Unified Maximizing Q-Learning and Behavior Cloning for Offline RL

Anonymous Authors[1]

Abstract

Offline RL aims to improve target policy based on pre-collected datasets. A major problem of offline RL is the distribution shift, The behavioral cloning algorithms try to constrain the target policy to be close to the offline data. Though this constraint can reduce extrapolation error of out-of-distribution actions, but it will make the learned policy to be conservative. The maximizing Q-Learning algorithms try to learn a perfect Q-value function according to Bellman eqution, then optimize the policy to generate better action to maximize the Q-value function. In this work, we unify maximizing Q-learning and behavioral cloning by implicit and explicit way to leverage their advantages. For implicit way, we propose to constrain the generated actions to be close to offline data by GAN; For explicit way, we first map the states to actions in offline data, then we propose multiple importance sampling to learning a weight for different state-action pairs.

It's more problematic under high-dimension continuous action space, where the offline data only cover a small part of action space. In this work, we firstly propose multiple importance sampling to better fit the offline data and stabilize the learning process based on DICE. After obtaining the importance weight from aforementioned step, we propose to use latent variables to extract the target policy. Our latent variables based method replaces traditional behavioral cloning methods that map states to actions directly by neural networks. We conducted extensive experiments on the D4RL dataset and use tSNE to visualize the actions generated by out latent variables method. Experiment results show out method improves the cumulative returns and exploration ability simultaneously compared to standard policy extraction algorithms, especially on the high-dimension action space, such as Adroit [1] of D4RL dataset, which makes a well trade-off between exploitation and exploration.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Offline reinforcement learning has wide applications where online interactions with real environment is costly or dangerous, such as autonomous driving and medical experiments. We can only train our model based on prior data. But this poses a major problem that the learned agent tends to be a copy of the behavior of prior data. Existing algorithms all amount to this principle. The regularization-type algorithms measure the discrepancy between target policy and behavioral policy, trying to make them close to each other. The value regularization algorithms, such as pessimistic, try to avoid distribution shift by regularizing the action value function, assigning low values to unobserved actions, and high values to observed actions in offline data. This kind of algorithms will face severe problems in high dimension domains, where observed actions occupy a small part of the action space. The regularization algorithms make the agent very conservative, performing similar actions to offline data in certain states. In this work, we propose to use latent variables based on VAE and GAN to map states to actions. Different from traditional methods that map states to actions explicitly by neural networks. The VAE-type latent variable method maps state to a latent variable and then map the latent variable to state, which is similar to encoder-decoder architecture. At the same time, there will be a constraint, such as f-divergence, to make the latent variable follow standard normal distribution, which aims to make the encoder mapping correctly. The GAN-type latent variable method replaces the constraint by a generator and a discriminator. The discriminator tries to discriminate the latent variable and samples from standard normal distribution, while the generator tries to generate more plausible latent variables to fool the discriminator. By introducing this intermediary latent variable, experimental results show it improves the returns greatly, especially in high dimensional occasions. What's more, we demonstrate by tSNE that it improves the diversity of the generated actions, which is important for offline RL. The intermediary latent variables make the agent less dependent on the trajectories seen on offline data.

On the other hand, existing algorithms assume the offline data is homogeneous. But in practical, the data may be collected under various sceneries and follows diverse distributions. Some data may be a mixture of human demonstration or hand-designed controllers, the data has different level of optimality. What's more, it's also possible that the data is non-Markovian, which is impossible to represent the data using Markovian policy. Such as human demonstrations that depends on external knowledge. Although such data is collected, but when training, our agent doesn't know the external human knowledge, making it hard to fit behavioral policy. When we use importance sampling algorithms, it's hard to estimate the distribution of behavioral data, resulting biased estimation of importance weight.

reminiscent of KL divergence

For suboptimal or random trajectories, if we fit the offline data by behavioral cloning without using importance sampling, the learned target policy would also be suboptimal or random.

Considering the suboptimality and diversity of offline data, we propose multiple importance sampling to model the offline data to improve over the behavioral policy.

2. Methodology

We unify the maximizing Q-Learning with behavioral cloning in implicit way and explicit way. Without behavioral cloning, the policy will be wildly extrapolated on unseen actions. In addition, the Q-value function is penalized by uncertainty, for state-action pairs with high uncertainty, their Q values are smaller than those with low uncertainty, thus prompting a pessimistic policy against OOD actions. The behavioral cloning is weighted by advantage of the action, for actions with low advantage, their contributions to the objective will be down-weighted.

The performance of maximizing Q-Learning depends on how accurate the Q value estimator is. But we only have limited state-action observations, which incurs extrapolation errors from OOD data, especially for continuous action space. This makes the Q value estimator fall easily into suboptimal area.

Behavioral cloning also has its shortcoming. When the offline data form behavioral policy is biased, the target policy we learned is also biased. This means behavioral cloning methods depend heavily on the quality of training data.

2.1. Implicit Unification

For the implicit unification, the maximizing Q-Learning is unified with behavioral cloning by adverisally training. In particular, the actions generated by target policy are forced to be close to actions supported by offline data by generative adversarial networks(Goodfellow et al., 2014). The behavioral cloning item favors generating actions supported by the offline data. At first, we need to learn the Q-value function, which is an ensemble of multiple Q networks penalized by uncertainty.

$$Q_{\phi_i}(s,a) = r + \gamma(\min_{1\le j\le N} Q_{\phi_j'}(s',a') - \log \pi_\theta(a'|s')) \quad (1)$$

where N is ensemble size, ϕ_j and $\phi_j^{'}$ are parameters of j-th Q function and target Q function, $\log \pi_\theta(a'|s')$ is the entropy regularization from SAC(Haarnoja et al., 2018). The final Q-value would be the minimum of ensemble Q-values, which encourages pessimistic and mitigates over-estimation for out-of-distribution (OOD) actions. This pessimistic principle can also be seen as uncertainty penalization(An et al., 2021) where the Q-value of unstable state-action pairs with large variance would be penalized to have smaller values.

$$E[\min_{1\le j\le N} Q_{\phi_j'}(s,a)] \approx \mu(s,a) - \Phi^{-1}(\frac{N-\pi/8}{N-\pi/4+1})\sigma(s,a) \quad (2)$$

For the policy, an extra loss is introduced to enforce the generated actions to stay close to behavioral policy, this method is called behavioral cloning. The motivation of behavioral cloning is trying to mitigate over-estimation of OOD actions. The policy keeps conservative when generating new actions. The discriminator D loss is

$$\mathcal{L}_D = \mathop{E}_{a\sim p_{data}}[D(a)] - \mathop{E}_{\hat{a}\sim\pi_\theta(s)}[D(\hat{a})] \quad (3)$$

Policy objective for implicit way is

$$\begin{aligned}\mathcal{L}(\pi) =& \lambda\left[\log \pi_\theta(\hat{a}|s) - \min_{1\le j\le N} Q_{\phi_j}(s,\hat{a})\right] \\ &+ (1-\lambda) E_{\hat{a}\sim\pi_\theta(s)}[D(\hat{a})]\end{aligned} \quad (4)$$

where $\hat{a} = \pi_\theta(s)$, λ is used to control the degree of conservatism.

Policy Optimization The discriminator D tries its best to discriminate the actions from offline data and actions generated by policy. Whiel the policy can be seen as the generator of GAN. The policy optimizater only optimizes parameters of generator and tries to generate actions to fool the discriminator, i.e., generating actions that can't be distinguished by discriminator. Different from mapping states to actions directly, this kind of adversarial training implicitly enforce generated actions to keep close to offline data, so we call it **Im**plicit **B**ehavioral **C**loning (**ImBC**). At the same time, assuming we have learned a perfect Q-value estimator, maximizing the lower bound $\min_{1\le j\le N} Q_{\phi_j}(s,\hat{a})$ will force the policy to generate optimal action $\hat{a}$ under state s.The entropy regularization Therefore, the unification leverages the merits of maximizing Q-Learning and behavioral cloning.

Algorithm 1 Unified Maximizing Q-Learning with Implicit BC

repeat
  Sample mini-batch data $(s, a, r, s\prime)$ from $\mathcal{D}$
  for i **to** n time steps **do**
    Train Q with (5)
    Train discriminator D with (3)
    Train policy with (4)
  end for
until converge

Q Optimization In the policy loss, the policy parameters π_θ are optimized to maximize the Q value $Q_\phi(s, \pi_\theta(s))$ without modifying the parameters ϕ of Q networks. In order to guide the update of policy parameters θ, the Q function should be accurate to estimate the Q value for (s, a) pair. In theory, $Q(s, a)$ function is perfect when it can predict the cumulative reward of real environment when performing action a under state s. The parameters of Q function are optimized by the observed offline data. To prevent the Q function wildly extrapolated on unseen state-action pairs, we use the pessimistic mechanism to penalize the Q value of state-action pair with high uncertainty, as is shown in (1). State-action pair with high uncertainty will have lower Q value, as is shown in (2). Pessimistic mechanism is especially important for continuous space, where the observed data occupies a small percentage compared with unseen data. Pessimistic mechanism has been adopted by prior works(Kumar et al., 2020; An et al., 2021).

In formula (4), considering the gradient of $\frac{\partial Q}{\partial \theta} = \frac{\partial Q}{\partial \hat{a}} \frac{\partial \hat{a}}{\partial \theta}$, to prevent the $\frac{\partial Q}{\partial \hat{a}}$ to dominate gradient, we add a regularization item $\left|\frac{\partial Q}{\partial \hat{a}}\right|^2$ in the loss to force it small, thus making $\frac{\partial \hat{a}}{\partial \theta}$ to dominate the gradient, so as to better optimize the policy parameters θ. The training objective for Q networks would be

$$\left(r + \gamma(\min_{1\le j\le N} Q_{\phi_j'}(s', a') - \log \pi_\theta(a'|s')) - Q_{\phi_i}(s, a)\right)^2 + \left|\frac{\partial Q}{\partial \hat{a}}\right|^2 \tag{5}$$

2.2. Explicit Way

For explicit way, we combine maximizing Q-Learning with a weighted neural network that maps states to actions in offline data directly. Different state-action pair will have different weight $w(s, a)$. The weight $w(s, a)$ is learned via multiple importance sampling. We want to maximize the cumulative reward on target policy, simultaneously constrain target policy to be close to behavioral policy. For the weighted behavioral cloning, we use following framework,

$$\max_{d^\pi} E_{(s,a)\sim d^\pi}\left[R(s, a)\right] - \alpha D_f(d^\pi || d^D) \tag{6}$$

$$s.t. \sum_a d^\pi(s, a) = (1-\gamma)\mu_0 + \gamma \mathcal{T}_* d(s),\ \forall s \in S \tag{7}$$

where $D_f(d^\pi||d^D)$ is f-divergence between d^π and d^D, f is convex; $\mathcal{T}_* d(s) = \sum_{\bar{s},\bar{a}} T(s|\bar{s}, \bar{a}) d(\bar{s}, \bar{a})$; The constraint equation (7) makes sure that $d^\pi(s, a)$ is the occupancy distribution of the target policy. The frist item in (6) maximize reward on target policy $\pi(a|s)$, while the second item in (6) constrain target policy to be close to behavioral policy. The α controls the strength of the constraint. We reformulate (6) and (7) by a Lagrangian multiplier $V(s)$:

$$\max_{d^\pi} \min_{V(s)} E_{s,a\sim d^\pi}[R(s, a)] - \alpha D_f(d^\pi||d^D) + \sum_s V(s)\left((1-\gamma)\mu_0 + \gamma T_* d(s) - \sum_a d^\pi(s, a)\right) \tag{8}$$

Note that $\mathcal{T}_*$ is adjoint of $\mathcal{T}$, we have,

$$\sum_s V(s) \cdot \mathcal{T}_* d(s) = \sum_{s,a} d(s, a) \mathcal{T} V(s, a)$$

So

$$\begin{aligned}
&E_{s,a\sim d^\pi}[R(s, a)] - \alpha D_f(d^\pi||d^D) \\
&+ \sum_s V(s)((1-\gamma)\mu_0 + \gamma \mathcal{T}_* d(s) - \sum_a d^\pi(s, a)) \\
=&E_{s,a\sim d^\pi}[R(s, a)] - \alpha E_{s,a\ d^D}[f(\frac{d^\pi(s, a)}{d^D(s, a)})] \\
&+ E_{s,a\ d^\pi}[\gamma \mathcal{T} V(s, a) - V(s)] + (1-\gamma) E_{s\sim\mu_0}[V(s)] \\
=&E_{s,a\sim d^D}[\frac{d^\pi(s, a)}{d^D(s, a)} A(s, a) - \alpha f(\frac{d^\pi(s, a)}{d^D(s, a)})] \\
&+ (1-\gamma) E_{s\sim\mu_0}[V(s)]
\end{aligned} \tag{9}$$

where $A(s, a) = R(s, a) + \gamma \mathcal{T} V(s, a) - V(s)$ is the advantage.

Since in real world, the offline data is complicated, we assume the data is heterogeneous. Therefore, we propose multiple importance sampling for the optimization of target policy. The behavioral policy is heterogeneous, so the corresponding target policy should also be heterogeneous. Assuming there are K distributions $(d^{\pi_1}, d^{\pi_2} ... d^{\pi_K})$ for target policy, each d^{π_k} has coefficient β_k its own Q-value and V-value function to fit. Then (9) would become

$$
\begin{aligned}
&E_{s,a\sim d^D}\left[\sum_{k=1}^{K}\left(\frac{\beta_k d^{\pi_k}}{d^D}A_k(s,a)-\alpha\beta_k f(\frac{d^{\pi_k}}{d^D})\right)\right] \\
&\qquad +(1-\gamma)\sum_{k=1}^{K}\beta_k E_{s\sim\mu_0}[V_k(s)] \\
&=E_{s,a\sim d^D}\left[\sum_{k=1}^{K}\left(\beta_k\omega_k A_k(s,a)-\alpha\beta_k f(\omega_k)\right)\right] \\
&\qquad +(1-\gamma)\sum_{k=1}^{K}\beta_k E_{s\sim\mu_0}[V_k(s)]
\end{aligned}
\tag{10}
$$

where $A_k(s,a)=R(s,a)+\gamma\mathcal{T}V_k(s,a)-V_k(s)$, and $\omega_k=\frac{d^{\pi_k}(s,a)}{d^D(s,a)}$. When we maximize (10), it is easy to get a closed-form for ω_k,

$$
\omega_k=(f')^{-1}\left(\frac{A_k(s,a)}{\alpha}\right)=f'_*\left(\frac{A_k(s,a)}{\alpha}\right) \tag{11}
$$

Finally, the overall weight is the mean of weights $w_1, w_2, ...w_K$. We will use this weight to differentially optimize different state-action pairs from the offline data. The (10) would become

$$
E_{(s,a)\sim d^D}\left[\sum_{k=1}^{K}\beta_k f_*\left(A_k(s,a)/\alpha\right)\right]+(1-\gamma)\sum_{k=1}^{K}\beta_k E_{s\ \mu_0}\left[V_k(s)\right] \tag{12}
$$

The motivation for multiple importance sampling is that if the original data (behavior policy d^D) is heterogeneous and follows multiple distributions, then the target policy d^π should also follow multiple distributions. What's more, multiple target policies allow the agent to perform multiple optimal actions under similar state s, which is the usual case in practice.

Policy Extraction The Policy Extraction module is weighted by the weight we learned from (11).

$$
\max_\theta \underset{(s,a)\sim d^\pi}{E}\log\pi_\theta(a|s)=\max_\theta \underset{(s,a)\sim d^D}{E}[w(s,a)\log\pi_\theta(a|s)] \tag{13}
$$

where $w(s,a)=\frac{1}{K}\sum_{k=1}^{K}w_k$. For each state-action pair, there would be a specific weight $w(s,a)$, this weight has nothing to do with policy parameter θ, $w(s,a)$ controls the gradient to the policy loss. In (11), for state-action pair that has large advantage, it will get a larger weight, thus contributing more to the gradient; For state-action pair that has small or negative advantage, it will contribute less to the gradient.

The policy objective for explicit way is

$$
\begin{aligned}
\mathcal{L}(\pi)=&\lambda\left[\log\pi_\theta\left(\hat{a}|s\right)-\min_{1\le j\le N}Q_{\phi_j}(s,\hat{a})\right] \\
&+(1-\lambda)w(s,a)\log\pi_\theta(a|s)
\end{aligned}
\tag{14}
$$

Algorithm 2 Unified Maximizing Q-Learning with Explicit BC

repeat
  Sample mini-batch data $(s,a,r,s\prime)$ from $\mathcal{D}$
  for i **to** n time steps **do**
    Train Q with (5)
    Train weight D with (12)
    Train policy with (14)
  end for
until converge

where λ is the same as (4). Different from the implicit behavioral cloning in (4), here we map states to actions directly, and assign different weights to different state-action pairs, so as to better optimize the objective for different state-action pairs.

3. Experiments

We evaluate our proposed approaches against existing approaches on the D4RL benchmark(Fu et al., 2020) with various continuous environments. For the MUJOCO data, since we evaluate on the v2 version, some methods that are early developed, like CQL, BCQ, may not be comparable with this approach. Implementation details are in the appendix. Similar to previous literatures, we use the normalized average rewards as the evaluating metric. The average rewards are obtained from 10 runs in each evaluation.

3.1. Performance on MuJoCo dataset

The MuJoCo data consists three environments (halfcheetah, hopper, walker2d), each environment consists 5 types (expert, medium-expert, medium-replay, medium, random). In particular, the *expert* data means the The *medium* data contains 1M samples generated from early-stoping SAC(Haarnoja et al., 2018) policy, where the samples are collected before the policy reaches optimal. The *expert* data contains samples from fully trained online SAC policy. The *medium-expert* means mixing equal amount samples of medium and expert policy. The *medium-replay* contains samples from the replay buffer when the policy reaches medium level. The *full-replay* contains samples from replay buffer of expert policy. The *random* contains samples of random policy.

Baselines We compare out approach with following baselines: EDAC(An et al., 2021), which belongs to the maximizing Q learning method, it adopts ensemble-diversified actor-critic to minimize the pairwise alignment (cosine distance) of the gradients for every pair Q-ensemble with regard to actions, thus reducing the ensemble size of Q function. IQL(Kostrikov et al., 2022), which belongs to pure behavioral cloning methods; IQL uses a state value function $V(s')$

to replace $Q(s', a')$ to learn the Q function, thus avoiding querying OOD action a', then it extracts target policy via advantage-weighted behavioral cloning. RORL(Yang et al., 2022), an updated EDAC with noise, making the policy more robust.

Experiment results of our approach and prior baselines are in Table 1. It can be seen our approach **ImBC** outperforms pure maximizing Q learning method EDAC and pure behavioral cloning method IQL across random to expert data. The performance gap is larger in the expert data, which can be interpreted in that the quality of expert data is higher so behavioral cloning may play an important role for these data compared with pure maximizing Q learning methods. The performance of ImBC and ExBC is different for different domain. For halfcheetah and walker2d, ImBC performs better than ExBC; For hopper data, ExBC performs better. The interpretation might be that hopper data suffers from misestimation due to distributional shift more than halfcheetah and walker2d. Since ExBC has stronger constraint on the generated actions of policy, thus suffering less from distributional shift of actions.

We also investigate the effect of behavioral cloning item in ImBC and ExBC. Figure **??** shows the results of different λ in (4) and (14). It can be seen that the performance varies with regard to different λ. With larger λ, the agent have stronger constraint to force the policy to generate actions close to offline data.

Our ImBC can be seen as a combination of these two strategies.

As for the overall perforance, we surpass prior methods

Experiment results are shown in Table 1.

3.2. Performance on Adroit dataset

3.3. Ablation Experiments

The unified maximizing Q-learning and behavioral cloning makes the training converge faster and has better performance compared with only behavioral cloning or maximizing Q-learning. We show the behavioral cloning item is necessary for the performance by running ImBC and ExBC with various λ. When $\lambda = 0$, the algorithm degrades into pure maximizing Q value. We also eliminate maximizing Q learning item in (4) and (14), then the algorithm degrades into pure behavioral cloning. The hyperparameter λ can be seen as the degree of conservatism. Larger λ means more conservative.

3.4. Visualization of Implicit/Explicit Behavioral Cloning

Paper Deadline: The deadline for paper submission that is advertised on the conference website is strict. If your full, anonymized, submission does not reach us on time, it will not be considered for publication.

Anonymous Submission: ICML uses double-blind review: no identifying author information may appear on the title page or in the paper itself. Section 3.8 gives further details.

Simultaneous Submission: ICML will not accept any paper which, at the time of submission, is under review for another conference or has already been published. This policy also applies to papers that overlap substantially in technical content with conference papers under review or previously published. ICML submissions must not be submitted to other conferences and journals during ICML's review period. Informal publications, such as technical reports or papers in workshop proceedings which do not appear in print, do not fall under these restrictions.

Authors must provide their manuscripts in **PDF** format. Furthermore, please make sure that files contain only embedded Type-1 fonts (e.g., using the program `pdffonts` in linux or using File/DocumentProperties/Fonts in Acrobat). Other fonts (like Type-3) might come from graphics files imported into the document.

Authors using **Word** must convert their document to PDF. Most of the latest versions of Word have the facility to do this automatically. Submissions will not be accepted in Word format or any format other than PDF. Really. We're not joking. Don't send Word.

Those who use **LaTeX** should avoid including Type-3 fonts. Those using `latex` and `dvips` may need the following two commands:

```
dvips -Ppdf -tletter -G0 -o paper.ps paper.dvi
ps2pdf paper.ps
```

It is a zero following the "-G", which tells dvips to use the config.pdf file. Newer TeX distributions don't always need this option.

Using `pdflatex` rather than `latex`, often gives better results. This program avoids the Type-3 font problem, and supports more advanced features in the `microtype` package.

Graphics files should be a reasonable size, and included from an appropriate format. Use vector formats (.eps/.pdf) for plots, lossless bitmap formats (.png) for raster graphics with sharp lines, and jpeg for photo-like images.

The style file uses the `hyperref` package to make clickable links in documents. If this causes problems for you,

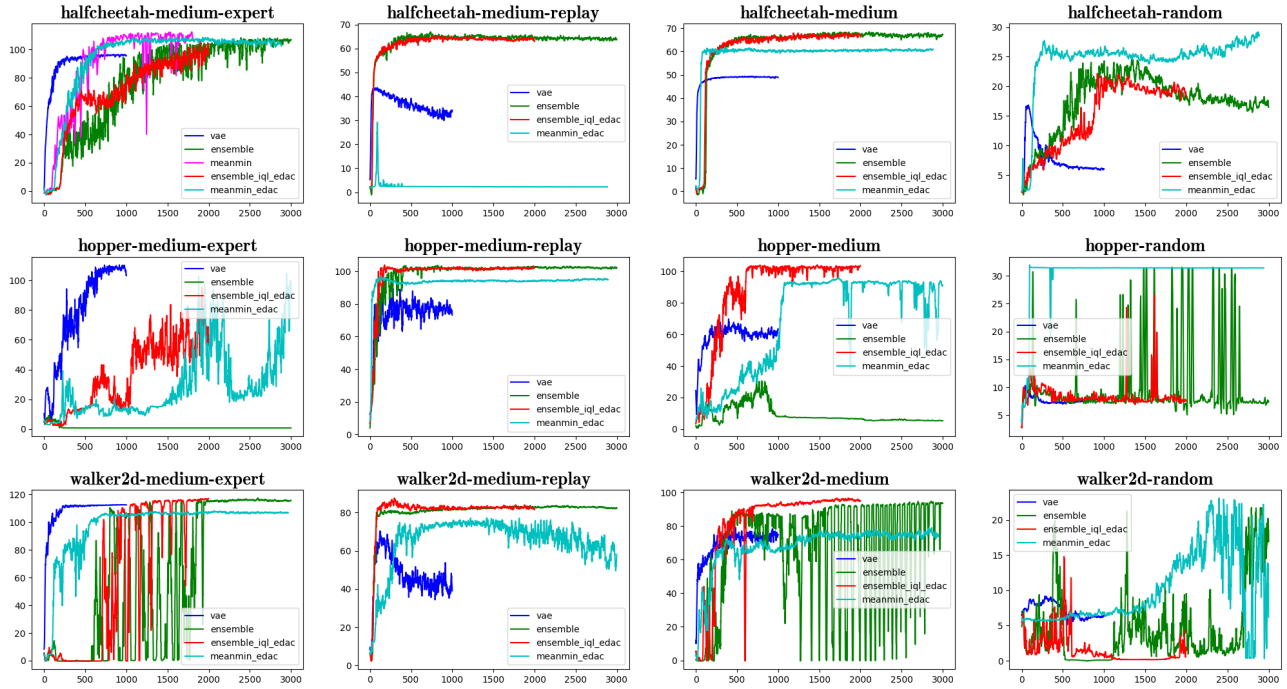

Figure 1. The performance and convergence speed of unified maximizing Q-learning and behavioral cloning

add `nohyperref` as one of the options to the `icml2022` usepackage statement.

3.5. Submitting Final Camera-Ready Copy

The final versions of papers accepted for publication should follow the same format and naming convention as initial submissions, except that author information (names and affiliations) should be given. See Section 3.8.2 for formatting instructions.

The footnote, "Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute." must be modified to "*Proceedings of the 39^{th} International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s)."

For those using the LaTeX style file, this change (and others) is handled automatically by simply changing `\usepackage{icml2022}` to

```
\usepackage[accepted]{icml2022}
```

Authors using **Word** must edit the footnote on the first page of the document themselves.

Camera-ready copies should have the title of the paper as running head on each page except the first one. The running title consists of a single line centered above a horizontal rule which is 1 point thick. The running head should be centered, bold and in 9 point type. The rule should be 10 points above the main text. For those using the LaTeX style file, the original title is automatically set as running head using the `fancyhdr` package which is included in the ICML 2022 style file package. In case that the original title exceeds the size restrictions, a shorter form can be supplied by using

```
\icmltitlerunning{...}
```

just before `\begin{document}`. Authors using **Word** must edit the header of the document themselves.

All submissions must follow the specified format.

3.6. Dimensions

The text of the paper should be formatted in two columns, with an overall width of 6.75 inches, height of 9.0 inches, and 0.25 inches between the columns. The left margin should be 0.75 inches and the top margin 1.0 inch (2.54 cm). The right and bottom margins will depend on whether you print on US letter or A4 paper, but all final versions must be produced for US letter size. Do not write anything on the margins.

The paper body should be set in 10 point type with a vertical spacing of 11 points. Please use Times typeface throughout the text.

Table 1. Comparison of our method to prior methods.

Task Name	CQL	ATAC	IQL	EDAC	RORL	ImBC	ExBC
halfcheetah-medium-expert	95.9± 0.2	95.5± 0	86.7± 5.3	106.3± 1.9	107.8± 1.1	**113.0± 0.4**	96.7± 0.2
halfcheetah-medium-replay	83.3± 0.6	49.5± 0	44.2± 1.2	61.3± 1.9	61.9± 1.5	**64.0± 0.2**	96.7± 0.2
halfcheetah-medium	61.9± 1.4	54.3± 0	47.4± 0.2	65.9± 0.6	66.8± 0.7	**72.0± 0.3**	96.7± 0.2
halfcheetah-random	74.8± 0.5	4.8± 0	-	28.4± 1.0	28.5± 0.8	**30.0± 0.2**	96.7± 0.2
hopper-medium-expert	73.3± 0.9	112.6± 0	91.5± 14.3	110.7± 0.1	112.7± 0.2	90.7± 0.2	**112.1± 0.3**
hopper-medium-replay	67.1± 0.6	102.8± 0	94.7± 8.6	101.0± 0.5	102.8± 0.5	104.5± 0.1	103.6± 0.2
hopper-medium	67.1± 0.6	102.8± 0	66.2± 5.7	101.6± 0.6	104.8± 0.1	58.0± 0.2	65.7± 0.2
hopper-random	67.1± 0.6	31.8± 0	-	25.3± 10.4	31.4± 0.1	32.1± 1.6	10.8± 0.2
walker2d-medium-expert	73.3± 0.9	116.3± 0	109.6± 1.0	114.7± 0.9	121.2± 1.5	118.0± 0.2	112.2± 0.8
walker2d-medium-replay	67.1± 0.6	94.1± 0	73.8± 7.1	87.1± 2.3	90.4± 0.5	88.0± 0.2	**95.1± 0.4**
walker2d-medium	67.1± 0.6	91.0± 0	78.3± 8.7	92.5± 0.8	102.4± 1.4	100.0± 0.2	96.7± 0.2
walker2d-random	67.1± 0.6	8.0± 0	-	16.6± 7.0	21.4± 0.2	**24.0± 0.2**	6.8± 0.2

3.7. Title

The paper title should be set in 14 point bold type and centered between two horizontal rules that are 1 point thick, with 1.0 inch between the top rule and the top edge of the page. Capitalize the first letter of content words and put the rest of the title in lower case.

3.8. Author Information for Submission

ICML uses double-blind review, so author information must not appear. If you are using LaTeX and the `icml2022.sty` file, use `\icmlauthor{...}` to specify authors and `\icmlaffiliation{...}` to specify affiliations. (Read the TeX code used to produce this document for an example usage.) The author information will not be printed unless `accepted` is passed as an argument to the style file. Submissions that include the author information will not be reviewed.

3.8.1. Self-Citations

If you are citing published papers for which you are an author, refer to yourself in the third person. In particular, do not use phrases that reveal your identity (e.g., "in previous work (Langley, 2000), we have shown ...").

Do not anonymize citations in the reference section. The only exception are manuscripts that are not yet published (e.g., under submission). If you choose to refer to such unpublished manuscripts (Author, 2021), anonymized copies have to be submitted as Supplementary Material via CMT. However, keep in mind that an ICML paper should be self contained and should contain sufficient detail for the reviewers to evaluate the work. In particular, reviewers are not required to look at the Supplementary Material when writing their review (they are not required to look at more than the first 8 pages of the submitted document).

3.8.2. Camera-Ready Author Information

If a paper is accepted, a final camera-ready copy must be prepared. For camera-ready papers, author information should start 0.3 inches below the bottom rule surrounding the title. The authors' names should appear in 10 point bold type, in a row, separated by white space, and centered. Author names should not be broken across lines. Unbolded superscripted numbers, starting 1, should be used to refer to affiliations.

Affiliations should be numbered in the order of appearance. A single footnote block of text should be used to list all the affiliations. (Academic affiliations should list Department, University, City, State/Region, Country. Similarly for industrial affiliations.)

Each distinct affiliations should be listed once. If an author has multiple affiliations, multiple superscripts should be placed after the name, separated by thin spaces. If the authors would like to highlight equal contribution by multiple first authors, those authors should have an asterisk placed after their name in superscript, and the term "*Equal contribution" should be placed in the footnote block ahead of the list of affiliations. A list of corresponding authors and their emails (in the format Full Name <email@domain.com>) can follow the list of affiliations. Ideally only one or two names should be listed.

A sample file with author names is included in the ICML2022 style file package. Turn on the `[accepted]` option to the stylefile to see the names rendered. All of the guidelines above are implemented by the LaTeX style file.

3.9. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of

11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase.

3.10. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

3.10.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

3.10.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

3.11. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction,

[1]Footnotes should be complete sentences.

[2]Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

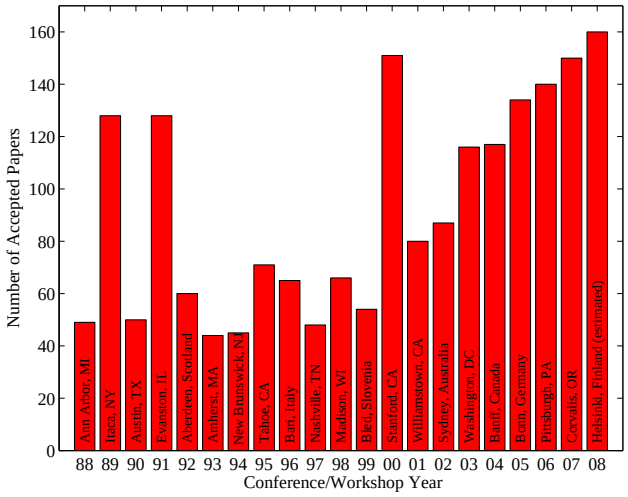

Figure 2. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure 2. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX). Always place two-column figures at the top or bottom of the page.

3.12. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 3 shows an example.

3.13. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the

Algorithm 3 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

table with at least 0.1 inches of space before the title and the same after it, as in **??**. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

3.14. Theorems and such

The preferred way is to number definitions, propositions, lemmas, etc. consecutively, within sections, as shown below.

Definition 3.1. A function $f : X \to Y$ is injective if for any $x, y \in X$ different, $f(x) \neq f(y)$.

Using Definition 3.1 we immediate get the following result:

Proposition 3.2. *If f is injective mapping a set X to another set Y, the cardinality of Y is at least as large as that of X*

Proof. Left as an exercise to the reader. □

Lemma 3.3 stated next will prove to be useful.

Lemma 3.3. *For any $f : X \to Y$ and $g : Y \to Z$ injective functions, $f \circ g$ is injective.*

Theorem 3.4. *If $f : X \to Y$ is bijective, the cardinality of X and Y are the same.*

An easy corollary of Theorem 3.4 is the following:

Corollary 3.5. *If $f : X \to Y$ is bijective, the cardinality of X is at least as large as that of Y.*

Assumption 3.6. The set X is finite.

Remark 3.7. According to some, it is only the finite case (cf. Assumption 3.6) that is interesting.

3.15. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2022.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (1959). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (Samuel, 1959). List multiple references separated by semicolons (Kearns, 1989; Samuel, 1959; Mitchell, 1980). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (Michalski et al., 1983).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to Section 3.8 for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (Samuel, 1959), conference publications (Langley, 2000), book chapters (Newell & Rosenbloom, 1981), books (Duda et al., 2000), edited volumes (Michalski et al., 1983), technical reports (Mitchell, 1980), and dissertations (Kearns, 1989).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

Please put some effort into making references complete, presentable, and consistent, e.g. use the actual current name of authors. If using bibtex, please protect capital letters of names and abbreviations in titles, for example, use {B}ayesian or {L}ipschitz in your .bib file.

Accessibility

Authors are kindly asked to make their submissions as accessible as possible for everyone including people with disabilities and sensory or neurological differences. Tips of how to achieve this and what to pay attention to will be provided on the conference website `http://icml.cc/`.

Software and Data

If a paper is accepted, we strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, **do not** include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the CMT reviewing system. Note that reviewers are not required to look at this material when writing their review.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

An, G., Moon, S., Kim, J. H., and Song, H. O. Uncertainty-based offline reinforcement learning with diversified q-ensemble. In *Proceedings of the 34th Advances in neural information processing systems*, pp. 7436–7447, 2021.

Author, N. N. Suppressed for anonymity, 2021.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4rl: Datasets for deep data-driven reinforcement learning. In *arXiv preprint arXiv:2004.07219*, 2020.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870, 2018.

Kearns, M. J. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (eds.). *Machine Learning: An Artificial Intelligence Approach, Vol. I.* Tioga, Palo Alto, CA, 1983.

Mitchell, T. M. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

Newell, A. and Rosenbloom, P. S. Mechanisms of skill acquisition and the law of practice. In Anderson, J. R. (ed.), *Cognitive Skills and Their Acquisition*, chapter 1, pp. 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):211–229, 1959.

Yang, R., Bai, C., Ma, X., Wang, Z., Zhang, C., and Han, L. RORL: Robust offline reinforcement learning via conservative smoothing. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one, even using the one-column format.