

Algorithm 1 BCQ

Input: Batch \mathcal{B} , horizon T , target network update rate τ , mini-batch size N , max perturbation Φ , number of sampled actions n , minimum weighting λ .

Initialize Q-networks $Q_{\theta_1}, Q_{\theta_2}$, perturbation network ξ_ϕ , and VAE $G_\omega = \{E_{\omega_1}, D_{\omega_2}\}$, with random parameters $\theta_1, \theta_2, \phi, \omega$, and target networks $Q_{\theta'_1}, Q_{\theta'_2}, \xi_{\phi'}$ with $\theta'_1 \leftarrow \theta_1, \theta'_2 \leftarrow \theta_2, \phi' \leftarrow \phi$.

for $t = 1$ **to** T **do**

 Sample mini-batch of N transitions (s, a, r, s') from \mathcal{B}

$\mu, \sigma = E_{\omega_1}(s, a), \quad \tilde{a} = D_{\omega_2}(s, z), \quad z \sim \mathcal{N}(\mu, \sigma)$

$\omega \leftarrow \operatorname{argmin}_\omega \sum (a - \tilde{a})^2 + D_{\text{KL}}(\mathcal{N}(\mu, \sigma) || \mathcal{N}(0, 1))$

 Sample n actions: $\{a_i \sim G_\omega(s')\}_{i=1}^n$

 Perturb each action: $\{a_i = a_i + \xi_\phi(s', a_i, \Phi)\}_{i=1}^n$

 Set value target y (Eqn. 13)

$\theta \leftarrow \operatorname{argmin}_\theta \sum (y - Q_\theta(s, a))^2$

$\phi \leftarrow \operatorname{argmax}_\phi \sum Q_{\theta_1}(s, a + \xi_\phi(s, a, \Phi)), a \sim G_\omega(s)$

 Update target networks: $\theta'_i \leftarrow \tau\theta + (1 - \tau)\theta'_i$

$\phi' \leftarrow \tau\phi + (1 - \tau)\phi'$

end for
