May 7, 2024

Meta-Learning Priors Using Unrolled Proximal Networks

Yilang Zhang

Dept. of ECE, University of Minnesota

Aknowledgement: Prof. Georgios B. Giannakis

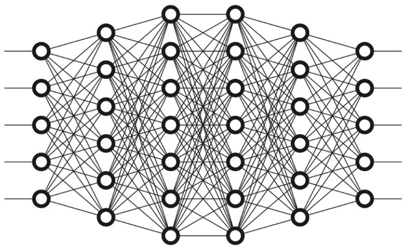
NSF grants 2102312, 2103256, 2128593, 2126052, and 2212318

Motivating context of meta-learning

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

Ex. ResNet-50 [He et al'15] HE-vs-MPM dataset [Han et al'23]

>23M parameters 116 breast cancer images



vs.

- Conventional supervised learning

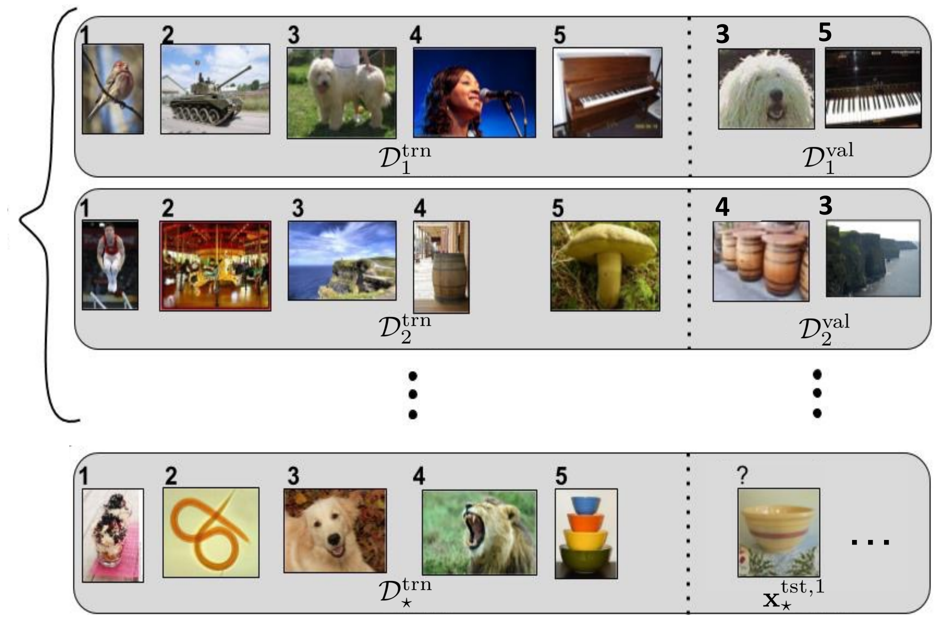
$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}^{\mathrm{trn}}) + \mathcal{R}(\boldsymbol{\theta})$$

- Model parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, training data $\mathcal{D}^{\mathrm{trn}} = \{(\mathbf{x}^n, y^n)\}_{n=1}^{N^{\mathrm{trn}}}$
- Bayesian view: $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}^{\mathrm{trn}}) = -\log p(\mathbf{y}^{\mathrm{trn}} | \boldsymbol{\theta}; \mathbf{X}^{\mathrm{trn}}) := \mathcal{L}^{\mathrm{trn}}(\boldsymbol{\theta}),\ \mathcal{R}(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta})$
- Overfitting if $d \gg N^{\mathrm{trn}}$ ➢ Rely on informative $\mathcal{R}(\boldsymbol{\theta})$

Remedy: extract and transfer task-invariant prior from related tasks

Meta-learning in a nutshell

- Supervised meta-learning
  - Given:
    - Tasks $t = 1, \ldots, T$, each with $\mathcal{D}_t^{\text{trn}}, \mathcal{D}_t^{\text{val}}$
    - New task $\star$ w/ limited $\mathcal{D}_\star^{\text{trn}}$ and test inputs $\{\mathbf{x}_\star^{\text{tst},n}\}_{n=1}^{N_\star^{\text{tst}}}$
  - To-do: predict labels $\{y_\star^{\text{tst},n}\}_{n=1}^{N_\star^{\text{tst}}}$



✓ **Goal:** learn task-invariant prior; solve new task via $\min_{\boldsymbol{\theta}_\star} \mathcal{L}_\star^{\text{trn}}(\boldsymbol{\theta}_\star) + \mathcal{R}(\boldsymbol{\theta}_\star)$

➢ Bilevel problem: task-specific model-param $\boldsymbol{\theta}_t \in \mathbb{R}^d$, task-invariant meta-param $\boldsymbol{\theta} \in \mathbb{R}^D$

$$\min_{\boldsymbol{\theta}} \sum_{t=1}^{T} \mathcal{L}_t^{\text{val}}(\boldsymbol{\theta}_t^*(\boldsymbol{\theta}))$$ meta/outer-level

$$\text{s.t. } \boldsymbol{\theta}_t^*(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}_t} \mathcal{L}_t^{\text{trn}}(\boldsymbol{\theta}_t; \boldsymbol{\theta}), \ \forall t$$ task/inner-level (ideal)

$$\boldsymbol{\theta}_t^*(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}_t} \mathcal{L}_t^{\text{trn}}(\boldsymbol{\theta}_t) + \mathcal{R}(\boldsymbol{\theta}_t; \boldsymbol{\theta}), \ \forall t$$ task/inner-level (general)

S. Ravi, and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. ICLR*, 2017.

Expressiveness challenge in prior selection

Q. How to parameterize $\mathcal{R}(\boldsymbol{\theta}_t; \boldsymbol{\theta})$?

- Implicit prior via initialization
  - Model-agnostic meta-learning (MAML) [Finn et al'17]:
    - Task-invariant initialization: $\boldsymbol{\theta}_t^0 = \boldsymbol{\theta}^{\text{init}} = \boldsymbol{\theta}, \ \forall t$
    - Task-specific optimization: $\boldsymbol{\theta}_t^k = \boldsymbol{\theta}_t^{k-1} - \alpha \nabla \mathcal{L}_t^{\text{trn}}(\boldsymbol{\theta}_t^{k-1}), \ k = 1, \ldots, K$

Lemma [Grant et al'18]. *Under second-order approximation, MAML satisfies*

$$\boldsymbol{\theta}_t^K(\boldsymbol{\theta}) \approx \boldsymbol{\theta}_t^*(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}_t} \mathcal{L}_t^{\text{trn}}(\boldsymbol{\theta}_t) + \frac{1}{2}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}\|_{\boldsymbol{\Lambda}_t}^2$$

where $\boldsymbol{\Lambda}_t$ *is determined by* $\alpha, K, \nabla^2 \mathcal{L}_t^{\text{trn}}(\boldsymbol{\theta})$.

  - ➢ Limited-step GD ≈ implicit Gaussian prior $p(\boldsymbol{\theta}_t; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Lambda}_t^{-1})$
  - Other ex: diag. Gaussian [Li et al'17], block-diag. Gaussian [Park et al'19], …
- Explicit prior via regularization
  - Isotropic Gaussian [Rajeswaran et al'19] $\mathcal{R}(\boldsymbol{\theta}_t; \boldsymbol{\theta}) = \frac{\lambda}{2}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^{\text{init}}\|_2^2, \ \boldsymbol{\theta} := \{\boldsymbol{\theta}^{\text{init}}, \lambda\}$
  - Sparse [Tian et al'20], factorable + degenerate [Bertinetto et al'18, Lee et al'19], …

Challenge: preselected prior have limited expressiveness

C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. ICML*, 2017.
E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths, "Recasting gradient-based meta-learning as hierarchical Bayes," *ICLR,* 2018.

Meta-learning priors with unrolled proximal NNs

✓ Key idea: learn the form of $\mathcal{R}(\boldsymbol{\theta}_t;\boldsymbol{\theta})$ by unrolling proximal GD

❑ Proximal GD (PGD) recap

$$\boldsymbol{\theta}_t^*(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}_t} \mathcal{L}_t^{\text{trn}}(\boldsymbol{\theta}_t) + \mathcal{R}(\boldsymbol{\theta}_t;\boldsymbol{\theta})$$

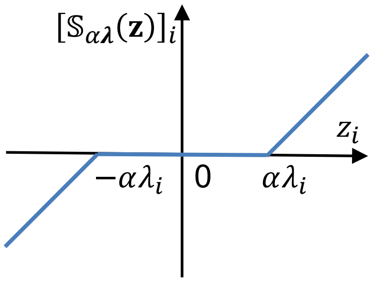
- Descend wrt $\mathcal{L}_t^{\text{trn}}$: $\mathbf{z}_t^k = \boldsymbol{\theta}_t^{k-1} - \alpha\nabla\mathcal{L}_t^{\text{trn}}(\boldsymbol{\theta}_t^{k-1})$
- Calibrate via $\mathcal{R}$: $\boldsymbol{\theta}_t^k = \arg\min_{\boldsymbol{\theta}_t} \frac{1}{2\alpha}\|\mathbf{z}_t^k - \boldsymbol{\theta}_t\|_2^2 + \mathcal{R}(\boldsymbol{\theta}_t;\boldsymbol{\theta}) := \text{prox}_{\alpha,\mathcal{R}}(\mathbf{z}_t^k)$

Ex 1. Diag. Gaussian $\mathcal{R}(\boldsymbol{\theta}_t;\boldsymbol{\theta}) = \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^{\text{init}}\|^2_{\text{diag}(\boldsymbol{\lambda})}$

➢ $\text{prox}_{\alpha,\mathcal{R}}(\mathbf{z}) = (\mathbf{z} - \boldsymbol{\theta}^{\text{init}})/(\mathbf{1}_d + \alpha\boldsymbol{\lambda}) + \boldsymbol{\theta}^{\text{init}}$

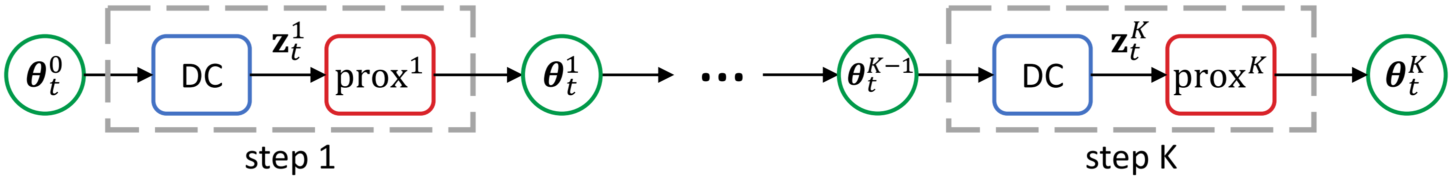
Ex 2. (Shifted) sparse $\mathcal{R}(\boldsymbol{\theta}_t;\boldsymbol{\theta}) = \|\text{diag}(\boldsymbol{\lambda})(\boldsymbol{\theta}_t - \boldsymbol{\theta}^{\text{init}})\|_1$

➢ $\text{prox}_{\alpha,\mathcal{R}}(\mathbf{z}) = \mathbb{S}_{\alpha\boldsymbol{\lambda}}(\mathbf{z} - \boldsymbol{\theta}^{\text{init}}) + \boldsymbol{\theta}^{\text{init}}$



❑ Learning prior via algorithm unrolling

❖ Optimal $\mathcal{R}$ unknown ➢ Learn per-step $\{\text{prox}^k\}_{k=1}^K$ from data

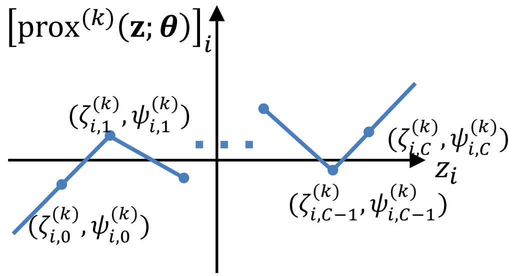


Y. Zhang, and G. B. Giannakis, "Meta-Learning Priors Using Unrolled Proximal Networks," in *Proc. ICLR*, Vienna, Austria, May 7-11, 2024.

Model proximal map via piecewise linear functions

Our idea: model $\{\text{prox}^{(k)}\}_{k=1}^{K}$ via dimension-wise piecewise linear functions (PLFs)

$$[\text{prox}^k(\mathbf{z};\boldsymbol{\theta})]_i = \begin{cases} \frac{\psi^k_{i,0}(\zeta^k_{i,1}-z_i)+\psi^k_{i,1}(z_i-\zeta^k_{i,0})}{\zeta^k_{i,1}-\zeta^k_{i,0}}, & z_i < \zeta^k_{i,1} \\ \frac{\psi^k_{i,c-1}(\zeta^k_{i,c}-z_i)+\psi^k_{i,c}(z_i-\zeta^k_{i,c-1})}{\zeta^k_{i,c}-\zeta^k_{i,c-1}}, & \zeta^k_{i,c-1} \le z_i < \zeta^k_{i,c} \text{ and } c = 2,\dots C-1 \\ \frac{\psi^k_{i,C}(\zeta^k_{i,C+1}-z_i)+\psi^k_{i,C+1}(z_i-\zeta^k_{i,C})}{\zeta^k_{i,C+1}-\zeta^k_{i,C}}, & z_i \ge \zeta^k_{i,C-1} \end{cases}$$



- Fix $\zeta^k_{i,c} = (\frac{2c}{C}-1)A,\ \forall c,i,k$, where $A>0$ is a constant
- Learning per-step PLFs reduces to optimizing $\boldsymbol{\psi}^k := [\psi^k_{1,0},\dots,\psi^k_{d,C+1}]^\top \in \mathbb{R}^{(C+2)d}$
- Meta-parameter $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{\text{init}},\boldsymbol{\psi}^1,\dots,\boldsymbol{\psi}^K\}$

Q. How good is the PLF-based $\text{prox}^{(k)}(\cdot;\boldsymbol{\theta})$ compared to the oracle $\text{prox}_{\alpha,\mathcal{R}}(\cdot)$?

Theorem 1. *Let $\hat{\boldsymbol{\theta}}_t(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\theta}}_t$ be the K-step PGD outputs with $\text{prox}^k(\cdot;\boldsymbol{\theta})$ and $\text{prox}_{\alpha,\mathcal{R}}(\cdot)$, respectively. Under mild assumptions, it holds for any $\text{prox}_{\alpha,\mathcal{R}} \in \mathcal{C}^1([-A,A]^d)$ that*

$$\min_{\boldsymbol{\theta}} \|\hat{\boldsymbol{\theta}}_t(\boldsymbol{\theta}) - \tilde{\boldsymbol{\theta}}_t\|_2 = \mathcal{O}(\frac{1}{C^2}).$$

Theorem 2. *Under mild assumptions, it holds for any $\text{prox}_{\alpha,\mathcal{R}} \in \mathcal{C}^0([-A,A]^d)$ that*

$$\min_{\boldsymbol{\theta}} \|\hat{\boldsymbol{\theta}}_t(\boldsymbol{\theta}) - \tilde{\boldsymbol{\theta}}_t\|_2 = \mathcal{O}(\frac{1}{C}).$$

Y. Zhang, and G. B. Giannakis, “Meta-Learning Priors Using Unrolled Proximal Networks,” in *Proc. ICLR*, Vienna, Austria, May 7-11, 2024.

Numerical experiments

- Few-shot classification on miniImageNet dataset

Method	Prior	5-class miniImageNet 1-shot (%)	5-class miniImageNet 5-shot (%)	5-class TieredImageNet 1-shot (%)	5-class TieredImageNet 5-shot (%)
LSTM (Ravi & Larochelle, 2017)	RNN-based	$43.44_{\pm 0.77}$	$60.60_{\pm 0.71}$	–	–
MAML (Finn et al., 2017)	implicit Gaussian	$48.70_{\pm 1.84}$	$63.11_{\pm 0.92}$	$51.67_{\pm 1.81}$	$70.30_{\pm 1.75}$
ProtoNets (Snell et al., 2017)	shifted sparse	$49.42_{\pm 0.78}$	$68.20_{\pm 0.66}$	$53.31_{\pm 0.87}$	$72.69_{\pm 0.74}$
R2D2 (Bertinetto et al., 2019)	shifted sparse	$51.8_{\pm 0.2}$	$68.4_{\pm 0.2}$	–	–
MC (Park & Oliva, 2019)	block-diag. Gaussian	$54.08_{\pm 0.93}$	$67.99_{\pm 0.73}$	–	–
L2F (Baik et al., 2020)	implicit Gaussian	$52.10_{\pm 0.50}$	$69.38_{\pm 0.46}$	$54.40_{\pm 0.50}$	$\mathbf{73.34}_{\pm 0.44}$
KML (Abdollahzadeh et al., 2021)	shifted sparse	$54.10_{\pm 0.61}$	$68.07_{\pm 0.45}$	$54.67_{\pm 0.39}$	$72.09_{\pm 0.27}$
MeTAL (Baik et al., 2021)	implicit Gaussian	$52.63_{\pm 0.37}$	$70.52_{\pm 0.29}$	$54.34_{\pm 0.31}$	$70.40_{\pm 0.21}$
MinimaxMAML (Wang et al., 2023)	inverted nlp	$51.70_{\pm 0.42}$	$68.41_{\pm 1.28}$	–	–
MetaProxNet+MAML	unrolling-based	$53.70_{\pm 1.40}$	$70.08_{\pm 0.69}$	$54.56_{\pm 1.44}$	$71.80_{\pm 0.73}$
MetaProxNet+MC	unrolling-based	$\mathbf{55.94}_{\pm 1.39}$	$\mathbf{71.97}_{\pm 0.67}$	$\mathbf{57.34}_{\pm 1.42}$	$\mathbf{73.38}_{\pm 0.73}$

  - Superior performance due to enhanced prior expressiveness

- Check our paper/poster for ablation tests and visualization of PLFs

Thank You!