On the Stability-Plasticity Dilemma in Continual Meta-Learning: Theory and Algorithm

Qi Chen [1] Changjian Shui [2] Ligong Han [3] Mario Marchand [1]

[1]Laval University, Canada, [2]McGill University, Canada, [3]Rutgers University, USA

Main Contribution

- **Unified theoretical framework** for continual meta-learning in both static and shifting task environments.
- Formal analysis of the **bi-level learning-forgetting trade-off**
- **Theoretically grounded algorithm**

Problem Setup

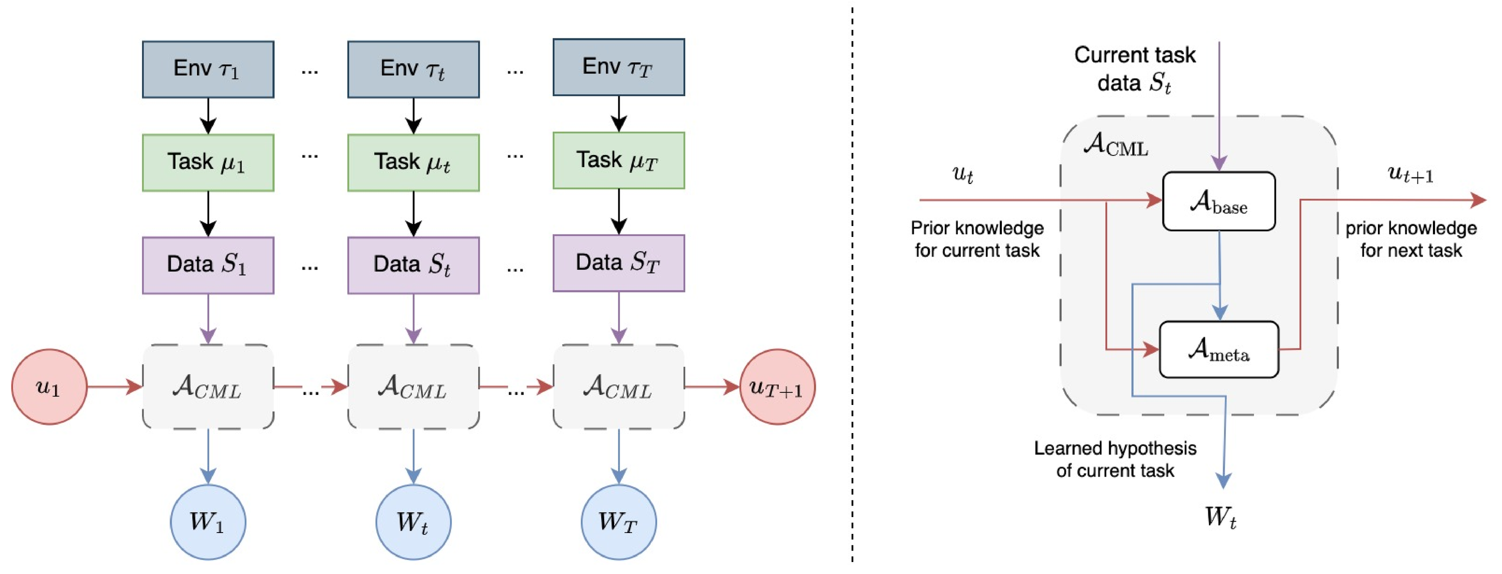

Figure 1. Illustration of Continual Meta-Learning (CML) process

- **Base learner** — batch learning algorithms $W_t = \mathcal{A}(u_t, S_t), u_t \in \mathcal{U}$
  - Excess Risk: $$R_{\text{excess}}(\mathcal{A}, u_t) \overset{\text{def}}{=} \mathbb{E}_{S_t}\mathbb{E}_{W_t\sim P_{W_t|S_t,u_t}}\left[\mathcal{L}_{\mu_t}(W_t) - \mathcal{L}_{\mu_t}(w_t^*)\right], w_t^* = \arg\min_{w\in\mathcal{W}} \mathcal{L}_{\mu_t}(w).$$
  - meta-parameter $u_t = (\beta_t, \phi_t)$, true risk $\mathcal{L}_{\mu_t}(w) \overset{\text{def}}{=} \mathbb{E}_{Z\sim\mu_t}\ell(w, Z)$
  - Assume $\mathcal{L}_{\mu_t}(w)$ has quadratic growth, we have the unified form of excess risk upper bound: $$f_t(u_t) = \kappa_t\left(a\beta_t + \frac{b\|\phi_t - w_t\|^2 + \epsilon_t + \epsilon_0}{\beta_t} + \Delta_t\right), \kappa_t, \epsilon_t, \beta_t, \Delta_t \in \mathbb{R}^+, \forall t \in [T], a, b, \epsilon_0 > 0.$$
  - f_t is **convex**, valide for $\mathcal{A} \in$ {Gibbs, RLM, SGD, SGLD}
- **Meta learner** — online learning algorithms
  - dynamic regret for N static slots $$R_T^{\text{dynamic}}(u_{1:N}^*) \overset{\text{def}}{=} \sum_{n=1}^N \sum_{k=1}^{M_n}\left[f_{n,k}(u_{n,k}) - f_{n,k}(u_n^*)\right], u_n^* \overset{\text{def}}{=} \arg\min_u \frac{1}{M_n}\sum_{k=1}^{M_n} f_{n,k}(u).$$
- **Continual meta-learning objective**

  — select $u_{1:T}$ to minimize the Average Excess Risk (AER):

$$\text{AER}_{\mathcal{A}}^T \overset{\text{def}}{=} \frac{1}{T}\sum_{t=1}^T R_{\text{excess}}(\mathcal{A}, u_t) \leq \frac{1}{T}R_T^{\text{dynamic}}(u_{1:N}^*) + \frac{1}{T}\sum_{n=1}^N\sum_{k=1}^{M_n} f_{n,k}(u_n^*).$$

DCML Algorithm

For $t = 1, ..., T$
- Sample task distribution $\mu_t \sim \tau_t$, Sample dataset $S_t \sim \mu_t^{m_t}$;
- Get meta parameter $u_t = (\beta_t, \phi_t)$, learn base parameter $w_t = \mathcal{A}(u_t, S_t)$, estimate $f_t(u_t)$
- Adjust the learning rate of the meta-parameter (γ_t) with the following strategy:
  - When an environment change is detected, γ_t is set to a large hopping rate $\gamma_t = \rho$
  - For k-th task inside the n-th environment (slot), $\gamma_t = \gamma_0/\sqrt{k}$
- Update meta parameter $u_{t+1} = \Pi_{\mathcal{U}}(u_t - \gamma_t\nabla f_t(u_t))$, *i.e.*, $$\phi_{t+1} = \left(1 - \frac{2b\kappa_t\gamma_t}{\beta_t}\right)\phi_t + \frac{2b\kappa_t\gamma_t}{\beta_t}w_t, \beta_{t+1} = \beta_t - \gamma_t\left(a\kappa_t - \frac{\kappa_t(b\|\phi_t - w_t\|^2 + \epsilon_t + \epsilon_0)}{{\beta_t}^2}\right)$$

Contact Information

- **Email:** qi.chen.1@ulaval.ca
- **Code:** https://github.com/livreQ/DynamicCML

Bi-level Learning-Forgetting Trade-off

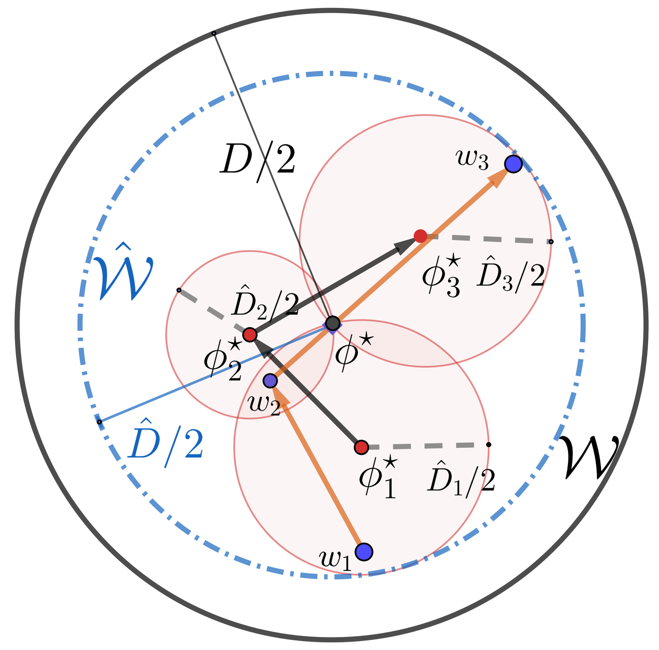

Figure 2. Illustration of a shifting environment.

- task-level learning-forgetting trade-off
  - two tasks $\|w_1 - w_2\| > \hat{D}_1/2$
  - directly adapt

    => catastrophic forgetting
  - keep the optimal prior ϕ_n^*

    => no forgetting

    $\|\phi_1^* - w_1\| < \hat{D}_1/2$ and $\|\phi_1^* - w_2\| < \hat{D}_1/2$
  - $\uparrow \hat{D}_n$ => $\uparrow$ number of examples needed to recover performance with no forgetting
- meta-level learning-forgetting trade-off
  - large environment shift $\hat{D} \gg \hat{D}_n$
  - $\uparrow$ learning rate of meta-knowledge
  - $\uparrow$ forgetting meta-knowledge inside slots

Main Theorem

Theorem 1 (Simplified) Consider both **static** and **shifting** environments. If the excess risk's upper bound of the base learner $\mathcal{A}(u_t, S_t)$ can be formulated as the unified form, then, the **AER** of **DCML** is upper bounded by:

$$\text{AER}_{\mathcal{A}}^T \leq \underbrace{\frac{2}{T}\sum_{n=1}^N \sqrt{a(bV_n^2 + \epsilon_n + \epsilon_0)}\kappa_n + \frac{\Delta_n}{2}}_{\text{optimal trade-off in hindsight}} + \underbrace{\frac{3}{2T}\sum_{n=1}^N \tilde{D}_n G_n\sqrt{M_n - 1}}_{\text{average regret over slots}} + \underbrace{\frac{\tilde{D}_{\max}}{T}\sqrt{2P^*\sum_{n=1}^N G_n^2}}_{\text{regret w.r.t environment shift}}$$

where $V_n^2 = \sum_{k=1}^{M_n}\frac{\kappa_{n,k}}{\kappa_n}\|\phi_n^* - w_{n,k}\|_2^2$ with $\phi_n^* = \sum_{k=1}^{M_n}\frac{\kappa_{n,k}}{\kappa_n}w_{n,k}$, G_n is the upper bound of the cost function gradient norm, and $\tilde{D}_n$ is the diameter of meta-parameters in n-th slot. The path length of N slots is $P^* = \sum_{n=1}^{N-1}\|u_n^* - u_{n+1}^*\| + 1$.

- Optimal trade-off <=> average of slot variances V_n^2 (task similarities)
- Task-level regret <=> slot diameters $\tilde{D}_n$ (task similarities)
- Environment-level regret <=> path length P^* (environment similarities, non-stationarity)

AER Bounds of Specific Base Learners

Theorem 2 (Gibbs Algorithm, simplified) Apply Gibbs algorithm as the base learner in **DCML** and further assume that each slot has equal length M and each task uses the sample number m. Then, the AER can be bounded by:

$$\text{AER}_{\text{Gibbs}}^T \leq \mathcal{O}\left(1 + \bar{V} + \frac{\sqrt{MN} + \sqrt{P^*}}{M\sqrt{N}}\right)\frac{1}{m^{\frac{1}{4}}}, \bar{V} \overset{\text{def}}{=} \frac{1}{N}\sum_{n=1}^N V_n.$$

- Single-task learning $\mathcal{O}\left((D+1)m^{-1/4}\right)$
- Static environments $\mathcal{O}((V+1)m^{-1/4})$ with rate $\mathcal{O}(1/\sqrt{T})$,
- Shifting environments $\mathcal{O}((\bar{V}+1)m^{-1/4})$ with rate $\mathcal{O}(1/\sqrt{M})$, $\bar{V} \leq V \leq \hat{D} \leq D$
- => Same AER with a smaller M, *i.e.*, faster-constructing meta-knowledge in new environments.

Theorem 3 (Stochastic Gradient Descent(SGD), simplified) Apply SGD as the base learner in **DCML** and further assume that each slot has equal length M and each task uses the sample number m. Then, the AER can be bounded by:

$$\text{AER}_{\text{SGD}}^T \leq \mathcal{O}\left(\bar{V} + \frac{\sqrt{MN} + \sqrt{P^*}}{M\sqrt{N}}\right)\sqrt{\frac{1}{K} + \frac{1}{m}}, \bar{V} \overset{\text{def}}{=} \frac{1}{N}\sum_{n=1}^N V_n.$$

- Static environment
  - $N = 1$, $P^* = 1, M = T$
  - recover static regret $\mathcal{O}(V + \frac{1}{\sqrt{T}})\sqrt{\frac{1}{K} + \frac{1}{m}}$
- Shifting environment
  - When N is small and P^* is large
  - better than $\mathcal{O}(\bar{V} + \frac{1}{\sqrt{M}} + \sqrt{\frac{P^*}{NM}})$

Experiments

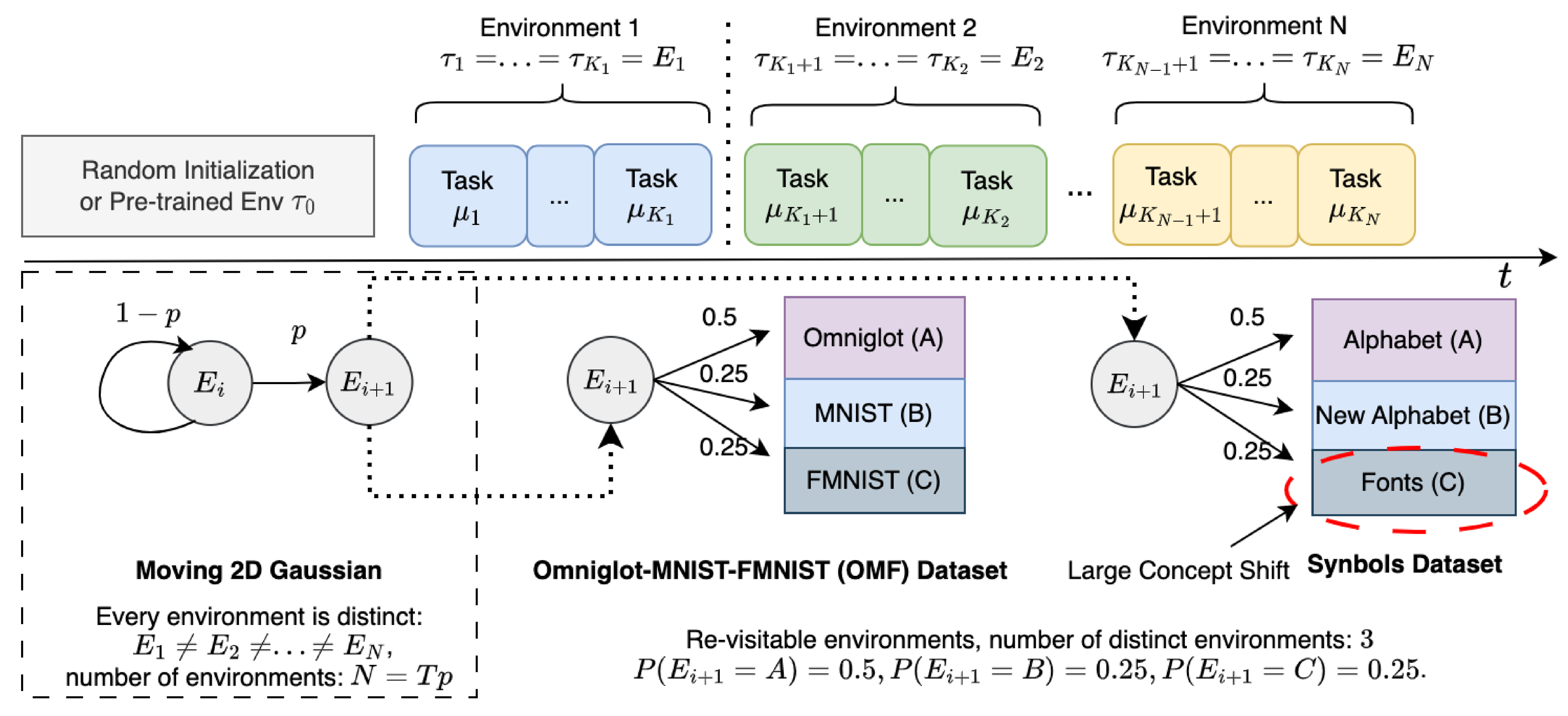

Figure 3. Illustration of the CML experimental setting on synthetic and real datasets. At each time t, the environment changes with probability p. If current environment is $\tau_t = E_i$, the next environment $P(\tau_{t+1} = E_{i+1}) = p, P(\tau_{t+1} = E_i) = 1 - p$.

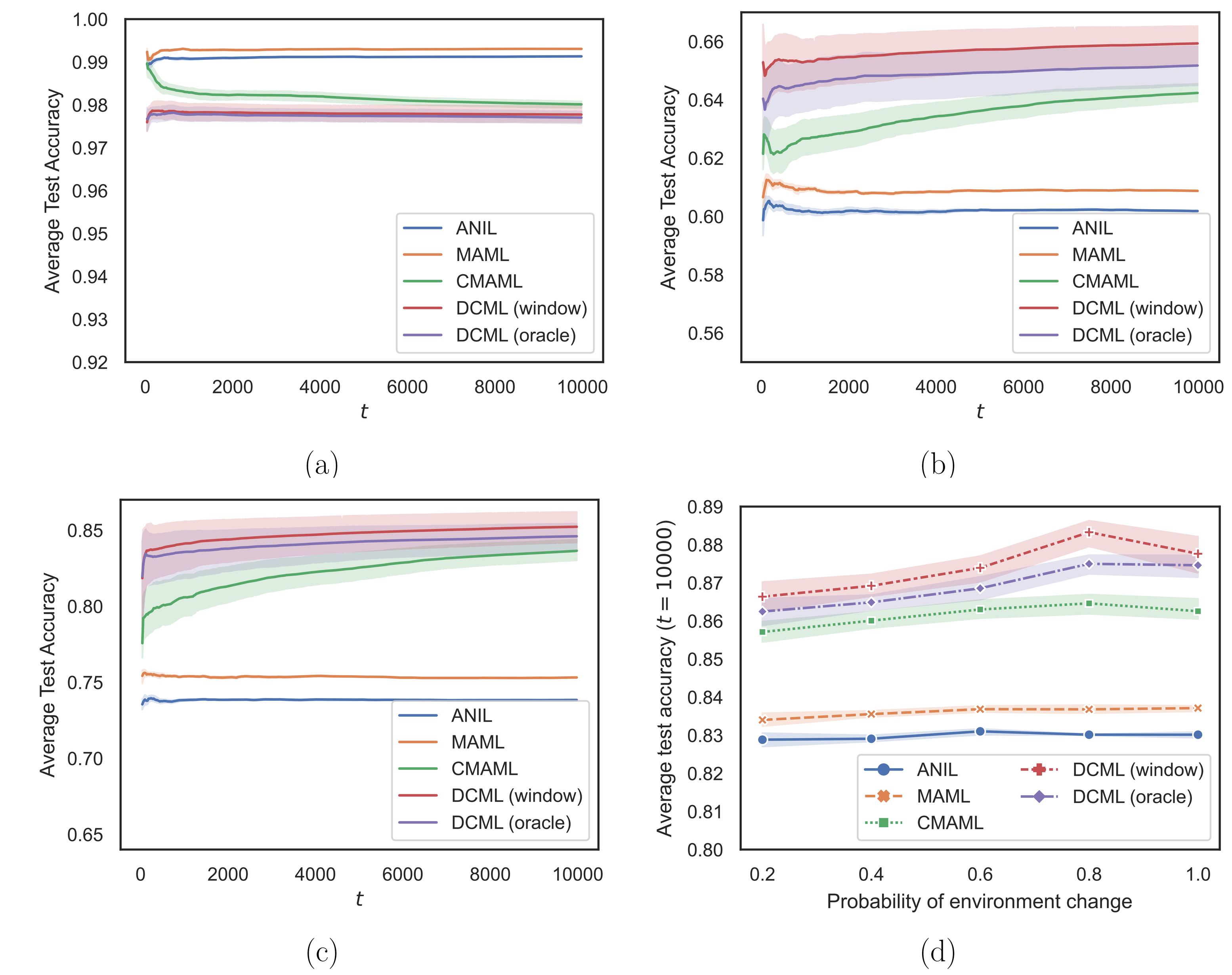

Figure 4. Running time average test accuracy on OMF dataset of OSAKA benchmark: **(a)** Omniglot (pre-trained environment), **(b)** FashionMNIST and **(c)** MNIST, where the environment shifts with probability $p = 0.2$. **(d)** Average test accuracy on overall environment at final step $t = 10000$ w.r.t p.

- better overall learning-forgetting trade-off
- stable to different levels of non-stationarity
- no need for precise environment shifts detection