TD-CD-SAC Online Training Algorithm

Algorithm Implementation

February 5, 2026

1 Notation

- Q_{θ_j}: j-th critic network with parameters θ_j
- π_ϕ: Actor network with parameters ϕ
- $\bar{\theta}_j$: Target network parameters for critic j
- α: Temperature parameter for entropy regularization
- $\mathcal{D}$: Replay buffer
- c_t: Constraint violation at time t
- δ_t: Soft termination probability at time t
- γ_t: Discount factor at time t
- $c_{\max}^{seen}$: Maximum constraint violation seen so far
- $c_{\max}^{ema}$: Exponential moving average of maximum constraint violation

2 Algorithm Description

This algorithm implements TD-CD-SAC (Time-Dependent Constraint-Directed Soft Actor-Critic), an online reinforcement learning method that handles constraints through time-dependent discount factors. The key innovation is the use of soft termination probabilities δ_t that modify the effective discount factor γ_t based on constraint violations.

The algorithm operates in two main phases:

1. **Data Collection**: Actions are sampled from the policy and executed in the environment to collect transitions.
2. **Network Updates**: The actor and critic networks are updated using the collected data, with modified target values that account for constraint violations.

The constraint handling mechanism works by:

- Detecting constraint violations through environment information
- Computing soft termination probabilities based on violation severity
- Modifying discount factors to penalize constraint-violating trajectories
- Updating the policy to avoid constraint violations while maximizing rewards

Algorithm 1 TD-CD-SAC Online Training (cd_sac_ball)

Require: Target g, constrained environment $\mathcal{E}$ with bounds $(v_{\max}, a_{\max})$
Require: Discount factor $\gamma \in (0,1)$, soft-update parameter τ, batch size B
Require: TD-CD hyperparameters $p^{\max} \in [0,1]$, EMA factor $\tau_c \in [0,1]$
Require: Switch `use_amount` $\in \{0,1\}$ (binary/continuous constraint signal)

Initialize SAC networks $(Q_{\theta_1}, Q_{\theta_2}, \pi_\phi)$ and targets $\bar{\theta}_1 \leftarrow \theta_1,\ \bar{\theta}_2 \leftarrow \theta_2$
Initialize replay buffer $\mathcal{D} \leftarrow \emptyset$
Initialize TD-CD stats: $c_{\max}^{seen} \leftarrow 0$, $c_{\max}^{ema} \leftarrow 1$
Sample initial state $s_0 \sim \mathcal{E}$.reset()
for $t = 0, 1, 2, \ldots, T-1$ **do**
    Select action $u_t \sim \begin{cases} \mathcal{U}([-1,1]^m) & t < t_{\text{start}} \\ \pi_\phi(\cdot \mid s_t) & \text{otherwise} \end{cases}$
    Execute in environment: $(s_{t+1}, r_t, d_t, \text{info}_t) \leftarrow \mathcal{E}.\text{step}(u_t)$
    ▷ Time-limit truncation is treated as non-terminal
    $d_t^{buf} \leftarrow d_t$; **if** $\text{info}_t.\texttt{time\_limit} = 1$ **then** $d_t^{buf} \leftarrow 0$
    if $d_t^{buf} = 1$ **then**
        $\gamma_t \leftarrow 0$
    else
        ▷ TD-CD Equation (7): build soft termination δ_t
        if `use_amount`=1 **then**
            $c_t \leftarrow \text{info}_t.\texttt{vel\_violation\_amount}$ ▷ $c_t \geq 0$
            $c_{\max}^{seen} \leftarrow \max(c_{\max}^{seen}, |c_t|)$
            $\delta_t \leftarrow p^{\max} \cdot \text{clip}\Big(\frac{|c_t|}{c_{\max}^{ema}},\ 0,\ 1\Big)$
        else
            $c_t \leftarrow \mathbb{I}[\text{info}_t.\texttt{constraint\_violation} = 1]$
            $\delta_t \leftarrow p^{\max} \cdot c_t$
        end if
        ▷ TD-CD Equation (9): per-step discount
        $\gamma_t \leftarrow \gamma\,(1 - \delta_t)$
    end if
    Store transition: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, u_t, r_t, s_{t+1}, d_t^{buf}, \gamma_t)\}$
    $s_t \leftarrow s_{t+1}$
    if $t \geq t_{\text{update}}$ and $|\mathcal{D}| \geq B$ **then**
        for $k = 1$ to K **do** ▷ K updates per environment step
            Sample minibatch $\{(s_i, u_i, r_i, s_i', d_i, \gamma_i)\}_{i=1}^{B} \sim \mathcal{D}$
            Sample next actions $a_i' \sim \pi_\phi(\cdot \mid s_i')$ and log-probs $\log \pi_\phi(a_i' \mid s_i')$
            Compute soft value target:

$$V(s_i') = \min_{j \in \{1,2\}} Q_{\bar{\theta}_j}(s_i', a_i') - \alpha \log \pi_\phi(a_i' \mid s_i')$$

            ▷ Key change vs vanilla SAC: use per-transition γ_i
            Critic target: $y_i \leftarrow r_i + \gamma_i\, V(s_i')$
            Update critics by MSE: $\theta_j \leftarrow \arg\min_{\theta_j} \frac{1}{B} \sum_i (Q_{\theta_j}(s_i, u_i) - y_i)^2$ for $j = 1, 2$
            Update actor (standard SAC):

$$\phi \leftarrow \arg\min_{\phi} \frac{1}{B} \sum_i \big(\alpha \log \pi_\phi(a_i \mid s_i) - \min_j Q_{\theta_j}(s_i, a_i)\big)$$

            (Optional) Update temperature α by entropy tuning
            Soft-update targets: $\bar{\theta}_j \leftarrow \tau\theta_j + (1-\tau)\bar{\theta}_j$ for $j = 1, 2$
        end for
    end if
    if $t > 0$ and $t \bmod N_{\text{eval}} = 0$ **then**
        ▷ TD-CD Equation (8): EMA update once per window
        $c_{\max}^{ema} \leftarrow \tau_c\, c_{\max}^{ema} + (1 - \tau_c)\, \max(c_{\max}^{seen}, \epsilon)$
        $c_{\max}^{seen} \leftarrow 0$
        (Optional) evaluate policy and save checkpoint
    end if
end for