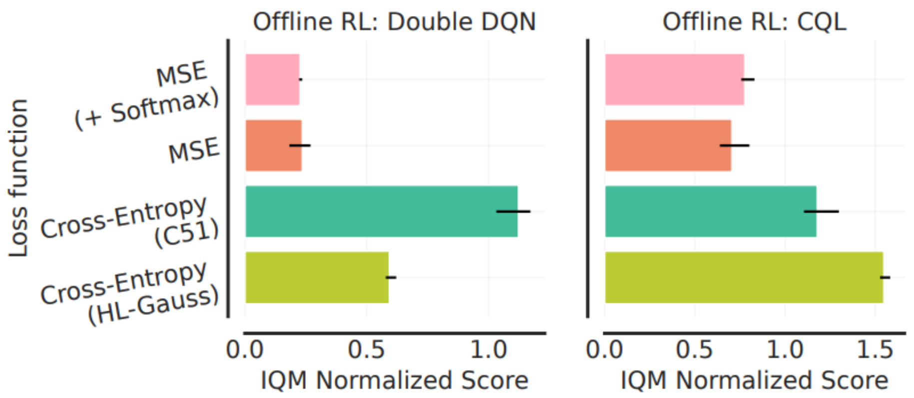
Chapter III

Why is HL-Gauss So Good?

What Component Matters in HL-Gauss?

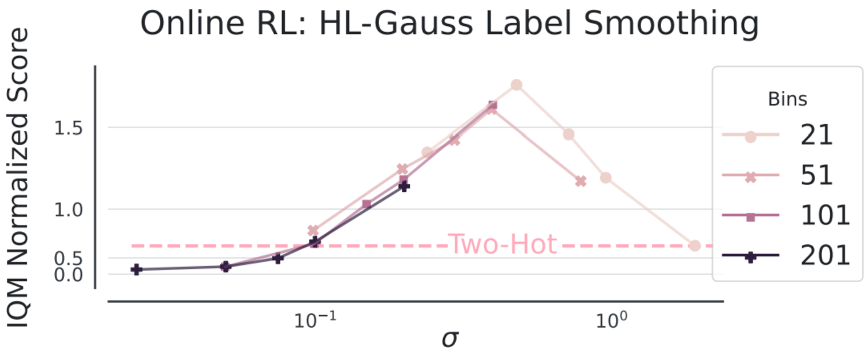
- SoftMoE + MSE worked well. Could it be the softmax operation?
  - MSE + Softmax doesn't work (both online & offline)…





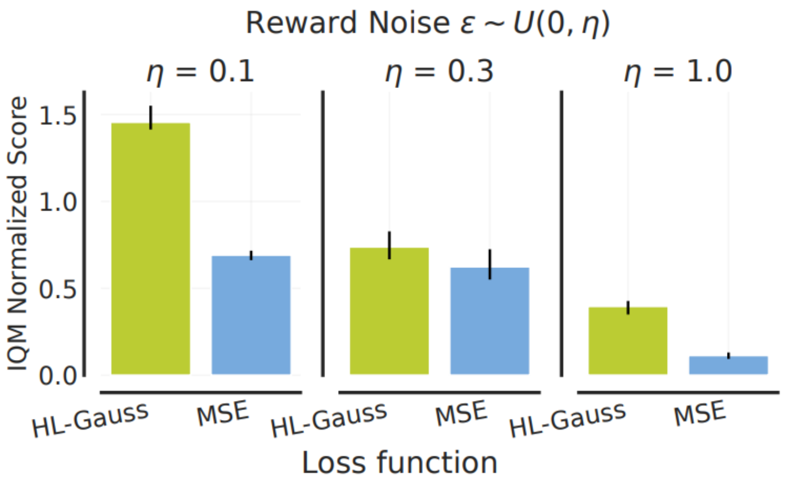
What Component Matters in HL-Gauss?

- Preventing overfitting is important. Could it the 'label smoothing' effect?
  - Ablation on <u># of bins</u> and <u>std of Gaussian</u>
  - Wide range of σ outperformed two-hot → Preventing overfitting does help, but…
  - Best performing σ was independent to # of bins
    → Degree of label smoothing did NOT matter (Note: same σ + larger # bins = stronger smoothing)
    → Preventing overfitting cannot be the only reason!
  - Exploitation of the ordinal structure is just as important.

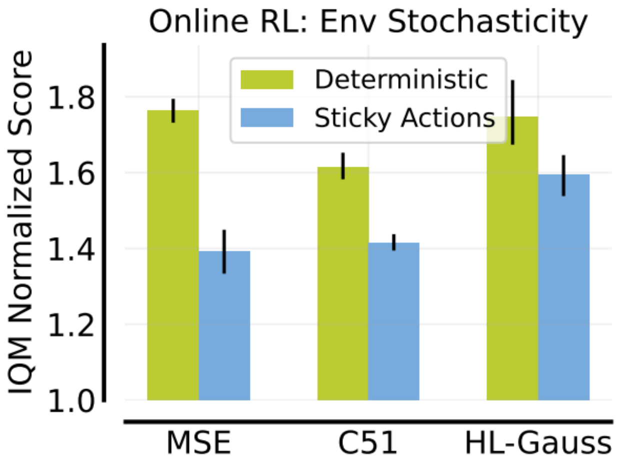


Benefits of Classification

- Overfits less to noisy labels and stochastic dynamics
  - Can be mitigated by 'label smoothing' and distributional modeling.
  - HL-Gauss is more robust(?) to artificial reward noise.
  - MSE and HL-Gauss perform similarly in deterministic dynamics.

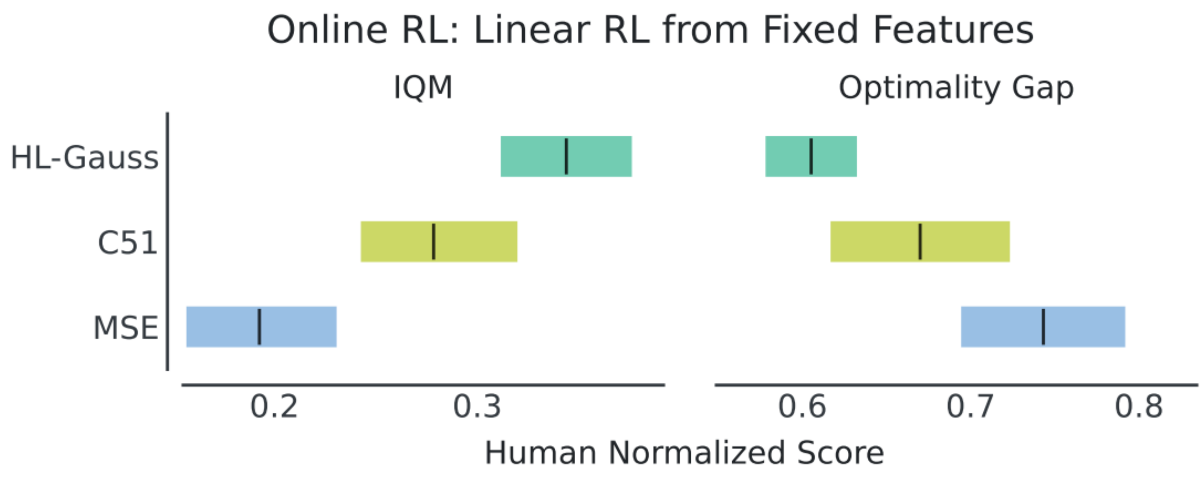


(* Atari)



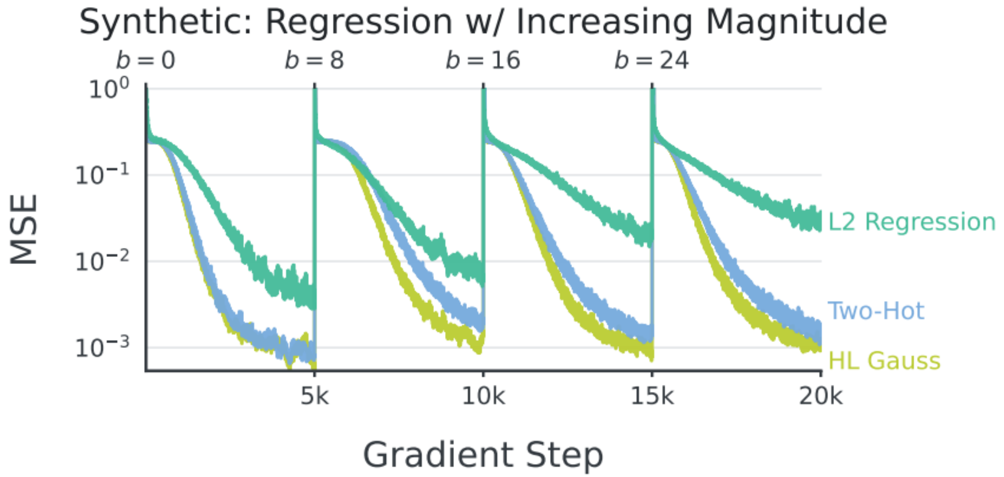
Benefits of Classification

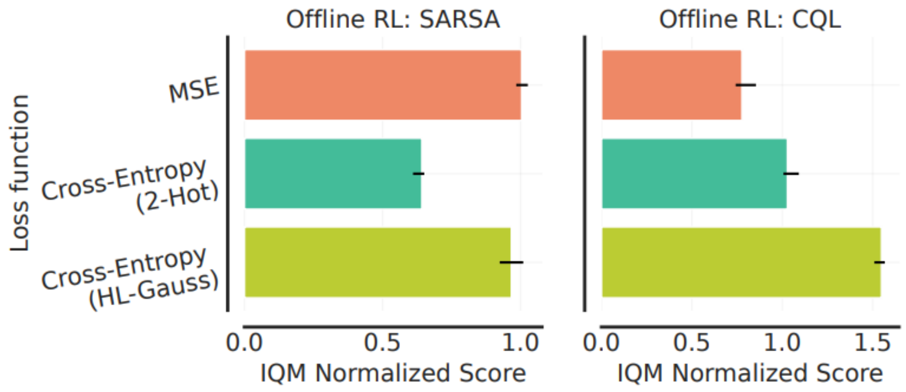
- Better representations
  - Less overfitting = Retain the representational power to model other value functions.
  - Higher linear proving RL performance.



Benefits of Classification

- More robust to non-stationary targets (better plasticity)
  - Hypothesize by C51 authors, but wasn't empirically shown since.
  - Synthetic setup: Regression target changes every 5k steps.
  - Offline RL setup: SARSA(stationary) vs CQL(non-stationary)





Summary

- The success of HL-Gauss can be attributed to:
  1. Preventing overfitting by spreading probability mass to neighbors (‘label smoothing’)
  2. Exploits the ordinal structure of regression task (unlike two-hot)
- The benefits of using classification instead of regression are:
  1. Robustness against noisy labels and stochastic dynamics
  2. Better representations
  3. Robustness against non-stationary targets

~ Fin ~