

Theoretical Framework for Prior Knowledge Transfer in Deep Learning

Thesis Defense Presentation

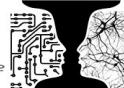
Qi Chen

11 January, 2024

Committee:
Prof. Yongyi Mao (University of Ottawa)
Prof. Pascal Germain
Prof. Audrey Durand
Supervisor:
Prof. Mario Marchand
President:
Prof. Brahim Chaib-draa



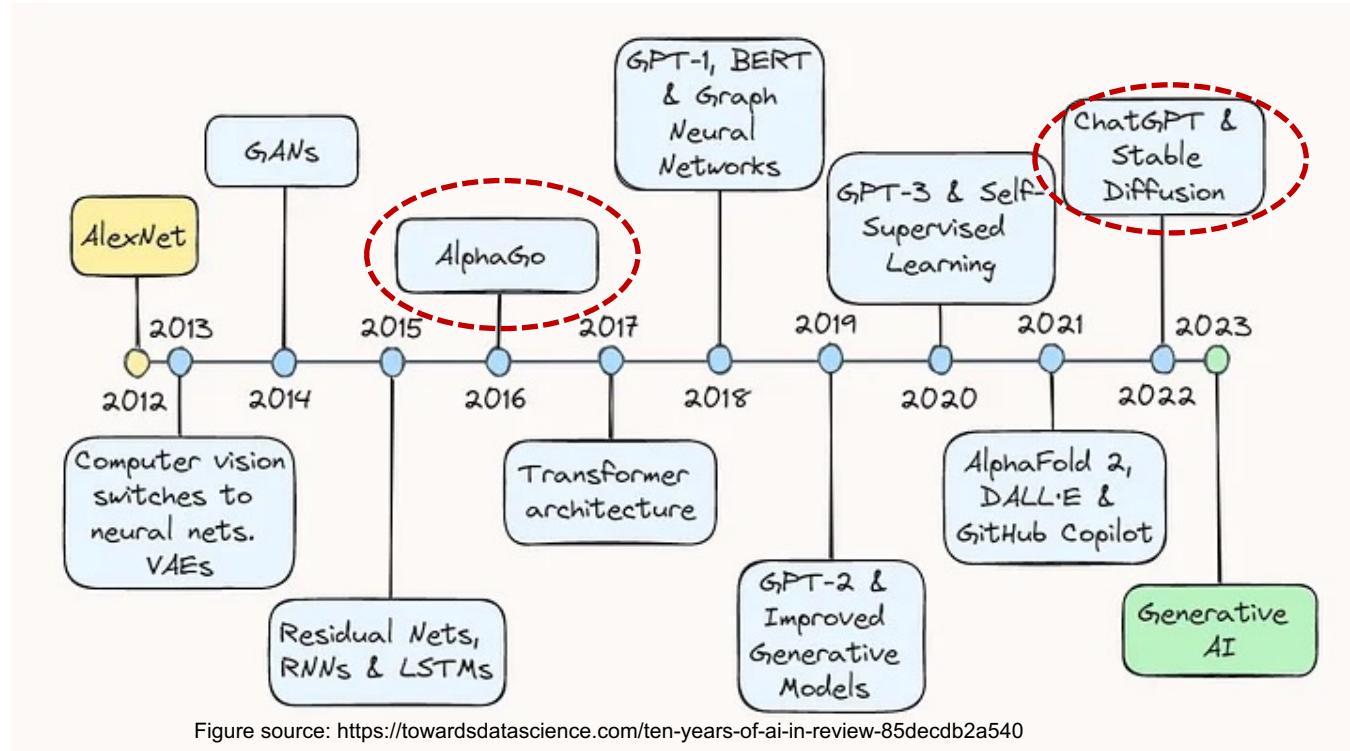
Groupe de
Recherche en
Apprentissage
Automatique de
Laval



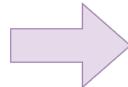
bdrc.ul
UNIVERSITÉ Laval
BIG DATA
RESEARCH CENTER



The Decade of Deep Learning



Large Model Scale



Training from scratch becomes infeasible for most researchers:
Limited data + high computational costs

Need prior knowledge transfer!!!!

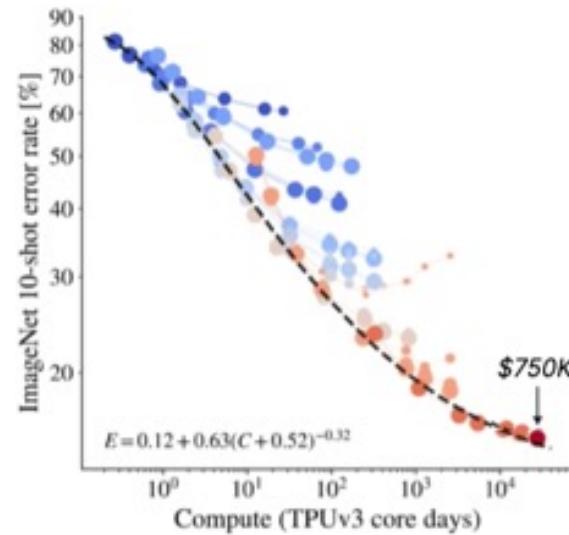
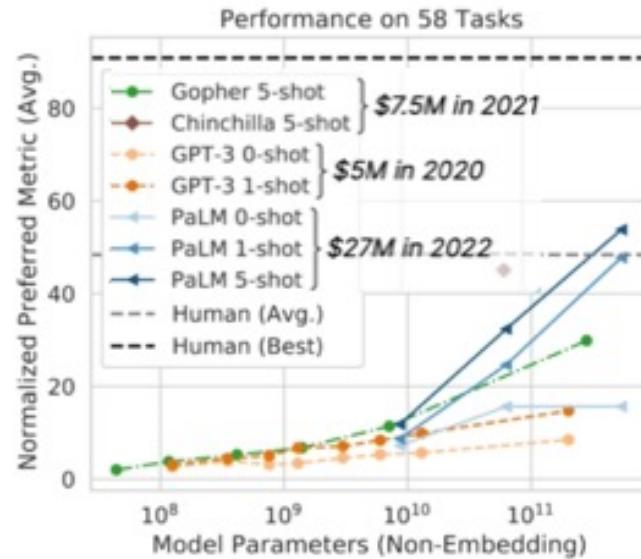


Figure source: <http://colinraffel.com/talks/faculty2023collaborative.pdf>

Prior Knowledge Transfer has been Ubiquitous

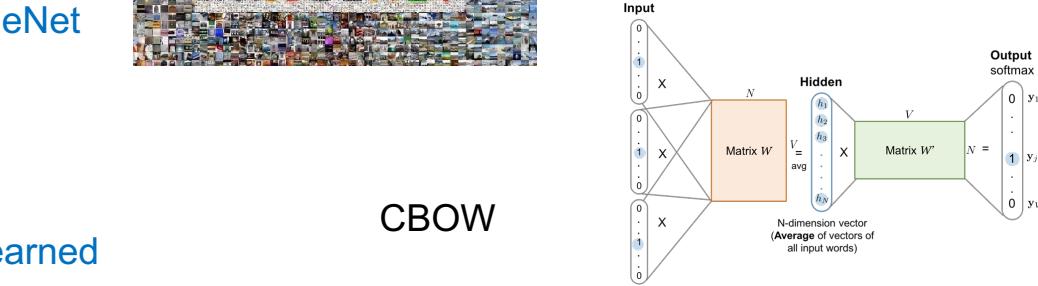
Computer vision downstream tasks:

Finetune on the pre-trained model for ImageNet



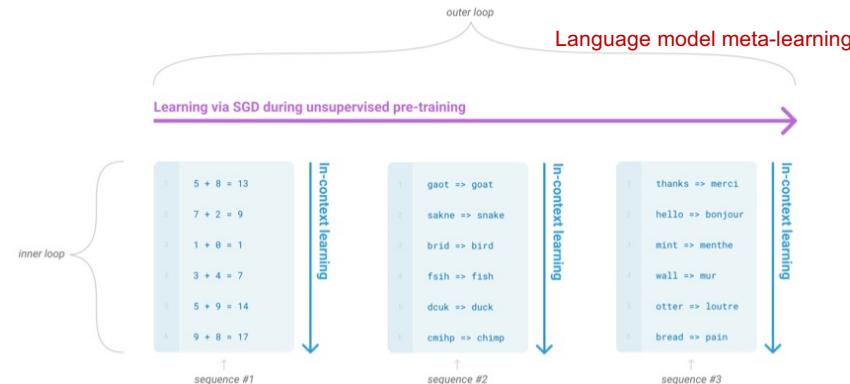
NLP downstream tasks:

Transfer unsupervised or self-supervised learned feature representation (word embedding)

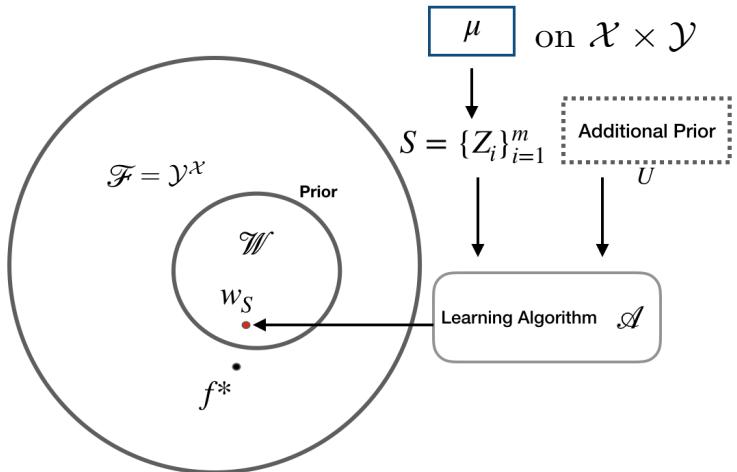


In-context learning:

Transfer knowledge from LLM



Prior Knowledge for Learning A Single Task



$$R_\mu(w) = \mathbb{E}_{Z \sim \mu} \ell(w, Z), R_S(w) = \frac{1}{m} \sum_{i=1}^m \ell(w, Z_i)$$

$$W_S = \mathcal{A}(U, S)$$

- **Objective of Learning**

$$\text{Small } R_\mu(w_S) - R^* = R_\mu(w_S) - \inf_{w \in \mathcal{W}} R_\mu(w) + \underbrace{\inf_{w \in \mathcal{W}} R_\mu(w) - R^*}_{\text{approximation error - bias}}$$

estimation error - complexity

Prior knowledge:

① Selection of \mathcal{W}

② Hyperparameters of $\mathcal{A} - U$

Contributions

Theoretical Frameworks for...

- Domain Adaptation

[Algorithm-Dependent Bounds for Representation Learning of Multi-Source Domain Adaptation. AISTATS 2023.](#)

- Meta-learning

[Generalization Bounds for Meta-Learning: An Information-Theoretic Analysis. NeurIPS 2021.](#)

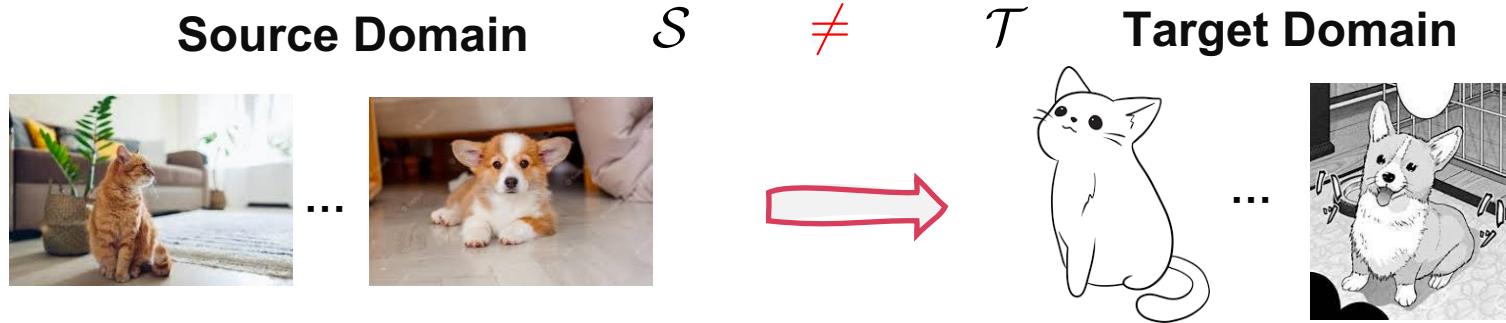
- Continual Meta-learning

[On the Stability-Plasticity Dilemma in Continual Meta-Learning: Theory and Algorithm. NeurIPS 2023.](#)

**Start from a simple prior knowledge
transfer concept:
Domain Adaptation**

Domain Adaptation (DA):

Infer Three Other Related Aspects from One Concept (举一反三)



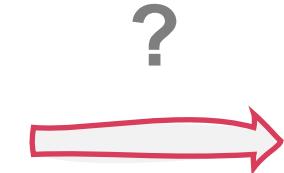
- Supervised DA (SDA)

$$S = \{(X_j^s, Y_j^s)\}_{j=1}^{m_s}, (X_j^s, Y_j^s) \sim \mathcal{S} \quad + \quad T = \{(X_j^t, Y_j^t)\}_{j=1}^{m_t}, (X_j^t, Y_j^t) \sim \mathcal{T}$$

- Unsupervised DA (UDA)

$$S = \{(X_j^s, Y_j^s)\}_{j=1}^{m_s}, (X_j^s, Y_j^s) \sim \mathcal{S} \quad + \quad T_X = \{X_j^t\}_{j=1}^{m_t}, X_j^t \sim \mathcal{T}(X)$$

Similarity



#Q1 How to measure the similarity between domains?

d: \mathcal{H} -divergence, f -divergence, IPM, discrepancy ...

Conditions for successful DA

#Q2 Under what conditions can we achieve successful DA?

- Supervised DA

$$R_{\mathcal{T}}(h) \leq \underbrace{R_{\mathcal{S}}(h) + d(\mathcal{S}, \mathcal{T})}$$

All the terms are small: a sufficient condition

- Unsupervised DA

$$R_{\mathcal{T}}(h) \leq R_{\mathcal{S}}(h) + d(\mathcal{S}(X), \mathcal{T}(X)) + \lambda^* \quad \text{Ideal joint error}$$

Covariate shift $\mathcal{S}(Y|X) = \mathcal{T}(Y|X)$, $\mathcal{S}(X) \neq \mathcal{T}(X)$



$\lambda^* = \min_{h \in \mathcal{H}} [R_{\mathcal{S}}(h) + R_{\mathcal{T}}(h)]$ is small
(necessary in the context of UDA)



$f_s \in \mathcal{H}_s, f_t \in \mathcal{H}_t, f_s \neq f_t$



$f_s = f_t$



$f_s, f_t \notin \mathcal{H}$

[1] David, Shai Ben, et al. "Impossibility theorems for domain adaptation." AISTATS 2010.

Invariant Representation Learning

Spurious correlation



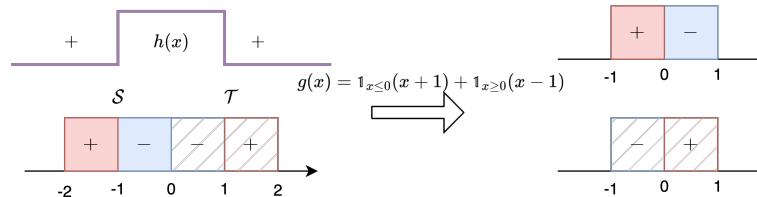
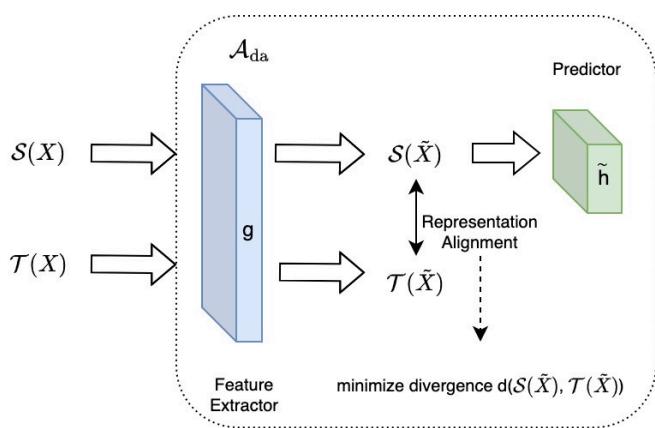
$Y=\{+ \text{ (cat)}, - \text{ (dog)}\}$ X_b – background, $X_{\setminus b}$ – other features

$$\mathcal{X} \xrightarrow{g} \tilde{\mathcal{X}} \xrightarrow{\tilde{h}} \mathcal{Y}$$

$$\exists g(X) = g(X_b, X_{\setminus b}) = X_{\setminus b} \quad \text{s.t. } R_S(h) + R_T(h) \leq \gamma, h = \tilde{h} \circ g \Rightarrow \lambda^* \leq \gamma$$

Q3 Is representation learning always beneficial for DA? No

Fail of Marginal Representation Alignment



After marginal representation alignment,

$\nexists \tilde{h} \circ g$ s.t. $R_{\mathcal{T}}(\tilde{h} \circ g) \leq \gamma$
 even $R_{\mathcal{S}}(\tilde{h} \circ g) < \gamma, d(\mathcal{S}(\tilde{X}), \mathcal{T}(\tilde{X})) = 0, \lambda^* = 0$

#P1 Previous bounds (e.g. [1-4]) cannot be directly used (No representation analysis) !!!

#P2 Marginal alignment cannot remove spurious correlation!!!

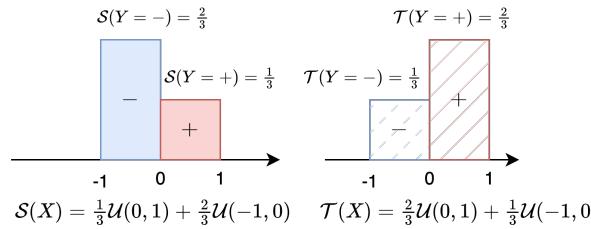
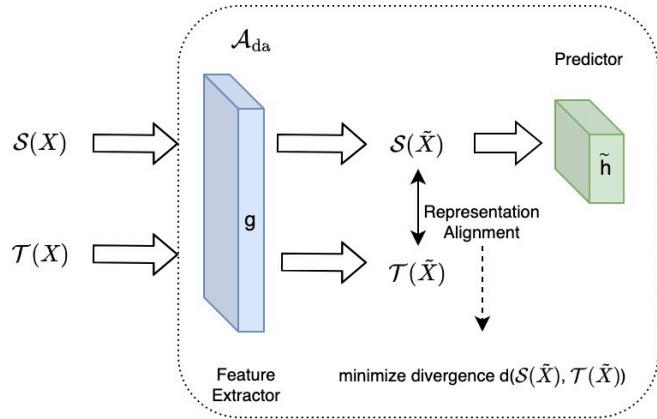
[1] Ben-David, Shai, et al. "A theory of learning from different domains." (ML 2010)

[2] Zhao, et al. "On learning invariant representations for domain adaptation." (ICML 2019)

[3] Zhang, Yuchen, et al. "Bridging theory and algorithm for domain adaptation." (ICML 2019)

[4] Acuna, David, et al. "f-domain adversarial learning: Theory and algorithms." (ICML 2021)

Fail of Marginal Representation Alignment



$$\mathcal{S}(Y) \neq \mathcal{T}(Y)$$

Lower bound ([1,2]):

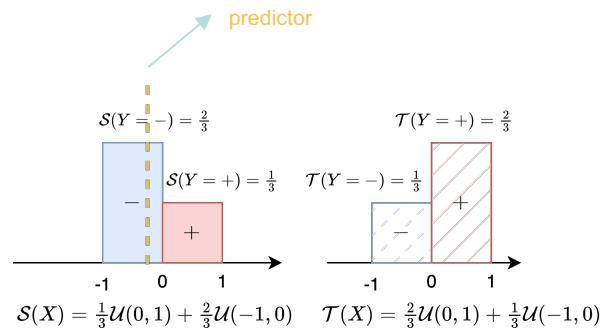
#P3 Under large target shift, $\downarrow d(\mathcal{S}(\tilde{X}), \mathcal{T}(\tilde{X}))$, Negative transfer !!!

$$R_{\mathcal{T}}(\tilde{h} \circ g) \geq \frac{1}{2} \left(\sqrt{D_{JS}(\mathcal{S}(Y) \parallel \mathcal{T}(Y))} - \sqrt{D_{JS}(\mathcal{S}(\tilde{X}) \parallel \mathcal{T}(\tilde{X}))} \right)^2 - R_{\mathcal{S}}(\tilde{h} \circ g) \quad \text{when} \quad \sqrt{D_{JS}(\mathcal{S}(Y) \parallel \mathcal{T}(Y))} \geq \sqrt{D_{JS}(\mathcal{S}(\tilde{X}) \parallel \mathcal{T}(\tilde{X}))}$$

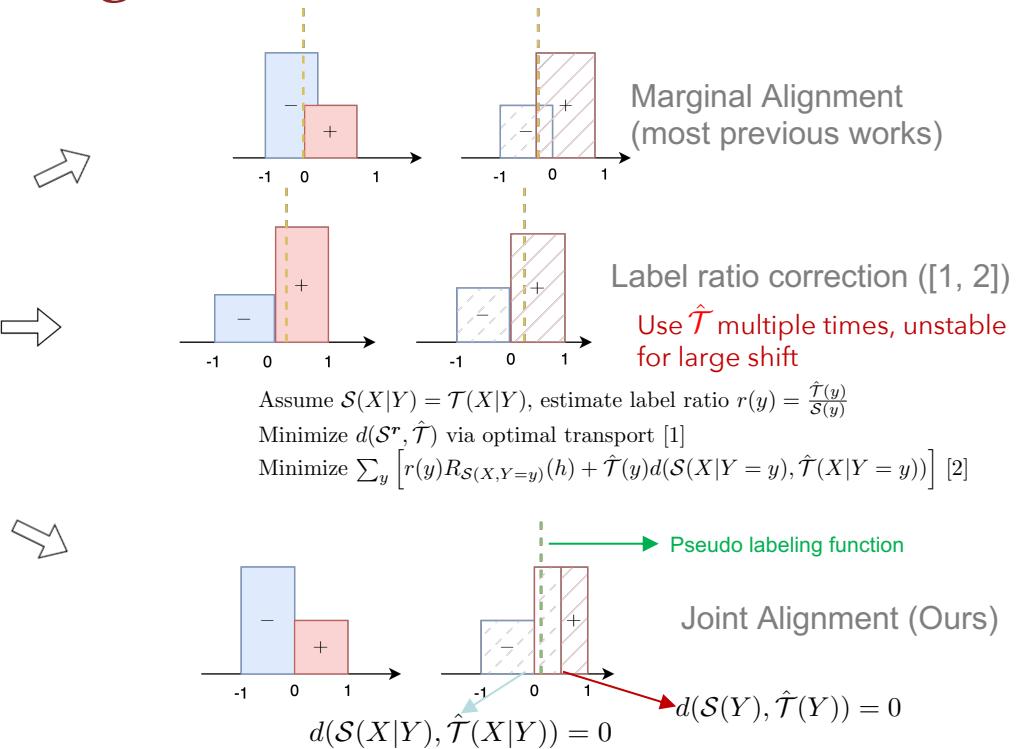
[1] Zhao, et al. "On learning invariant representations for domain adaptation." (ICML 19)

[2] Combes, et al. "Domain adaptation with conditional distribution matching and generalized label shift." (NeurIPS 20)

Joint Representation Alignment



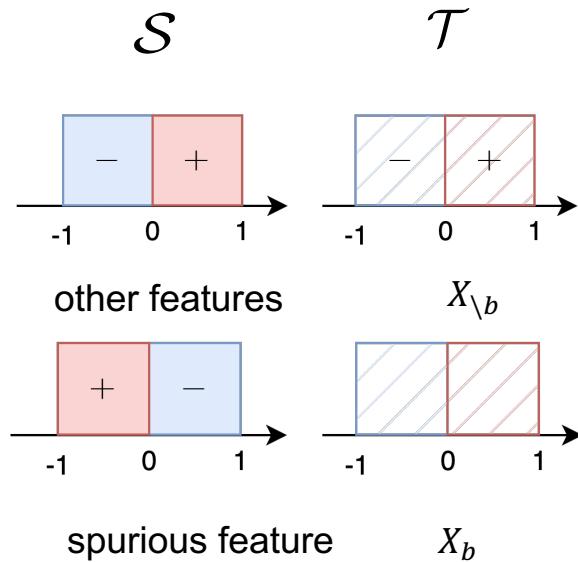
More stable to target shift



[1] Ievgen Redko et al. "Optimal transport for multi-source domain adaptation under target shift". (AISTATS 19)

[2] Changjian Shui et al. "Aggregating from multiple target-shifted sources". In: International Conference on Machine Learning. (ICML 21)

Joint Representation Alignment



$$d(\mathcal{S}(X_{\setminus b}, Y), \hat{\mathcal{T}}(X_{\setminus b}, Y)) = d(\mathcal{S}(X_{\setminus b}, Y), \mathcal{T}(X_{\setminus b}, Y)) = 0$$

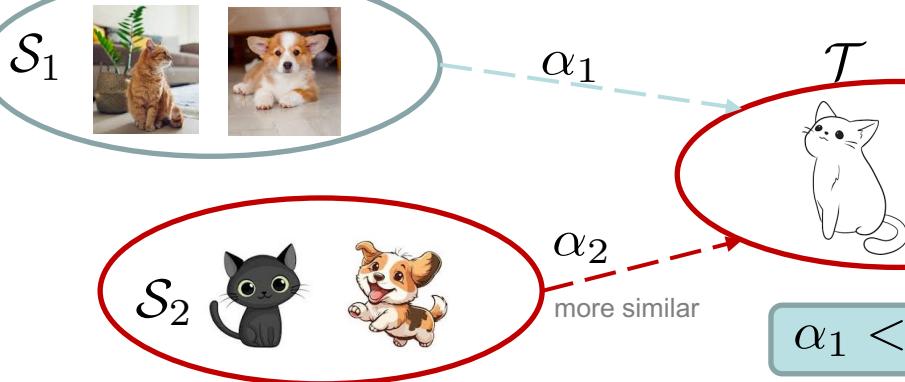
Pseudo label is credible

$$\begin{aligned} d(\mathcal{S}(X_b), \mathcal{T}(X_b)) &= 0 \\ d(\mathcal{S}(X_b, Y), \hat{\mathcal{T}}(X_b, Y)) &\text{ is large} \end{aligned}$$

Joint alignment can remove spurious feature

Multi-source Domain Adaptation (MDA)

Select most relevant sources via similarity



Combined source distribution:

$$\mathcal{S}^\alpha := \sum_{i=1}^N \alpha_i \mathcal{S}_i, \Delta_N := \{\alpha : \alpha_i \geq 0, \sum_{i=1}^N \alpha_i = 1\}$$

Target and pseudo target distributions:

$$\mathcal{T} \text{ and } \hat{\mathcal{T}}$$

Joint alignment for MDA

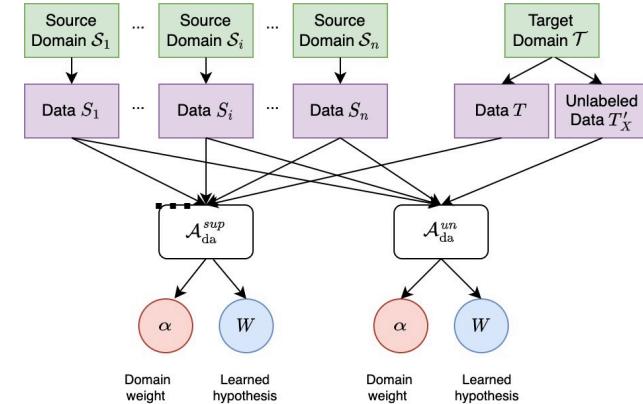
$$\min_{\alpha} d(\mathcal{S}^\alpha, \hat{\mathcal{T}})$$

Algorithm-Dependent Bounds for Representation Learning of Multi-Source Domain Adaptation. AISTATS 2023.

Previous works:

$$\sum_{i=1}^N \alpha_i [R_{\mathcal{S}_i}(h) + d(\mathcal{S}_i(X), \mathcal{T}(X))] \quad \text{Zhao (2018), Wen (2020) ...}$$

$$\sum_{i=1}^N \alpha_i \left[\sum_y r_i(y) R_{\mathcal{S}_i(X, Y=y)}(h) + \hat{\mathcal{T}}(y) d(\mathcal{S}_i(X|Y=y), \mathcal{T}(X|Y=y)) \right] \quad \text{Shui (2021)}$$

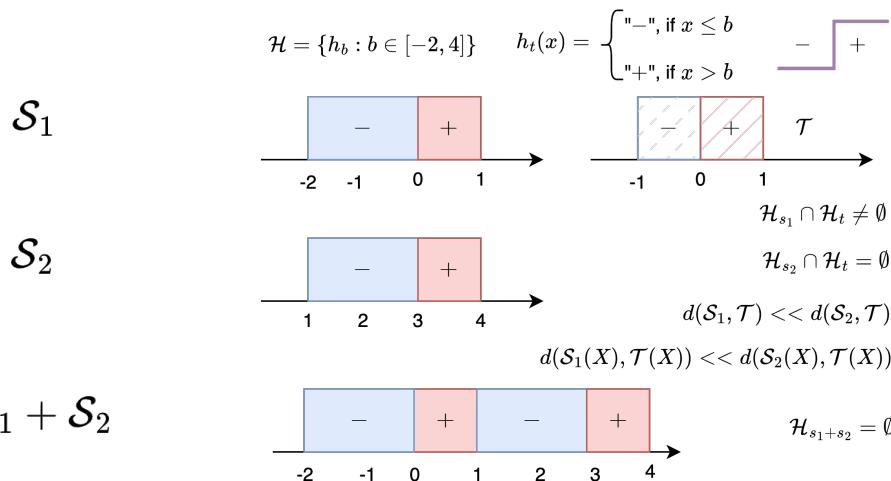


Multi-source Domain Adaptation (MDA)

Q4 Why is domain weight relevant for MDA?

Q5 Why not adapt from an equally mixed source?

Let $\mathcal{H}_{s_1} := \{h \in \mathcal{H} : R_{\mathcal{S}_1}(h) \leq \gamma\}$, $\mathcal{H}_{s_2} := \{h \in \mathcal{H} : R_{\mathcal{S}_2}(h) \leq \gamma\}$
and $\mathcal{H}_t := \{h \in \mathcal{H} : R_{\mathcal{T}}(h) \leq \gamma\}$



Successful adaptation with

$$\alpha_1 = 1, \alpha_2 = 0$$

We can also construct

$$\mathcal{H}_{s_1} \cap \mathcal{H}_t \neq \emptyset, \mathcal{H}_{s_2} \cap \mathcal{H}_t \neq \emptyset$$

However, $\mathcal{H}_{s_1} \cap \mathcal{H}_{s_2} \cap \mathcal{H}_t = \emptyset$

Theoretical Results (Simplified): Unsupervised MDA

$h \circ g$ pseudo labeling function (parametrized by u, v)

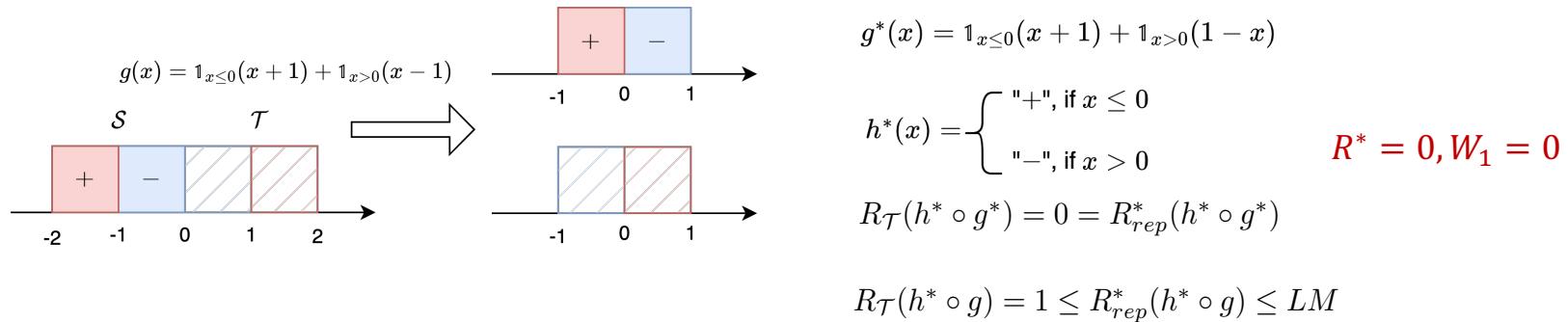
$$R_{\mathcal{T}}(u, v) \leq \mathbf{W}_1(\tilde{\mathcal{T}}_{u,v}, \tilde{\mathcal{S}}_u^\alpha) + R_{rep}^*(u, v) + R^*$$

$$R^* = R_{\mathcal{S}^\alpha}(u^*, v^*) + R_{\mathcal{T}}(u^*, v^*)$$

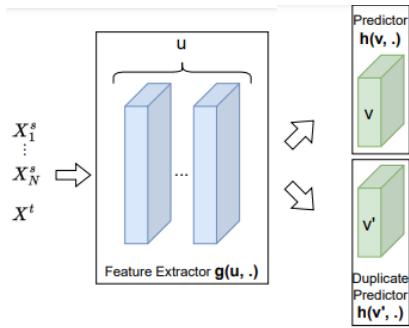
$$u^*, v^* = \arg \min_{u,v} R_{\mathcal{S}^\alpha}(u, v) + R_{\mathcal{T}}(u, v)$$

$$R_{rep}^*(u, v) \stackrel{\text{def}}{=} LM \int_{\mathcal{Z} \times \mathcal{Z}} [\rho_{\tilde{x}}(g(u^*, x), g(u^*, x')) - \rho_{\tilde{x}}(g(u, x), g(u, x'))] d\gamma_{u,v}^*(\hat{z}, z')$$

Lipschitz constants



Theoretical Results (Simplified): Unsupervised MDA



$$\hat{R}_{\mathcal{S}^\alpha}(u, \mathbf{v}') \stackrel{\text{def}}{=} \sum_{i=1}^N \frac{\alpha_i}{m_i} \sum_{j=1}^{m_i} \ell(h(\mathbf{v}', g(u, X_{i,j}^s)), Y_{i,j}^s)$$

$$\hat{R}_{\mathcal{T}_{u,v}}(u, v, v') \stackrel{\text{def}}{=} \frac{1}{m_t} \sum_{j=1}^{m_t} \ell(h(v', g(u, X_j^t)), h(v, g(u, X_j^t)))$$

Disagreement

$$U, V = \mathcal{A}_{un}(S^\alpha, T_X) \quad \alpha \text{ will be optimized}$$

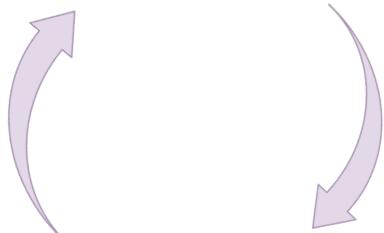
Expected target true risk

$$\mathbb{E}_{U,V,S^\alpha,T}[R_{\mathcal{T}}(U, V)] \leq \mathbb{E}_{U,V,S^\alpha,T} [\hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^\alpha)] + \sqrt{2\sigma'^2 \left(\sum_{i=1}^N \frac{\alpha_i^2}{m_i} + \frac{1}{m'_t} \right) I(U, V; S^\alpha, T'_X)} + \mathbb{E}_{U,V} R_{rep}^*(U, V) + R^*$$

- Mitigate target shift #P3
- Remove spurious correlation #P2
- Generalization for Few-shot Adaptation
- Equal importance $\alpha_i = \frac{1}{N}$, sample size $m_i = m$
- $\frac{1}{m'_t} \ll \sum_{i=1}^N \frac{\alpha_i^2}{m_i}$, $\mathcal{O}\left(\sqrt{\sum_{i=1}^N \frac{\alpha_i^2}{m_i}}\right) = \mathcal{O}\left(\frac{1}{\sqrt{Nm}}\right)$
- few-shot adaptation with $N \rightarrow \infty$
- Theoretic gap: suboptimal representation #P1

IMDA Algorithm (part for unsupervised MDA)

① fix α , $\min_{u,v} \max_{v'} [\hat{R}_{\mathcal{T}_{u,v}}(u, v, v') - \hat{R}_{\mathcal{S}^\alpha}(u, v')]$



② fix u, v optimize the domain weights α , $\forall C_0, C_1 > 0$:

$$\min_{\alpha} \left(C_0 \hat{R}_{\mathcal{S}^\alpha}(u, v) - \hat{R}_{\mathcal{S}^\alpha}(u, v') + C_1 \sqrt{\delta_u + \delta_v} R(\alpha) \right),$$

$$R(\alpha) = \sqrt{\sum_{i=1}^N \frac{\alpha_i^2}{m_i}}, \text{ s.t. } \forall i \in [N], \alpha_i \geq 0, \sum_{i=1}^N \alpha_i = 1,$$

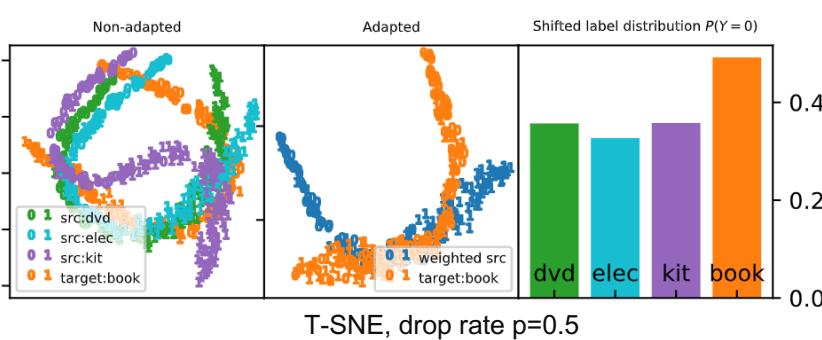
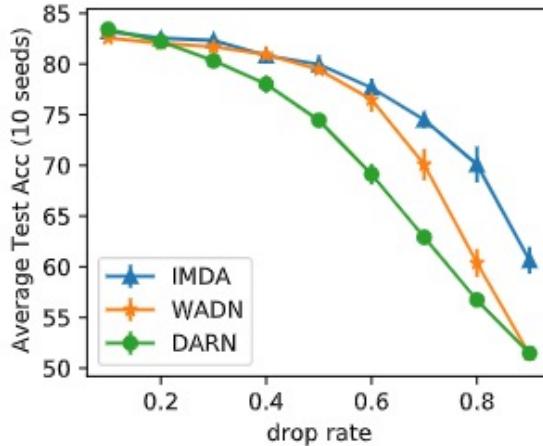
Approximation of

$$R_{\mathcal{S}^\alpha}(u^*, v^*)$$

Accumulate gradient norm for u, v
(gradient norm estimation of mutual information term
for SGLD, omitted in presentation)

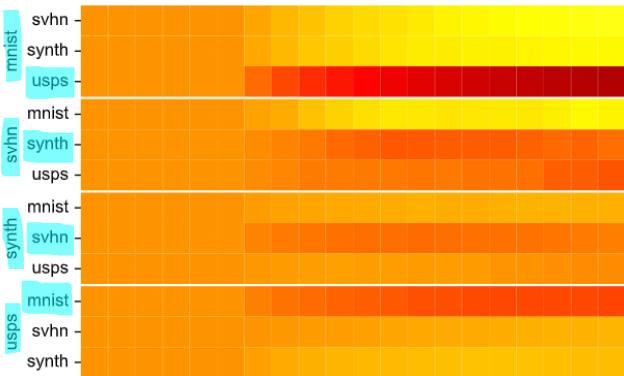
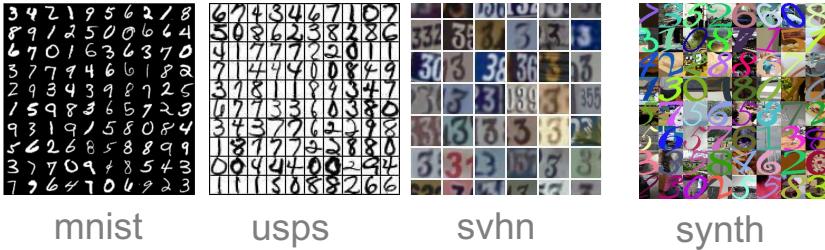
Experimental Results (Amazon review)

IMDA(Ours, joint alignment)
WADN(label ratio correction)
DARN(marginal alignment)

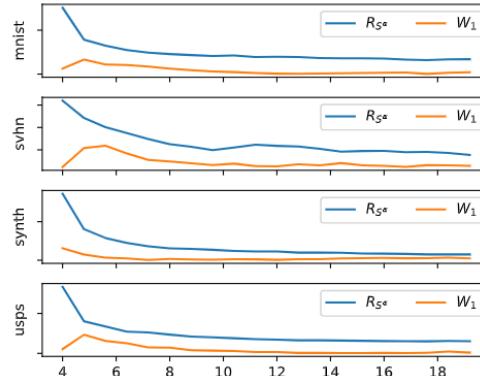


Randomly drop (p)% negative reviews ($Y=0$) in source domain

Experimental Results (Digits)



(a) Evolution of Domains Weights α



(b) Evolution of $\hat{R}_{S^\alpha}(u, v)$ and $\hat{W}_1(\tilde{T}_{u,v}, \tilde{\mathcal{S}}_u)$

Summary of Contribution

- Unified approach for both **supervised** and **unsupervised** MDA
 - Formal theoretical analysis on **representation learning** of MDA
In contrast to [1], we provide explicit bounds.
 - Mitigate target shift through **joint alignment**
 - **Memory efficient** algorithm
- 1 duplicate predictor VS N domain discriminator and $N|\mathcal{Y}|$ centroids [2]
- **Fully algorithm-dependent bounds** via information-theoretic learning

The bounds in [3,4,5] contain non-optimizable KL divergence, where the generalization gap of KL is not algorithm-dependent.

[1] Zhao, et al. "On learning invariant representations for domain adaptation." (ICML 2019)

[2] Changjian Shui et al. "Aggregating from multiple target-shifted sources". In: International Conference on Machine Learning." (ICML 21)

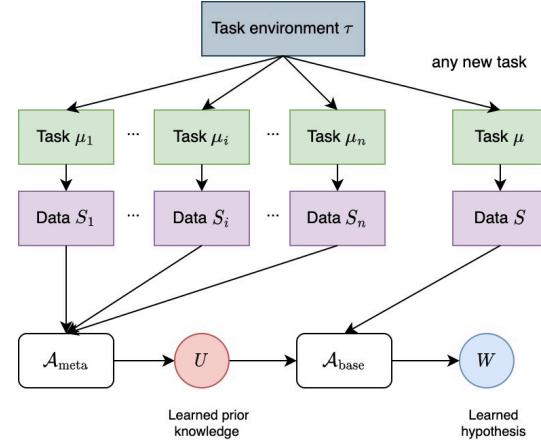
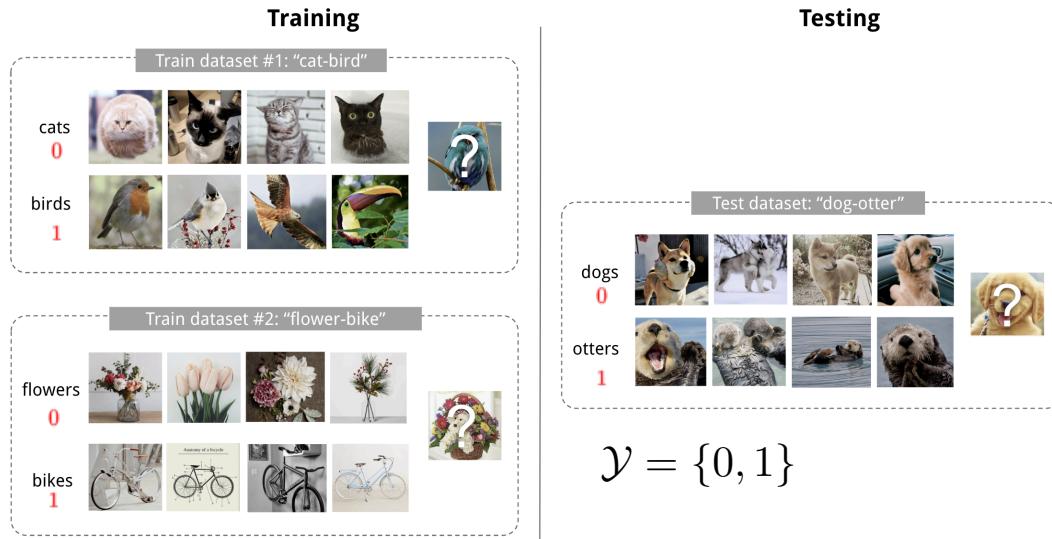
[3] Wu, Xuetong, Jonathan H Manton, Uwe Aickelin, and Jingge Zhu. "An Information-Theoretic Analysis for Transfer Learning: Error Bounds and Applications" [2022]

[4] Wu, Xuetong, Jonathan H Manton, Uwe Aickelin, and Jingge Zhu. "Information-theoretic analysis for transfer learning". [ISIT 2020]

[5] Wang, Ziqiao, and Yongyi Mao. "Information-Theoretic Analysis of Unsupervised Domain Adaptation." [ICLR 2023]

Closer to real intelligence: Learning to Learn (Meta-Learning)

Meta-learning/Learning to Learn



Problem formulation

Figure source: <https://lilianweng.github.io/posts/2018-11-30-meta-learning/>

Learn higher level knowledge: how to classify objects

Joint-training

- Meta learner $\mathcal{A}_{\text{meta}}$: $U = \mathcal{A}_{\text{meta}}(S_{1:n}) \sim P_{U|S_{1:n}}$
- Base learner $\mathcal{A}_{\text{base}}$: $W = \mathcal{A}_{\text{base}}(U, S) \sim P_{W|U,S}$

True meta risk:

$$R_\tau(U) \stackrel{\text{def}}{=} \mathbb{E}_{S \sim \mu_m, \tau} \mathbb{E}_{W \sim P_{W|S,U}} [R_\mu(W)] = \mathbb{E}_{\mu \sim \tau} \mathbb{E}_{S \sim \mu^m} \mathbb{E}_{W \sim P_{W|S,U}} [R_\mu(W)]$$

Empirical meta risk:

$$R_{S_{1:n}}(U) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W_i \sim P_{W_i|S_i,U}} [R_{S_i}(W_i)]$$

Meta generalization gap:

$$\begin{aligned} |\text{gen}_{\text{meta}}^{\text{joi}}(\tau, \mathcal{A}_{\text{meta}}, \mathcal{A}_{\text{base}})| &\stackrel{\text{def}}{=} \mathbb{E}_{U, S_{1:n}} [R_\tau(U) - R_{S_{1:n}}(U)]_{\text{environment-level uncertainty}} \\ &\leq \sqrt{\frac{2\sigma^2}{nm} I(U, W_{1:n}; S_{1:n})} \leq \underbrace{\sqrt{\frac{2\sigma^2}{mn} I(U; S_{1:n})}}_{T_1} + \underbrace{\sqrt{\frac{2\sigma^2}{mn} \sum_{i=1}^n I(W_i; S_i|U)}}_{T_2} \end{aligned}$$

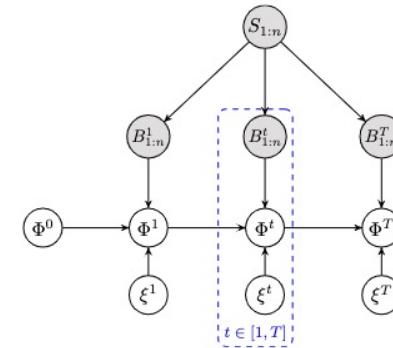
Assume bounded MI, improved rate over [1]

$$\mathcal{O}\left(\frac{1}{\sqrt{nm}} + \frac{1}{\sqrt{m}}\right) \text{ VS. } \mathcal{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)$$

- Few-shot learning $n \rightarrow \infty, m$ small: $T_1 \rightarrow 0, T_2 > 0$ Biased estimation, consistent with [2]
- No need of transfer $m \rightarrow \infty, n$ is finite: $T_1 \rightarrow 0, T_2 \rightarrow 0$

[1] Amit, Ron and Ron Meir. "Meta-learning by adjusting priors based on extended PAC-Bayes theory" (ICML 2018)

[2] Bai, Yu, Minshuo Chen, Pan Zhou, Tuo Zhao, Jason Lee, Sham Kakade, Huan Wang, and Caiming Xiong. "How Important is the Train-Validation Split in Meta-Learning?" (ICML 2021)



Jointly optimize $\Phi = (U, W_{1:n})$ [1]

Alternate-training

- Meta learner $\mathcal{A}_{\text{meta}}$: $U = \mathcal{A}_{\text{meta}}(S_{1:n}) \sim P_{U|S_{1:n}}$
- Base learner $\mathcal{A}_{\text{base}}$: $W = \mathcal{A}_{\text{base}}(U, S^{tr}) \sim P_{W|U, S^{tr}}$

True meta risk:

$$R_\tau(U) \stackrel{\text{def}}{=} \mathbb{E}_{S \sim \mu_m, \tau} \mathbb{E}_{W \sim P_{W|S^{tr}, U}} [R_\mu(W)] = \mathbb{E}_{\mu \sim \tau} \mathbb{E}_{S | \mu \sim \mu^m} \mathbb{E}_{W \sim P_{W|S^{tr}, U}} [R_\mu(W)]$$

Empirical meta risk:

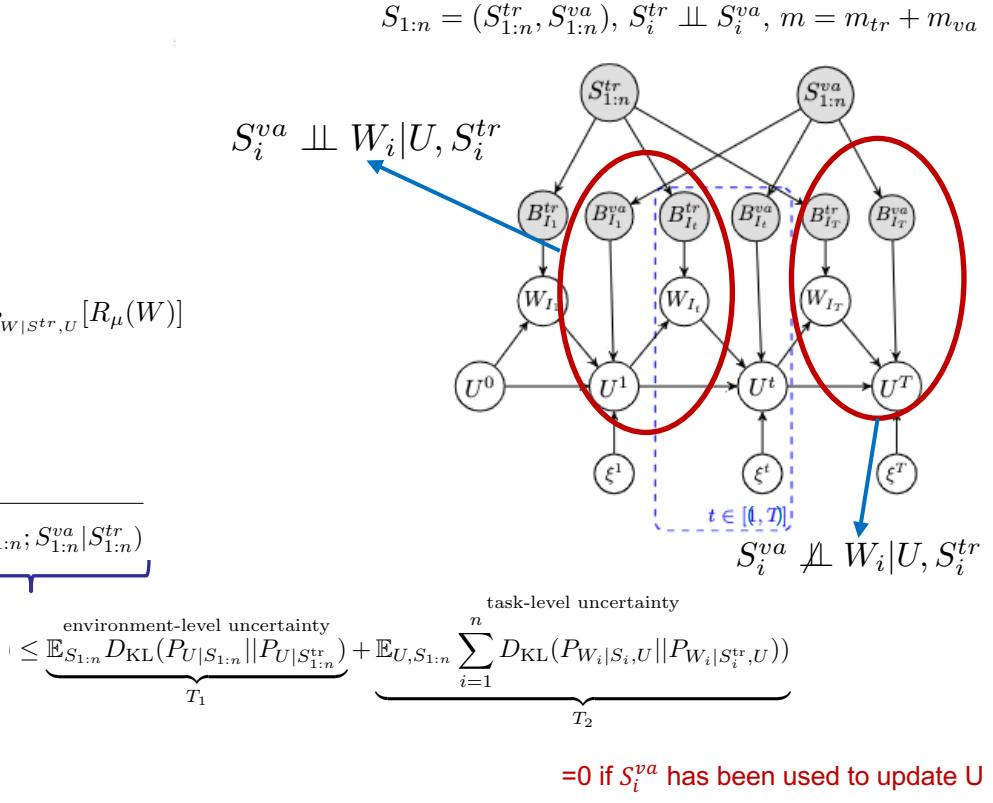
$$\tilde{R}_{S_{1:n}}(U) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W_i \sim P_{W_i|S_i^{tr}, U}} [R_{S_i^{va}}(W_i)]$$

Meta generalization gap:

$$\text{gen}_{\text{meta}}^{\text{alt}}(\tau, \mathcal{A}_{\text{meta}}, \mathcal{A}_{\text{base}}) \stackrel{\text{def}}{=} \mathbb{E}_{U, S_{1:n}} [R_\tau(U) - \tilde{R}_{S_{1:n}}(U)] \leq \sqrt{\frac{2\sigma^2}{nm^{va}} I(U, W_{1:n}; S_{1:n}^{va} | S_{1:n}^{tr})}$$

Train-validation split trade-off, $\uparrow m_{va} \iff \downarrow m_{tr}$:

- $\downarrow \frac{2\sigma^2}{nm^{va}}$
- $\uparrow D_{\text{KL}}(P_{W_i|S_i, U} || P_{W_i|S_i^{tr}, U})$ Similar form to chaser loss [1]
- $\uparrow D_{\text{KL}}(P_{U|S_{1:n}} || P_{U|S_{1:n}^{tr}})$



[1] Yoon, Jaesik, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. "Bayesian model-agnostic meta-learning". (NeurIPS 2018)

Non-vacuous bound for few-shot learning

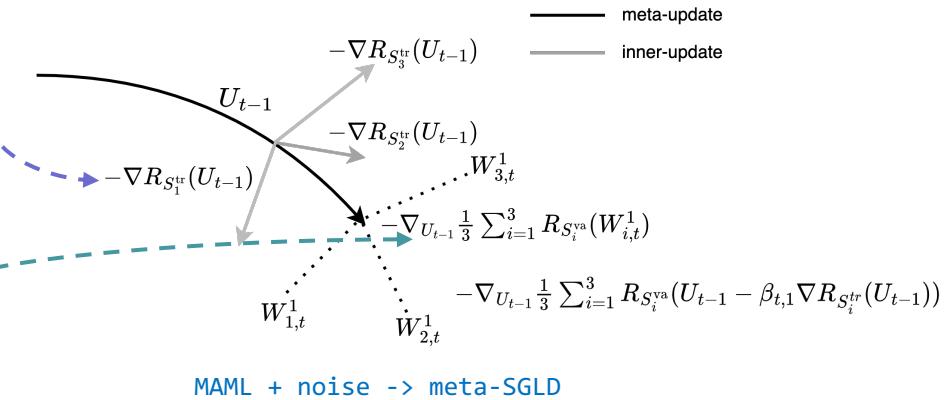
Inner update:

$$W_{i,t}^0 = U^{t-1}$$

$$W_{i,t}^k = W_{i,t}^{k-1} - \beta_{t,k} \nabla R_{B_{i,t,k}^{\text{tr}}}(W_{i,t}^{k-1}) + \zeta^{t,k}, \zeta^{t,k} \sim \mathcal{N}(\mathbf{0}, \frac{2\beta_{t,k}}{\gamma_{t,k}} \mathbb{I}_d)$$

meta update:

$$U^t = U^{t-1} - \eta_t \nabla \tilde{R}_{B_{I_t}^{\text{ya}}}(U^{t-1}) + \xi_t, \xi_t \sim \mathcal{N}(\mathbf{0}, \frac{2\eta_t}{\gamma_t} \mathbb{I}_d)$$



Bound through Gradient Estimation:

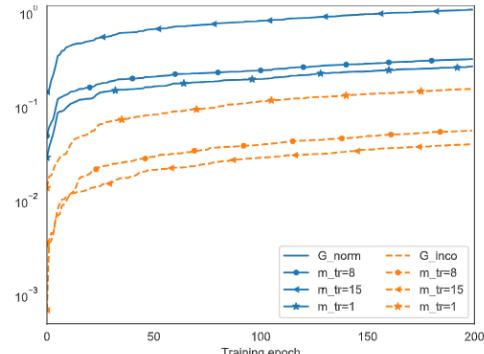
$$|\text{gen}_{\text{meta}}^{\text{alt}}(\tau, \text{SGLD}, \text{SGLD})| \leq \sqrt{\frac{2\sigma^2 I(U, W_{1:n}; S_{1:n}^{\text{va}} | S_{1:n}^{\text{tr}})}{nm_{\text{va}}}} \leq \frac{\sigma}{\sqrt{nm_{\text{va}}}} \sqrt{\sum_{t=1}^T \mathbb{E}_{B_{I_t}^{\text{va}}, B_{I_t}^{\text{tr}}, W_{I_t}, U^{t-1}} \frac{\eta_t \gamma_t \|\epsilon_t^u\|_2^2}{2} + \sum_{t=1}^T \sum_{i=1}^{|I_t|} \sum_{k=1}^K \mathbb{E}_{B_{i,t,k}^{\text{va}}, B_{i,t,k}^{\text{tr}}, W_{i,t}^{k-1}} \frac{\beta_{t,k} \gamma_{t,k} \|\epsilon_{t,i,k}^w\|_2^2}{2}}$$

Gradient incoherence: $\epsilon_t^u \stackrel{\text{def}}{=} \nabla \tilde{R}_{B_{I_t}}(U^{t-1}) - \nabla \tilde{R}_{B_{I_t}^{\text{tr}}}(U^{t-1})$

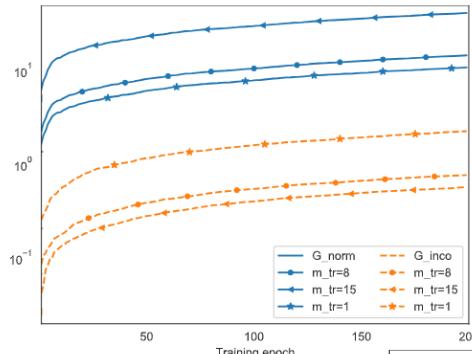
Gradient norm: $\epsilon_t^u = \nabla \tilde{R}_{B_{I_t}^{\text{ya}}}(U^{t-1})$

$\epsilon_{t,i,k}^w \stackrel{\text{def}}{=} \nabla R_{B_{i,t,k}}(W_{i,t}^{k-1}) - \nabla R_{B_{i,t,k}^{\text{tr}}}(W_{i,t}^{k-1})$

Non-vacuous bound for few-shot learning



(a) Bound of U

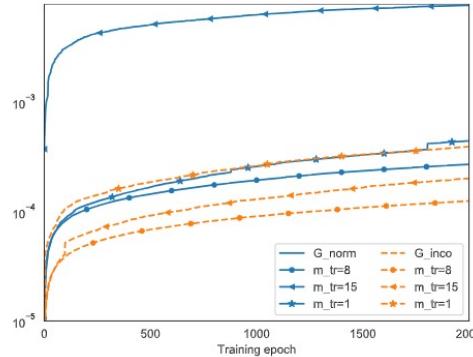


(b) Bound of W

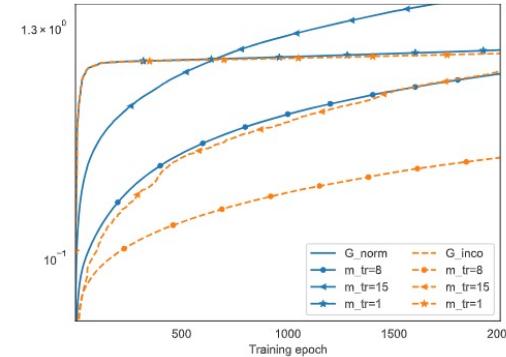
$$m_{tr} + m_{va} = 16$$

Omniglot →

- $m_{tr} = 15$, $G_{norm} \gg G_{inco}$
- $m_{tr} = 8$, tightest bound among three settings, optimal split
- $m_{tr} = 1$, $G_{norm} \approx G_{inco}$



(a) Bound of U



(b) Bound of W

← 2D Gaussian

Consistent with [1]

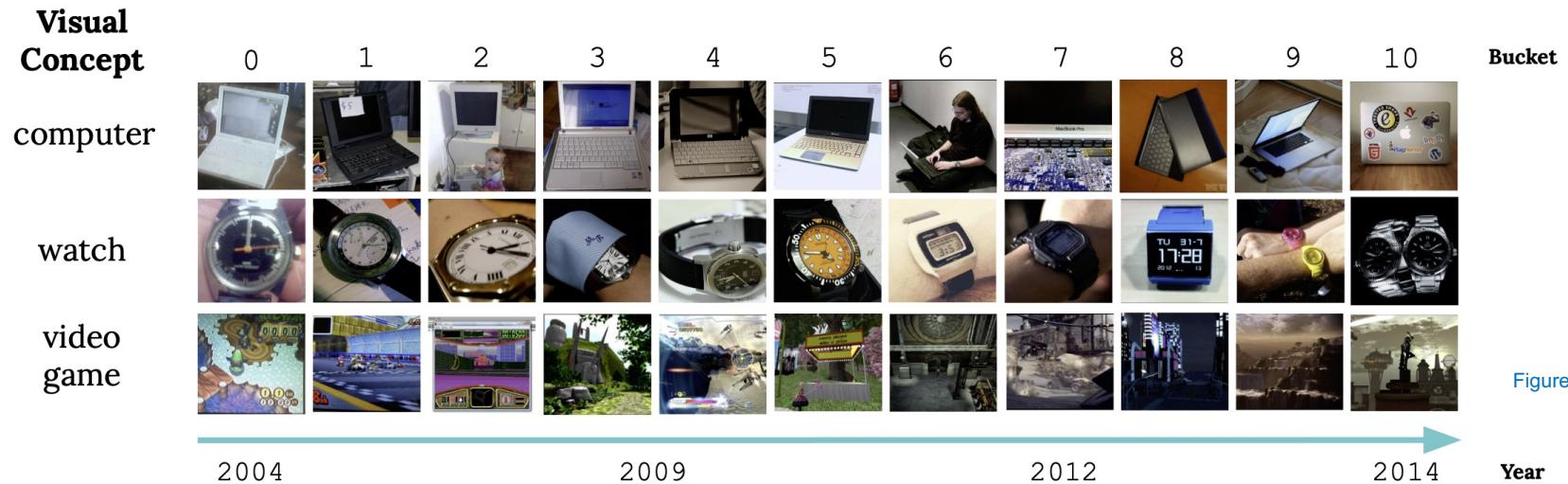
- $m_{tr} = 15$: tightest bound, largest G_{norm}/G_{inco} gap, optimal split
- $m_{tr} = 1$ (1-shot learning): poorest estimation, largest G_{inco} bound,

Summary of Contribution

- Unified theoretical analysis
 - analysis for both **joint training and alternate training**
 - interpretable theoretical results
- Flexible bounds
 - **data**-dependent, **algorithm**-dependent
 - valid for **non-convex loss**, not constrained in **linear** models
- Non-vacuous bound for gradient-based few-shot learning
 - orders of **magnitude tighter** in most situations
 - empirically validated the **train-validation split trade-off**

A More Challenging Scenario: Continual Meta-Learning

The World is Changing...



2004

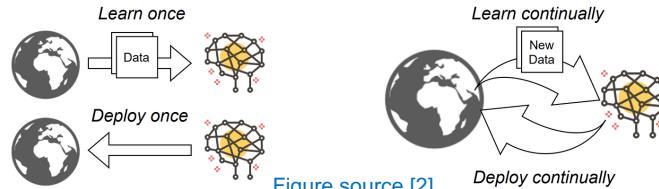
2009

2012

2014

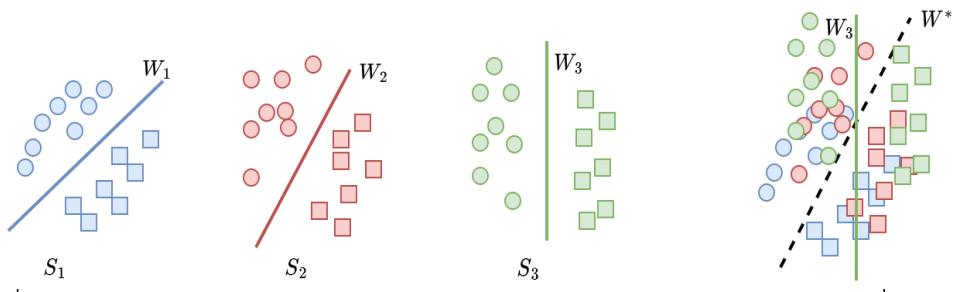
Static ML

Adaptive ML



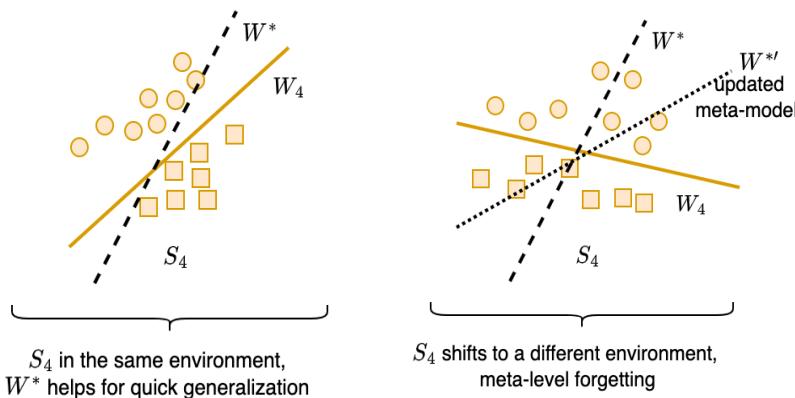
- [1] Lin, Zhiqiu, et al. "The clear benchmark: Continual learning on real-world imagery." NeurIPS dataset (2021).
[2] <https://ai.kuleuven.be/stories/post/2021-05-10-continual-learning/>

Bi-level Learning-forgetting Trade-off



Catastrophic forgetting

- negative backward transfer

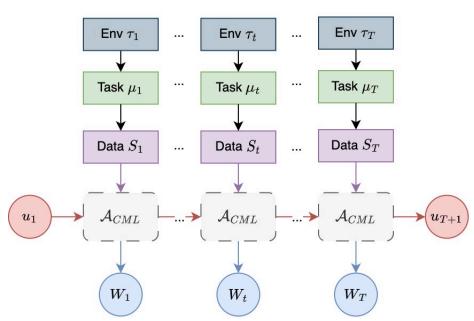


Quick generalization

- positive forward transfer

Bi-level learning-forgetting trade-off...

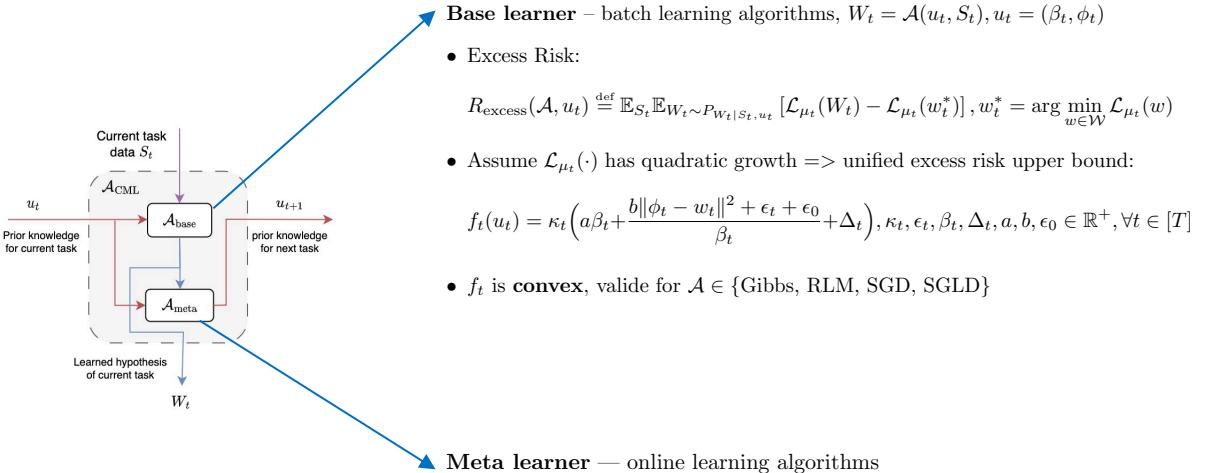
Continual Meta-Learning (CML)



Continual meta-learning objective

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

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



Base learner — batch learning algorithms, $W_t = \mathcal{A}(u_t, S_t), u_t = (\beta_t, \phi_t)$

- Excess Risk:

$$R_{\text{excess}}(\mathcal{A}, u_t) \stackrel{\text{def}}{=} \mathbb{E}_{S_t} \mathbb{E}_{W_t \sim P_{W_t|S_t,u_t}} [\mathcal{L}_{\mu_t}(W_t) - \mathcal{L}_{\mu_t}(w_t^*)], w_t^* = \arg \min_{w \in \mathcal{W}} \mathcal{L}_{\mu_t}(w)$$

- Assume $\mathcal{L}_{\mu_t}(\cdot)$ has quadratic growth => unified excess risk upper bound:

$$f_t(u_t) = \kappa_t \left(a\beta_t + \frac{b\|\phi_t - w_t\|^2 + \epsilon_t + \epsilon_0}{\beta_t} + \Delta_t \right), \kappa_t, \epsilon_t, \beta_t, \Delta_t, a, b, \epsilon_0 \in \mathbb{R}^+, \forall t \in [T]$$

- f_t is **convex**, valide for $\mathcal{A} \in \{\text{Gibbs, RLM, SGD, SGLD}\}$

Meta learner — online learning algorithms

- dynamic regret for N static slots

$$R_T^{\text{dynamic}}(u_{1:N}^*) \stackrel{\text{def}}{=} \sum_{n=1}^N \sum_{k=1}^{M_n} [f_{n,k}(u_{n,k}) - f_{n,k}(u_n^*)], u_n^* \stackrel{\text{def}}{=} \arg \min_u \frac{1}{M_n} \sum_{k=1}^{M_n} f_{n,k}(u)$$

Dynamic CML (DCML)

K-step SGD

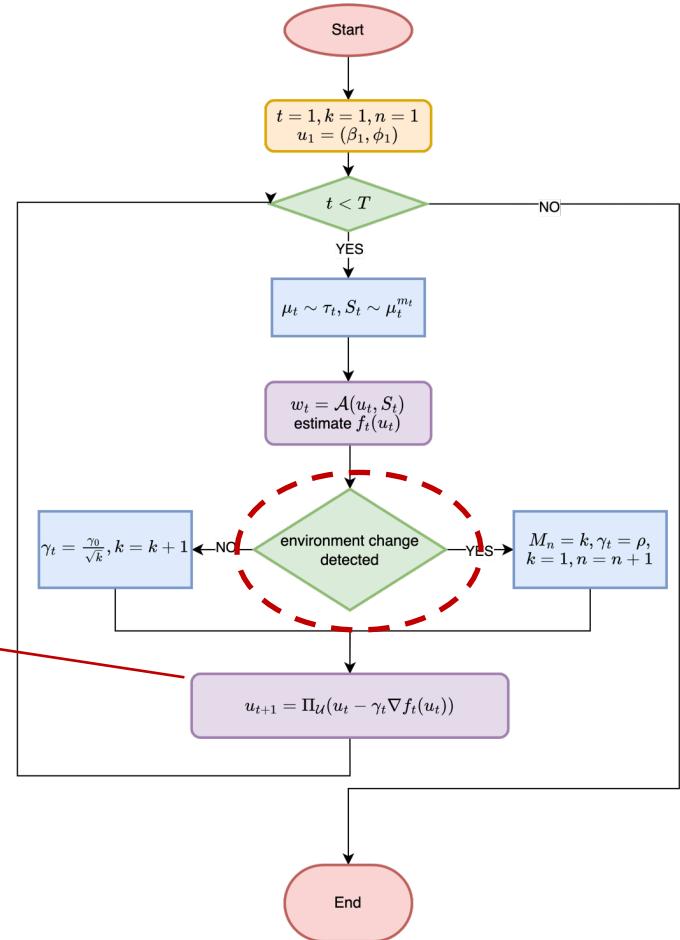
$$\kappa_t = \sqrt{\frac{1}{K\alpha} + \frac{2}{m_t}\alpha L}$$

control meta level trade-off

$$\phi_{t+1} = (1 - \frac{2b\kappa_t\gamma_t}{\beta_t})\phi_t + \frac{2b\kappa_t\gamma_t}{\beta_t}w_t$$

$$\beta_{t+1} = \beta_t - \gamma_t(a\kappa_t - \frac{\kappa_t(b\|\phi_t - w_t\|^2 + \epsilon_t + \epsilon_0)}{\beta_t^2})$$

control task level trade-off



AER Bound for DCML

Meta-param (k-th task in n-th slot): $u_{n,k} = (\beta_{n,k}, \phi_{n,k}) \in \hat{\mathcal{U}}_n \stackrel{\text{def}}{=} \hat{\mathcal{W}}_n \times \hat{\mathcal{B}}_n$

Slot optimal $\phi_n^* = \sum_{k=1}^{M_n} \frac{\kappa_{n,k}}{\kappa_n} w_{n,k}$

Slot variance $V_n^2 = \sum_{k=1}^{M_n} \frac{\kappa_{n,k}}{\kappa_n} \|\phi_n^* - w_{n,k}\|_2^2$

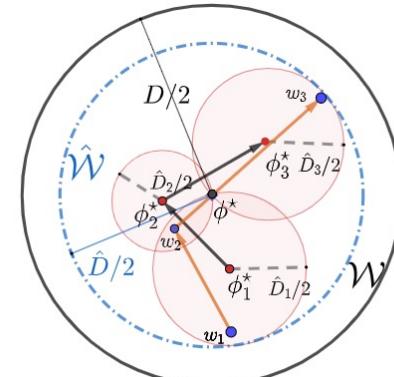
max gradient norm
of f in n-th slot

$$P^* = \sum_{n=1}^{N-1} \|u_n^* - u_{n+1}^*\| + 1$$

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

diameter of $\hat{\mathcal{U}}_n$
Intra-slot task similarity

path length
Inter-slot task (environment) similarity
+ non-stationarity



AER Bound for DCML

Gibbs base learner:

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

- Single-task learning $\mathcal{O}((D+1)m^{-1/4})$
- 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 $\mathcal{O}(1/\sqrt{M})$
- $\bar{V} \leq V \leq \hat{D} \leq D$

SGD base learner:

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

● 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 $\mathcal{O}(\bar{V} + \frac{1}{\sqrt{M}} + \sqrt{\frac{P^*}{NM}})$ [1]

[1] Khodak et.al . “Adaptive gradient-based meta-learning methods”. (NeurIPS 2019.)

Experimental Setting

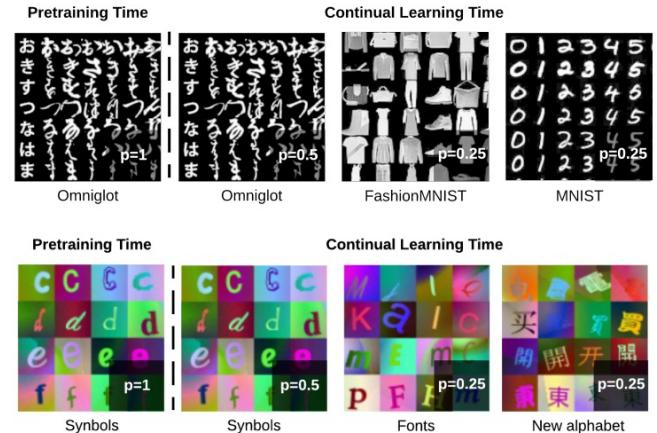
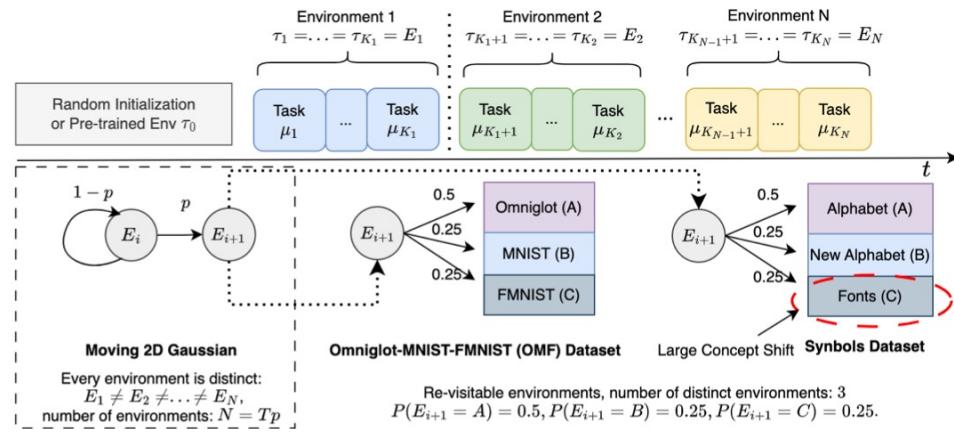
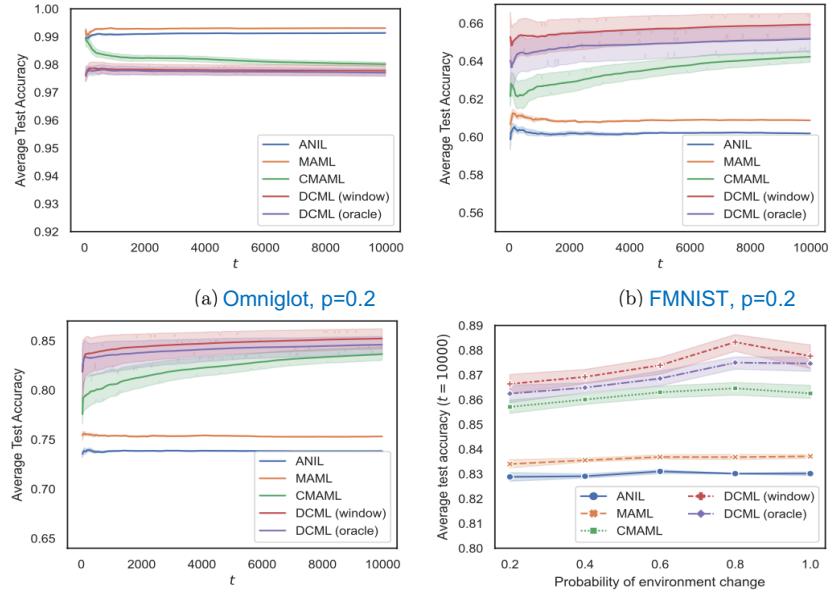


Figure source [1]

[1] Caccia, Massimo, et al. "Online fast adaptation and knowledge accumulation (osaka): a new approach to continual learning". NeurIPS (2020)

Experimental Results

- better overall learning-forgetting trade-off
- stable to different levels of non-stationarity



MODEL	SYMBOLS, $p = 0.2$			
	ALL ENV.	ALPHA.	NEW ALPHA.	FONT
FINE TUNING	25.3 ± 0.7	25.4 ± 0.9	25.1 ± 0.4	25.2 ± 0.5
METACOG [9]	25.3 ± 0.8	25.7 ± 0.9	25.0 ± 0.8	25.1 ± 1.0
METABGD [9]	29.9 ± 6.9	31.8 ± 9.6	28.1 ± 4.8	27.5 ± 3.6
ANIL [4]	60.5 ± 0.4	77.6 ± 0.6	53.6 ± 0.2	32.7 ± 0.4
MAML [4]	76.7 ± 0.4	96.8 ± 0.1	73.1 ± 0.1	40.5 ± 0.7
CMAML [43]	61.9 ± 2.5	76.2 ± 2.1	56.8 ± 3.2	37.9 ± 2.4
DCML(ORACLE)	77.2 ± 0.5	96.0 ± 0.3	73.4 ± 0.1	42.4 ± 0.6
DCML(WINDOW)	77.4 ± 0.4	96.2 ± 0.2	73.8 ± 0.2	42.4 ± 0.7

catastrophic forgetting

use variance to control forgetting

no learning, no forgetting

tend to track change

better trade-off

(c) MNIST, $p=0.2$

(d) All envs

Large concept shift

Summary of Contribution

- Unified theoretical framework for continual meta-learning
valid for both **static and shifting** task environments
- Formal analysis of the **bi-level learning-forgetting trade-off**
- Theoretically grounded algorithm
 - improved rate** in bounds
 - improved empirical performance** on real datasets

Conclusion

Limitation and Future Work

- Limitation
 - 1. For MDA, the **pseudo-labeling** process will fail under a **large concept shift**, and there possibly exists **error accumulation**.
 - 2. For Meta-learning, we only analyzed the generalization gap, not **excess risk**.
 - 3. For CML, only a **single meta-model** was considered.
- Future work
 - 1. Investigate more on pseudo labeling and its relation to **noisy labels**.
 - 2. Prove **general excess risk** bound for meta-learning with non-convex loss.
 - 3. Explore the **theoretical limits of memory-based** CML.

Thanks!