Learning-To-Learn Stochastic Gradient Descent with Biased Regularization

Giulia Denevi [1,2]
Carlo Ciliberto [3,4]
Riccardo Grazzi [1,4]
Massimiliano Pontil [1,4]

1 Istituto Italiano di Tecnologia, Genoa, Italy
2 University of Genoa, Genoa, Italy
3 Imperial College of London, London, United Kingdom
4 University College of London, London, United Kingdom

36th International Conference on Machine Learning, Long Beach, 2019

The Learning-To-Learn Problem

- $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ data space, $\mathcal{X} = \big\{x \in \mathbb{R}^d : \|x\| \leq R\big\}$, $\mathcal{Y} \subseteq \mathbb{R}$
- A **meta-distribution** ρ sampling tasks μ over $\mathcal{Z}$, parametrized by

$$w_\mu \in \operatorname*{argmin}_{w \in \mathbb{R}^d} \mathcal{R}_\mu(w) \qquad \mathcal{R}_\mu(w) = \mathbb{E}_{(x,y)\sim\mu}\, \ell\big(\langle x, w\rangle, y\big)$$

$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ s.t. $\ell(\cdot, y)$ is L-Lipschitz and convex for any $y \in \mathcal{Y}$

- A parametrized family of **online algorithms**

$$\mathcal{A}_\mathcal{H} = \Big\{Z = (z_k)_{k=1}^n \sim \mu^n \mapsto A_h(Z) \in \mathbb{R}^d, h \in \mathcal{H}\Big\}$$

- Aim: find $A_h \in \mathcal{A}_\mathcal{H}$ with small expected excess risk

$$\mathcal{E}(A_h) - \mathcal{E}_\rho = \mathbb{E}_{\mu\sim\rho}\, \mathbb{E}_{Z\sim\mu^n} \big[\mathcal{R}_\mu\big(A_h(Z)\big) - \mathcal{R}_\mu(w_\mu)\big]$$

The Family of Learning Algorithms

We consider $A_h = \bar{w}_h$, the average of the iterations of SGD applied to

$$\mathcal{R}_\mu(w) + \frac{\lambda}{2}\left\|w - h\right\|^2$$

Algorithm 1 Within-Task Algorithm

Input $\lambda > 0$ regularization parameter, h bias, μ task

Initialization ${w_h}^{(1)} = h$

For $k = 1$ to n

Receive $z_k = (x_k, y_k) \sim \mu$

Compute $s_k = u_k + \lambda\left({w_h}^{(k)} - h\right)$, $u_k \in \partial\ell(\langle x_k, {w_h}^{(k)}\rangle, y_k)$ and $\gamma_k = \dfrac{1}{k\lambda}$

Update ${w_h}^{(k+1)} = {w_h}^{(k)} - \gamma_k s_k$

Return $({w_h}^{(k)})_{k=1}^{n+1}$, $\bar{w}_h = \dfrac{1}{n}\sum_{k=1}^{n} {w_h}^{(k)}$

Advantage of Using the Right Bias

Theorem 1. *Let $\bar{w}_h$ be the output of* **Alg. 1** *with an appropriate λ. Then,*

$$\mathcal{E}(\bar{w}_h) - \mathcal{E}_\rho \leq \mathrm{Var}_h \, 2RL \sqrt{\frac{2\big(\log(n) + 1\big)}{n}}$$

$$\mathrm{Var}_h^2 = \frac{1}{2} \, \mathbb{E}_{\mu\sim\rho} \, \big\| w_\mu - h \big\|^2$$

- Advantage of $h = \mathrm{m} = \mathbb{E}_{\mu\sim\rho} w_\mu$ (oracle) w.r.t. $h = 0$ (ITL) when

$$\mathrm{Var}_{\mathrm{m}}^2 = \frac{1}{2} \, \mathbb{E}_{\mu\sim\rho} \, \| w_\mu - \mathrm{m} \|^2 \ll \frac{1}{2} \, \mathbb{E}_{\mu\sim\rho} \, \| w_\mu \|^2 = \mathrm{Var}_0^2$$

- Since m is not available in practice, we estimate it from the data

A Proxy to Estimate the Bias

Proxy: $\mathcal{E}(\bar{w}_h) = \mathbb{E}_{\mu\sim\rho}\, \mathbb{E}_{Z\sim\mu^n}\, \mathcal{R}_\mu\big(\bar{w}_h(Z)\big) \approx \hat{\mathcal{E}}(h) = \mathbb{E}_{\mu\sim\rho}\, \mathbb{E}_{Z\sim\mu^n}\, \mathcal{L}_Z(h)$

$$\mathcal{L}_Z(h) = \min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{k=1}^{n} \ell\big(\langle x_k, w\rangle, y_k\big) + \frac{\lambda}{2}\|w-h\|^2 \tag{1}$$

- $\mathcal{L}_Z$ is a convex and Lipschitz upper bound

$$\mathbb{E}_{Z\sim\mu^n}\, \mathcal{R}_\mu\big(\bar{w}_h(Z)\big) \leq \mathbb{E}_{Z\sim\mu^n}\, \mathcal{L}_Z(h) + \frac{2R^2L^2\big(\log(n)+1\big)}{\lambda n}$$

- Denoting by $\hat{w}_h(Z)$ the minimizer in Eq. (1),

$$\nabla\mathcal{L}_Z(h) = -\lambda\big(\hat{w}_h(Z) - h\big)$$

A Fully Online Meta-Algorithm

We apply SGD to $\hat{\mathcal{E}}$ **approximating the gradients** by the last inner iterate

$$\hat{\nabla}^{(t)} = -\lambda\Big(w_{h^{(t)}}^{(n+1)}(Z^{(t)}) - h^{(t)}\Big) \approx \nabla \mathcal{L}_{Z^{(t)}}\big(h^{(t)}\big) \qquad (2)$$

Algorithm 2 Meta-Algorithm

Input $\gamma > 0$ step-size, $\lambda > 0$ regularization parameter, ρ meta-distribution

Initialization $h^{(1)} = 0 \in \mathbb{R}^d$

For $t = 1$ to T

Receive $\mu_t \sim \rho$, $Z^{(t)} \sim \mu_t^n$

Run **Alg. 1** with $Z^{(t)}$ and $h^{(t)}$ and compute $\hat{\nabla}^{(t)}$ as in Eq. (2)

Update $h^{(t+1)} = h^{(t)} - \gamma \hat{\nabla}^{(t)}$

Return $(h^{(t)})_{t=1}^{T+1}$ and $\bar{h}_T = \frac{1}{T}\sum_{t=1}^{T} h^{(t)}$

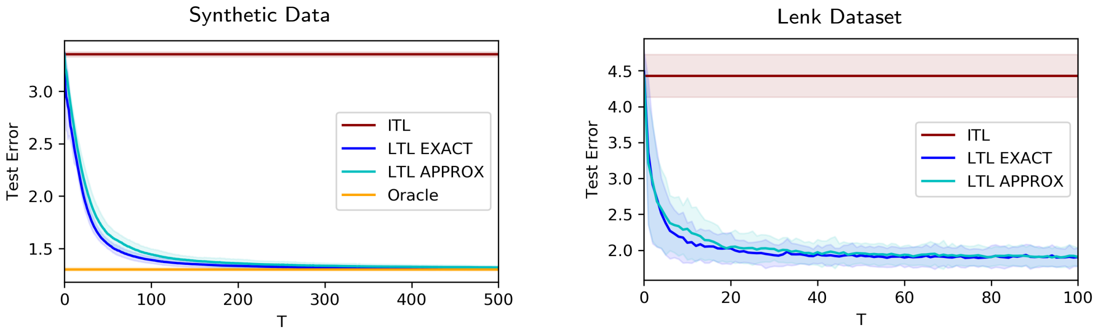
Theoretical Guarantees

Theorem 2. *Let* $\bar{h}_T$ *be the output of* **Alg. 2** *with an appropriate* γ. *Let* $\bar{w}_{\bar{h}_T}$ *be the output of* **Alg. 1** *with bias* $h = \bar{h}_T$ *and an appropriate* λ. *Then, in expectation w.r.t. the sampling of the datasets*

$$\mathbb{E}\left[\mathcal{E}(\bar{w}_{\bar{h}_T})\right] - \mathcal{E}_\rho \leq \mathrm{Var}_{\mathrm{m}}\, 4RL\sqrt{\frac{\log(n)+1}{n}} + \|\mathrm{m}\| RL\sqrt{2\Big(1+\frac{4\big(\log(n)+1\big)}{n}\Big)\frac{1}{T}}$$

- When $T \to \infty$, we retrieve the oracle ($\implies$ advantage w.r.t. ITL)
- The approximation error on the meta-gradients does not affect

Experiments



Averaged test performance vs. the number of training tasks

MORE AT THE POSTER #257