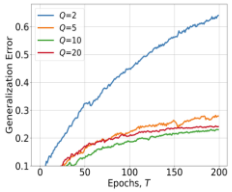
On the Stability and Generalization of Meta-Learning: the Impact of Inner-Levels

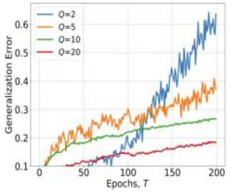
Wenjun Ding[1], Jingling Liu[1], Lixing Chen[2], Xiu Su[1], Tao Sun[3], Fan Wu[1], Zhe Qu[1]

[1]Central South University [2]Shanghai Jiao Tong University [3]National University of Defense Technology

1.Observations on Generalization Error



(a) MAML.



(b) Meta-MinibatchProx.

$$w_{\mathcal{T}_i} = w - \alpha \sum_{q=0}^{Q-1} \nabla \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}^q, \mathcal{D}_i) = \arg\min_{w_{\mathcal{T}_i}} \{ \langle \sum_{q=0}^{Q-1} \nabla \mathcal{L}_i(w_{\mathcal{T}_i}^q, \mathcal{D}_i), w_{\mathcal{T}_i} - w \rangle + \frac{1}{2\alpha} \|w_{\mathcal{T}_i} - w\|_2^2 \}$$

$$w_{\mathcal{T}_i} = \mathrm{argmin}_{w_{\mathcal{T}_i}} \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i}, \mathcal{D}_i) + \frac{\lambda}{2} \|w_{\mathcal{T}_i} - w\|^2$$

It can be observed that, under two different inner-level optimization processes—**gradient descent and proximal descent**—the generalization error exhibits **two distinct dependencies** on the number of inner-task update steps Q.

2.Contributions

- We summarize six mainstream meta-learning algorithms and extract their structural features. Based on their inner-processes, we classify these algorithms into two frameworks: **GDF and PDF**, and develop **two definitions of on-average stability**, respectively. Accordingly, we establish a quantitative relationship between **innerlevels and the generalization error** in convex and non-convex settings.
- Our results reveal the influence of the inner-levels Q on generalization error. In particular, we identify a **trade-off relationship** in GDF, whereas PDF demonstrates a **beneficial relationship** in its generalization bound. The primary reason for this difference lies in the term introduced by the inner-process. For example, in convex setting, the term for GDF, $O(\frac{TQ}{mn^{tr}})$, increases with Q, whereas the term for PDF, $O(\frac{T}{mCQ})$, decreases with Q. These findings help to design a more efficient inner-process of meta-learning.
- Based on the generalization results of GDF and PDF, we further derive the generalization bounds for **six meta-learning algorithms** and analyze their implications. In general, note that the meta-objective $F(w)$ plays a crucial role in reducing the generalization bound. Motivated by this, we propose **a new meta-objective** $F_{new}(w)$ and prove $F_{new}(w) < F(w)$, thereby enhancing generalization performance. Extensive experiments confirm the efficiency of the proposed objective.

3.Results

Algorithm 1 GDF and PDF

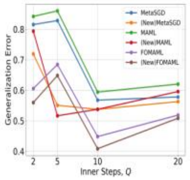
1: The set of datasets $\mathcal{S} = \{S_i\}_{i=1}^m$, outer iterations T, inner-levels Q, regulation λ.
2: Choose arbitrary initial point $w^0 \in W$;
3: **for** $t = 0$ **to** $T-1$ **do**
4: Randomly choose the task i.
5: Inner-Level: $w_{\mathcal{T}_i,0}^t = w_t$
6: **for** $q = 0, 1, ..., Q-1$ **do**
7: $w_{\mathcal{T}_i,q+1}^t = w_{\mathcal{T}_i,q}^t - \alpha \nabla \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,q}^t, S_i^{\mathrm{tr}})$;
8: $w_{\mathcal{T}_i,q+1}^t = w_{\mathcal{T}_i,q}^t - \alpha \nabla \widehat{\mathcal{K}}_i(w_{\mathcal{T}_i,q}^t, S_i)$;
9: **end for**
10: $w^{t+1} := w^t - \eta_t \nabla_w \widehat{\mathcal{L}}_i(w_{\mathcal{T}_i,Q}^t, S_i^{\mathrm{ts}})$
11: $w^{t+1} := w^t - \eta_t \lambda (w^t - w_{\mathcal{T}_i,Q}^t)$
12: **end for**
13: w^T and $\overline{w}^T := \frac{1}{T+1} \sum_{t=0}^T w^t$;

Analysis.

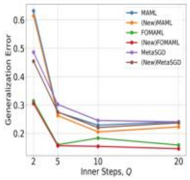
- In GDF, the first term of the bound increases with Q, worsening generalization, while the second term decreases with Q,indicating a trade-off.
- In PDF, the adverse effect of Qon the first term disappears, whereas its beneficial effect on the second term remains.

Frame.	Algorithm	Convex	Non-convex
GDF	MAML	$\mathcal{O}(\sum_{t=0}^{T-1} \eta_t (1+\alpha L)^{Q-1} \frac{(6QG+Q^2\alpha^2G^2\rho)}{mn^{\mathrm{tr}}} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{c\gamma}}{m}(1+(1+\alpha L)^{2(Q-1)} \frac{(6QG+Q^2\alpha^2G^2\rho)}{n^{\mathrm{tr}}}))^{\frac{1}{c\gamma}} (F(w^0)T)^{\frac{c\gamma}{1+c\gamma}})$
	FOMAML	$\mathcal{O}(\sum_{t=0}^{T-1} \eta_t \frac{2Q\alpha LG}{mn^{\mathrm{tr}}} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{c\gamma}}{m}(1+\frac{2Q(1+\alpha L)^Q\alpha LG}{n^{\mathrm{tr}}}))^{\frac{1}{c\gamma}} (F(w^0)T)^{\frac{c\gamma}{1+c\gamma}})$
	Meta-SGD	$\mathcal{O}(\sum_{t=0}^{T-1} \eta_t (1+\widehat{\alpha}_t L)^{Q-1} \frac{(6QG+Q^2\widehat{\alpha}_t^2G^2\rho)}{mn^{\mathrm{tr}}} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{c\gamma}}{m}(1+(1+\widehat{\alpha} L)^{2(Q-1)} \frac{(6QG+Q^2\widehat{\alpha}^2G^2\rho)}{n^{\mathrm{tr}}}))^{\frac{1}{c\gamma}} (F(w^0)T)^{\frac{c\gamma}{1+c\gamma}})$
PDF	iMAML	$\mathcal{O}(\sum_{t=0}^{T-1} \frac{2L\eta_t(G^2+G)}{\lambda mn^{\mathrm{tr}}} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{c\gamma}}{m}(1+\frac{2L(G^2+G)}{(\lambda-L)n^{\mathrm{tr}}}))^{\frac{1}{c\gamma}} (F(w^0)T)^{\frac{c\gamma}{1+c\gamma}})$
	Meta-MinibatchProx	$\mathcal{O}(\sum_{t=0}^{T-1} \frac{2\eta_t\lambda}{mCQ} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{c\gamma}}{m}(1+\frac{G(\lambda+L)}{CQ}))^{\frac{1}{c\gamma}} (F(w^0)T)^{\frac{c\gamma}{1+c\gamma}})$
	Fo-MuML	$\mathcal{O}(\sum_{t=0}^{T-1} \frac{2\eta_t G^2}{\lambda mn^{\mathrm{tr}}} + \frac{\mathcal{Q}(F(w^0))}{m})$	$\mathcal{O}(\frac{1+\frac{1}{c\gamma}}{m}(1+\frac{2G^2}{(\lambda-L)n^{\mathrm{tr}}})^{\frac{1}{c\gamma}} (F(w^0)T)^{\frac{c\gamma}{1+c\gamma}})$

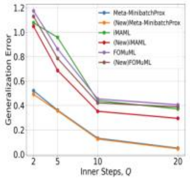
4.Experiment



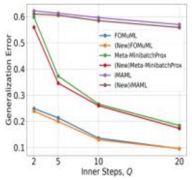
(a) Convex GDF



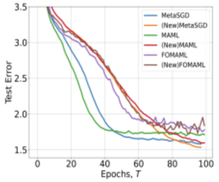
(b) Non-Convex GDF



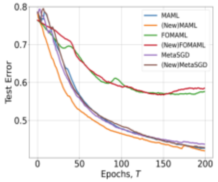
(c) Convex PDF



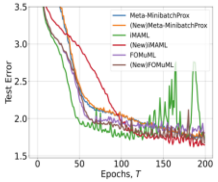
(d) Non-Convex PDF



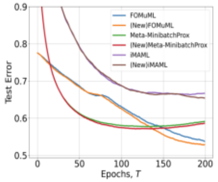
(a) Convex GDF.



(b) Non-Convex GDF.



(c) Convex PDF.



(d) Non-Convex PDF.