References

Alekh Agarwal, Akshay Krishnamurthy, John Langford, Haipeng Luo, et al. Open problem: First-order regret bounds for contextual bandits. In *Conference on Learning Theory*, pages 4–7. PMLR, 2017.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Zeyuan Allen-Zhu, Sébastien Bubeck, and Yuanzhi Li. Make the minority great again: First-order regret bound for contextual bandits. In *International Conference on Machine Learning*, pages 186–194. PMLR, 2018.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Gabriel Barth-Maron, Matthew W. Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva TB, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributional policy gradients. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyZipzbCb.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017.

Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

Cristian Bodnar, Adrian Li, Karol Hausman, Peter Pastor, and Mrinal Kalakrishnan. Quantile qt-opt for risk-aware vision-based robotic grasping. *Robotics: Science and Systems*, 2020.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Jonathan Chang, Kaiwen Wang, Nathan Kallus, and Wen Sun. Learning bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning*, pages 2938–2971. PMLR, 2022.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pages 3852–3878. PMLR, 2022.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Chris Dann, Yishay Mansour, Mehryar Mohri, Ayush Sekhari, and Karthik Sridharan. Guarantees for epsilon-greedy reinforcement learning with function approximation. In *International Conference on Machine Learning*, pages 4666–4689. PMLR, 2022.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems*, 31, 2018.

Mónika Farsang, Paul Mineiro, and Wangda Zhang. Conditionally risk-averse contextual bandits. *arXiv preprint arXiv:2210.13573*, 2022.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34: 18907–18919, 2021.

Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*, 29, 2016.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, pages 3489–3489. PMLR, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Audrey Huang, Jinglin Chen, and Nan Jiang. Reinforcement learning in low-rank mdps with density features. *arXiv preprint arXiv:2302.02252*, 2023.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020b.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021a.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021b.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, pages 10598–10632. PMLR, 2022.

Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443, 2020.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33: 21810–21823, 2020.

J Kolter. The fixed points of off-policy td. *Advances in Neural Information Processing Systems*, 24, 2011.

Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. 2009.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Shiau Hong Lim and Ilyas Malik. Distributional reinforcement learning for risk-sensitive policies. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=wSVEd3Ta42m.

Thodoris Lykouris, Karthik Sridharan, and Éva Tardos. Small-loss bounds for online learning with partial information. *Mathematics of Operations Research*, 47(3):2186–2218, 2022.

Clare Lyle, Marc G Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4504–4511, 2019.

Yecheng Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 34:19235–19247, 2021.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*, 2021.

Anna Montoya, BigJek14, Bull, denisedunleavy, egrad, FleetwoodHack, Imbayoh, PadraicS, Pru_Admin, tpitman, and Will Cukierski. Prudential life insurance assessment, 2015. URL https://kaggle.com/competitions/prudential-life-insurance-assessment.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, pages 1360–1375. PMLR, 2015.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 29–37. PMLR, 2018.

Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G Bellemare, and Will Dabney. An analysis of quantile temporal-difference learning. *arXiv preprint arXiv:2301.04462*, 2023.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Flemming Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on information theory*, 46(4):1602–1609, 2000.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=tyrJsbKAe6.

Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198. URL http://doi.acm.org/10.1145/2641190.2641198.

István Vincze. On the concept and measure of information contained in an observation. In *Contributions to Probability*, pages 207–214. Elsevier, 1981.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pages 22384–22429. PMLR, 2022.

Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with cvar. *International Conference on Machine Learning*, 2023.

Ruosong Wang, Dean Foster, and Sham M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=30EvkP2aQLD.

Ruosong Wang, Yifan Wu, Ruslan Salakhutdinov, and Sham Kakade. Instabilities of offline rl with pre-trained neural representation. In *International Conference on Machine Learning*, pages 10948–10960. PMLR, 2021b.

Yuanhao Wang, Ruosong Wang, and Sham Kakade. An exponential lower bound for linearly realizable mdp with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 34:9521–9533, 2021c.

Runzhe Wu, Masatoshi Uehara, and Wen Sun. Distributional offline policy evaluation with predictive error guarantees. *International Conference of Machine Learning*, 2023.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=LQIjzPdDt3q.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pages 26447–26466. PMLR, 2022.

Tong Zhang. *Mathematical Analysis of Machine Learning Algorithms*. 2023. http://www.tongzhang-ml.org/lt-book.html.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023.

Appendices

A Notations

Table 2: List of Notations

$\mathcal{S}, \mathcal{A}, A$	State and action spaces, and $A = \lvert\mathcal{A}\rvert$.
$\Delta(S)$	The set of distributions supported by S.
$\bar{d}$	The expectation of any real-valued distribution d, *i.e.*, $\bar{d} = \mathbb{E}_{y\sim d}[y]$.
$[N]$	$\{1, 2, \dots, N\}$ for any natural number N.
$Z_h^\pi(x,a)$	Distribution of $\sum_{t=h}^H c_t$ given $x_h = x, a_h = a$ rolling in from π.
$Q_h^\pi(x,a), V_h^\pi(x)$	$Q_h^\pi(x,a) = \bar{Z}_h^\pi(x,a)$ and $V_h^\pi = \mathbb{E}_{a\sim\pi(x)}[Q_h^\pi(x,a)]$.
$\pi^\star$	Optimal policy, *i.e.*, $\pi^\star = \arg\min_\pi V_1^\pi(x_1)$.
	Without loss of optimality, we take $\pi^\star : \mathcal{X} \to \mathcal{A}$ to be Markov & deterministic.
$Z_h^\star, Q_h^\star, V_h^\star$	$Z_h^\pi, Q_h^\pi, V_h^\pi$ with $\pi = \pi^\star$, the optimal policy.
$\mathcal{T}_h^\pi, \mathcal{T}_h^\star$	The Bellman operators that act on functions.
$\mathcal{T}_h^{\pi,D}, \mathcal{T}_h^{\star,D}$	The distributional Bellman operators that act on conditional distributions.
$V^\pi, Z^\pi, V^\star, Z^\star$	$V^\pi = V_1^\pi(x_1)$, $Z^\pi = Z_1^\pi(x_1)$. $V^\star, Z^\star$ are defined similarly with $\pi^\star$.
$d_h^\pi(x,a)$	The probability of π visiting (x,a) at time h.
$C^{\tilde{\pi}}$	Coverage coefficient $\max_h \left\lVert \mathrm{d}d_h^{\tilde{\pi}}/\mathrm{d}\nu_h \right\rVert_\infty$.
$D_\triangle(f \parallel g)$	Triangular discrimination between f, g.
$H(f \parallel g)$	Hellinger distance between f, g.
$D_{KL}(f \parallel g)$	KL divergence between f, g.

A.1 Statistical Distances

Let f, g be distributions over $\mathcal{Y}$. Then,

$$
\begin{aligned}
D_\triangle(f \parallel g) &= \sum_y \frac{\left(f(y) - g(y)\right)^2}{f(y) + g(y)}, \\
H(f \parallel g) &= \sqrt{\frac{1}{2}\sum_y \left(\sqrt{f(y)} - \sqrt{g(y)}\right)^2}, \\
D_{KL}(f \parallel g) &= \sum_y f(y)\log(f(y)/g(y)), \\
D_{TV}(f \parallel g) &= \frac{1}{2}\sum_y |f(y) - g(y)|.
\end{aligned}
$$

The following standard inequalities will be helpful:

$$
\begin{aligned}
&H^2 \le D_{TV} \le \sqrt{2}H, \\
&2H^2 \le D_\triangle \le 4H^2, \qquad \text{(Lemma A.1)} \\
&H \le \sqrt{D_{KL}}.
\end{aligned}
$$

Lemma A.1. *For any distributions f, g, we have $2H^2(f \parallel g) \le D_\triangle(f \parallel g) \le 4H^2(f \parallel g)$.*

Proof. Recall that

$$
D_\triangle(f \parallel g) = \int_y \left(\frac{f(y) - g(y)}{\sqrt{f(y) + g(y)}}\right)^2.
$$

Applying $\frac{1}{\sqrt{f(y)}+\sqrt{g(y)}} \le \frac{1}{\sqrt{f(y)+g(y)}} \le \frac{\sqrt{2}}{\sqrt{f(y)}+\sqrt{g(y)}}$ concludes the proof. □

B Omitted Algorithms

In this section, we present the O-DISCO algorithm with Uniform Action Exploration (UAE), as described in Section 5.2. We also present versions of O-DISCO and P-DISCO for the reward-maximizing setting (instead of the cost-minimizing setting studied throughout the paper); if SMALLRETURN is turned on, we can derive small-return bounds in Appendix I.

Algorithm 4 O-DISCO (with UAE and small return)

Input: number of episodes K, distribution function class $\mathcal{F}$, threshold β, flag UAE, flag SMALLRETURN.
Initialize $\mathcal{D}_{h,0} \leftarrow \emptyset$ for all $h \in [H]$, and set $\mathcal{F}_0 = \mathcal{F}$.
Set $\mathrm{op} = \max$ if SMALLRETURN else $\mathrm{op} = \min$.
for episode $k = 1, 2, \ldots, K$ **do**
  Set $f^{(k)} = \arg\mathrm{op}_{f \in \mathcal{F}_{k-1}} \mathrm{op}_a \bar{f}_1(x_1, a)$.
  Set $\pi_h^k(x) = \arg\mathrm{op}_a \bar{f}_h^{(k)}(x, a)$.
  if UAE **then**
    For each $h \in [H]$, collect $x_{h,k} \sim d_h^{\pi^k}, a_{h,k} \sim \mathrm{unif}(\mathcal{A}), c_{h,k} \sim C_h(x_{h,k}, a_{h,k}), x'_{h,k} \sim P_h(x_{h,k}, a_{h,k})$, and augment the dataset $\mathcal{D}_{h,k} = \mathcal{D}_{h,k-1} \cup \left\{(x_{h,k}, a_{h,k}, c_{h,k}, x'_{h,k})\right\}$.
  else
    Roll out π^k and obtain a trajectory $x_{1,k}, a_{1,k}, c_{1,k}, \ldots, x_{H,k}, a_{H,k}, c_{H,k}$.
    For each $h \in [H]$, augment the dataset $\mathcal{D}_{h,k} = \mathcal{D}_{h,k-1} \cup \{(x_{h,k}, a_{h,k}, c_{h,k}, x_{h+1,k})\}$.
  end if
  For all $(h, f) \in [H] \times \mathcal{F}$, sample $y_{h,i}^f \sim f_{h+1}(x'_{h,i}, a')$ and $a' = \arg\mathrm{op}_a \bar{f}_{h+1}(x'_{h,i}, a)$, where $(x_{h,i}, a_{h,i}, c_{h,i}, x'_{h,i})$ is the i-th datapoint of $\mathcal{D}_{h,k}$. Also, set $z_{h,i}^f = c_{h,i} + y_{h,i}^f$ and define the confidence set,

$$\mathcal{F}_k = \left\{ f \in \mathcal{F} : \sum_{i=1}^{k} \log f_h(z_{h,i}^f \mid x_{h,i}, a_{h,i}) \geq \max_{\widetilde{f} \in \mathcal{F}} \sum_{i=1}^{k} \log \widetilde{f}_h(z_{h,i}^f \mid x_{h,i}, a_{h,i}) - 7\beta, \forall h \in [H] \right\}.$$

end for
Output: $\bar{\pi} = \mathrm{unif}(\pi^{1:K})$.

Algorithm 5 P-DISCO (with small return)

Input: datasets $\mathcal{D}_1, \ldots, \mathcal{D}_H$, distribution function class $\mathcal{F}$, threshold β, policy class Π, flag SMALLRETURN.
For all $(h, f, \pi) \in [H] \times \mathcal{F} \times \Pi$, sample $y_{h,i}^{f,\pi} \sim f_{h+1}(x'_{h,i}, \pi_{h+1}(x'_{h,i}))$, where $(x_{h,i}, a_{h,i}, c_{h,i}, x'_{h,i})$ is the i-th datapoint of $\mathcal{D}_h$. Then, set $z_{h,i}^{f,\pi} = c_{h,i} + y_{h,i}^{f,\pi}$ and define the confidence set,

$$\mathcal{F}_\pi = \left\{ f \in \mathcal{F} : \sum_{i=1}^{N} \log f_h(z_{h,i}^{f,\pi} \mid x_{h,i}, a_{h,i}) \geq \max_{\widetilde{f} \in \mathcal{F}} \sum_{i=1}^{N} \log \widetilde{f}_h(z_{h,i}^{f,\pi} \mid x_{h,i}, a_{h,i}) - 7\beta, \forall h \in [H] \right\}.$$

Set $\mathrm{op} = \max$ if SMALLRETURN else $\mathrm{op} = \min$.
For each $\pi \in \Pi$, define the pessimistic estimate $f^\pi = \arg\mathrm{op}_{f \in \mathcal{F}_\pi} \mathbb{E}_{a \sim \pi(x_1)}\left[\bar{f}_1(x_1, a)\right]$.
Output: $\widehat{\pi} = \arg\mathrm{op}_{\pi \in \Pi} \mathbb{E}_{a \sim \pi(x_1)}\left[\bar{f}_1^\pi(x_1, \pi)\right]$.

C Proofs for DISTCB

Lemma C.1 (Azuma). *Let $\{X_i\}_{i\in[N]}$ be a sequence of random variables supported on $[0,1]$, adapted to filtration $\{\mathcal{F}_i\}_{i\in[N]}$. For any $\delta\in(0,1)$, we have w.p. at least $1-\delta$,*

$$\sum_{t=1}^{N}\mathbb{E}[X_t\mid\mathcal{F}_{t-1}]\le\sum_{t=1}^{N}X_t+\sqrt{N\log(2/\delta)},\qquad\text{(Standard Azuma)}$$

$$\sum_{t=1}^{N}\mathbb{E}[X_t\mid\mathcal{F}_{t-1}]\le 2\sum_{t=1}^{N}X_t+2\log(1/\delta).\qquad\text{(Multiplicative Azuma)}$$

Proof. For standard Azuma, see Zhang [2023, Theorem 13.4]. For multiplicative Azuma, apply [Zhang, 2023, Theorem 13.5] with $\lambda=1$. The claim follows, since $\frac{1}{1-\exp(-\lambda)}\le 2$. □

Theorem 4.1. *Fix any $\delta\in(0,1)$ and set $\gamma=10A\vee\sqrt{\frac{40A(C^\star+\log(1/\delta))}{112\left(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\right)}}$. Then, w.p. at least $1-\delta$,* DISTCB *satisfies,*

$$\mathrm{Regret}_{\textsc{DistCB}}(K)\le 232\sqrt{AC^\star\,\mathrm{Regret}_{\log}(K)\log(1/\delta)}+2300A\big(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\big),$$

where $C^\star=\sum_{k=1}^{K}\min_{a\in\mathcal{A}}\bar{C}(x_k,a)$ is the cumulative cost of the optimal policy.

Proof of Theorem 4.1. First, recall the per-step inequality of ReIGW Foster and Krishnamurthy [2021, Theorem 4], which states: for any $\widehat{f}$ and $\gamma\ge 2A$, if we set $p=\mathrm{ReIGW}_\gamma(\widehat{f},\gamma)$, then, for all $f\in[0,1]^A$, we have

$$\textstyle\sum_a p(a)(f(a)-f(a^\star))\le\frac{5A}{\gamma}\sum_a p(a)f(a)+7\gamma\sum_a p(a)\frac{\left(\widehat{f}(a)-f(a)\right)^2}{\widehat{f}(a)+f(a)},$$

where $a^\star=\arg\min_a f(a)$. For any $k\in[K]$, applying this to $\widehat{f}=\bar{f}^{(k)}(s_k,\cdot)$, $p=p_k$ and $f=\bar{C}(s_k,\cdot)$, we have

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}_{a_k}\left[\bar{C}(s_k,a_k)-\bar{C}(s_k,\pi^\star(s_k))\right]&\le\sum_{k=1}^{K}\mathbb{E}_{a_k}\left[\frac{5A}{\gamma}\bar{C}(s_k,a_k)+7\gamma\frac{\left(\bar{f}^{(k)}(s_k,a_k)-\bar{C}(s_k,a_k)\right)^2}{\bar{f}^{(k)}(s_k,a_k)+\bar{C}(s_k,a_k)}\right]\\&\le\sum_{k=1}^{K}\mathbb{E}_{a_k}\left[\frac{5A}{\gamma}\bar{C}(s_k,a_k)+7\gamma D_{\triangle}(f^{(k)}(s_k,a_k)\,\|\,C(s_k,a_k))\right]\end{aligned}$$

(Eq. ($\triangle_1$))

Since $D_\triangle\le 4H^2$, we have

$$\begin{aligned}&\sum_{k=1}^{K}\mathbb{E}_{a_k}\left[D_{\triangle}(f^{(k)}(s_k,a_k)\,\|\,C(s_k,a_k))\right]\\&\le 4\sum_{k=1}^{K}\mathbb{E}_{a_k}\left[H^2\Big(C(s_k,a_k)\,\|\,f^{(k)}(s_k,a_k)\Big)\right]\\&\le 8\sum_{k=1}^{K}H^2\Big(C(s_k,a_k)\,\|\,f^{(k)}(s_k,a_k)\Big)+8\log(1/\delta)\quad\text{(Multiplicative Azuma, since } H^2\in[0,1])\\&\le 8\,\mathrm{Regret}_{\log}(K)+10\log(1/\delta).\qquad\text{(Foster et al. [2021, Lemma A.14])}\end{aligned}$$

Hence, we have

$$\sum_{k=1}^{K}\mathbb{E}_{a_k}\left[\bar{C}(s_k,a_k)-\bar{C}(s_k,\pi^\star(s_k))\right]\le\frac{5A}{\gamma}\sum_{k=1}^{K}\mathbb{E}_{a_k}\left[\bar{C}(s_k,a_k)\right]+70\gamma\big(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\big).$$

Finally, recalling that $1/(1-\varepsilon) \le 1+2\varepsilon$ when $\varepsilon \le \frac{1}{2}$, and the fact that $\frac{5A}{\gamma} \le \frac{1}{2}$, we have

$$\sum_{k=1}^{K} \mathbb{E}_{a_k}\big[\bar{C}(s_k,a_k) - \bar{C}(s_k,\pi^\star(s_k))\big] \le \frac{10A}{\gamma}\sum_{k=1}^{K}\mathbb{E}_{a_k}\big[\bar{C}(s_k,\pi^\star(s_k))\big] + 140\gamma\big(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\big).$$

By Azuma's inequality, we have

$$\begin{aligned}
&\sum_{k=1}^{K} \bar{C}(s_k,a_k) - \bar{C}(s_k,\pi^\star(s_k)) \\
&\le 2\sum_{k=1}^{K} \mathbb{E}_{a_k}\big[\bar{C}(s_k,a_k) - \bar{C}(s_k,\pi^\star(s_k))\big] + 2\log(1/\delta) \\
&\le \frac{20A}{\gamma}\sum_{k=1}^{K}\mathbb{E}_{a_k}\big[\bar{C}(s_k,\pi^\star(s_k))\big] + 140\gamma\big(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\big) + 2\log(1/\delta) \\
&\le \frac{40A}{\gamma}(C^\star + \log(1/\delta)) + 140\gamma\big(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\big) + 2\log(1/\delta).
\end{aligned}$$

Now set $\gamma = \sqrt{\frac{40A(C^\star+\log(1/\delta))}{140\left(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\right)}} \vee 10A$.

Case 1 is when $\sqrt{\frac{40A(C^\star+\log(1/\delta))}{140\left(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\right)}} \le 10A$, *i.e.*, $(C^\star+\log(1/\delta)) \le 280A\big(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\big)$, we have the above is at most

$$\begin{aligned}
&4(C^\star+\log(1/\delta)) + 1120A\big(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\big) + 2\log(1/\delta) \\
&\le 2240A\big(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\big) + 2\log(1/\delta).
\end{aligned}$$

Case 2 is when the left term dominates, then the bound is,

$$\begin{aligned}
&2\sqrt{4480A(C^\star+\log(1/\delta))\big(\mathrm{Regret}_{\log}(K)+\log(1/\delta)\big)} + 2\log(1/\delta) \\
&\le 2\sqrt{13440AC^\star\,\mathrm{Regret}_{\log}(K)\log(1/\delta) + 4480A\log^2(1/\delta)} + 2\log(1/\delta) \\
&\le 232\sqrt{AC^\star\,\mathrm{Regret}_{\log}(K)\log(1/\delta)} + 134\sqrt{A}\log(1/\delta) + 2\log(1/\delta).
\end{aligned}$$

Putting these two cases together, we have the result. □

D Placeholder

This section used to contain information that is no longer needed. We kept this placeholder section to ensure the main text's references to the appendix are consistent.

E Maximum Likelihood Estimation

This section reviews generalization bounds for the maximum likelihood estimator (MLE). We adopt the same sequential condition probability estimation setup as in Agarwal et al. [2020, Appendix E], which we now recall for completeness. Let $\mathcal{X}$ be the context/feature space and $\mathcal{Y}$ be the label space, and we are given a dataset $D = \{(x_i, y_i)\}_{i\in[n]}$ from a martingale process: for $i = 1, 2, ..., n$, sample $x_i \sim \mathcal{D}_i(x_{1:i-1}, y_{1:i-1})$ and $y_i \sim p(\cdot \mid x_i)$. Let $f^\star(x, y) = p(y \mid x)$ and we are given a realizable, *i.e.*, $f^\star \in \mathcal{F}$, function class $\mathcal{F} : \mathcal{X} \times \mathcal{Y} \to \Delta(\mathbb{R})$ of distributions. The MLE is an estimate for $f^\star$ that maximizes the log-likelihood objective over our dataset:

$$\widehat{f}_{\text{MLE}} = \arg\max_{f\in\mathcal{F}} \sum_{i=1}^{n} \log f(x_i, y_i).$$

For our guarantees to hold for general hypotheses classes $\mathcal{F}$, we use the bracketing number to quantify the statistical complexity of $\mathcal{F}$ [van de Geer, 2000].

Definition E.1 (Bracketing Number)**.** Let $\mathcal{G}$ be a set of functions mapping $\mathcal{X} \to \mathbb{R}$. Given two functions l, u such that $l(x) \le u(x)$ for all $x \in \mathcal{X}$, the bracket $[l, u]$ is the set of functions $g \in \mathcal{G}$ such that $l(x) \le g(x) \le u(x)$ for all $x \in \mathcal{X}$. We call $[l, u]$ an ε-bracket if $\|u - l\| \le \varepsilon$. Then, the ε-bracketing number of $\mathcal{G}$ with respect to $\|\cdot\|$, denoted by $N_{[]}(\varepsilon, \mathcal{G}, \|\cdot\|)$ is the minimum number of ε-brackets needed to cover $\mathcal{G}$.

Since the triangular discrimination is equivalent to squared Hellinger up to universal constants, we now prove MLE generalization bounds in terms of squared Hellinger.

Lemma E.2. *Let $f_1 : \mathcal{X} \to \Delta(\mathcal{Y})$ and $f_2 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ satisfying $\sup_{x\in\mathcal{X}} \int_{\mathcal{Y}} f_2(x, y)\mathrm{d}y \le s$, then for any distribution $\mathcal{D} \in \Delta(\mathcal{X})$, we have*

$$\mathbb{E}_{x\sim\mathcal{D}}\big[H^2(f_1(x) \parallel f_2(x, \cdot))\big] \le (s - 1) - 2\log \mathbb{E}_{x\sim\mathcal{D}, y\sim f_1(x)} \exp\left(-\frac{1}{2}\log(f_1(x, y)/f_2(x, y))\right).$$

Proof. This follows from the proof of Wu et al. [2023, Lemma C.1]. □

Lemma E.3. *Fix $\delta \in (0, 1)$. Then w.p. at least $1 - \delta$, for any $f \in \mathcal{F}$, we have*

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}_{x\sim\mathcal{D}_i}\big[H^2(f(x, \cdot) \parallel f^\star(x, \cdot))\big] \\
&\quad\le 6n\epsilon|\mathcal{Y}| + 2\sum_{i=1}^{n} \log\big(f^\star(x_i, y_i)/f(x_i, y_i)\big) + 8\log\big(N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)/\delta\big). \qquad (2)
\end{aligned}$$

Rearranging, we also have

$$\sum_{i=1}^{n} \log\big(f(x_i, y_i)/f^\star(x_i, y_i)\big) \le 3n\epsilon|\mathcal{Y}| + 4\log\big(N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)/\delta\big). \qquad (3)$$

Proof. We take an ϵ-bracketing of $\mathcal{F}$, $\{[l_i, u_i] : i = 1, 2, \dots\}$, and denote $\widetilde{\mathcal{F}} = \{u_i : i = 1, 2, \dots\}$. Applying Lemma 24 of Agarwal et al. [2020] to function class $\widetilde{\mathcal{F}}$ and using Chernoff method, w.p. at least $1 - \delta$, for all $\tilde{f} \in \widetilde{\mathcal{F}}$, we have

$$\underbrace{-\log \mathop{\mathbb{E}}_{D'} \exp(L(\tilde{f}(D), D'))}_{\text{(i)}} \le \underbrace{-L(\tilde{f}(D), D) + 2\log\big(N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)/\delta\big)}_{\text{(ii)}}. \qquad (4)$$

Now, fix any $f \in \mathcal{F}$ and pick $\tilde{f} \in \widetilde{\mathcal{F}}$ as the upper bracket, *i.e.*, $f \le \tilde{f}$. Now set $L(f, D) = \sum_{i=1}^{n} -1/2 \log(f^\star(x_i, y_i)/f(x_i, y_i))$. Then the right hand side of (4) is

$$\begin{aligned}
\text{(ii)} =& \frac{1}{2}\sum_{i=1}^{n} \log(f^\star(x_i, y_i)/\tilde{f}(x_i, y_i)) + 2\log\big(N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)/\delta\big) \\
\le& \frac{1}{2}\sum_{i=1}^{n} \log(f^\star(x_i, y_i)/f(x_i, y_i)) + 2\log\big(N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)/\delta\big).
\end{aligned}$$

On the other hand, since H is a metric, we have

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} H^2\left(f(x,\cdot), f^\star(x,\cdot)\right) \le \sum_{i=1}^{n} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left(H\left(f(x,\cdot), \tilde{f}(x,y)\right) + H\left(\tilde{f}(x,y), f^\star(x,\cdot)\right)\right)^2$$
$$\le \underbrace{2\sum_{i=1}^{n} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} H^2\left(f(x,\cdot), \tilde{f}(x,y)\right)}_{\text{(iii)}} + \underbrace{2\sum_{i=1}^{n} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} H^2\left(\tilde{f}(x,y), f^\star(x,\cdot)\right)}_{\text{(iv)}}.$$

For (iii), by the definition, we have $\tilde{f}(x,y) - f(x,y) \in [0,\epsilon]$ for all x, so

$$\text{(iii)} = \sum_{i=1}^{n} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} H^2\left(f(x,\cdot), \tilde{f}(x,y)\right) \le \sum_{i=1}^{n} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} 2\int_y \left|f(x,y) - \tilde{f}(x,y)\right| \mathrm{d}y \le 2n\epsilon|\mathcal{Y}|.$$

For (iv), we apply Lemma E.2 with $f_1 = f^\star$ and $f_2 = \tilde{f}$ (thus $s = 1 + \epsilon|\mathcal{Y}|$) and get

$$\begin{aligned}
\text{(iv)} =& n\epsilon|\mathcal{Y}| - 2\sum_{i=1}^{n} \log \mathop{\mathbb{E}}_{x,y\sim f^\star(x,\cdot)} \exp\left(-\frac{1}{2}\log\left(f^\star(x,y)/\tilde{f}(x,y)\right)\right) \\
=& n\epsilon|\mathcal{Y}| - 2\sum_{i=1}^{n} \log \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_i} \exp\left(-\frac{1}{2}\log\left(f^\star(x,y)/\tilde{f}(x,y)\right)\right) \\
=& n\epsilon|\mathcal{Y}| - 2\log \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}'} \left[\exp\left(\sum_{i=1}^{n} -\frac{1}{2}\log\left(f^\star(x,y)/\tilde{f}(x,y)\right)\right)\middle| D\right] \\
=& n\epsilon|\mathcal{Y}| + 2\cdot\text{(i)}.
\end{aligned}$$

By plugging (iii) and (iv) back we get

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} H^2\left(f(x,\cdot), f^\star(x,\cdot)\right) \le 6n\epsilon|\mathcal{Y}| + 4\cdot\text{(i)}.$$

Notice that (i) $\le$ (ii), so we complete the proof by plugging (ii) into the above. □

We first state the MLE generalization result for finite $\mathcal{F}$.

Theorem E.4. *Suppose $\mathcal{F}$ is finite. Fix any $\delta \in (0,1)$, set $\beta = \log(|\mathcal{F}|/\delta)$ and define*

$$\widehat{\mathcal{F}} = \left\{f \in \mathcal{F} : \sum_{i=1}^{n} \log f(x_i,y_i) \ge \max_{\tilde{f}\in\mathcal{F}} \sum_{i=1}^{n} \tilde{f}(x_i,y_i) - 4\beta\right\}.$$

Then w.p. at least $1-\delta$, the following holds:

(1) The true distribution is in the version space, i.e., $f^\star \in \widehat{\mathcal{F}}$.

(2) Any function in the version space is close to the ground truth data-generating distribution, i.e., for all $f \in \widehat{\mathcal{F}}$

$$\sum_{i=1}^{n} \mathbb{E}_{x\sim\mathcal{D}_i}\left[H^2(f(x,\cdot) \,\|\, f^\star(x,\cdot))\right] \le 22\beta.$$

Proof. These two claims follow from Lemma E.3 with $\epsilon = 0$, and so $N_{[]}(\epsilon,\mathcal{F},\|\cdot\|_\infty) = |\mathcal{F}|$. For (1), apply Eq. (3) to $f = \widehat{f}_{\text{MLE}}$ to see that $f^\star \in \widehat{\mathcal{F}}$. For (2), apply Eq. (2) and note that the sum term is at most 4β. Thus, the right hand side of Eq. (2) is at most $(6+8+8)\beta = 22\beta$. □

We now state the result for infinite $\mathcal{F}$ using bracketing entropy.

Theorem E.5. *Fix any $\delta \in (0,1)$, set $\beta = \log(N_{[]}((n|\mathcal{Y}|)^{-1}, \mathcal{F}, \|\cdot\|_\infty)/\delta)$ and define*

$$\widehat{\mathcal{F}} = \left\{ f \in \mathcal{F} : \sum_{i=1}^{n} \log f(x_i, y_i) \geq \max_{\widetilde{f} \in \mathcal{F}} \sum_{i=1}^{n} \widetilde{f}(x_i, y_i) - 7\beta \right\}.$$

Then w.p. at least $1-\delta$, the following holds:

(1) The true distribution is in the version space, i.e., $f^\star \in \widehat{\mathcal{F}}$.

(2) Any function in the version space is close to the ground truth data-generating distribution, i.e., for all $f \in \widehat{\mathcal{F}}$

$$\sum_{i=1}^{n} \mathbb{E}_{x \sim \mathcal{D}_i}\left[H^2(f(x,\cdot) \,\|\, f^\star(x,\cdot))\right] \leq 28\beta.$$

Proof. These two claims follow from Lemma E.3 with $\epsilon = 1/n|\mathcal{Y}|$. For (1), apply Eq. (3) to $f = \widehat{f}_{\mathrm{MLE}}$ to see that $f^\star \in \widehat{\mathcal{F}}$. For (2), apply Eq. (2) and note that the sum term is at most 7β. Thus, the right hand side of Eq. (3) is at most $(6+14+8)\beta = 28\beta$. □

F Confidence set construction with general function class

In this section, we extend the confidence set construction of O-DISCO and P-DISCO to general $\mathcal{F}$, which can be infinite. Our procedure constructs the confidence set by performing the thresholding scheme on an ε-net of $\mathcal{F}$. While constructing an ε-net for $\mathcal{F}$ is admittedly a computationally hard procedure, this is still information theoretically possible and our focus in O-DISCO and P-DISCO is to show that distributional RL information-theoretically leads to small-loss bounds.

We first define some notations. Let $\mathcal{F}^\downarrow$ and $\mathcal{F}^\uparrow$ denote a lower and upper ε-bracketing of $\mathcal{F}$, *i.e.*, for any $f \in \mathcal{F}$, there exists an ε-bracket $[f^\downarrow, f^\uparrow]$ such that for all h, $f_h^\downarrow \leq f_h \leq f_h^\uparrow$ with $f^\downarrow \in \mathcal{F}^\downarrow, f^\uparrow \in \mathcal{F}^\uparrow$. Recall that a lower bracket $g \in \mathcal{F}^\downarrow$ may not be a valid distribution, but since elements of $\mathcal{F}$ map to non-negative values, we can assume g has non-negative entires as well. Also, we have $\alpha_h^g(x,a) := \int g_h(z \mid x,a) \geq 1-\varepsilon$, so for ε small enough, g is normalizable. Hence, define $\widetilde{g}(z \mid x,a) = \alpha_h^g(x,a)^{-1} g(z \mid x,a)$ as the normalized version, which is a valid distribution that we can sample from.

Now, consider any martingale $\{x_{h,i}, a_{h,i}, c_{h,i}\}_{i\in[n],h\in[H]}$, which could be the online data up to episode k or the offline data (consisting of N i.i.d. samples). We define the MLE with respect to a lower bracket element as follows. For any $h \in [H], g \in \mathcal{F}^\downarrow, \pi \in \Pi$, sample $y_{h,i}^{g,\pi} \sim \widetilde{g}_{h+1}(x'_{h,i}, \pi(x'_{h,i}))$, and $z_{h,i}^{g,\pi} = c_{h,i} + y_{h,i}^{g,\pi}$, define the MLE solution for (g,π) at time h as,

$$\textsc{Mle}_h^{g,\pi} = \arg\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f_h(z_{h,i}^{g,\pi} \mid x_{h,i}, a_{h,i}).$$

Also, define the version space with respect to the above MLE as,

$$\mathcal{F}_{g,\pi,h} = \left\{ f \in \mathcal{F} : \sum_{i=1}^{n} \log f_h(z_{h,i}^{g,\pi} \mid x_{h,i}, a_{h,i}) \geq \sum_{i=1}^{n} \log \textsc{Mle}_h^{g,\pi}(z_{h,i}^{g,\pi} \mid x_{h,i}, a_{h,i}) - \beta \right\}.$$

We now prove a key result that implies that $\mathcal{T}_h^\pi f_{h+1}^\downarrow$ falls into the confidence set $\mathcal{F}_{f^\downarrow,\pi,h}$.

Theorem F.1. *For any $\delta \in (0,1)$ and suppose $n \geq 2$. Then, w.p. at least $1-\delta$, for any $h \in [H], g \in \mathcal{F}, f^\downarrow \in \mathcal{F}^\downarrow, \pi \in \Pi$, we have*

$$\sum_{i=1}^{n} \log g_h(z_{h,i}^{f^\downarrow,\pi} \mid x_{h,i}, a_{h,i}) - \log \mathcal{T}_h^\pi f_{h+1}^\downarrow(z_{h,i}^{f^\downarrow,\pi} \mid x_{h,i}, a_{h,i}) \leq \log(e^4 N_{[]}(n^{-1}, \mathcal{F}, \|\cdot\|_\infty)^2 |\Pi|/\delta).$$

where $z_{h,i}^{f^\downarrow,\pi} = c_{h,i} + y_{h,i}^{f^\downarrow,\pi}$ and $y_{h,i}^{f^\downarrow,\pi} \sim \widetilde{f}_{h+1}^\downarrow(\cdot \mid x'_{h,i}, \pi_{h+1}(x'_{h,i}))$.

Proof of Theorem F.1. Consider a ε-bracketing of $\mathcal{F}$ where $\varepsilon \le 1/n \le 1/2$; we will study each element and conclude with a union bound. For any lower bracket l and upper bracket u in the bracketing (note l, u need not correspond to the same bracket). Recall that $\alpha^l_{h+1}(x,a) := \int l_{h+1}(z \mid x,a)$, so we have $1-\varepsilon \le \alpha^l_{h+1} \le 1$ since l is a lower ε-bracket of distributions. Therefore, we have

$$\mathbb{E}\left[\exp\sum_{i=1}^n \log\left(\frac{u_h(z^{l,\pi}_{h,i} \mid x_{h,i},a_{h,i})}{\mathcal{T}^\pi_h l_{h+1}(z^{l,\pi}_{h,i} \mid x_{h,i},a_{h,i})}\right)\right] = \prod_{i=1}^n \mathbb{E}_{\nu_{h,i}}\left[\frac{u_h(z^{l,\pi}_{h,i} \mid x_{h,i},a_{h,i})}{\mathcal{T}^\pi_h l_{h+1}(z^{l,\pi}_{h,i} \mid x_{h,i},a_{h,i})}\right],$$

where $\nu_{h,i}$ is the distribution of data from i-th round and time h. Note that $\nu_{h,i}(x,a,c,x') = d_{h,i}(x,a)C_h(c \mid x,a)P_h(x' \mid x,a)$ for some distribution $d_{h,i}(x,a)$. Now focus on each i, so for all i, we have

$$\begin{aligned}
&\mathbb{E}_{\nu_{h,i}}\left[\frac{u_h(z^{l,\pi}_{h,i} \mid x_{h,i},a_{h,i})}{\mathcal{T}^\pi_h l_{h+1}(z^{l,\pi}_{h,i} \mid x_{h,i},a_{h,i})}\right] \\
&= \int_{x,a,c,x',y} \nu_{h,i}(x,a,c,x')\widetilde{l}_{h+1}(y \mid x',\pi(x'))\frac{u_h(c+y \mid x,a)}{\int_{c,x'}\nu_{h,i}(c,x' \mid x,a)l_{h+1}(y \mid x',\pi(x'))} \\
&= \int_{x,a,z} d_{h,i}(x,a)\int_z u_h(z \mid x,a) \\
&\times \int_{c,x'} \nu_{h,i}(c,x' \mid x,a)\widetilde{l}_{h+1}(z-c \mid x',\pi(x'))\frac{1}{\int_{c,x'}\nu_{h,i}(c,x' \mid x,a)l_{h+1}(z-c \mid x',\pi(x'))} \\
&= \int_{x,a,z} d_{h,i}(x,a)\int_z u_h(z \mid x,a)\alpha^l_{h+1}(x,a)^{-1} \\
&\le \frac{1+\varepsilon}{1-\varepsilon} = 1 + \frac{2\varepsilon}{1-\varepsilon} \le 1 + \frac{4}{n}.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\exp\sum_{i=1}^n \log\left(\frac{u_h(z^{l,\pi}_{h,i} \mid x_{h,i},a_{h,i})}{\mathcal{T}^\pi_h l_{h+1}(z^{l,\pi}_{h,i} \mid x_{h,i},a_{h,i})}\right)\right] \le (1+4/n)^n \le e^4.$$

Thus, by Markov's inequality, w.p. at least $1-\delta$, we have

$$\sum_{i=1}^n \log\left(\frac{u_h(z^{l,\pi}_{h,i} \mid x_{h,i},a_{h,i})}{\mathcal{T}^\pi_h l_{h+1}(z^{l,\pi}_{h,i} \mid x_{h,i},a_{h,i})}\right) \le \ln(e^4/\delta).$$

To conclude, apply union bound to get this result for all brackets. □

For the remainder of this section, we assume the policy class Π is finite. However, it is possible to extend our results using policy covers in the Hamming distance; in that case, $\log|\Pi|$ would be replaced by the log covering number or entropy integral of Π [as in Zhou et al., 2023, Kallus et al., 2022]. We note that for the *online* case, we rely on the assumption that for any $f \in \mathcal{F}$ we have $\pi^f \in \Pi$, where recall that $\pi^f_h(x) = \arg\min_a \bar{f}_h(x,a)$. This is because $\mathcal{T}^{\star,D}$ is not a contraction so we cannot operate with $\mathcal{T}^{\star,D}$ directly and instead operate with $\mathcal{T}^{\pi^f,D}$. We highlight that this assumption is automatically satisfied in tabular MDPs, since the whole policy space is finite, and $\log|\Pi| = \mathcal{O}(X\log(A))$ is lower order compared to log of the bracketing entropy of $\mathcal{F}_{tab}$, which is $\mathcal{O}(X^2A^2)$. In contrast, in non-distributional methods such as GOLF, the regular Bellman optimality operator is a contraction so standard Lipschitz arguments for covering go through. We note that it is also possible to construct covers of $\mathcal{F}$ in the Hellinger distance, but the metric entropy of $\mathcal{F}_{tab}$ seems to be on the same order as its bracketing entropy.

We now describe the version space construction for general $\mathcal{F}$, first for the online setting. Fix any k, and define the set

$$\mathcal{F}_{f^\downarrow,\pi,h} = \left\{f \in \mathcal{F} : \sum_{i=1}^k \log f_h(z^{f^\downarrow,\pi}_{h,i} \mid x_{h,i},a_{h,i}) \ge \sum_{i=1}^k \log \textsc{Mle}^{f^\downarrow,\pi}_h(z^{f^\downarrow,\pi}_{h,i} \mid x_{h,i},a_{h,i}) - \beta\right\}$$

Then, construct the version space as

$$\mathcal{F}_k = \left\{f \in \mathcal{F} : f_h \in \mathcal{F}_{f^\downarrow,\pi^f,h}, \forall h \in [H]\right\}.$$

Theorem F.2. *Fix any $\delta \in (0,1)$ and suppose Assumption 5.1. Set $\beta = \log(KH \cdot N_{[]}(K^{-1}, \mathcal{F}, \|\cdot\|_\infty)|\Pi|/\delta)$. Then, w.p. at least $1-\delta$, the following holds:*

(1) The optimal cost distribution is in the version space, i.e., $Z^\star \in \mathcal{F}_k$.

(2) For all $f \in \mathcal{F}_k$ and $h \in [H]$,

$$\sum_{i=1}^{k} \mathbb{E}_{\pi^i}\left[H^2(f_h(x_h,a_h) \,\|\, \mathcal{T}_h^{\star,D} f_{h+1}(x_h,a_h))\right] \le 60\beta.$$

Proof. First, we want to verify that $Z^\star \in \mathcal{F}_k$. Let $f^\downarrow$ be the lower bracket of $Z^\star$ and set $g = \textsc{Mle}_h^{f^\downarrow,\pi^\star} \in \mathcal{F}$; note $\pi^\star = \pi^{Z^\star}$. By Theorem F.1, we have $\sum_{i=1}^k \log \textsc{Mle}_h^{f^\downarrow,\pi^\star}(z_{h,i}^{f^\downarrow,\pi^\star} \mid x_{h,i},a_{h,i}) - \log \mathcal{T}_h^{\pi^\star,D} f_{h+1}^\downarrow(z_{h,i}^{f^\downarrow,\pi^\star} \mid x_{h,i},a_{h,i}) \le \mathcal{O}(\beta)$. Therefore, noting that $Z_h^\star = \mathcal{T}_h^{\pi^\star,D} Z_{h+1}^\star \ge \mathcal{T}_h^{\pi^\star,D} f_{h+1}^\downarrow$ shows that $Z_h^\star \in \mathcal{F}_{f^\downarrow,\pi^\star,h}$ for every h, implying that $Z^\star \in \mathcal{F}_k$.

For the second claim, fix any $f \in \mathcal{F}_k$ and $h \in [H]$. Then,

$$\begin{aligned}
&\sum_{i=1}^{k} \mathbb{E}_{\pi^i}\left[H^2(f_h(x_h,a_h) \,\|\, \mathcal{T}_h^{\star,D} f_{h+1}(x_h,a_h))\right] \\
&= \sum_{i=1}^{k} \mathbb{E}_{\pi^i}\left[H^2(f_h(x_h,a_h) \,\|\, \mathcal{T}_h^{\pi^f,D} f_{h+1}(x_h,a_h))\right] \\
&\le 2\sum_{i=1}^{k} \mathbb{E}_{\pi^i}\left[H^2(f_h(x_h,a_h) \,\|\, \mathcal{T}_h^{\pi^f,D} \widetilde{f}_{h+1}^\downarrow(x_h,a_h)) + H^2(\mathcal{T}_h^{\pi^f,D} \widetilde{f}_{h+1}^\downarrow(x_h,a_h) \,\|\, \mathcal{T}_h^{\pi^f,D} f_{h+1}(x_h,a_h))\right] \\
&\le 2(28\beta + 3k\varepsilon).
\end{aligned}$$

The β comes from Theorem E.5, and for ε, we used the fact that $H^2 \le H \le TV$, and

$$\begin{aligned}
&\sum_{i=1}^{k} \mathbb{E}_{\pi^i}\left[TV(\mathcal{T}_h^{\pi^f,D} \widetilde{f}_{h+1}^\downarrow(x_h,a_h) \,\|\, \mathcal{T}_h^{\pi^f,D} f_{h+1}(x_h,a_h))\right] \\
&= \sum_{i=1}^{k} \mathbb{E}_{\pi^i} \int_z \left|\mathcal{T}_h^{\pi^f,D} \widetilde{f}_{h+1}^\downarrow(z \mid x_h,a_h) - \mathcal{T}_h^{\pi^f,D} f_{h+1}(z \mid x_h,a_h))\right| \\
&= \sum_{i=1}^{k} \mathbb{E}_{\pi^i} \int_z \sum_{c,x'} \nu(c,x' \mid x_h,a_h) \left|\widetilde{f}_{h+1}^\downarrow(z-c \mid x', \pi^f(x')) - f_{h+1}(z-c \mid x', \pi^f(x'))\right| \\
&\le \sum_{i=1}^{k} 3\varepsilon = 3k\varepsilon,
\end{aligned}$$

since for any x,a, we have $\int_z \left|\widetilde{f}_{h+1}^\downarrow(z \mid x,a) - f_{h+1}(z \mid x,a)\right| \le 3\varepsilon$. There are two cases. If $\widetilde{f}_{h+1}^\downarrow(z \mid x,a) \ge f_{h+1}(z \mid x,a)$, then $\widetilde{f}_{h+1}^\downarrow(z \mid x,a) - f_{h+1}(z \mid x,a) \le (1-\varepsilon)^{-1} f_{h+1}^\downarrow(z \mid x,a) - f_{h+1}(z \mid x,a) \le 2\varepsilon f_{h+1}(z \mid x,a)$ since $(1-\varepsilon)^{-1} \le 1+2\varepsilon$. If $\widetilde{f}_{h+1}^\downarrow(z \mid x,a) < f_{h+1}(z \mid x,a)$, then $f_{h+1}(z \mid x,a) - \widetilde{f}_{h+1}^\downarrow(z \mid x,a) \le f_{h+1}(z \mid x,a) - f_{h+1}^\downarrow(z \mid x,a) \le \varepsilon$. Thus, $\int_z \max(2\varepsilon f_{h+1}(z \mid x,a), \varepsilon) \le \int_z 2\varepsilon f_{h+1}(z \mid x,a) + \varepsilon = 3\varepsilon$. Thus, setting $\varepsilon = 1/K$ gives

$$\sum_{i=1}^{k} \mathbb{E}_{\pi^i}\left[H^2(f_h(x_h,a_h) \,\|\, \mathcal{T}_h^{\star,D} f_{h+1}(x_h,a_h))\right] \le 59\beta.$$

□

For the offline setting, fix any π and define its general version space as,

$$\mathcal{F}_\pi = \left\{f \in \mathcal{F} : f_h \in \mathcal{F}_{f^\downarrow,\pi,h}, \forall h \in [H]\right\}.$$

Theorem F.3. *Fix any $\delta \in (0,1)$ and suppose Assumption 5.1. Set $\beta = \log(H|\Pi| \cdot N_{[]}((n|\mathcal{Y}|)^{-1}, \mathcal{F}, \|\cdot\|_\infty)/\delta)$. Then, w.p. at least $1-\delta$, the following holds for all policies $\pi \in \Pi$:*

(1) The policy cost distribution is in the version space, i.e., $Z^\pi \in \mathcal{F}_\pi$.

(2) Any function in the version space has bounded triangular discrimination with the ground truth data-generating distribution, i.e., for all $f \in \mathcal{F}_\pi$ and $h \in [H]$,

$$\mathbb{E}_{\nu_h}\left[H^2(f_h(x_h,a_h) \,\|\, \mathcal{T}_h^{\pi,D} f_{h+1}(x_h,a_h))\right] \le 60\beta N^{-1}.$$

Proof. The proof is the same as in Theorem F.2, but instead of π^f, we fix any π. □

G Proofs for Online RL

G.1 Preliminary Lemmas

Lemma G.1. *For any policy π, conditional distribution d and $h \in [H]$, we have*

$$\overline{\mathcal{T}_h^{\pi,D} d(x,a)} = \mathcal{T}_h^{\pi} \bar{d}(x,a),$$
$$\overline{\mathcal{T}_h^{\star,D} d(x,a)} = \mathcal{T}_h^{\star} \bar{d}(x,a).$$

Proof.

$$\begin{aligned}
\overline{\mathcal{T}_h^{\pi,D} d(x,a)} &= \mathbb{E}_{y \sim \mathcal{T}_h^{\pi,D} d(x,a)}[y] \\
&= \mathbb{E}_{c \sim C_h(x,a), x' \sim P_h(x,a), a' \sim \pi_{h+1}(x'), y' \sim d(x',a')}[c + y'] \\
&= \bar{C}_h(x,a) + \mathbb{E}_{x' \sim P_h(x,a), a' \sim \pi_{h+1}(x'), y' \sim d(x',a')}[y'] \\
&= \bar{C}_h(x,a) + \mathbb{E}_{x' \sim P_h(x,a), a' \sim \pi_{h+1}(x')}\left[\bar{d}(x',a')\right] \\
&= \mathcal{T}_h^{\pi} \bar{d}(x,a).
\end{aligned}$$

$$\begin{aligned}
\overline{\mathcal{T}_h^{\star,D} d(x,a)} &= \mathbb{E}_{y \sim \mathcal{T}_h^{\star,D} d(x,a)}[y] \\
&= \mathbb{E}_{c \sim C_h(x,a), x' \sim P_h(x,a), a' = \arg\min_{\widetilde{a}} \bar{d}(x',\widetilde{a}), y' \sim d(x',a')}[c + y'] \\
&= \bar{C}_h(x,a) + \mathbb{E}_{x' \sim P_h(x,a), a' = \arg\min_{\widetilde{a}} \bar{d}(x',\widetilde{a}), y' \sim d(x',a')}[y'] \\
&= \bar{C}_h(x,a) + \mathbb{E}_{x' \sim P_h(x,a), a' = \arg\min_{\widetilde{a}} \bar{d}(x',\widetilde{a})}\left[\bar{d}(x',a')\right] \\
&= \bar{C}_h(x,a) + \mathbb{E}_{x' \sim P_h(x,a)}\left[\min_{a'} \bar{d}(x',a')\right] \\
&= \mathcal{T}_h^{\star} \bar{d}(x,a).
\end{aligned}$$

□

Lemma G.2 (Performance Difference Lemma (PDL))**.** *For any $f : (\mathcal{X} \times \mathcal{A} \to \mathbb{R})^H$ and policies π, π', we have*

$$V^{\pi} - \mathbb{E}_{a \sim \pi'(x_1)}[f_1(x_1,a)] = \sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\mathcal{T}_h^{\pi'} f_{h+1}(x_h,a_h) - f_h(x_h,\pi')\right]. \tag{5}$$

Proof. We proceed by inducting on the following claim: for all $h = H+1, H, \ldots, 1$,

$$V_h^{\pi}(x_h) - f_h(x_h,\pi') = \sum_{t=h}^{H} \mathbb{E}_{\pi,x_h}\left[\mathcal{T}_t^{\pi'} f_{t+1}(x_t,a_t) - f_t(x_t,\pi')\right].$$

The base case of $H+1$ is trivially true as everything is 0. Now fix any h and suppose the IH at $h+1$ is true. Then

$$\begin{aligned}
&V_h^{\pi}(x_h) - f_h(x_h,\pi') \\
&= \mathbb{E}_{\pi,x_h}\left[c_h + V_{h+1}^{\pi}(x_{h+1}) - f_{h+1}(x_{h+1},\pi') + f_{h+1}(x_{h+1},\pi') - f_h(x_h,\pi')\right] \\
&= \mathbb{E}_{\pi,x_h}\left[V_{h+1}^{\pi}(x_{h+1}) - f_{h+1}(x_{h+1},\pi')\right] + \mathbb{E}_{\pi,x_h}[c_h + f_{h+1}(x_{h+1},\pi') - f_h(x_h,\pi')].
\end{aligned}$$

By the IH, the first term is equal to $\sum_{t=h+1}^{H} \mathbb{E}_{\pi,x_h}\left[\mathcal{T}_t^{\pi'} f_{t+1}(x_t,a_t) - f_t(x_t,\pi')\right]$. The second term is exactly $\mathbb{E}_{\pi,x_h}\left[\mathcal{T}_h^{\pi'} f_{h+1}(x_h,a_h) - f_h(x_h,\pi')\right]$, which concludes the proof. □

G.2 General Regret and PAC Bounds

For our analysis, we define a complexity measure inspired by the Sequential Extrapolation Coefficient (SEC) of Xie et al. [2023]. The SEC measures how well a function can be extrapolated on the k-th episode, using data from the first $k-1$ episodes, and has interesting connections to the coverability of the MDP. Recall the definition of SEC for function class Ψ, distribution class $\mathcal{D}$, both indexed by h, and number of episodes K:

$$\mathrm{SEC}(\Psi, \mathcal{D}, K) = \max_{\forall k: f^{(k)} \in \Psi, d^{(k)} \in \mathcal{D}} \sum_{k=1}^{K} \frac{\left(\mathbb{E}_{d^{(k)}}\left[f^{(k)}(z)\right]\right)^2}{1 \vee \sum_{i<k} \mathbb{E}_{d^{(i)}}\left[f^{(k)}(z)^2\right]}.$$

Xie et al. [2023] showed that the regret of standard (non-distributional) GOLF can be captured by the SEC. However, for our distributional algorithm, we need to define a slightly different term, which we call the *Linear SEC* (LSEC):

$$\mathrm{LSEC}(\Psi, \mathcal{D}, K) := \max_{\forall k: f^{(k)} \in \Psi, d^{(k)} \in \mathcal{D}} \sum_{k=1}^{K} \frac{\mathbb{E}_{d^{(k)}}\left[f^{(k)}(z)\right]}{1 \vee \sum_{i<k} \mathbb{E}_{d^{(i)}}\left[f^{(k)}(z)\right]}. \tag{6}$$

The difference with the SEC is that our quantity does not have squares, hence we call it "linear". By Jensen's inequality, we have $\mathrm{SEC}(\{f^2 : f \in \Psi\}, \mathcal{D}, K) \le \mathrm{LSEC}(\Psi, \mathcal{D}, K)$, which shows that our LSEC is in general a larger quantity. Nonetheless, we will show that it is controlled for tabular MDPs. For our regret bound, the function class and distribution class are instantiated as, for each h,

$$\begin{aligned} \mathcal{D}_h(\Pi) &= \{z \mapsto d^\pi(z) : \pi \in \Pi\} \\ \Psi_h &= \left\{z \mapsto D_\triangle(f(z) \,\|\, \mathcal{T}^{\star,D} f(z)) : f \in \mathcal{F}\right\}, \end{aligned} \tag{7}$$

where $z = (s, a)$. So let us denote $\mathrm{LSEC}(K) = \max_h \mathrm{LSEC}(\Psi_h, \mathcal{D}_h(\Pi), K)$. This quantity will appear in our small-loss regret bounds.

We can also define V-type analogs of LSEC, which we will use for obtaining small-loss PAC bounds for latent variable models. The key difference in the V-type LSEC is that the distributions in $\mathcal{D}_h(\Pi)$ are in the form $d^\pi(s) \cdot \mathrm{unif}(a)$,*i.e.*,

$$\begin{aligned} \mathcal{D}_{h,v}(\Pi) &= \{(s,a) \mapsto d^\pi(s)/A : \pi \in \Pi\} \\ \mathrm{LSEC}_v(\Psi, \mathcal{D}, K) &= \max_h \mathrm{LSEC}(\Psi_h, \mathcal{D}_{h,v}(\Pi), K). \end{aligned} \tag{8}$$

We now prove the our main regret bound.

Theorem G.3. *Assume Assumption 5.1. Fix any $\delta \in (0,1)$ and set $\beta = \log(HK|\mathcal{F}|/\delta)$ and $\beta' = 60\beta$. Then, w.p. at least $1-\delta$, running* O-DISCO *(Algorithm 4) with* UAE $=$ FALSE *yields the following small-loss regret bound,*

$$\mathrm{Regret}_{\text{O-DISCO}}(K) \le 5H\sqrt{KV^\star \,\mathrm{LSEC}(K)\beta'} + 18H^2\,\mathrm{LSEC}(K)\beta'.$$

If instead UAE $=$ TRUE*, the outputted policy $\bar{\pi}$ enjoys the following small-loss PAC bound,*

$$V^{\bar{\pi}} - V^\star \le 5H\sqrt{\frac{AV^\star\,\mathrm{LSEC}_v(K)\beta'}{K}} + 18H^2\frac{A\,\mathrm{LSEC}_v(K)\beta'}{K}.$$

Proof. We first prove the regret bound (UAE $=$ FALSE); the PAC bound follows from the same argument. For shorthand, let $\delta_{h,k}(x,a) := D_\triangle(f_h^{(k)}(x,a) \,\|\, \mathcal{T}_h^{\star,D} f_{h+1}^{(k)}(x,a))$ and $\Delta_k := \sum_{h=1}^H \mathbb{E}_{\pi^k}[\delta_{h,k}(x_h,a_h)]$. Notice that since $\pi_{h+1}^k(x) = \arg\min_a \bar{f}_{h+1}^{(k)}(x,a)$, we have $\mathcal{T}_h^{\pi^k,D} f_{h+1}^{(k)}(x,a) = \mathcal{T}_h^{\star,D} f_{h+1}^{(k)}(x,a)$, so $\delta_{h,k}(x,a) = D_\triangle(f_h^{(k)}(x,a) \,\|\, \mathcal{T}_h^{\pi^k,D} f_{h+1}^{(k)}(x,a))$ as well.

By Theorem F.2, we have the following two facts for all $k \in [K]$,

(i) Optimism: $\min_a \bar{f}_1^{(k)}(x_1,a) \le V^\star$ (since $Z^\star \in \mathcal{F}_k$) and

(ii) $\sum_{i<k} \mathbb{E}_{\pi^i}[\delta_{h,k}(s_h,a_h)] \le \beta'$ for all h. If UAE=TRUE, then a_h is sampled from $\mathrm{unif}(\mathcal{A})$ rather than π^i, *i.e.*, we have $\sum_{i<k} \mathbb{E}_{s_h \sim \pi^i, a_h \sim \mathrm{unif}(\mathcal{A})}[\delta_{h,k}(s_h,a_h)] \le \beta'$, where $\beta' \lesssim \beta$. Theorem F.2 and the fact that $D_\triangle \le 4H^2$ certifies that $\beta' = 240\beta$ is sufficient.

Now, fix any episode $k \in [K]$.

$$
\begin{aligned}
& V^{\pi^k} - V^\star \\
& \le V^{\pi^k} - \min_a \bar{f}_1^{(k)}(x_1, a) && \text{(Fact (i))} \\
& = \sum_{h=1}^{H} \mathbb{E}_{\pi^k}\left[\mathcal{T}_h^{\pi^k} \bar{f}_{h+1}^{(k)}(x_h, a_h) - \bar{f}_h^{(k)}(x_h, \pi_h^k(x_h))\right] && \text{(PDL Lemma G.2)} \\
& = \sum_{h=1}^{H} \mathbb{E}_{\pi^k}\left[\overline{\mathcal{T}_h^{\pi^k, D} f_{h+1}^{(k)}}(x_h, a_h) - \bar{f}_h^{(k)}(x_h, a_h)\right] && \text{(Lemma G.1)} \\
& \le \sum_{h=1}^{H} \sqrt{\mathbb{E}_{\pi^k}\left[4\bar{f}_h^{(k)}(x_h, a_h) + \delta_{h,k}(x_h, a_h)\right]} \cdot \sqrt{\mathbb{E}_{\pi^k}[\delta_{h,k}(x_h, a_h)]} && \text{(Eq. } (\triangle_2)\text{)} \\
& \le \sum_{h=1}^{H} \sqrt{4eV^{\pi^k} + 17H \sum_{t=h}^{H} \mathbb{E}_{\pi^k}[\delta_{t,k}(x_t, a_t)]} \cdot \sqrt{\mathbb{E}_{\pi^k}[\delta_{h,k}(x_h, a_h)]} \\
& && \text{(Lemma G.4 and } \mathbb{E}_\pi[Q_h^\pi(s_h, a_h)] \le V^\pi\text{)} \\
& \le \sqrt{4eV^{\pi^k} + 17H\Delta_k} \cdot \sqrt{H\Delta_k} && (\bigstar) \\
& \le \sqrt{4eHV^{\pi^k}\Delta_k} + 5H\Delta_k \\
& \le 2\sqrt{H}\eta^{-1}V^{\pi^k} + 2\sqrt{H}\eta\Delta_k + 5H\Delta_k.
\end{aligned}
$$

In $\bigstar$, we used Cauchy Schwartz. Setting $\eta = 4\sqrt{H}$ and rearranging, we have

$$V^{\pi^k} \le 2V^\star + 16H\Delta_k + 10H\Delta_k \le 2V^\star + 26H\Delta_k.$$

Plugging this into $\bigstar$, and noting $104e + 17 \le 300$, we have

$$V^{\pi^k} - V^\star \le \sqrt{8eV^\star + 300H\Delta_k}\sqrt{H\Delta_k}.$$

Thus, summing the instantaneous regrets over all episodes, we get

$$
\begin{aligned}
\sum_{k=1}^{K} V^{\pi^k} - V^\star &\le \sum_{k=1}^{K} \sqrt{8eV^\star + 300H\Delta_k}\sqrt{H\Delta_k} \\
&\le \sqrt{8eKV^\star + 300H\sum_k \Delta_k}\sqrt{H\sum_k \Delta_k} && \text{(Cauchy-Schwartz)} \\
&\le 5\sqrt{HKV^\star \sum_k \Delta_k} + 18H\sum_k \Delta_k.
\end{aligned}
$$

Finally it remains the bound the sum of Δ_k,

$$
\begin{aligned}
\sum_{k=1}^{K} \Delta_k &= \sum_{h=1}^{H}\sum_{k=1}^{K} \frac{\mathbb{E}_{\pi^k}[\delta_{h,k}(x_h, a_h)]}{1 \vee \sum_{i=1}^{k-1} \mathbb{E}_{\pi^i}[\delta_{h,k}(s_h, a_h)]} \cdot \left(1 \vee \sum_{i=1}^{k-1} \mathbb{E}_{\pi^i}[\delta_{h,k}(s_h, a_h)]\right) \\
&\le H \,\mathrm{LSEC}(K) \cdot \beta'. && \text{(Fact (ii))}
\end{aligned}
$$

If UAE=TRUE, we instead bound the sum of Δ_k using the V-type LSEC:

$$
\begin{aligned}
\sum_{k=1}^{K} \Delta_k &\le \sum_{h=1}^{H}\sum_{k=1}^{K} \frac{\mathbb{E}_{\pi^k}[\delta_{h,k}(x_h, a_h)]}{1 \vee \sum_{i=1}^{k-1} \mathbb{E}_{s_h \sim \pi^i, a_h \sim \mathrm{unif}(\mathcal{A})}[\delta_{h,k}(s_h, a_h)]} \cdot \left(1 \vee \sum_{i=1}^{k-1} \mathbb{E}_{s_h \sim \pi^i, a_h \sim \mathrm{unif}(\mathcal{A})}[\delta_{h,k}(s_h, a_h)]\right) \\
&\le AH \,\mathrm{LSEC}_v(K) \cdot \beta'. && \text{(Fact (ii))}
\end{aligned}
$$

This concludes the proof for both the regret and PAC bounds. $\square$

Lemma G.4 (Self-bounding lemma)**.** *Let $f \in \mathcal{F}$ and let π be any policy. Let us denote $\delta_h(x,a) := D_{\triangle}(f_h(x,a) \,\|\, \mathcal{T}_h^{\pi,D} f_{h+1}(x,a))$. Then, for all $h \in [H]$, for all x_h, a_h, we have*

$$\bar{f}_h(x_h,a_h) \le eQ_h^{\pi}(x_h,a_h) + 4H \sum_{t=h}^{H} \mathbb{E}_{\pi,x_h,a_h}[\delta_t(x_t,a_t)].$$

Proof. We prove the following refined subclaim inductively: for all $h \in [H]$, for all x_h, a_h, we have

$$\bar{f}_h(x_h,a_h) \le \sum_{t=h}^{H} \left(1+\frac{1}{H}\right)^{t-h} \mathbb{E}_{\pi,x_h,a_h}[\bar{c}_t(x_t,a_t) + 2H\delta_t(x_t,a_t)]. \tag{IH}$$

For $H+1$ this is trivially true. Now fix any h and suppose IH is true for $h+1$. By Eq. ($\triangle_2$), for any h, x_h, a_h, we have,

$$\begin{aligned}
\bar{f}_h(x_h,a_h) - \mathcal{T}_h^{\pi}\bar{f}_{h+1}(x_h,a_h) &\le \sqrt{4\mathcal{T}_h^{\pi}\bar{f}_{h+1}(x_h,a_h) + \delta_h(x_h,a_h)}\sqrt{\delta_h(x_h,a_h)} \\
&\le \sqrt{4\mathcal{T}_h^{\pi}\bar{f}_{h+1}(x_h,a_h)\delta_h(x_h,a_h)} + \delta_h(x_h,a_h) \\
&\le \frac{1}{H}\mathcal{T}_h^{\pi}\bar{f}_{h+1}(x_h,a_h) + (H+1)\delta_h(x_h,a_h). \qquad \text{(AM-GM)}
\end{aligned}$$

In particular, we have that

$$\begin{aligned}
&\bar{f}_h(x_h,a_h) \\
&\le \left(1+\frac{1}{H}\right)\mathcal{T}_h^{\pi}\bar{f}_{h+1}(x_h,a_h) + 2H\delta_h(x_h,a_h) \\
&= \left(1+\frac{1}{H}\right)\left(\bar{c}_h(x_h,a_h) + \mathbb{E}_{x_{h+1}\sim P_h^{\star}(x_h,a_h)}\left[\bar{f}_{h+1}(x_{h+1},\pi)\right]\right) + 2H\delta_h(x_h,a_h) \\
&\le \left(1+\frac{1}{H}\right)\left(\bar{c}_h(x_h,a_h) + \mathbb{E}_{x_{h+1}\sim P_h^{\star}(x_h,a_h)}\left[\sum_{t=h+1}^{H}\left(1+\frac{1}{H}\right)^{t-h-1}\mathbb{E}_{\pi,x_{h+1}}[\bar{c}_t(x_t,a_t) + 2H\delta_t(x_t,a_t)]\right]\right) \qquad \text{(IH)} \\
&\quad + 2H\delta_h(x_h,a_h),
\end{aligned}$$

which proves the inductive claim. Noting that $\sum_{t=1}^{H}(1+1/H)^t \le e$, we have proven the lemma. □

G.3 Bounding the LSEC

In this section, we show that the LSEC quantity is bounded for tabular MDPs and latent variable models. First, recall the notion of Coverability from Xie et al. [2023],

$$C_{\mathrm{Cov}} := \inf_{\mu} \max_{\pi} \max_{h,x,a} \frac{d_h^{\pi}(x,a)}{\mu_h(x,a)}.$$

Let $\mu^{\star}$ be the measure that realizes this infimum. C_{Cov} was shown to be equivalent to $\max_h \sum_{x,a} \sup_{\pi} d_h^{\pi}(x,a)$ by Xie et al. [2023, Lemma 3]. For example, in tabular MDPs with X states and A actions, we have $C_{\mathrm{Cov}} \le XA$, and in low-rank MDPs (and hence latent variable models) with rank d, we have $C_{\mathrm{Cov}} \le d$ [Huang et al., 2023, Proposition 3].

Bounding the LSEC in Tabular MDPs First, consider any function class Ψ and distribution class $\mathcal{D}$. For all k, let $f^{(k)} \in \Psi$ and $d^{(k)} \in \mathcal{D}$. Define $\widetilde{d}^{(k)} = \sum_{i<k} d^{(i)}$ and $\tau(z) := \min\left\{k \mid \widetilde{d}^{(k)}(z) \ge C_{\mathrm{Cov}}\mu^{\star}(z)\right\}$. Then, for any $f \in \Psi$ and $d \in \mathcal{D}$, we have

$$\begin{aligned}
&\sum_{k=1}^{K} \frac{\mathbb{E}_{d^{(k)}}\left[f^{(k)}\right]}{1 \vee \sum_{i<k}\mathbb{E}_{d^{(i)}}\left[f^{(k)}\right]} \\
&= \underbrace{\sum_{k=1}^{K} \frac{\mathbb{E}_{d^{(k)}}\left[f^{(k)}(z)\mathbb{I}\left[k < \tau(z)\right]\right]}{1 \vee \sum_{i<k}\mathbb{E}_{d^{(i)}}\left[f^{(k)}\right]}}_{\text{Term 1}} + \underbrace{\sum_{k=1}^{K} \frac{\mathbb{E}_{d^{(k)}}\left[f^{(k)}(z)\mathbb{I}\left[k \ge \tau(z)\right]\right]}{1 \vee \sum_{i<k}\mathbb{E}_{d^{(i)}}\left[f^{(k)}\right]}}_{\text{Term 2}}.
\end{aligned}$$

Focusing on Term 1, we have it is at most,

$$\sum_{k=1}^{K} \mathbb{E}_{d^{(k)}}\left[f^{(k)}(z)\mathbb{I}\left[k < \tau(z)\right]\right] \leq \sum_{k=1}^{K} \mathbb{E}_{d^{(k)}}\left[\mathbb{I}\left[k < \tau(z)\right]\right] \leq 2C_{\text{Cov}},$$

by the proof of Proposition 13 of Xie et al. [2023].

For Term 2, we need to specialize $\mathcal{D}$. If the MDP is tabular, we can set $\mathcal{D}$ as defined in Eq. (7). Then, for $z = (x, a)$,

$$\begin{aligned}
&\sum_{k=1}^{K} \frac{\mathbb{E}_{d^{(k)}}\left[f^{(k)}(z)\mathbb{I}\left[k \geq \tau(z)\right]\right]}{\sum_{i<k}\mathbb{E}_{d^{(i)}}\left[f^{(k)}\right]} \\
&= \sum_{k=1}^{K}\sum_{z} \frac{d^{(k)}(z)f^{(k)}(z)\mathbb{I}\left[k \geq \tau(z)\right]}{\sum_z \widetilde{d}^{(k)}(z)f^{(k)}(z)} \\
&\leq \sum_{k=1}^{K}\sum_{z} \frac{d^{(k)}(z)f^{(k)}(z)\mathbb{I}\left[k \geq \tau(z)\right]}{\widetilde{d}^{(k)}(z)f^{(k)}(z)} && \text{(terms are non-negative)} \\
&= \sum_{k=1}^{K}\sum_{z} \frac{d^{(k)}(z)\mathbb{I}\left[k \geq \tau(z)\right]}{\widetilde{d}^{(k)}(z)} \\
&\leq 2\sum_{z}\sum_{k=1}^{K} \frac{d^{(k)}(z)\mathbb{I}\left[k \geq \tau(z)\right]}{\widetilde{d}^{(k)}(z) + C_{\text{Cov}}\mu^{\star}(z)} \\
&\leq 2\sum_{z} 2\log(K+1) && \text{(Xie et al. [2023, Lemma 4])} \\
&= 4Z\log(K+1).
\end{aligned}$$

Since the MDP is tabular we have $Z = XA$. We have proven the following lemma,

Lemma G.5. *Suppose the MDP is tabular. Then, for any Ψ, K, we have*

$$\text{LSEC}(\Psi, \mathcal{D}_h(\Pi), K) \in \mathcal{O}(XA\log(K)).$$

Combining this with Theorem G.3 directly implies Theorem 5.2.

Bounding V-type LSEC in Latent Variable Models Now suppose the MDP is a latent variable model (LVM), *i.e.*, an MDP with small non-negative rank d Modi et al. [2021]. The sampling procedure for latent variable model is, start with a distribution over d latent states p_1, sample an unobserved latent state $s_1 \sim p_1$, observe $x_1 \sim o(s_1)$, take action $a_1 \sim \pi_1(s_1)$ and transition to the next distribution of latent states p_2. This process repeats H times. Note that the observation set $\mathcal{X}$ can be very large or infinite, so instead of having a bound that depends on X, we'd like to depend on the number of latent states S. To do so, we make a simple modification to our previous argument.

Set $\tau(s, a) = \min\left\{k \mid \widetilde{d}^{(k)}(s, a) \geq C_{\text{Cov}}\mu^{\star}(s, a)\right\}$, where we've abused notation to use s as input instead of x, denoting that we are considering distributions over latent states rather than observations. For any distribution, we have $d(x, a) = o(x \mid s)d(s, a)$ where s is the encoded latent state corresponding to x. Crucially, τ depends on s rather than x.

In this case, we can take $\mathcal{D}$ as the V-type distributions $\mathcal{D}_{h,v}(\Pi)$. So $d^{(k)}(s,a) = d^{\pi^k}(s)/A$ and we can bound Term 2 as follows,

$$
\begin{aligned}
&\sum_{k=1}^{K} \frac{\mathbb{E}_{d^{(k)}}\left[f^{(k)}(x,a)\mathbb{I}\left[k \geq \tau(s,a)\right]\right]}{\sum_{i<k}\mathbb{E}_{d^{(i)}}\left[f^{(k)}\right]} \\
&= \sum_{k=1}^{K}\sum_{s,a} \frac{d^{(k)}(s,a)\mathbb{E}_{x\sim o(s)}\left[f^{(k)}(x,a)\right]\mathbb{I}\left[k \geq \tau(s,a)\right]}{\sum_{s,a}\widetilde{d}^{(k)}(s,a)\mathbb{E}_{x\sim o(s)}\left[f^{(k)}(x,a)\right]} \\
&\leq \sum_{k=1}^{K}\sum_{s,a} \frac{d^{(k)}(s,a)\mathbb{E}_{x\sim o(s)}\left[f^{(k)}(x,a)\right]\mathbb{I}\left[k \geq \tau(s,a)\right]}{\widetilde{d}^{(k)}(s,a)\mathbb{E}_{x\sim o(s)}\left[f^{(k)}(x,a)\right]} && \text{(terms are non-negative)} \\
&= \sum_{k=1}^{K}\sum_{s,a} \frac{d^{(k)}(s,a)\mathbb{I}\left[k \geq \tau(s,a)\right]}{\widetilde{d}^{(k)}(s,a)} \\
&\leq 2\sum_{s,a}\sum_{k=1}^{K} \frac{d^{(k)}(s,a)\mathbb{I}\left[k \geq \tau(s,a)\right]}{\widetilde{d}^{(k)}(s,a) + C_{\text{Cov}}\mu^\star(s,a)} \\
&\leq 2\sum_{s,a} 2\log(K+1) && \text{(Xie et al. [2023, Lemma 4])} \\
&= 4SA\log(K+1).
\end{aligned}
$$

We highlight that this argument only works for the V-type LSEC, since the uniform action a does not depend on the observation generating process, $x \sim o(s)$, while the action from the Q-type LSEC does. This dependence in the Q-type LSEC is what prevents us from doing the decomposition in the first step. This is why uniform action exploration is needed for our theory to extend to latent variable models. Thus, we've shown the following lemma,

Lemma G.6. *Suppose the MDP is a latent variable model. Then, for any Ψ, K, we have*

$$\text{LSEC}_v(\Psi, \mathcal{D}_{h,v}(\Pi), K) \in \mathcal{O}(SA\log(K)).$$

Combining this with Theorem G.3 directly implies Theorem 5.4.

H Proofs for Offline RL

Theorem 6.1 (Small-Loss PAC bound for P-DISCO)**.** *Assume Assumption 5.1. Fix any $\delta \in (0,1)$ and set $\beta = \log(H|\Pi||\mathcal{F}|/\delta)$. Then, w.p. at least $1-\delta$,* P-DISCO *learns a policy $\widehat{\pi}$ such that for any comparator policy $\widetilde{\pi} \in \Pi$, we have*

$$V^{\widehat{\pi}} - V^{\widetilde{\pi}} \le 9H\sqrt{\frac{C^{\widetilde{\pi}} V^{\widetilde{\pi}} \beta}{N}} + \frac{30H^2 C^{\widetilde{\pi}} \beta}{N}.$$

Proof of Theorem 6.1. For shorthand, let $\delta_h^\pi(x,a) = D_\triangle(f_h^\pi(x,a) \,\|\, \mathcal{T}_h^{\pi,D} f_{h+1}^\pi(x,a))$ and $\Delta^\pi = \sum_{h=1}^H \mathbb{E}_\pi[\delta_h^\pi(x_h,a_h)]$. Also, let $f(x,\pi) = \mathbb{E}_{a\sim\pi(x)}[f(x,a)]$.

By Theorem F.3, we have the following two facts, for all $\pi \in \Pi$,
(i) Pessimism: $V^\pi \le \bar{f}_1^\pi(x_1,\pi)$ (since $Z^\pi \in \mathcal{F}_\pi$) for all $\pi \in \Pi$, and
(ii) $\mathbb{E}_{\nu_h}[\delta_h^\pi(x_h,a_h)] \le \beta' N^{-1}$ for all h where Theorem F.3 and the fact that $D_\triangle \le 4H^2$ certifies that $\beta' = 240\beta$ is sufficient.

With these two facts, we can bound the suboptimality of $\widehat{\pi}$ as follows:

$$\begin{aligned}
&V^{\widehat{\pi}} - V^{\widetilde{\pi}} \\
&\le \bar{f}_1^{\widehat{\pi}}(x_1,\widehat{\pi}) - V^{\widetilde{\pi}} && \text{(Fact (i))} \\
&\le \bar{f}_1^{\widetilde{\pi}}(x_1,\widetilde{\pi}) - V^{\widetilde{\pi}} && \text{(Policy selection scheme in Algorithm 3 (Line 4))} \\
&= \sum_{h=1}^H \mathbb{E}_{\widetilde{\pi}}\left[\bar{f}_h^{\widetilde{\pi}}(x_h,\widetilde{\pi}) - \mathcal{T}_h^{\widetilde{\pi}} \bar{f}_{h+1}^{\widetilde{\pi}}(x_h,a_h)\right] && \text{(PDL Lemma G.2)} \\
&\le \sum_{h=1}^H \sqrt{\mathbb{E}_{\widetilde{\pi}}\left[4\bar{f}_h^{\widetilde{\pi}}(x_h,a_h) + \delta_h^{\widetilde{\pi}}(x_h,a_h)\right]} \sqrt{\mathbb{E}_{\widetilde{\pi}}\left[\delta_h^{\widetilde{\pi}}(x_h,a_h)\right]} && \text{(Eq. } (\triangle_2)\text{)} \\
&\le \sum_{h=1}^H \sqrt{4eV^{\widetilde{\pi}} + 17H\sum_{t=h}^H \mathbb{E}_{\widetilde{\pi}}\left[\delta_t^{\widetilde{\pi}}(x_t,a_t)\right]} \sqrt{\mathbb{E}_{\widetilde{\pi}}\left[\delta_h^{\widetilde{\pi}}(x_h,a_h)\right]} && \text{(Lemma G.4)} \\
&\le \sqrt{4eV^{\widetilde{\pi}} + 17H\Delta^{\widetilde{\pi}}}\sqrt{H\Delta^{\widetilde{\pi}}} \\
&\le 4\sqrt{HV^{\widetilde{\pi}}\Delta^{\widetilde{\pi}}} + 5H\Delta^{\widetilde{\pi}}.
\end{aligned}$$

Finally, we can bound $\Delta^{\widetilde{\pi}}$ by a change of measure,

$$\begin{aligned}
\Delta^{\widetilde{\pi}} &= \sum_{h=1}^H \mathbb{E}_{\widetilde{\pi}}\left[\delta_h^{\widetilde{\pi}}(x_h,a_h)\right] \\
&\le C^{\widetilde{\pi}} \sum_{h=1}^H \mathbb{E}_{\nu_h}[\delta_h(x_h,a_h)] \\
&\le C^{\widetilde{\pi}} H \cdot \beta' N^{-1}. && \text{(Fact (ii))}
\end{aligned}$$

Therefore,

$$V^{\widehat{\pi}} - V^{\widetilde{\pi}} \le 4H\sqrt{\frac{C^{\widetilde{\pi}} V^{\widetilde{\pi}} \beta'}{N}} + \frac{5H^2 C^{\widetilde{\pi}} \beta'}{N}.$$

□

I Extension: Small-Return Bounds

In this section, we show that O-DISCO and P-DISCO can also be used to obtain small-return bounds. Compared to the algorithms presented in the main text for minimizing cost, we simply have to replace $\min$ with $\max$ (and vice versa) for maximizing reward, *i.e.*, see Appendix B and enable the SMALLRETURN flag. The proofs are also largely the same, with slight changes to the first few steps.

Theorem I.1. *Assume Assumption 5.1 and suppose we want to maximize returns (instead of minimize cost), so enable the* SMALLRETURN *flag. Fix any $\delta \in (0,1)$ and set $\beta = \log(HK|\mathcal{F}|/\delta)$ and $\beta' = 60\beta$. Then, w.p. at least $1-\delta$, running* O-DISCO *(Algorithm 4) with* UAE $=$ FALSE *yields the following small-loss regret bound,*

$$\mathrm{Regret}_{\text{O-DISCO}}(K) \le 5H\sqrt{KV^\star \,\mathrm{LSEC}(K)\beta'} + 18H^2\,\mathrm{LSEC}(K)\beta'. \tag{9}$$

If instead UAE $=$ TRUE*, the outputted policy $\bar{\pi}$ enjoys the following small-loss PAC bound,*

$$V^\star - V^{\bar{\pi}} \le 5H\sqrt{\frac{AV^\star\,\mathrm{LSEC}_v(K)\beta'}{K}} + 18H^2\frac{A\,\mathrm{LSEC}_v(K)\beta'}{K}.$$

Proof. Adopt the same notation as in the proof of Theorem G.3. By Theorem F.2, we have the following two facts for all $k \in [K]$,

(i) Optimism: $V^\star \le \max_a \bar{f}_1^{(k)}(x_1,a)$ (since $Z^\star \in \mathcal{F}_k$) and

(ii) $\sum_{i<k}\mathbb{E}_{\pi^i}[\delta_{h,k}(s_h,a_h)] \le \beta'$ for all h. If UAE=TRUE, then a_h is sampled from $\mathrm{unif}(\mathcal{A})$ rather than π^i, *i.e.*, we have $\sum_{i<k}\mathbb{E}_{s_h\sim\pi^i,a_h\sim\mathrm{unif}(\mathcal{A})}[\delta_{h,k}(s_h,a_h)] \le \beta'$, where $\beta' \lesssim \beta$. Theorem F.2 certifies that $\beta' = 60\beta$ is sufficient.

Fix any episode $k \in [K]$. Then,

$$\begin{aligned}
&V^\star - V^{\pi^k} \\
&\le \max_a \bar{f}_1^{(k)}(x_1,a) - V^{\pi^k} && \text{(Fact (i))} \\
&= \sum_{h=1}^H \mathbb{E}_{\pi^k}\left[\bar{f}_h^{(k)}(x_h,\pi_h^k(x_h)) - \mathcal{T}_h^{\pi^k}\bar{f}_{h+1}^{(k)}(x_h,a_h)\right] && \text{(PDL Lemma G.2)} \\
&= \sum_{h=1}^H \mathbb{E}_{\pi^k}\left[\bar{f}_h^{(k)}(x_h,a_h) - \overline{\mathcal{T}_h^{\pi^k,D} f_{h+1}^{(k)}}(x_h,a_h)\right] && \text{(Lemma G.1)} \\
&\le \sum_{h=1}^H \sqrt{\mathbb{E}_{\pi^k}\left[4\bar{f}_h^{(k)}(x_h,a_h) + \delta_{h,k}(x_h,a_h)\right]}\cdot\sqrt{\mathbb{E}_{\pi^k}[\delta_{h,k}(x_h,a_h)]} && \text{(Eq. } (\triangle_2)) \\
&\le \sum_{h=1}^H \sqrt{4eV^{\pi^k} + 17H\sum_{t=h}^H \mathbb{E}_{\pi^k}[\delta_{t,k}(x_t,a_t)]}\cdot\sqrt{\mathbb{E}_{\pi^k}[\delta_{h,k}(x_h,a_h)]} \\
& && \text{(Lemma G.4 and } \mathbb{E}_\pi[Q_h^\pi(s_h,a_h)] \le V^\pi) \\
&\le \sqrt{4eV^{\pi^k} + 17H\Delta_k}\cdot\sqrt{H\Delta_k} && (\clubsuit) \\
&\le \sqrt{4eV^\star + 17H\Delta_k}\cdot\sqrt{H\Delta_k}
\end{aligned}$$

Thus, summing the instantaneous regrets over all episodes, we get

$$\begin{aligned}
\sum_{k=1}^K V^{\pi^k} - V^\star &\le \sum_{k=1}^K \sqrt{4eV^\star + 17H\Delta_k}\sqrt{H\Delta_k} \\
&\le \sqrt{4eKV^\star + 17H\sum_k \Delta_k}\sqrt{H\sum_k \Delta_k} && \text{(Cauchy-Schwartz)} \\
&\le 5\sqrt{HKV^\star\sum_k \Delta_k} + 18H\sum_k \Delta_k.
\end{aligned}$$

The bounds for Δ_k are the same as in Theorem G.3. □

In some sense, the proof for the small-returns bound is actually easier than the small-loss bound. Recall that in the cost-minimizing setting, we needed to perform a crucial Cauchy-Schwartz step to rearrange terms at the step labelled ♣. However, in the reward-maximizing setting, we simply bound $V^{\pi^k} \le V^\star$, without needing to rearrange terms.

Theorem I.2. *Assume Assumption 5.1 and suppose we want to maximize returns (instead of minimize cost), so enable the* SMALLRETURN *flag. Fix any $\delta \in (0,1)$ and set $\beta = \log(H|\Pi||\mathcal{F}|/\delta)$. Then, w.p. at least $1-\delta$,* P-DISCO *(Algorithm 4) learns a policy $\widehat{\pi}$ such that for any comparator policy $\widetilde{\pi} \in \Pi$, we have*

$$V^{\widetilde{\pi}} - V^{\widehat{\pi}} \le 9H\sqrt{\frac{C^{\widetilde{\pi}} V^{\widetilde{\pi}} \beta}{N}} + \frac{30H^2 C^{\widetilde{\pi}} \beta}{N}.$$

Proof of Theorem I.2. Adopt the same notation as in the proof of Theorem 6.1. By Theorem F.3, we have the following two facts, for all $\pi \in \Pi$,
(i) Pessimism: $\bar{f}_1^{\pi}(x_1, \pi) \le V^{\pi}$ (since $Z^{\pi} \in \mathcal{F}_{\pi}$) for all $\pi \in \Pi$, and
(ii) $\mathbb{E}_{\nu_h}[\delta_h^{\pi}(x_h, a_h)] \le \beta' N^{-1}$ for all h where $\beta' \le 60\beta$.

With these two facts, we can bound the suboptimality of $\widehat{\pi}$ as follows:

$$\begin{aligned}
&V^{\widetilde{\pi}} - V^{\widehat{\pi}} \\
&\le V^{\widetilde{\pi}} - \bar{f}_1^{\widehat{\pi}}(x_1, \widehat{\pi}) && \text{(Fact (i))} \\
&\le V^{\widetilde{\pi}} - \bar{f}_1^{\widetilde{\pi}}(x_1, \widetilde{\pi}) && \text{(Policy selection rule in Line 5)} \\
&= \sum_{h=1}^{H} \mathbb{E}_{\widetilde{\pi}}\left[\mathcal{T}_h^{\widetilde{\pi}} \bar{f}_{h+1}^{\widetilde{\pi}}(x_h, a_h) - \bar{f}_h^{\widetilde{\pi}}(x_h, \widetilde{\pi})\right] && \text{(PDL Lemma G.2)} \\
&\le \sum_{h=1}^{H} \sqrt{\mathbb{E}_{\widetilde{\pi}}\left[4\bar{f}_h^{\widetilde{\pi}}(x_h, a_h) + \delta_h^{\widetilde{\pi}}(x_h, a_h)\right]} \sqrt{\mathbb{E}_{\widetilde{\pi}}\left[\delta_h^{\widetilde{\pi}}(x_h, a_h)\right]}. && \text{(Eq. } (\triangle_2)\text{)}
\end{aligned}$$

From here, the same argument in the proof of Theorem 6.1 finishes the proof. □

J Experiment Details

Experiment Settings

In our experiments, as outlined in Foster and Krishnamurthy [2021], our γ learning rate at each time step t is set to $\gamma_t = \gamma_0 t^p$ where γ_0 and p are hyperparameters. We use batch sizes of 32 samples per episode, and the King County and Prudential experiments run for $5,000$ episodes while the CIFAR-100 experiment runs for $15,000$.

For each dataset, we select the hyperparameter configuration with the best performance for each algorithm. As we report two metrics, performance over the last 100 episodes and over all episodes, we choose the best hyperparameters for each metric as well. While it is often the same hyperparameters that give the best last 100 episodes and all episodes results for a model, that is not always the case. We use the WandB (Weights and Biases) library to run sweeps over hyperparameters.

Oracles

For our regression oracles, we use ResNet18 [He et al., 2016], with a modified output layer (so that the output is suited for 100 prediction classes) for CIFAR-100, and a simple 2 hidden-layer neural network for the Prudential Life Insurance and King's County Housing datasets. For DistCB, the oracle's output layer has size AC where A is the number of actions and C is the number of potential costs. This is reshaped so that for each action, there are predictions associated with each potential cost, which then have a softmax function applied to them to represent cost probabilities. For SquareCB and FastCB, the output size is A because there is just a single prediction associated with each action. As per Foster and Krishnamurthy [2021], a sigmoid function is applied to this output layer. All experiments were implemented using PyTorch.

Datasets

We now provide an overview table as well as additional details and context to our setups for each dataset. Note that the number of items in each dataset in the table is the count after preprocessing.

Datasets			
Dataset	Items	Number of Actions	Number of Costs
CIFAR-100	$50,000$	100	3
Prudential Life Insurance	$59,381$	8	9
King County Housing	$20,148$	100	101

Table 3: Overview of the three datasets and their experimental setups

Prudential Life Insurance This dataset is from the Prudential Life Insurance Kaggle competition [Montoya et al., 2015]. It is featured in Farsang et al. [2022], which inspires our experimental setup. The risk level in $[8]$ directly determines the price charged to the customer. Thus, we can consider the chosen risk level as the action taken. If the model overpredicts the risk level, we get a cost of 1.0 because this is considered over charging the customer and not getting a sale. Otherwise, the model's prediction is charging too little for the customer. To reiterate, the cost in this case is $.1 * (y - \hat{y})$ where y is the actual risk level, and $\hat{y}$ is the predicted risk level.

King County Housing The King County housing dataset is also used in Farsang et al. [2022]. An interesting part of the setup is that the cost construction in the case of not overpredicting differs from the Prudential experiment, even though they're both effectively about predicting a price point. Here, the model's chosen price is considered the gain, which is why the cost is 1.0 minus the chosen price. On the other hand, in the Prudential experiment, the cost is a linear function of the difference between the chosen value and the actual value.

CIFAR-100 For the CIFAR-100 experiment, we use the training dataset of $50,000$ images as our dataset. The inclusion of the superclass is critical, as it lets us delineate 3 possible costs that DistCB can learn. Without the super class, the cost construction would be a pure binary of correct vs. incorrect. If this were the case, the ability to test the effectiveness of learning the distribution would be nullified. The distribution would just be whether an action is correct or not, which means our algorithm would essentially be predicting the mean directly.

Results

The largest advantages DISTCB had over the next best algorithm were in the Prudential experiment, with DISTCB having a $.086$ advantage over the last 100 episodes and a $.045$ advantage over all episodes. While the gaps were not as large for the other two datasets, they are still statistically significant and further showcase the benefit of distribution learning.