Safe Offline Reinforcement Learning using Diffusion Policies

Extended Abstract

Ankita Kushwaha
IIIT Hyderabad
Hyderabad, India
ankita.kushwaha@research.iiit.ac.in

Kiran Ravish
IIIT Hyderabad
Hyderabad, India
kiran.ravish@research.iiit.ac.in

Preeti
IIIT Hyderabad
Hyderabad, India
preeti.preeti@research.iiit.ac.in

Pawan Kumar
IIIT Hyderabad
Hyderabad, India
pawan.kumar@iiit.ac.in

Anuj Mahajan*
Meta
New York, USA
anuj.mahajan.phd@gmail.com

ABSTRACT

Diffusion models have shown great promise for offline RL by capturing complex data distributions, yet their standard formulations lack explicit safety mechanisms. We propose Safe Diffusion Q-learning, which extends Diffusion-QL by integrating a cost critic and a direct penalty term into the diffusion policy objective to enforce constraint satisfaction during action generation. Evaluated on the DSRL benchmark, our method achieves near-zero constraint violation on challenging BulletSafetyGym and SafetyGym tasks while maintaining competitive reward performance. These results demonstrate that expressive diffusion policies can be robustly constrained, enabling their use in safety-critical offline RL.

KEYWORDS

Safe Offline RL; Diffusion Policies; Q-Learnig; Single Agent

ACM Reference Format:
Ankita Kushwaha, Kiran Ravish, Preeti, Pawan Kumar, and Anuj Mahajan. 2026. Safe Offline Reinforcement Learning using Diffusion Policies: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 4 pages. https://doi.org/10.65109/QVZU7986

1 INTRODUCTION

Safe RL incorporated safety constraints, typically formulated as a constrained Markov decision process (CMDP) to ensure safety [1, 3, 6, 13, 23, 24, 31–33, 37, 38]. The safety constarints can be hard constraints (zero voilation at every step) [12, 40] or soft constarints (expected total cost below a threshold) [15, 36]. We adopt the soft-constraint framework.

Safe offline RL faces two primary challenges: the out of distribution (OOD) problem [5, 25], and the difficulty of learning policies that are both high reward and safe [28, 39]. Previous work on offline RL generally addressed the OOD problem by regularizing how far the policy can deviate from the behavior policy [9, 10, 22, 29, 30, 35]. Recently, diffusion models have shown great success in tackling the OOD problem by generating highly expressive policies [2, 4, 16, 18, 21, 34, 41] through an iterative denoising process. Their ability to accurately model the complex distributions found in offline datasets [7, 27] naturally helps to regularize the policy. However, safety still remains a crucial aspect that these models do not inherently address. In this paper, we introduce Safe Diffusion Q-learning (Safe-DQL), which extends Diffusion-QL [34] by using the expressive diffusion model as the policy and explicitly incorporating safety constraints. We introduce a dedicated cost critic to estimate future risk and a direct penalty term in the policy objective to discourage unsafe actions. This design cleanly separates the two core challenges of safe offline RL: the out-of-distribution (OOD) problem is mitigated by diffusion-based behavior regularization, while in-distribution safety is enforced through the explicit penalty. As shown in Table 1, Safe-DQL achieves strong performance on the DSRL benchmark [27], attaining near-zero constraint violation while maintaining competitive rewards.

*Work done outside of Meta.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/QVZU7986

2 SAFE DIFFUSION Q-LEARNING

Our approach builds on Diffusion-QL [34], which represents the policy $\pi_\theta(a \mid s)$ as the reverse process of a conditional denoising diffusion model [19], enabling expressive and multi-modal action generation. The diffusion-based behavior cloning objective implicitly constrains the policy to the data manifold, mitigating extrapolation errors from out-of-distribution (OOD) actions in offline RL. To incorporate safety, we extend the critic architecture with two parallel cost critics that estimate the expected cost-to-go for each state-action pair. The critics are trained via standard Bellman error minimization [10, 26] with a double Q-learning trick [17], where the reward target takes the minimum over two critics [11], while the cost critics use the maximum to enforce pessimism and avoid underestimating safety risks. Safety is further enforced at the policy level by augmenting the Diffusion-QL objective with an explicit penalty term that discourages actions whose predicted cost exceeds a predefined threshold κ. The policy π_θ is trained by minimizing the following objective:

$$\begin{aligned}\mathcal{L}(\theta) &= \mathcal{L}_d(\theta) - \eta\, \mathbb{E}_{s\sim\mathcal{D},\, a_0\sim\pi_\theta}[Q_r(s,a_0)] \\ &+ \lambda \mathbb{E}_{s\sim\mathcal{D},\, a_0\sim\pi_\theta}[\mathrm{ReLU}(\tilde{Q}^c_{\max}(s,a_0)-\kappa)]. \qquad (1)\end{aligned}$$

where $\mathcal{L}_d(\theta)$ is the diffusion behavior-cloning loss, $Q_r = \min_{i=1,2} Q^r_{\phi_i}$ denotes the reward critic, and $\tilde{Q}^c_{\max} = \max_{i=1,2} Q^c_{\psi_i}$ is the pessimistic cost estimate obtained from the cost-critic ensemble.

Table 1: We evaluate Safe-DQL on Safety-Gymnasium [8, 20] and Bullet-Safety-Gym [14] tasks on DSRL benchmark [27]. Normalized reward (↑) and cost (↓) (cost threshold = 1) are averaged over 20 evaluation episodes and 3 random seeds. Bold: Safe agents (normalized cost < 1). Gray: Unsafe agents. Blue: Safe agent with the highest reward. We compare against the representative baselines [10, 24, 28, 36] and behavior cloning variants (BC-All and BC-Safe).

Task	BC-All		BC-Safe		CDT		BCQ-Lag		CPQ		COptiDICE		Safe-DQL (ours)	
	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓
PointCircle1	**0.23**	**0.79**	**0.33**	**0.12**	**0.49**	**0.06**	0.19	1.06	**0.42**	**0.09**	0.71	7.18	**0.46**	**0.00**
PointCircle2	0.64	7.02	**0.23**	**0.00**	**0.54**	**0.40**	0.51	3.12	**-0.02**	**0.40**	0.80	13.04	**0.42**	**0.17**
CarGoal1	**0.36**	**0.80**	**0.23**	**0.43**	0.57	1.83	0.51	1.85	**-0.71**	**0.27**	**0.44**	**0.72**	**0.29**	**0.45**
PointGoal1	0.59	1.58	**0.23**	**0.56**	0.48	1.49	0.72	1.87	**-0.41**	**0.06**	0.36	1.80	**0.75**	**0.23**
PointPush1	0.18	1.83	**0.21**	**0.96**	0.20	1.68	0.17	1.74	**-0.02**	**0.30**	**0.12**	**0.95**	**0.23**	**0.05**
PointPush2	**0.21**	**0.86**	0.10	1.42	**0.15**	**0.82**	**0.08**	**0.81**	**-0.01**	**0.55**	0.04	1.33	**0.22**	**0.24**
AntCircle	0.60	3.37	**0.41**	**0.65**	0.49	2.07	0.54	2.10	**0.01**	**0.87**	0.09	3.91	**0.56**	**0.34**
CarCircle	0.65	11.16	**0.33**	**0.53**	0.71	2.19	0.68	8.84	0.49	4.48	0.44	7.73	**0.72**	**0.00**
BallCircle	0.72	31.31	0.37	2.35	0.62	11.23	0.67	24.05	**0.66**	**0.00**	0.61	32.60	**0.73**	**0.06**
AntRun	0.66	3.38	**0.61**	**0.35**	0.63	4.90	**0.46**	**0.74**	**0.06**	**0.00**	**0.41**	0.68	**0.54**	**0.30**
CarRun	0.96	1.88	**0.96**	**0.01**	0.99	1.10	0.84	2.51	1.06	10.49	**0.92**	**0.00**	**0.73**	**0.20**
BallRun	0.67	11.38	**0.04**	**0.00**	**0.32**	**0.45**	0.43	6.25	0.85	13.67	0.55	11.32	**0.25**	**0.00**

3 THEORETICAL RESULTS

3.1 Assumptions (Max-aggregator)

Our theoretical results rely on the following assumptions, which formalize boundedness, coverage, and critic approximation conditions under which the safety guarantees of Safe-DQL hold.

Assumption 1 (Bounded signals and critics). *Rewards and costs are bounded:* $|r(s,a)| \le R_{\max}$ *and* $0 \le c(s,a) \le C_{\max}$ *for all* (s,a). *Consequently, for any policy* π,

$$|Q_r^\pi(s,a)| \le B_r := \frac{R_{\max}}{1-\gamma}, \qquad 0 \le Q_c^\pi(s,a) \le B_c := \frac{C_{\max}}{1-\gamma}.$$

We assume the learned reward and cost critics are uniformly bounded (e.g., via clipping) by the same constants, i.e., $|Q_r(s,a)| \le B_r$ *and* $|Q_{\psi_i}^c(s,a)| \le B_c$ *for* $i \in \{1,2\}$.

Assumption 2 (Coverage / concentrability). *There exists a constant* $C_\infty \ge 1$ *such that for any non-negative measurable function* f *and any policy* π *in the class,*

$$\mathbb{E}_{d_\pi}[f] \;\le\; C_\infty\, \mathbb{E}_{d_\mu}[f],$$

where d_μ *and* d_π *denote the discounted occupancy measures of the behavior policy and* π, *respectively.*

Assumption 3 (Critic approximation). *There exist* $\varepsilon_{c,1}, \varepsilon_{c,2}, \varepsilon_r \ge 0$, *such that* $\sup_{s,a}\left|Q_{\psi_i}^c(s,a) - Q_c^{\pi_\theta}(s,a)\right| \le \varepsilon_{c,i}$ *for* $i \in \{1,2\}$, *and for all policies* π *appearing in the reward analysis (i.e.,* $\pi \in \{\pi_\theta, \pi^\dagger\}$*),* $\sup_{s,a}\left|Q_r(s,a) - Q_r^\pi(s,a)\right| \le \varepsilon_r$.

Assumption 4 (Conservatism for safety, a weak version). *At least one cost critic is uniformly conservative at* π_θ*: there exists* $i^\star \in \{1,2\}$ *such that* $Q_{c,\psi_{i^\star}}(s,a) \ge Q_c^{\pi_\theta}(s,a) - \varepsilon_{c,i^\star}$ *for all* (s,a).

Assumption 5 (BC controls divergence). *There exists a constant* $c_d > 0$ *such that* $\mathbb{E}_{s\sim D}[\mathrm{KL}(\pi_\theta(\cdot \mid s) \,\|\, \mu(\cdot \mid s))] \le c_d\, L_d(\theta)$.

Assumption 6 (Feasible reference policy). *There exists a reference policy* π_0 *such that* $J_c(\pi_0) \le \kappa$ *and*

$$\mathbb{E}_{D,\pi_0}\left[\mathrm{ReLU}\left(\tilde{Q}_c^{\max}(s,a) - \kappa\right)\right] = 0, \quad \tilde{Q}_c^{\max}(s,a) := \max_{i\in\{1,2\}} Q_{\psi_i}^c(s,a).$$

3.2 Safety via the Max-aggregated Hinge Penalty

Since we use RELU function to enforce our constraints on the cost critic, following simple bound will be useful.

Lemma 7 (Hinge upper-bounds predicted cost). *For any real-valued random variable* X *and any threshold* $\kappa \in \mathbb{R}$,

$$X \le \kappa + \mathrm{ReLU}(X - \kappa).$$

Consequently, for any integrable $\tilde{Q}_c^{\max}$,

$$\mathbb{E}_{D,\pi}\left[\tilde{Q}_c^{\max}\right] \;\le\; \kappa + \mathbb{E}_{D,\pi}\left[\,\mathrm{ReLU}(\tilde{Q}_c^{\max} - \kappa)\right].$$

Lemma 8 (Predicted ⇒ true cost under weak conservatism). *Under Assumptions 3 and 4,*

$$\mathbb{E}_{D,\pi_\theta}\left[Q_c^{\pi_\theta}\right] \;\le\; \mathbb{E}_{D,\pi_\theta}\left[\tilde{Q}_c^{\max}\right] + \varepsilon_{c,i^\star},$$

where $i^\star$ *is the conservative critic in Assumption 4.*

Theorem 9 guarantees that the learned policy is nearly feasible: its true discounted cost exceeds the threshold κ by at most an explicit slack consisting of the conservative cost-critic error $\varepsilon_{c,i^\star}$, a penalty-controlled term $(\mathcal{L}_d(\pi_0) + 2\eta B_r)/\lambda$, and an initial-state mismatch term Δ_{occ}, which vanishes when $\mathcal{D}_S = \mu_0$. Consequently, increasing λ and improving the conservative cost critic directly tighten the safety guarantee.

Theorem 9 (Safety with explicit slack: max-aggregator). *Under Assumptions 1–6, any global minimizer* $\pi_{\hat{\theta}}$ *of* (1) *satisfies*

$$J_c(\pi_{\hat{\theta}}) \;\le\; \kappa \;+\; \varepsilon_{c,i^\star} \;+\; \frac{L_d(\pi_0)}{\lambda} \;+\; \frac{2\eta B_r}{\lambda} \;+\; \Delta_{\mathrm{occ}},$$

where $\Delta_{\mathrm{occ}} := 2B_c\,\mathrm{TV}(\mu_0, D_S)$ *converts* $\mathbb{E}_{D,\pi_{\hat{\theta}}}[Q_c^{\pi_{\hat{\theta}}}]$ *to* $J_c(\pi_{\hat{\theta}})$ *(*D_S *is the state marginal of* D*); thus* $\Delta_{\mathrm{occ}} = 0$ *when* $D_S = \mu_0$.

4 CONCLUSION

Safe Diffusion Q-learning integrates pessimistic cost estimation and explicit penalization into diffusion-based offline RL to address both OOD errors and constraint violations. The empirical results in Table 1, together with our theoretical guarantees, establish Safe-DQL as a practical and principled approach to safe offline RL.

ACKNOWLEDGMENTS

This research was financially supported by the University Grant Commission (UGC), Government of India, through the award of the Junior Research Fellowship (JRF) under the National Eligibility Test (NET) (Ref. no. 221610070470 & Ref. No. 221610037786). The work was also supported by the Council of Scientific and Industrial Research (CSIR), Government of India, through a CSIR Research Fellowship (09/0917(17235)/2023-EMR-I).

REFERENCES

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained Policy Optimization. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, Sydney, Australia, 22–31. https://proceedings.mlr.press/v70/achiam17a.html

[2] Suzan Ece Ada, Erhan Oztop, and Emre Ugur. 2024. Diffusion Policies for Out-of-Distribution Generalization in Offline Reinforcement Learning. *IEEE Robotics and Automation Letters* 9, 4 (2024), 3116–3123. https://doi.org/10.1109/LRA.2024.3363530

[3] Eitan Altman. 1999. *Constrained Markov Decision Processes*. Routledge. https://doi.org/10.1201/9781315140223

[4] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2024. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. arXiv:2303.04137 [cs.RO] https://arxiv.org/abs/2303.04137

[5] Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. 2024. A Survey on Offline Reinforcement Learning: Taxonomy, Review, and Open Problems. *IEEE Transactions on Neural Networks and Learning Systems* 35, 8 (Aug. 2024), 10237–10257. https://doi.org/10.1109/tnnls.2023.3250269

[6] Yannis Flet-Berliac and Debabrota Basu. 2022. SAAC: Safe Reinforcement Learning as an Adversarial Game of Actor-Critics. arXiv:2204.09424 [cs.LG] https://arxiv.org/abs/2204.09424

[7] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2021. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. arXiv:2004.07219 [cs.LG] https://arxiv.org/abs/2004.07219

[8] Scott Fujimoto, Edoardo Conti, Mohammad Ghavamzadeh, and Joelle Pineau. 2019. Benchmarking Batch Deep Reinforcement Learning Algorithms. arXiv:1910.01708 [cs.LG] https://arxiv.org/abs/1910.01708

[9] Scott Fujimoto and Shixiang (Shane) Gu. 2021. A Minimalist Approach to Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 20132–20145. https://proceedings.neurips.cc/paper_files/paper/2021/file/a8166da05c5a094f7dc03724b41886e5-Paper.pdf

[10] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-Policy Deep Reinforcement Learning without Exploration. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2052–2062. https://proceedings.mlr.press/v97/fujimoto19a.html

[11] Scott Fujimoto, Herke van Hoof, and David Meger. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 1587–1596. https://proceedings.mlr.press/v80/fujimoto18a.html

[12] Milan Ganai, Zheng Gong, Chenning Yu, Sylvia Herbert, and Sicun Gao. 2023. Iterative Reachability Estimation for Safe Reinforcement Learning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 69764–69797. https://proceedings.neurips.cc/paper_files/paper/2023/file/dca63f2650fe9e88956c1b68440b8ee9-Paper-Conference.pdf

[13] Javier García, Fern, and o Fernández. 2015. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research* 16, 42 (2015), 1437–1480. http://jmlr.org/papers/v16/garcia15a.html

[14] Sven Gronauer. 2022. *Bullet-Safety-Gym: A Framework for Constrained Reinforcement Learning*. Technical Report. mediaTUM. https://doi.org/10.14459/2022md1639974

[15] Junyu Guo, Zhi Zheng, Donghao Ying, Ming Jin, Shangding Gu, Costas Spanos, and Javad Lavaei. 2025. Don't Trade Off Safety: Diffusion Regularization for Constrained Offline RL. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[16] Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. 2023. IDQL: Implicit Q-Learning as an Actor-Critic Method with Diffusion Policies. arXiv:2304.10573 [cs.LG]

[17] Hado van Hasselt. 2010. Double Q-learning. In *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 2* (Vancouver, British Columbia, Canada) *(NIPS'10)*. Curran Associates Inc., Red Hook, NY, USA, 2613–2621.

[18] Shashank Hegde, Sumeet Batra, K.R. Zentner, and Gaurav Sukhatme. 2023. Generating Behaviorally Diverse Policies with Latent Diffusion Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 7541–7554. https://proceedings.neurips.cc/paper_files/paper/2023/file/180d4373aca26bd86bf45fc50d1a709f-Paper-Conference.pdf

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 574, 12 pages.

[20] Jiaming Ji, Jiayi Zhou, Borong Zhang, Juntao Dai, Xuehai Pan, Ruiyang Sun, Weidong Huang, Yiran Geng, Mickel Liu, and Yaodong Yang. 2024. OmniSafe: An Infrastructure for Accelerating Safe Reinforcement Learning Research. *Journal of Machine Learning Research* 25, 285 (2024), 1–6. http://jmlr.org/papers/v25/23-0681.html

[21] Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. 2023. Efficient Diffusion Policies For Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 67195–67212. https://proceedings.neurips.cc/paper_files/paper/2023/file/d45e0bfb5a39477d56b55c0824200008-Paper-Conference.pdf

[22] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. 2019. Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., 11784–11794. https://proceedings.neurips.cc/paper_files/paper/2019/file/c2073ffa77b5357a498057413bb09d3a-Paper.pdf

[23] Ankita Kushwaha, Kiran Ravish, Preeti Lamba, and Pawan Kumar. 2025. A Survey of Safe Reinforcement Learning and Constrained MDPs: A Technical Survey on Single-Agent and Multi-Agent Safety. arXiv:2505.17342 [cs.LG] https://arxiv.org/abs/2505.17342

[24] Jongmin Lee, Cosmin Paduraru, Daniel J. Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez. 2022. COptiDICE: Offline Constrained Reinforcement Learning via Stationary Distribution Correction Estimation. arXiv:2204.08957 [cs.LG] https://arxiv.org/abs/2204.08957

[25] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. arXiv:2005.01643 [cs.LG] https://arxiv.org/abs/2005.01643

[26] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2019. Continuous control with deep reinforcement learning. arXiv:1509.02971 [cs.LG] https://arxiv.org/abs/1509.02971

[27] Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, and Ding Zhao. 2024. Datasets and Benchmarks for Offline Safe Reinforcement Learning. *Journal of Data-centric Machine Learning Research* (2024).

[28] Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. 2023. Constrained Decision Transformer for Offline Safe Reinforcement Learning. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, Honolulu, Hawaii, USA, 21611–21630. https://proceedings.mlr.press/v202/liu23m.html

[29] Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. 2022. Mildly Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 1711–1724. https://proceedings.neurips.cc/paper_files/paper/2022/file/0b5669c3b07bb8429af19a7919376ff5-Paper-Conference.pdf

[30] Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. 2021. AWAC: Accelerating Online Reinforcement Learning with Offline Datasets. arXiv:2006.09359 [cs.LG] https://arxiv.org/abs/2006.09359

[31] Adam Stooke, Joshua Achiam, and Pieter Abbeel. 2020. Responsive Safety in Reinforcement Learning by PID Lagrangian Methods. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 9133–9143. https://proceedings.mlr.press/v119/stooke20a.html

[32] Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. 2018. Reward Constrained Policy Optimization. arXiv:1805.11074 [cs.LG] https://arxiv.org/abs/1805.11074

[33] Akifumi Wachi, Xun Shen, and Yanan Sui. 2024. A Survey of Constraint Formulations in Safe Reinforcement Learning. arXiv:2402.02025 [cs.LG] https://arxiv.org/abs/2402.02025

[34] Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. 2023. Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning. arXiv:2208.06193 [cs.LG] https://arxiv.org/abs/2208.06193

[35] Yifan Wu, George Tucker, and Ofir Nachum. 2019. Behavior Regularized Offline Reinforcement Learning. arXiv:1911.11361 [cs.LG] https://arxiv.org/abs/1911.11361

[36] Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. 2022. Constraints Penalized Q-learning for Safe Offline Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 8 (Jun. 2022), 8753–8760. https://doi.org/

10.1609/aaai.v36i8.20855

[37] Qisong Yang, Thiago D. Simão, Simon H Tindemans, and Matthijs T. J. Spaan. 2021. WCSAC: Worst-Case Soft Actor Critic for Safety-Constrained Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 12 (May 2021), 10639–10646. https://doi.org/10.1609/aaai.v35i12.17272

[38] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J. Ramadge. 2020. Projection-Based Constrained Policy Optimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rke3TJrtPS

[39] Linrui Zhang, Li Shen, Long Yang, Shixiang Chen, Xueqian Wang, Bo Yuan, and Dacheng Tao. 2022. Penalized Proximal Policy Optimization for Safe Reinforcement Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 3744–3750. https://doi.org/10.24963/ijcai.2022/520 Main Track.

[40] Yinan Zheng, Jianxiong Li, Dongjie Yu, Yujie Yang, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. 2024. Safe Offline Reinforcement Learning with Feasibility-Guided Diffusion Model. In *The Twelfth International Conference on Learning Representations*. Vienna, Austria.

[41] Zhengbang Zhu, Hanye Zhao, Haoran He, Yichao Zhong, Shenyu Zhang, Haoquan Guo, Tingting Chen, and Weinan Zhang. 2024. Diffusion Models for Reinforcement Learning: A Survey. arXiv:2311.01223 [cs.LG] https://arxiv.org/abs/2311.01223