Meta-Learning with Versatile Loss Geometries for Fast Adaptation Using Mirror Descent

Yilang Zhang, Bingcong Li, Georgios B. Giannakis
Department of ECE, University of Minnesota

Motivating context

Challenge in deep learning: large-scale model vs limited training data

Ex. ResNet-50 [He et al'15] — > 23M parameters; HE-vs-MPM dataset [Han et al'23] — only 116 breast cancer images

- Conventional supervised learning

$$\min_\phi \mathcal{L}(\phi;\mathcal{D}^{\text{trn}})+\mathcal{R}(\phi)$$

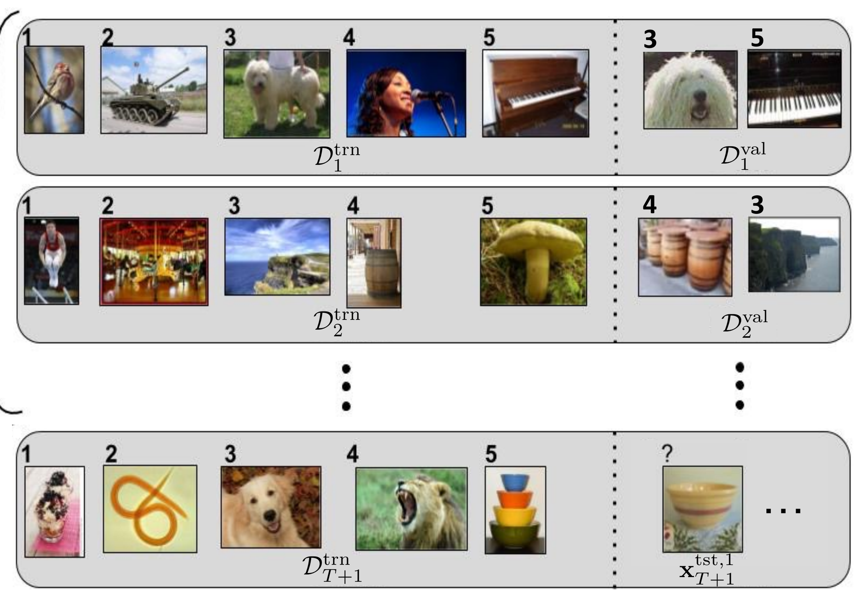
- Parameter $\phi\in\mathbb{R}^d$, training data $\mathcal{D}^{\text{trn}} := \{(\mathbf{x}^n, y^n)\}_{n=1}^{N^{\text{trn}}}$
- $\mathcal{L}(\phi;\mathcal{D}^{\text{trn}}) = -\log p(\mathbf{y}^{\text{trn}}|\phi;\mathbf{X}^{\text{trn}}) := \mathcal{L}^{\text{trn}}(\phi),\ \mathcal{R}(\phi) = -\log p(\phi)$
- Under-determinacy $d \gg N^{\text{trn}}$ ➤ Rely on informative $\mathcal{R}(\phi)$

Ex. Leverage Gaussian prior to cope with underdetermined regression

Meta-learning: learn a task-invariant prior from related tasks

Problem statement

- Supervised meta-learning
  - Given
    - Related tasks $t=1,\dots,T$, each with (limited) $\mathcal{D}_t^{\text{trn}}, \mathcal{D}_t^{\text{val}}$
    - New task with limited $\mathcal{D}_{T+1}^{\text{trn}}$ and test inputs $\{\mathbf{x}_{T+1}^{\text{tst},n}\}_{n=1}^{N_{T+1}^{\text{tst}}}$
  - Predict $\{y_{T+1}^{\text{tst},n}\}_{n=1}^{N_{T+1}^{\text{tst}}}$



✓ **Goal:** learn task-invariant prior from related tasks, and transfer to new task via $\min_{\phi_{T+1}} \mathcal{L}_{T+1}^{\text{trn}}(\phi_{T+1}) + \mathcal{R}(\phi_{T+1})$

➤ Bilevel learning: task-specific model-parameter $\phi_t\in\mathbb{R}^d$; task-invariant meta-parameter $\theta\in\mathbb{R}^D$

$$\min_{\theta} \sum_{t=1}^{T} \mathcal{L}_t^{\text{val}}(\phi_t^*(\theta)) \quad \text{meta-level}$$
$$\text{s.t. } \phi_t^*(\theta) = \arg\min_{\phi_t} \mathcal{L}_t^{\text{trn}}(\phi_t) + \mathcal{R}(\phi_t;\theta),\ \forall t \quad \text{task-level (explicit prior)}$$
$$\text{or } \phi_t^*(\theta) = \arg\min_{\phi_t} \mathcal{L}_t^{\text{trn}}(\phi_t;\theta),\ \forall t \quad \text{task-level (implicit prior)}$$

✓ Optimize via alternating solver

Bilevel optimization for meta-learning

- Model-agnostic meta-learning (MAML) [Finn et al'17]
  - Task-invariant initialization: $\phi_t^0 = \phi^0,\ \forall t,\ \theta := \{\phi^0\}$ $(d=D)$
  - Task-level iteration: $\phi_t^k(\theta) = \phi_t^{k-1}(\theta) - \alpha\nabla\mathcal{L}_t^{\text{trn}}(\phi_t^{k-1}(\theta)),\ k=1,\dots,K$; $\hat{\phi}_t(\theta) := \phi_t^K(\theta)$
  - After K iterations, iterative solver $\hat{\phi}_t(\theta)$ will approximate global optimum $\phi_t^*(\theta)$

> Lemma [Grant et al'18]. *Using 2nd-order Taylor approx. of the loss, MAML satisfies*
> $$\hat{\phi}_t(\theta) \approx \arg\min_{\phi_t} \mathcal{L}_t^{\text{trn}}(\phi_t) + \tfrac{1}{2}\|\phi_t-\theta\|^2_{\Lambda_t}$$
> *where* Λ_t *is a function of* $\alpha, K, \nabla^2\mathcal{L}_t^{\text{trn}}(\theta)$.

➤ Implicit Gaussian prior $p(\phi_t;\theta) = \mathcal{N}(\theta, \Lambda_t^{-1})$

- Accuracy versus complexity tradeoff with K
  - Converges to a stationary point $\|\hat{\phi}_t - \bar{\phi}_t\|_2 = \mathcal{O}(\frac{L}{K})$ — L: Lipschitz smoothness of loss
  - Grad. error $\|\nabla_\theta\mathcal{L}_t^{\text{val}}(\hat{\phi}_t(\theta)) - \nabla_\theta\mathcal{L}_t^{\text{val}}(\bar{\phi}_t(\theta))\|$ is linear with $\|\hat{\phi}_t-\bar{\phi}_t\|_2$ [Zhang et al'23]
  - Meta-level iteration $\theta^r = \theta^{r-1} - \beta\frac{1}{|\mathcal{T}^r|}\sum_{t\in\mathcal{T}^r}\nabla_\theta\mathcal{L}_t^{\text{val}}(\hat{\phi}_t(\theta^{r-1}))$ — $\mathcal{T}^r$: mini-batch of tasks
  - Overall complexity grows linearly with K

Prior art on accelerated task-specific optimization

- Gradient descent (GD) recap

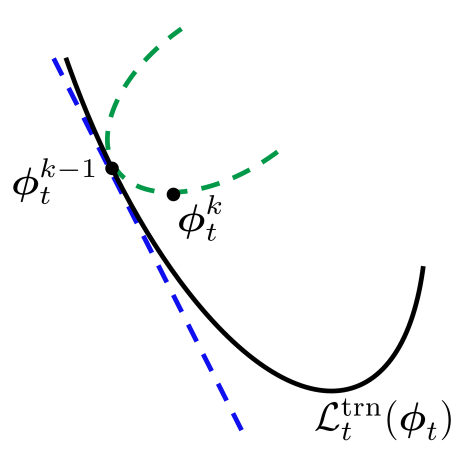
linearization $:= \text{lin}(\mathcal{L}_t^{\text{trn}}, \phi_t^{k-1})(\phi_t)$ — quadratic upper bound $0<\alpha\le\frac{1}{L}$

$$\phi_t^k = \arg\min_{\phi_t} \mathcal{L}_t^{\text{trn}}(\phi_t^{k-1}) + \nabla\mathcal{L}_t^{\text{trn}}(\phi_t^{k-1})(\phi_t-\phi_t^{k-1}) + \tfrac{1}{2\alpha}\|\phi_t-\phi_t^{k-1}\|_2^2 = \phi_t^{k-1} - \alpha\nabla\mathcal{L}_t^{\text{trn}}(\phi_t^{k-1})$$

❖ Bound not adaptive to k, t, and ϕ_t dimensions

- Preconditioned GD (PGD) with matrix $\mathbf{P}$ can accelerate GD

$$\phi_t^k = \arg\min_{\phi_t} \text{lin}(\mathcal{L}_t^{\text{trn}}, \phi_t^{k-1})(\phi_t) + \tfrac{1}{2\alpha}\|\phi_t-\phi_t^{k-1}\|^2_{\mathbf{P}^{-1}} = \phi_t^{k-1} - \alpha\mathbf{P}\nabla\mathcal{L}_t^{\text{trn}}(\phi_t^{k-1})$$



- Meta-learning with task-invariant preconditioner

$$\phi_t^k = \phi_t^{k-1} - \alpha\mathbf{P}(\theta_P)\nabla\mathcal{L}_t^{\text{trn}}(\phi_t^{k-1}),\ \theta := \{\phi^0, \theta_P\}$$

- $\mathbf{P}(\theta_P)$ choices: diag. [Li et al'17], block-diag. [Park et al'19], low-rank [Flennerhag et al'19],…

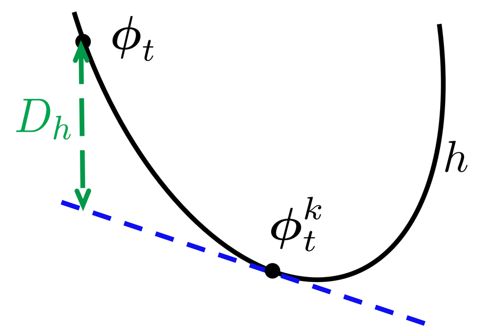
Learning loss geometry priors via mirror descent

Q. How about non-quadratic upper bounds ?

A. Bregman divergence $D_h(\phi_t,\phi_t^k) := h(\phi_t) - \text{lin}(h,\phi_t^k)(\phi_t)$

- Distance generating function h is strongly convex

Ex. If $h(\cdot) = \frac{1}{2}\|\cdot\|^2_{\mathbf{P}^{-1}}$, then $D_h(\phi_t,\phi_t^k) = \frac{1}{2}\|\phi_t-\phi_t^k\|^2_{\mathbf{P}^{-1}}$



- Mirror descent (MD) iteration

$$\phi_t^k = \arg\min_{\phi_t} \text{lin}(\mathcal{L}_t^{\text{trn}}, \phi_t^{k-1})(\phi_t) + \tfrac{1}{\alpha}D_h(\phi_t,\phi_t^{k-1}) = \nabla h^*\big(\nabla h(\phi_t^{k-1}) - \alpha\nabla\mathcal{L}_t^{\text{trn}}(\phi_t^{k-1})\big)$$

- Fenchel conjugate $h^*(\mathbf{z}) := \sup_\phi \phi^\top\mathbf{z} - h(\phi)$
- Properties: **P1.** h^* is convex and Lipschitz smooth; **P2.** if $h\in\mathcal{C}^1(\mathbb{R}^d)$, then $\nabla h^* = (\nabla h)^{-1}$
- MD with proper h, accelerates convergence rate/constant over GD

- Learnable loss geometries for meta-learning
  - MD can be viewed as optimization over dual variable $\mathbf{z}_t^k := \nabla h(\phi_t^k)$

$$\mathbf{z}_t^k = \mathbf{z}_t^{k-1} - \alpha\nabla\mathcal{L}_t^{\text{trn}}(\nabla h^*(\mathbf{z}_t^{k-1})),\ k=1,\dots,K, \qquad \hat{\phi}_t = \nabla h^*(\mathbf{z}_t^K)$$

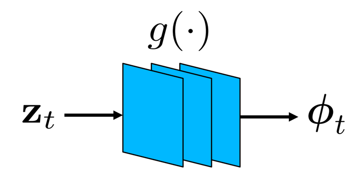
➤ In addition to initialization $\mathbf{z}^0 := \nabla h(\phi^0)$, we need to learn $\nabla h^*: \phi_t = \nabla h^*(\mathbf{z}_t)$

Key idea: learn a data-driven inverse mirror map $\nabla h^*: \mathbb{R}^d \mapsto \mathbb{R}^d$

➤ With h^* as in **P1**, ∇h^* is increasing and Lipschitz continuous

Learning inverse mirror map with a NN model

- Block-wise autoregression g as a candidate NN model
  - Let $\{\mathcal{B}_i\}_{i=1}^B$ be a partition of set $\{1,\dots,d\}$



$$[g(\mathbf{z}_t)]_{\mathcal{B}_i} = [\mathbf{z}_t]_{\mathcal{B}_i}\odot\sigma(\alpha_i)+\mu_i \qquad [\alpha_i,\mu_i] := d_i(\{e_j([\mathbf{z}_t]_{\mathcal{B}_j})\}_{j=1}^{i-1}),\ i=1,\dots,B$$

where σ positive and bounded (e.g., sigmoid); and $\{e_i,d_i\}_{i=1}^{B-1}$ multi-layer perceptrons

> Lemma. *For any partition* $\{\mathcal{B}_i\}_{i=1}^B$, g *is increasing and Lipschitz continuous.*

➤ $\nabla h^* = g$ is a desirable choice; meta-parameter $\theta := \{\mathbf{z}^0, \theta_g\}$ — θ_g: parameter of NN g

- Research outlook: model $h^*: \mathbb{R}^d\mapsto\mathbb{R}$, and analyze bilevel convergence

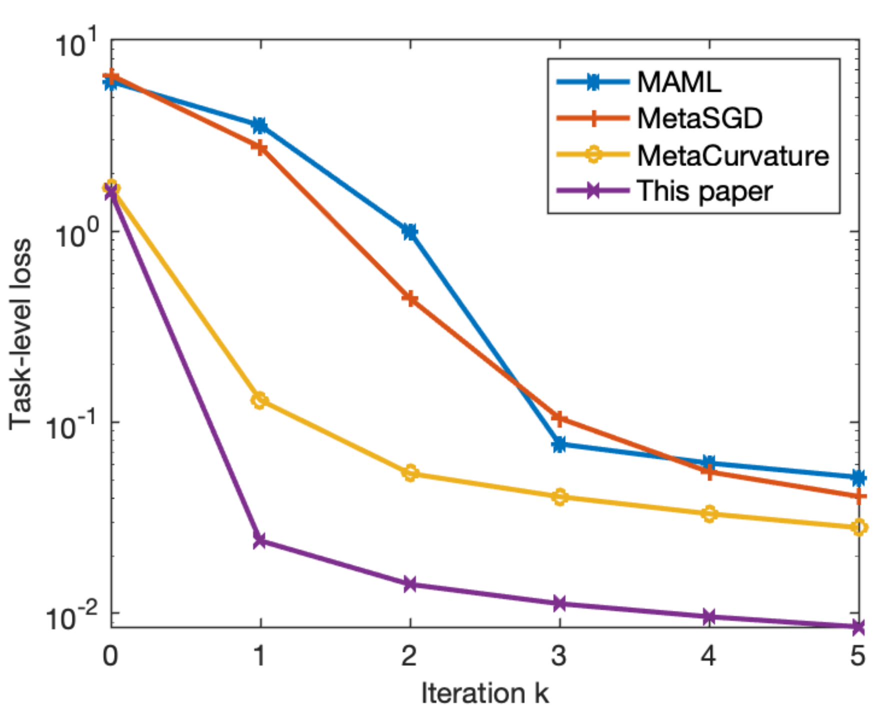
Numerical tests

- Comparison with existing loss geometry models on *miniImageNet* [Vinyals et al'16]
  - $\mathcal{D}_t^{\text{trn}}$: 1 or 5 images for each of 5 classes
  - Metric: mean accuracy ± 95% confidence interval on 600 new tasks
  - Deep learning architecture: Standard 4-layer 64-channel CNN [Ravi et al'16]

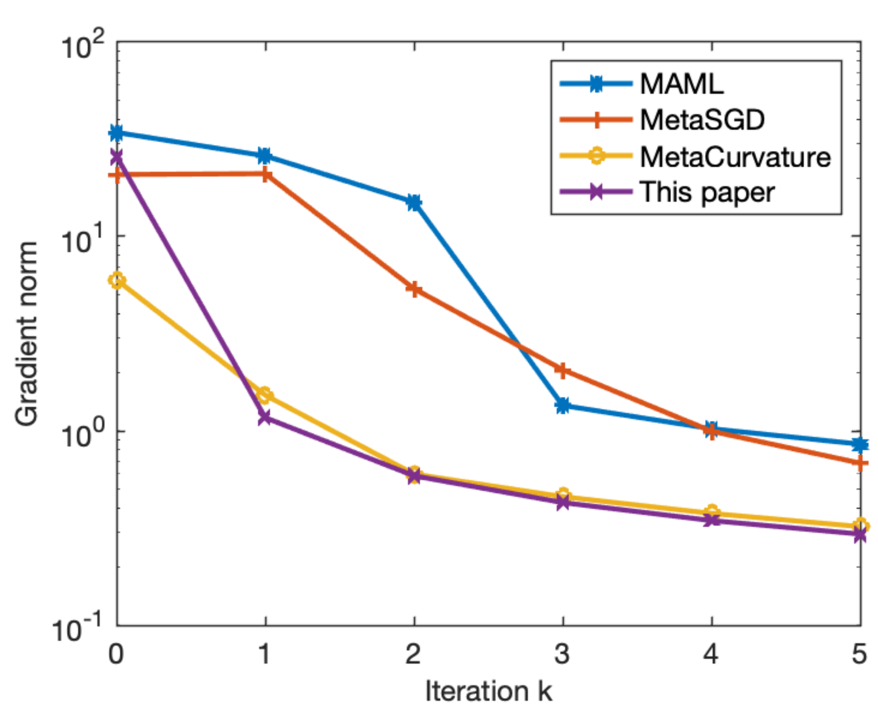
Method	Task-level optimizer	Loss geometry model	Avg. acc. ± 95% confid. interval 1-shot/class	5-shot/class
MAML [6]	GD	identity matrix	$48.70\pm1.84\%$	$63.11\pm0.92\%$
MetaSGD [11]	PGD	diagonal matrix	$50.47\pm1.87\%$	$64.03\pm0.94\%$
MT-net [14]	PGD	block diagonal matrix	$51.70\pm1.84\%$	–
WarpGrad [15]	PGD	NN-based low-rank matrix	$52.3\pm0.8\%$	$68.4\pm0.6\%$
MetaCurvature [13]	PGD	block diag. & Kron. (low-rank) matrix	$54.23\pm0.88\%$	$67.99\pm0.73\%$
MetaKFO [17]	NN-transformed GD	NN-based gradient transformation	–	64.9%
ECML [16]	PGD	Gauss-Newton approximation	$48.94\pm0.80\%$	$65.26\pm0.67\%$
This paper's method	MD	blockIAF-based mirror map	$\mathbf{56.10\pm1.43\%}$	$\mathbf{69.59\pm0.71\%}$

➤ Superior performance due to improved loss geometry model

- Versatile loss geometry model accelerates task-level convergence
  - $\mathcal{L}_t^{\text{trn}}(\phi_t^k)$ vs k



  - $\|\nabla\mathcal{L}_t^{\text{trn}}(\phi_t^k)\|_2$ vs k



➤ Better initialization and faster reduction ➤ Proximity to a stationary point