# Towards Stable and Storage-efficient Dataset Distillation: Matching Convexified Trajectory

## Supplementary Material

Table A. Dataset Configuration and Hyperparameter Settings

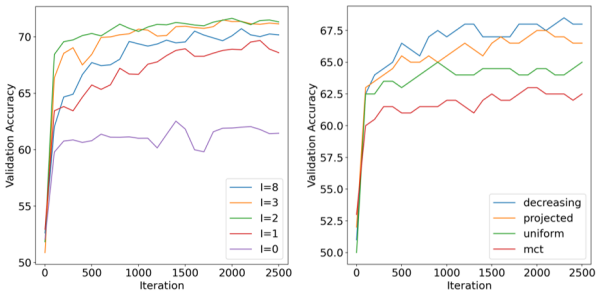
| Dataset | ipc | syn_step | exp_ep | max_ep | lr_img | lr_lr | lr_init |
|---|---|---|---|---|---|---|---|
| CIFAR-10 | 1 | 50 | 5 | 4 | 1e3 | 1e-7 | 1e-2 |
| | 10 | 40 | 5 | 10 | 1e3 | 1e-5 | 1e-2 |
| | 50 | 40 | 5 | 30 | 1e3 | 1e-5 | 1e-3 |
| CIFAR-100 | 1 | 40 | 5 | 10 | 1e3 | 1e-5 | 1e-2 |
| | 10 | 30 | 5 | 40 | 1e3 | 1e-5 | 1e-2 |
| | 50 | 50 | 5 | 40 | 1e3 | 1e-5 | 1e-2 |
| TinyImageNet | 1 | 20 | 5 | 5 | 1e4 | 1e-4 | 1e-2 |
| | 10 | 30 | 4 | 40 | 1e4 | 1e-4 | 1e-2 |
| | 50 | 50 | 4 | 40 | 1e4 | 1e-4 | 1e-2 |

**Algorithm A** Convexified Trajectory Construction

**Input:** Original trajectory $\tau_{\text{mtt}} = \{\theta_{\mathcal{T}}^0, \theta_{\mathcal{T}}^1, \ldots, \theta_{\mathcal{T}}^K\}$
**Input:** Waypoints position $\mathcal{P} = \{p_0, p_1, \ldots, p_I\}$, where $p_0 = 0, p_I = K$
Initialize waypoints $\tau_{\text{conv}} \leftarrow \{\theta_{\mathcal{T}}^0\}$
Initialize step sizes B $\leftarrow \emptyset$
**for** each waypoints position $i = 1$ **to** $I$ **do**
  Initialize weight: $\beta^{(p_{i-1})} \leftarrow \mathbf{0} \in \mathbb{R}^L$ ▷ $L$ = number of network layers
  Initialize segment steps: $\mathcal{S}_i \leftarrow \{\beta^{(p_i)}\}$
  **for** each inner point index $j = (p_{i-1} + 1)$ **to** $p_i$ **do**
    Compute weight delta: $\|\Delta\beta\| = \text{Norm}(\theta_{\mathcal{T}}^{(j+1)} - \theta_{\mathcal{T}}^{(j)})$
    Compute next weight: $\beta^{(j+1)} \leftarrow \beta^{(j)} + \|\Delta\beta\|$
    Append $\beta^{(j+1)}$ to $\mathcal{S}_i$
  **end for**
  Normalize: $\mathcal{S}_i \leftarrow \mathcal{S}_i / \beta^{(p_i)}$
  Append $\mathcal{S}_i$ to B
  Append $\theta_{\mathcal{T}}^{p_i}$ to $\tau_{\text{conv}}$
**end for**
**Output:** Convexified trajectory $\tau_{\text{conv}} = [\theta_{\mathcal{T}}^{p_0}, \theta_{\mathcal{T}}^{p_1} ..., \theta_{\mathcal{T}}^{p_I}]$
**Output:** Step sizes B $= [S_1, ..., S_I]$

## A. Implementation Details

**Algorithm and Hyperparameter Settings.** We illustrate our convexified trajectory construction algorithm in Algorithm A. The algorithm converts the original SGD trajectory into convexified trajectory with certain interpolation points. During the dataset distillation process, we can sample trajectories at arbitrary real-valued $c \in [0, K]$ position and reconstruct the model parameters $\hat{\theta}^{(c)}$ using continuous sampling strategy. Our optimal hyperparameter configurations across different baselines are summarized in Table A. All experiments utilized ZCA whitening.




(a) Different Numbers of $I$ (b) Different Calculation of $\beta$

Figure A. Visualization of synthetic dataset.

## B. Extensive Experiments

**Impact of different numbers of interpolation points.** An ablation study is conducted to investigate the impact of different number of interpolation points $I = \{0, 1, 2, 3, 8\}$ on the CIFAR-10 dataset with ipc $= 50$. For each parameter configuration, we uniformly sample the interpolation points (if had) along the original expert trajectory and generate our convexified trajectories as in Algorithm A. The experimental results (Fig. Aa) demonstrate that both insufficient ($I = \{0, 1\}$) and excessive ($I > 3$) interpolation points degrade distillation performance. The optimal performance is achieved around $I = 2$.

**Impact of different $\beta$ calculation.** We also conducted an ablation study to examine the impact of different $\beta$ calculation strategies on the CIFAR-10 dataset with ipc $= 50$. The experiment results are showed in Fig. Ab, where the $\beta^{(t)}$ in "decreasing" is defined as:

$$\beta^{(t)} = \frac{\sum_{b=0}^{t} -b + K}{\sum_{i=0}^{K} i}, \tag{A}$$

in "projected" is defined as:

$$\beta^{(t)} = \frac{\vec{V}_{\mathcal{T}}^{(0\to t)} \cdot \vec{V}_{\mathcal{T}}^{(0\to K)}}{\|\vec{V}_{\mathcal{T}}^{(0\to K)}\|_2^2}, \tag{B}$$

in "uniform" is defined as:

$$\beta^{(t)} = \sum_{i=0}^{t} \frac{1}{K}, \tag{C}$$

and in "mct" denotes the strategy in the paper. Through extensive experiments, we have identified potential improvements to the $\beta$ calculation method in MCT, which will be investigated as a key direction in our future work.