CBO + BFF applied to Q-control

Francesco Insulla, Yuhua Zhu, Lexing Ying

October 24, 2021

1 Introduction

Recently there has been progress in applying a novel gradient-free optimization heuristic to high dimensional machine learning, called Consensus Based Optimization (CBO) [1], and on a novel sampling method for Reinforcement Learning (RL), called Borrowing from Future (BFF) [2]. Given the advantages CBO and BFF individually, we explore if there is an advantage to using both. We first introduce Bellman residual minimization (BRM) for Q-control and outline CBO vs Stochastic Gradient Descent (SGD) and BFF vs UR (Unrealistic) vs DS (Double) sampling. Next we prove some results for the difference between UR and BFF in CBO. We then setup a numerical example is continuous and discrete statespace, outline a procedure, and analyze the results. Finally, we give an upper bound on the difference in CBO's parameter modification using BFF vs UR. While CBO+BFF appears advantageous in the discrete case, this is not the case for the continuous case, likely attributable to the higher dimensionality of the hyperparameter space for CBO and/or the highly non-convex loss function for the continuous case.

2 Models

In Q-control, the objective is to find a $Q^* : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ satisfying the Bellman equation $Q^*(s,a) = \mathbb{T}^{\pi_*} Q^*(s,a)$ where

$$\mathbb{T}^{\pi_*} Q^*(s,a) = \mathbb{E}\Big[r(s_{m+1}, s_m, a_m) + \gamma \max_{a'} Q^*(s_{m+1}, a'; \theta) \Big| (s_m, a_m) = (s,a)\Big]$$

One method of doing this is to solve the Bellman residual minimization (BRM) problem, which amounts to minimizing $J(\theta) = \mathbb{E}[\delta^2(s,a;\theta)]$ where $\delta(s,a;\theta) = \mathbb{T}^{\pi_*} Q^*(s,a) - Q^*(s,a)$, i.e.

$$\begin{aligned} J(\theta) &= \mathbb{E}[((\mathbb{T}^{\pi_*} - \mathbb{I})Q^*(s,a;\theta))^2] \\ &= \mathbb{E}[\mathbb{E}[j^{ctrl}(s_m, a_m, s_{m+1}; \theta)|s_m, a_m]^2] \\ j^{\text{ctrl}}(s_m, a_m, s_{m+1}; \theta) &= r(s_{m+1}, s_m, a_m) + \gamma \max_{a'} Q^*(s_{m+1}, a'; \theta) - Q^*(s_m, a_m; \theta) \end{aligned}$$

Solving this minimization problem requires an optimization scheme and a sampling method.

For the optimization schemes, we consider Stochastic Gradient Descent (SGD) and Consensus Based Optimization (CBO).

At iteration k, with batch B_k, samples $\{(s_m, a_m, s_{m+1}, s'_{m+1})\}_{m \in B_k}$, parameters θ_k, and learning rate τ_k, SGD is formulated by the following update equation

$$\begin{aligned} \hat{F}_k &= \frac{1}{|B_k|} \sum_{m \in B_k} j^{\text{ctrl}}(s_m, a_m, s_{m+1}; \theta_k) \nabla_\theta j^{\text{ctrl}}(s_m, a_m, s'_{m+1}; \theta_k) \\ \Delta\theta_k &= -\tau_k \hat{F}_k \end{aligned}$$

At iteration k, with batch B_k, samples $\{(s_m, a_m, s_{m+1}, s'_{m+1})\}_{m \in B_k}$, particle parameters $\{\theta_k\}_{j=1}^N$, learning rate η_k, diffusion rate τ_k, and inverse temperature β_k, CBO is formulated by the following update equation

$$
\begin{aligned}
\hat{J}_k^j &= \frac{1}{|B_k|}\sum_{m\in B_k} j^{\mathrm{ctrl}}(s_m, a_m, s_{m+1}; \theta_k^j)\, j^{\mathrm{ctrl}}(s_m, a_m, \tilde{s}_{m+1}; \theta_k^j) \\
\bar{\theta}_k &= \frac{\sum_{j=1}^N \theta_k^j e^{-\beta_k \hat{J}_k^j}}{\sum_{j=1}^N e^{-\beta_k \hat{J}_k^j}} \\
\Delta\theta_k^j &= (-\eta_k \lambda_k I + \tau_k \sqrt{\eta_k} Z_j)(\theta_k^j - \bar{\theta}_k); \quad Z_j \sim \mathrm{diag}(\mathcal{N}(0, I))
\end{aligned}
$$

additionally, the CBO algorithm checks if $\|\Delta\theta_k^j\| < \delta$ and adds $\tau_k \sqrt{\eta}_k Z_j'$ to $\Delta\theta_k^j$, with $Z_j' \sim \mathrm{diag}(\mathcal{N}(0, I))$.

Given (s_m, a_m, s_{m+1}) we need a sampling method to generate $\tilde{s}_{m+1}$, ideally from the dynamics of the model which are unknown. Assume the dynamics of the model have the form

$$
\Delta s_m = \mu(s_m, a_m)\epsilon + \sigma_m \sqrt{\epsilon} z_m
$$

with $a_m \sim \pi_b(\cdot|s_m)$ and $z_m \sim \mathcal{N}(0, I)$. Unrealistic sampling (UR/Uncorrelated Sampling) generates $\tilde{s}_{m+1}$ directly from the dynamics of the model, Double Sampling (DS/Sample Cloning) reuses the original sample setting $\tilde{s}_{m+1} = s_{m+1}$, and Borrowing From the Future (BFF) exploits a difference with a future sample setting $\tilde{s}_{m+1} = s_m + \Delta s_{m+1}$ where $\Delta s_{m+1} = s_{m+2} - s_{m+1}$.

3 Theoretical Results

3.1 Difference in UR vs BFF in CBO

We consider the CBO model with UR and BFF sampling and prove resulting relating to convergence. The approach is adapted from [2], where a similar analysis is done for SGD.

Taking $N \to \infty$, $\eta \to 0$, the mean field limit of CBO model is given by the following SDE with corresponding Fokker-Planck equation

$$
\begin{aligned}
d\Theta &= -\lambda(\Theta - \bar{\Theta})dt + \tau \sum_{i=1}^d e^{(i)}(\Theta - \bar{\Theta})_i dW_i \\
\bar{\Theta} &= \frac{\mathbb{E}_\Theta[\Theta e^{-\beta J(\Theta)}]}{\mathbb{E}_\Theta[e^{-\beta J(\Theta)}]} \\
\partial_t \rho &= \lambda \nabla \cdot ((\Theta - \bar{\Theta})\rho) + \frac{1}{2}\tau^2 \sum_{i=1}^d \partial_{ii}((\Theta - \bar{\Theta})_i \rho)
\end{aligned}
$$

Where $\Theta_k \approx \theta_{\eta k}$. As $\bar{\Theta}$ depends on the sample s, a, we approximate the process as

$$
\begin{aligned}
d\Theta &= -\lambda(\Theta - \mathbb{E}_{s,a}[\bar{\Theta}])dt + \tau \sum_{i=1}^d e^{(i)}(\Theta - \mathbb{E}_{s,a}[\bar{\Theta}])_i dW_i + \lambda \eta^{1/2}\, \mathbb{V}_{s,a}[\bar{\Theta}]^{1/2} dW_0 \\
\partial_t \rho &= \lambda \nabla \cdot ((\Theta - \mathbb{E}_{s,a}[\bar{\Theta}])\rho) + \frac{1}{2}\sum_{i=1}^d \partial_{ii}((\tau^2(\Theta - \mathbb{E}_{s,a}[\bar{\Theta}])_i^2 + \lambda^2 \eta \mathbb{V}_{s,a}[\bar{\Theta}]_{ii})\rho)
\end{aligned}
$$

We leave the proof of that the approximation is correct for future study. Intuitively, similarly to how SGD is approximated, $\bar{\Theta} = \mathbb{E}_{s,a}[\bar{\Theta}] + \eta^{1/2}\mathbb{V}_{s,a}[\bar{\Theta}]^{1/2}\frac{dW_0}{dt}$. Henceforth, we use δ denote any difference between a UR and BFF quantity, such as $\delta J(\Theta) = J(s, \Theta) - J(s', \Theta)$.

Lemma 1. *Fix t. Assume* $\sup_{s\in\mathbb{S},a\in\mathbb{A}} |\mathbb{E}_a[\mu(s,a)|s]-\mu(s,a)| \leq v$, $\sup_{s\in\mathbb{S}} \|\bar{\Theta}(s)\| \leq m$, $\sup_{s\in\mathbb{S}} \|\partial_s\bar{\Theta}(s)\| \leq m_s$. *almost surely. Then*

$$\|\delta\mathbb{E}_{s,a}[\bar{\Theta}]\| \leq \epsilon v m_s + o(\epsilon)$$
$$|\delta(\Theta - \mathbb{E}_{s,a}[\bar{\Theta}])_i^2| \leq \epsilon v m m_s + o(\epsilon)$$

Proof. Taylor expanding $\bar{\Theta}(\tilde{s})$ about s_m we find

$$\bar{\Theta}(\tilde{s}_{m+1}) = \bar{\Theta}(s_m) + (\tilde{s}_{m+1} - s_m)\partial_s\bar{\Theta}(s_m) + \tfrac{1}{2}(\tilde{s}_{m+1} - s_m)^2\partial_s^2\bar{\Theta}(s_m) + o(\epsilon)$$

Note that

$$\begin{aligned}\tilde{s}_{m+1} - s_m &= \mu(s_m, a_m)\epsilon + \sigma\epsilon^{1/2}Z_m \\ \tilde{s}'_{m+1} - s_m &= \mu(s_m, a_{m+1})\epsilon + \sigma\epsilon^{1/2}Z_{m+1}\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}[\bar{\Theta}(\tilde{s}_{m+1})|a_m, s_m] &= \bar{\Theta}(s_m) - \epsilon\mu(s_m, a_m)\partial_s\bar{\Theta}(s_m) + \tfrac{1}{2}(\sigma^2\epsilon)\partial_s^2\bar{\Theta}(s_m) + o(\epsilon) \\ \mathbb{E}[\bar{\Theta}(\tilde{s}'_{m+1})|a_m, s_m] &= \bar{\Theta}(s_m) - \epsilon\mathbb{E}[\mu(s_m, a_{m+1})|s_m]\partial_s\bar{\Theta}(s_m) + \tfrac{1}{2}(\sigma^2\epsilon)\partial_s^2\bar{\Theta}(s_m) + o(\epsilon) \\ \delta\mathbb{E}[\bar{\Theta}|a_m, s_m] &= \epsilon(\mathbb{E}[\mu(s_m, a_{m+1})|s_m] - \mu(s_m, a_m))\partial_s\bar{\Theta}(s_m) + o(\epsilon)\end{aligned}$$

Taylor expanding $\pi(a|s)$ about s_{m+1} we have

$$\begin{aligned}\mathbb{E}[\mu(s_m, a_{m+1})|s_m] &= \int da\, \mu(s_m, a)\pi(a|s_{m+1}) \\ &= \int da\, \mu(s_m, a)(\pi(a|s_m) + \partial_s\pi(a|s_m)(\mu(s_m, a_m)\epsilon + \sigma\epsilon^{1/2}Z_m)) + O(\epsilon) \\ &= \mathbb{E}[\mu(s_m, a_m)|s_m] + O(\epsilon^{1/2})\end{aligned}$$

Thus

$$\|\delta\mathbb{E}_{s,a}[\bar{\Theta}]\| \leq \epsilon v m_s + o(\epsilon)$$

Similarly we find

$$\delta\mathbb{E}[(\Theta - \bar{\Theta})_i|a_m, s_m]^2 = \epsilon(\Theta - \bar{\Theta}(s_m))_i(\mathbb{E}[\mu(s_m, a_{m+1})|s_m] - \mu(s_m, a_m))\partial_s\bar{\Theta}_i(s_m) + O(\epsilon^2)$$

Thus

$$|\delta(\Theta - \mathbb{E}_{s,a}[\bar{\Theta}])_i^2| \leq \epsilon v m m_s + o(\epsilon)$$

■

Suppose $\mathbb{V}_{s,a}[\bar{\Theta}] = \xi^2 I$ and $\tau = 0$. Note then that $\rho_\infty(\Theta) = \frac{1}{Z}e^{-\alpha\|r(\Theta)\|^2}$ is a steady state solution, where $r(\Theta) = \Theta - \mathbb{E}_{s,a}[\bar{\Theta}]$, $\alpha = \frac{1}{\lambda\eta\xi^2}$, and $Z = (\pi/\alpha)^{1/2}$ is the normalization constant. We also notice that the difference pdf $\delta\rho$ satisfies

$$\partial_t\delta\rho = \lambda\nabla\cdot((\Theta - \mathbb{E}_{s,a}[\bar{\Theta}])\delta\rho - \delta\mathbb{E}_{s,a}[\bar{\Theta}]\rho') + \frac{1}{2}\sum_{i=1}^{d}\partial_{ii}(\tau^2(\Theta - \mathbb{E}_{s,a}[\bar{\Theta}])_i^2\delta\rho + \tau^2\delta(\Theta - \mathbb{E}_{s,a}[\bar{\Theta}])_i^2\rho' + \eta\lambda^2\xi^2\delta\rho)$$

Consider the norm

$$\|\delta\rho\|_* = \int \frac{(\delta\rho)^2}{\rho_\infty}\, d\Theta$$

Note that $\lim_{|\Theta|\to\infty} \|r(\Theta)\|^2 = \infty$, $Z < \infty$, $\lim_{|\Theta|\to\infty}(\frac{1}{2}|\nabla\|r(\Theta)\|^2| - \Delta\|r(\Theta)\|^2) = \infty$. This guarantees that ρ_∞ satisfies the Poincaré inequality

$$\int f^2 \rho_\infty d\theta \le \zeta(\alpha) \int (\nabla f)^2 \rho_\infty d\theta, \quad \forall \int f d\theta = 0$$

where $\zeta(\alpha)$ is the Poincaré constant which typically has $\partial_\alpha \zeta < 0$.

$$\|\delta\rho(t)\|_* \le \|\rho(0) - \rho^\infty\|_* \ldots$$

Proof. ■

4 Numerical Example

4.1 Setup

We consider continuous and a discrete statespace examples. For the continuous case, $\mathbb{S} = (0, 2\pi]$, $\mathbb{A} = \{\pm 1\}$, the dynamics are governed by $\Delta s_m = a_m \epsilon + \sigma\sqrt{\epsilon} Z_m$, $a_m \sim \pi_b(\cdot|s_m)$ where $\pi_b(a|s) = \frac{1}{|\mathbb{A}|}$, $\varepsilon = \frac{2\pi}{32}$, $\sigma = 0.2$, $r(s_{m+1}, s_m, a_m) = \sin(s_{m+1}) + 1$, and we model Q^* using a ResNet

$$F_3^{50,2}\left(F_2^{50,50} \circ \cos \circ F_1^{2,50}(\sin(s), \cos(s)) + F_1^{2,50}(\sin(s), \cos(s))\right)[a]$$

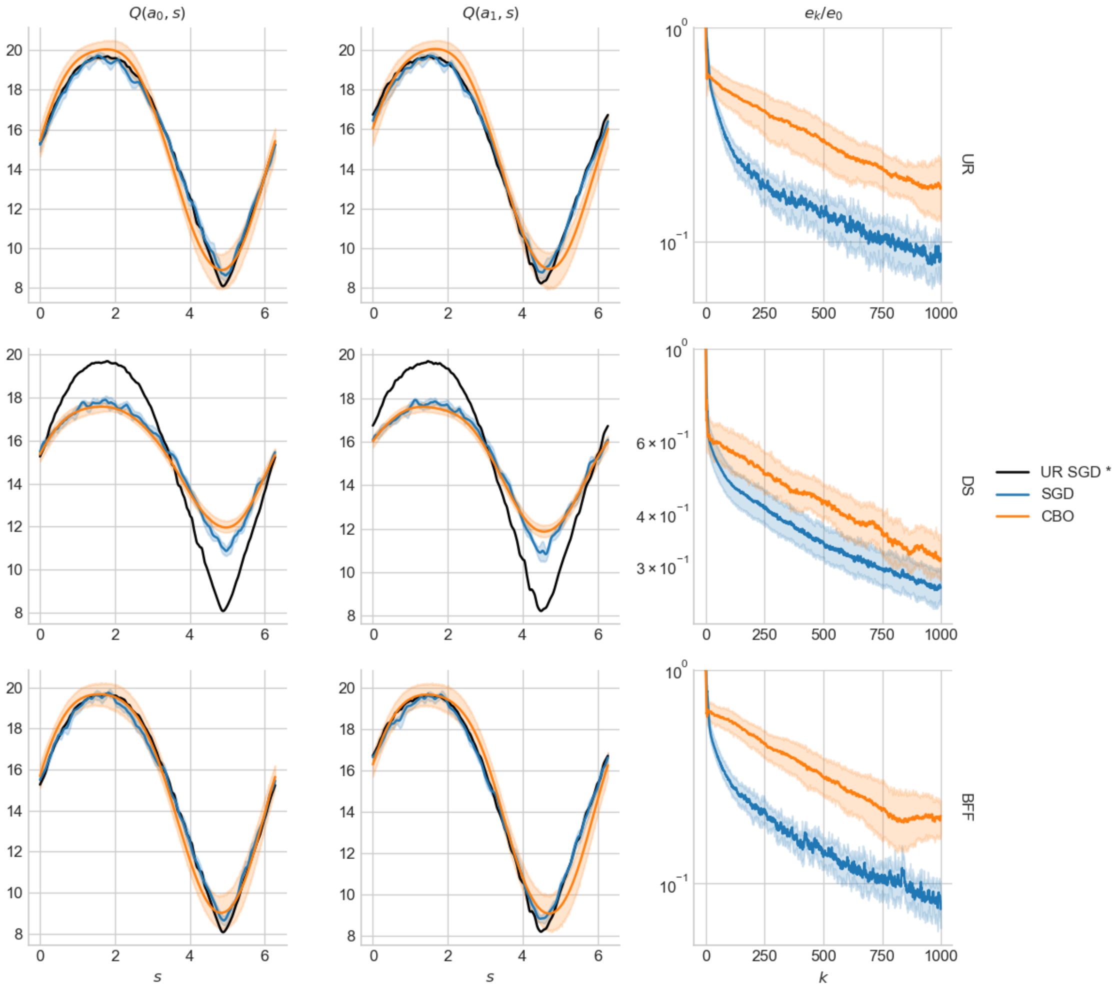
where $F_j^{n,m}(x) = W_j x + b_j$ with $x \in \mathbb{R}^n, b_j \in \mathbb{R}^m, W_j \in \mathbb{R}^{m\times n}$

For the discrete case, $\mathbb{S} = \{\frac{2\pi k}{n} : k \in \mathbb{Z} \cap [0, n-1]\}$, $\mathbb{A} = \{\pm 1\}$, the dynamics are governed by $\Delta s_m = \frac{2\pi}{n} a_m \epsilon + \sigma\sqrt{\epsilon} Z_m$, $a_m \sim \pi_b(\cdot|s_m)$, where $\pi_b(a|s) = \frac{1}{2} + a\sin(s)$, $n = 32$, $\varepsilon = 1$, $\sigma = 1$, $r(s_{m+1}, s_m, a_m) = \sin(s_{m+1}) + 1$, and we model Q^* tabularly, meaning $Q^*(s, a; \theta) = \theta[s, a]$.

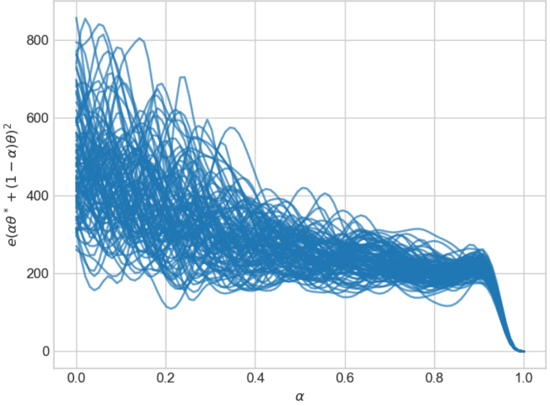
4.2 Procedure

- The fixed behavior policy π_b is sampled generating normal (10^6 steps) and long (10^7 steps) trajectories.
- A reference θ^* is computed by running UR SGD for 1 epoch based using the long trajectory with learning rate $\tau_k = \max(0.8 \cdot 0.9992^k, 0.3)$, batch size $M = 1000$, discount factor $\gamma = 0.9$.
- We apply hyperparameter optimization, using optuna, running 150 trials, on the normal trajectory. For SGD, we fix batch size $M = 1000$, 1 epoch, discount factor $\gamma = 0.9$, learning rate $\tau_k = \max(\tau_i \cdot \tau_r^k, \tau_i \cdot \tau_f)$ where τ_i, τ_r, τ_f are the hyperparameters to optimize. For CBO, we fix number of particles $N = 90$, batch size $m = 1000$, 1 epoch, discount factor $\gamma = 0.9$, $\lambda = 1$, threshold for brownian motion bump $\delta = 1 \times 10^{-5}$, learning rate $\eta_k = \max(\eta_i \cdot \eta_r^k, \eta_i \cdot \eta_f)$, exploration rate $\tau_k = \max(\tau_i \cdot \tau_r^k, \tau_i \cdot \tau_f)$, reciprocal of characteristic energy $\beta_k = \min(\beta_i \cdot \beta_r^k, \beta_i \cdot \beta_f)$ where $\eta_i, \eta_r, \eta_f, \tau_i, \tau_r, \tau_f, \beta_i, \beta_r, \beta_f$ are the hyperparameters to optimize.
- We average 10 instances using best hyperparameters and plot $Q(\cdot, \cdot; \theta)$, as well as log relative error $e_k = e(\theta_k) = \|Q(\cdot, \cdot; \theta^*) - Q(\cdot, \cdot; \theta_k)\|_{L^2(\mathbb{S}\times\mathbb{A})}$.
- The optimization landscape is also visualized by evaluating error on affine combination of parameters of θ^* and random initialization parameters.

4.3 Results

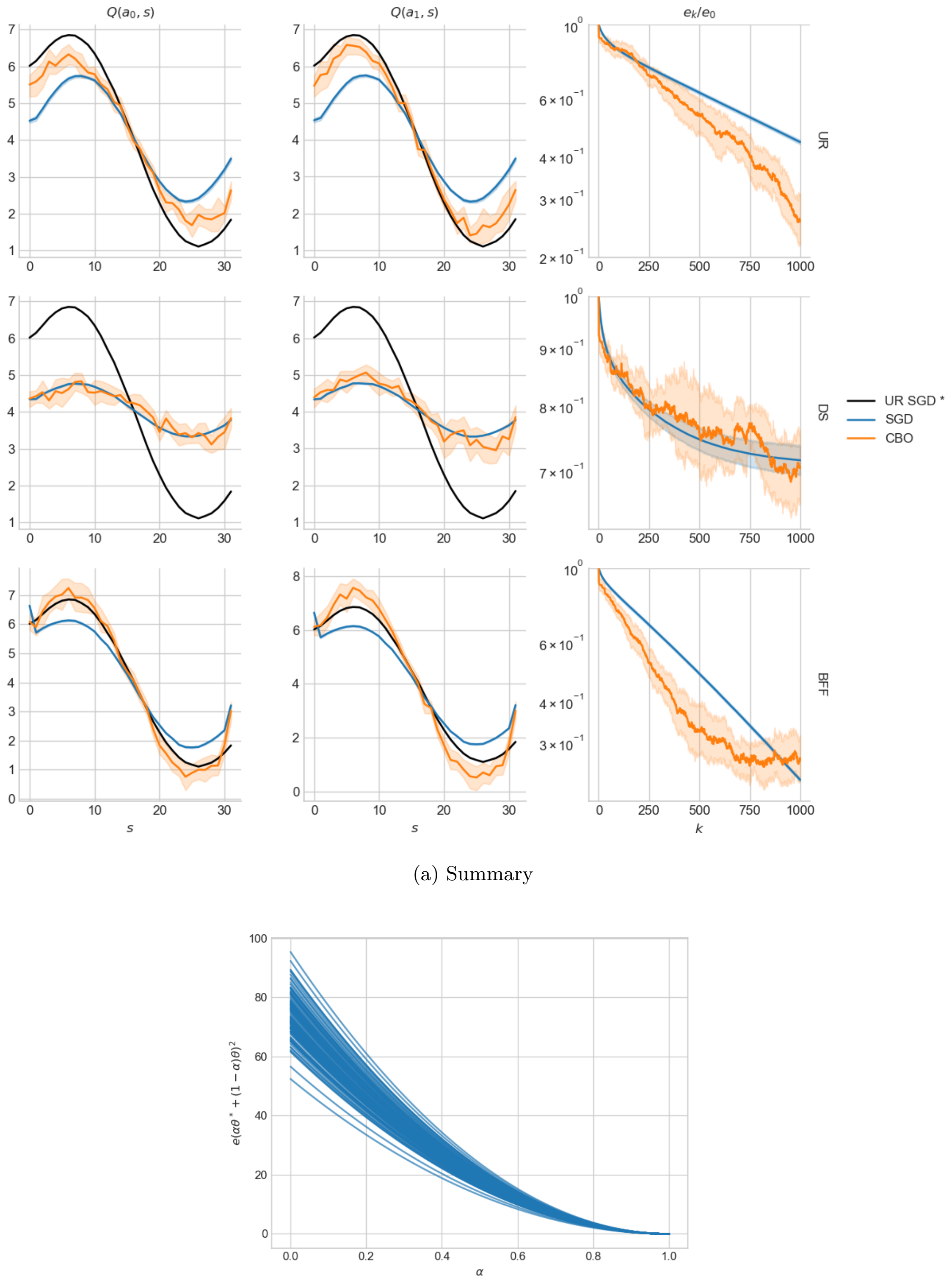


(a) Summary



(b) Error landscape

Figure 1: Continuous case



(a) Summary

(b) Error landscape

Figure 2: Discrete case

SGD	τ
i	0.08001579582322532
f	0.9835361410049764
r	0.950886309322411

Table 1: Optimal hyperparemters for continuous case, SGD

CBO	η	τ	β
i	0.27998130431694734	0.45180905444083275	8.51669145194007
f	0.5194195083343263	0.4287097014952387	1.7500535407136808
r	0.9698276350455655	0.9578569049519733	1.0213109427307054

Table 2: Optimal hyperparemters for continuous case, CBO

SGD	τ
i	4.703979337147098
f	0.654883440315296
r	0.9730502549231891

Table 3: Optimal hyperparemters for discrete case, SGD

CBO	η	τ	β
i	0.9785432879550536	0.893733865582011	12.055703082474112
f	0.8242055859864529	0.3129317992204857	2.6535471522264684
r	0.9827373829994701	0.9999130685913801	1.019158705231164

Table 4: Optimal hyperparemters for discrete case, CBO

5 Conclusion

To summarize the results, we can say that in the Discrete + Tabular case, CBO $\gg$ SGD, BFF $>$ UR while in the Continuous + ResNet case CBO $<$ SGD and BFF $\sim$ UR.

An important note is that the differences observed above could be related to the number of variables in the hyper-parameter search (3 for SGD, 9 for CBO), as well as the complexity/convexity of the problem, ie. if we performed more trials, it might be possible for CBO $>$ SGD and BFF $>$ UR in the continuous case, although we have not observed this. This note is also based on the fact that CBO $>$ SGD for V-eval without hyperparameter search.

References

[1] José A. Carrillo, Shi Jin, Lei Li, and Yuhua Zhu. A consensus-based global optimization method for high dimensional machine learning problems, 2020.

[2] Yuhua Zhu, Zach Izzo, and Lexing Ying. Borrowing from the future: Addressing double sampling in model-free control, 2020.

[3] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference*

on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pages 2101–2110. PMLR, 06–11 Aug 2017.

[4] Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent, 2018.