Provable Guarantees for Gradient-Based Meta-Learning

Mikhail Khodak, Maria-Florina Balcan, Ameet Talwalkar
Carnegie Mellon University

12 June 2019, Poster #253

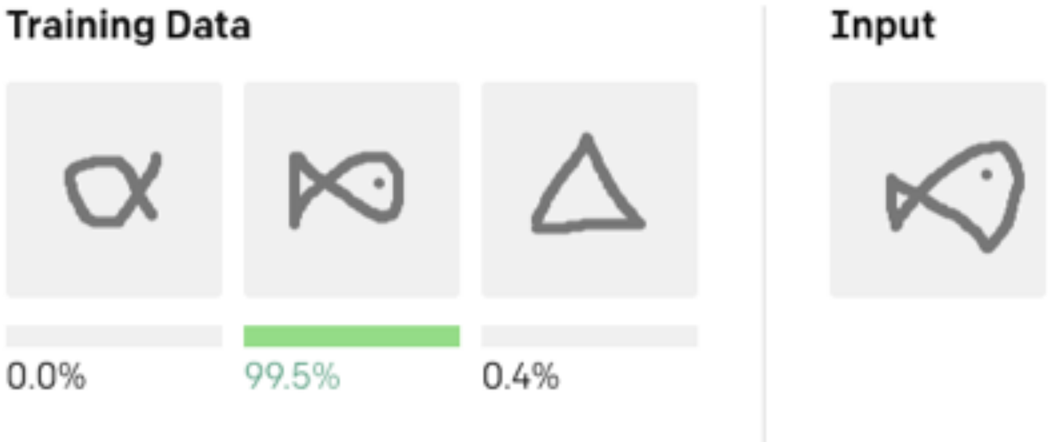
khodak@cmu.edu

Gradient-Based Meta-Learning: A simple but effective approach

Method: learn an initialization so gradient descent on a few samples from an unseen task returns a good model

Applications:

Meta-RL (MAML, [FAL'17])



Few-Shot Learning (Reptile, [NAS'18])

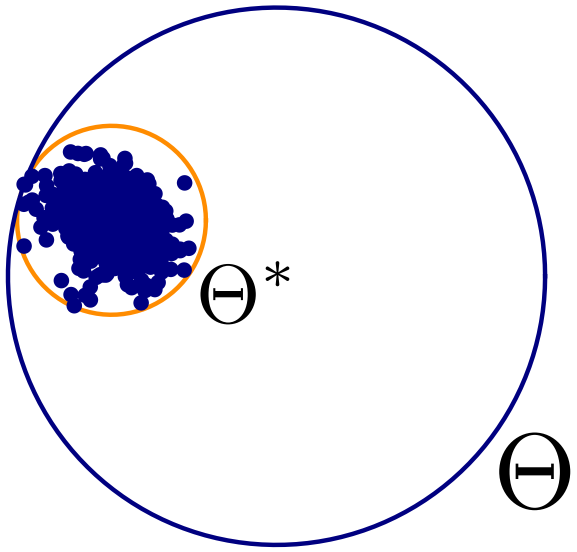
Gradient-Based Meta-Learning: Theoretical questions:

- What kinds of task-relationships can GBML exploit?

- Are we restricting ourselves by using such simple methods?

- How does GBML relate to classical multi-task methods?

Gradient-Based Meta-Learning: Our contributions:



- What kinds of task-relationships can GBML exploit?
  - better average performance per-task if optimal task-parameters are close together.
- Are we restricting ourselves by using such simple methods?
  - GBML is the best we can do without stronger task-similarity assumptions.
- How does GBML relate to classical multi-task methods?
  - natural connection to regularized multi-task learning (MTL), e.g. Evgeniou & Pontil [2004].

Connecting to online convex optimization (OCO)

generic GBML on parameter space Θ (given T tasks with m samples each):

pick first initialization $\phi_1 \in \Theta$

for task $t = 1, \dots, T$:

run descent method initialized at ϕ_t on m samples from task t

use resulting parameter $\hat{\theta}_t$ to set ϕ_{t+1}

return meta-initialization ϕ_{T+1}

Connecting to online convex optimization (OCO)

~~generic GBML~~ on parameter space Θ (given T tasks with m samples each):

Reptile [NAS'17]

pick first initialization $\phi_1 \in \Theta$

for task $t = 1, \ldots, T$:

~~run descent method initialized at~~ $\cancel{\phi_t}$ ~~on~~ $\cancel{m}$ ~~samples from task~~ $\cancel{t}$

run m steps of online gradient descent (OGD) initialized at ϕ_t

~~use resulting parameter~~ $\cancel{\hat{\theta}_t}$ ~~to set~~ $\cancel{\phi_{t+1}}$

update ϕ_{t+1} using OCO algorithm on the regret of OGD as function of ϕ_t

return meta-initialization ϕ_{T+1}

Connecting to online convex optimization (OCO)

Benefits:

- import OCO regret guarantees that naturally encode distance from initialization
- bound excess transfer risk at meta-test-time via online-to-batch conversion
- connect to regularized MTL through the Follow-the-Regularized-Leader meta-algorithm

Result: GBML reduces average regret

Assumption:

optimal parameters lie in subset Θ^* of radius $D^* \ll D$, the diameter of action-space Θ

- GBML guarantee (this work):

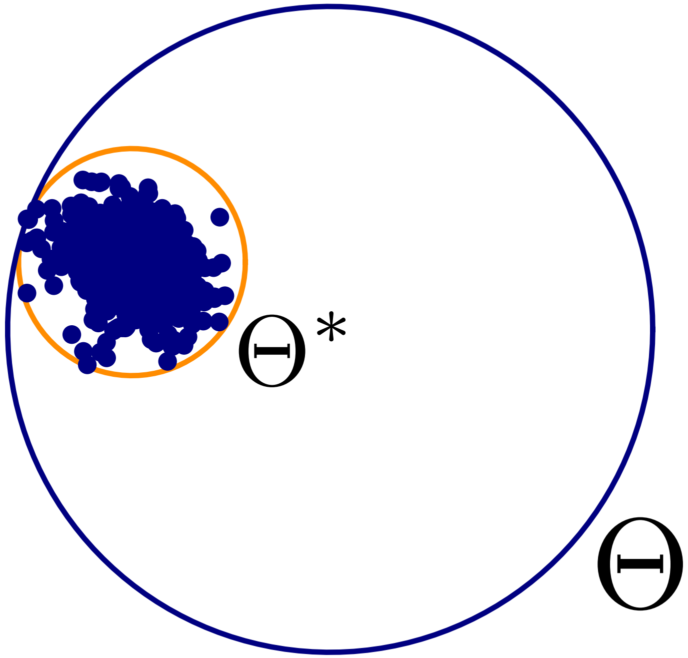
$$\text{average regret} = O\left(D^* + \frac{\log T}{T}\right)\sqrt{m}$$

- Minimax single-task guarantee [ABRT'08]:

$$\text{regret} = \Theta(D\sqrt{m})$$

- Multi-task lower bound (this work):

$$\text{average regret} = \Omega(D^*\sqrt{m})$$



Result: GBML reduces excess transfer risk

Run GBML to learn initialization over i.i.d. samples $(x_{t,i}, y_{t,i}) \sim P_t \sim Q$

When OGD is run on m samples are drawn from $P \sim Q$, the average iterate satisfies

$$\mathbb{E}_P \ell(\bar{\theta}) \quad = \quad \mathbb{E}_P \ell(\theta^*) \quad + \quad \frac{O(D^*)}{\sqrt{m}} \quad + \quad \sqrt{\frac{8}{T} \log \frac{1}{\delta}}$$

risk of learned model

minimum risk

small when tasks are similar

small with more task-samples

Come to poster 253 to discuss

- Details and proofs of theoretical results

- Generalizations
  - not using OGD within-task
  - the batch-within-online setting

- Connecting GBML to
  - federated learning
  - classical multi-task learning

- New adaptive and dynamic methods for practical GBML