Provable Guarantees for Gradient-Based Meta-Learning

Mikhail Khodak, Maria-Florina Balcan, Ameet Talwalkar
Carnegie Mellon University
khodak@cmu.edu

Meta-Learning a Model Bias

Parameter-transfer meta-learning is a popular approach for many tasks with few samples each. For example, one can learn a deep net initialization ϕ so a few gradient steps yield good weights $\hat{\theta}$ [2]:

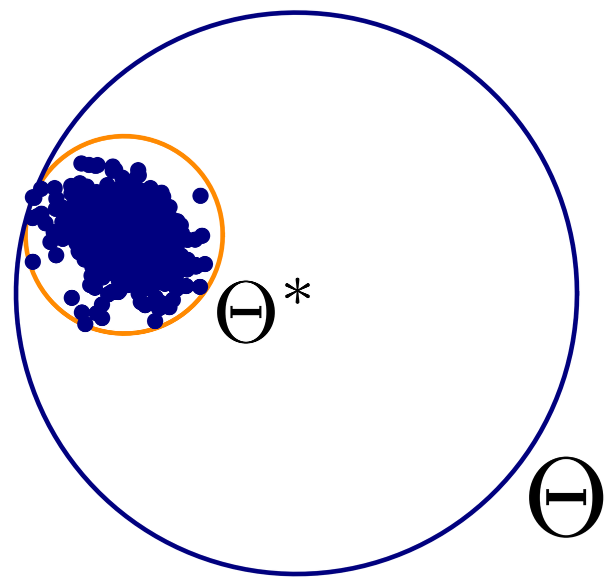
$$\hat{\theta} = \phi - \eta \sum_i \nabla L(f_\phi(x_i), y_i)$$

A more classical approach is to learn a ϕ for solving a biased ℓ_2-regularization problem to get $\hat{\theta}$ [1]:

$$\hat{\theta} = \arg\min_\theta \frac{\|\theta - \phi\|_2^2}{2\eta} + \sum_i L(f_\theta(x_i), y_i)$$

We study whether simple, scalable, gradient-based approaches, e.g. MAML [2] and Reptile [3], can learn a good parameter for similar tasks. Online convex optimization (OCO) enables a unified analysis of meta-initialization and meta-regularization through the equivalence between Follow-the-Regularized-Leader (FTRL) and Online Mirror Descent (OMD) [4].

Setup and Assumptions



Setting - Lifelong Optimization:

- Agent sees a sequence of tasks $t = 1, \dots, T$.
- At each task t, agent takes m actions $\theta_{t,1}, \dots, \theta_{t,m} \in \Theta$ and suffers losses $\ell_{t,1}(\theta_{t,1}), \dots, \ell_{t,m}(\theta_{t,m})$.

Goal - Low Task-Averaged Regret:

$$\bar{\mathbf{R}} = \frac{1}{T}\sum_{t=1}^T \left(\sum_{i=1}^m \ell_{t,i}(\theta_{t,i}) - \min_{\theta_t \in \Theta} \sum_{i=1}^m \ell_{t,i}(\theta_t) \right)$$

Task Similarity Assumption

The optimal task parameters θ_t^* are within some small subset $\Theta^* \subset \Theta$ with diameter $D^* \ll D$, the diameter of Θ.

Algorithm and Guarantees

Follow-the-Meta-Regularized-Leader:

1. start with initialization ϕ_1 and diameter guess D_1
2. for task $t \in [T]$:
3. set learning rate $\eta_t = \frac{D_t}{\sqrt{m}}$
4. run FTRL/OMD on task t:
$$\theta_{t,i} = \arg\min_{\theta \in \Theta} \frac{1}{2\eta_t}\|\theta - \phi_t\|_2^2 + \sum_{j<i} \ell_{t,j}(\theta)$$
5. double guess if violated, i.e. $\|\theta_t^* - \phi_t\|_2 > D_t$
6. update ϕ_{t+1} by adaptive OGD on $\frac{1}{2}\|\theta_t^* - \phi\|_2^2$

Algorithm can be extended to practical meta-learning [3] as a special case and to non-Euclidean geometries by replacing $\frac{1}{2}\|\theta - \phi_t\|_2^2$ with any Bregman divergence $B_R(\theta||\phi_t)$ for strongly convex R.

Theorem 1:
Average Regret Upper & Lower Bounds

(Upper Bound): For convex Lipschitz losses, Follow-the-Meta-Regularized-Leader achieves average regret

$$\bar{\mathbf{R}} \le \mathcal{O}\left(D^* + D \log T / T\right)\sqrt{m}$$

(Lower Bound): No algorithm can achieve average regret better than $\Omega(D^*\sqrt{m})$.

Proof Sketch of Upper Bound.

Assume known D^* (otherwise use doubling trick). Set $\eta_t = \frac{D^*}{\sqrt{m}}\ \forall\, t$.

$$\bar{\mathbf{R}} = \frac{1}{T}\sum_{t=1}^T \left(\sum_{i=1}^m \ell_{t,i}(\theta_{t,i}) - \min_{\theta_t \in \Theta} \sum_{i=1}^m \ell_{t,i}(\theta_t) \right)$$

$$\le \frac{1}{T}\sum_{t=1}^T \frac{1}{2\eta_t}\|\theta_t^* - \phi_t\|_2^2 + \eta_t m \qquad \textbf{(regret of FTRL/OMD)}$$

$$= \underbrace{\frac{\sqrt{m}}{T}\sum_{t=1}^T \frac{\|\theta_t^* - \phi_t\|_2^2 - \|\theta_t^* - \phi^*\|_2^2}{2D^*}}_{\textbf{(regret on strongly convex losses)}} + \underbrace{\frac{\sqrt{m}}{T}\sum_{t=1}^T \frac{\|\theta_t^* - \phi^*\|_2^2}{2D^*} + D^*}_{\textbf{(optimal task regret)}}$$

$$\le \quad \mathcal{O}\left(\frac{D \log T}{T}\right)\sqrt{m} \quad + \quad 2D^*\sqrt{m}$$

[1] Evgeniou & Pontil. *Regularized Multi-Task Learning.* KDD 2004.
[2] Finn, Abbeel, & Levine. *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks.* ICML 2018.
[3] Nichol, Achiam, & Schulman. *On First-Order Meta-Learning Algorithms.* arXiv 2018.
[4] Shalev-Shwartz. *Online Learning and Online Convex Optimization.* FTML 2011.

Approximate and Distributional Settings

Sometimes we cannot/do not want to compute the optimal parameter θ_t^* at each task t because:

1. Exact minimization is expensive but can use the last iterate.
2. θ_t^* is the true-risk minimizer in a batch-within-online setting, so we only have a stochastic approximation.

Theorem 2: Approximate Meta-Updates

In cases 1 and 2 above, if the loss functions satisfy a quadratic growth condition then (w.h.p. in case 2) Follow-the-Meta-Regularized-Leader achieves average regret

$$\bar{\mathbf{R}} \le \mathcal{O}\left(D^* + D \log T / T + D/\sqrt[6]{m}\right)\sqrt{m}$$

Quadratic growth is a data-dependent condition often satisfied for relevant problems such as few-shot logistic regression.

Online meta-learning algorithms often run in batch setting. Can we provide guarantees when tasks are drawn i.i.d. from a distribution?

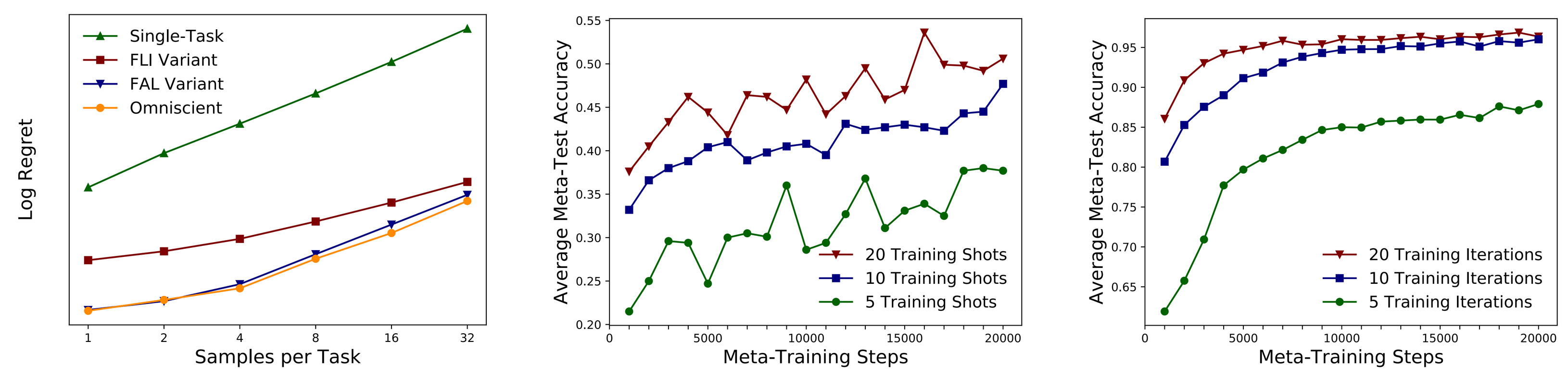
Theorem 3: Online-to-Batch Conversion

On a new task distribution $\mathcal{P} \sim \mathcal{Q}$, a learned meta-initialization gives in expectation the following generalization bound on the average iterate over m i.i.d. samples from $\mathcal{P}$ w.p. $1 - \delta$:

$$\mathop{\mathbb{E}}_{\mathcal{Q}} \ell_{\mathcal{P}}(\bar{\theta}) \quad \le \quad \mathop{\mathbb{E}}_{\mathcal{Q}} \ell_{\mathcal{P}}(\theta^*) \quad + \quad \frac{\bar{\mathbf{R}}}{m} \quad + \quad \sqrt{\frac{8}{T}\log\frac{1}{\delta}}$$

Thus excess transfer risk improves with lower task-averaged regret, which can be made small when tasks are similar.

Experiments



We use a new linear classification dataset, MiniWiki, for meta-learning experiments. On the left we see the average regret of our methods is very close to that of the omniscient algorithm at different task-sizes m. In few-shot learning, we see that better approximation at meta-training time leads to better performance at meta-test time on both 1-shot 5-way Mini-ImageNet (center) and 5-shot 20-way Omniglot (right).