1 Proof

Proposition 1. *Let $\eta_t = 1/t$. Assume $\gamma < 1/\sigma_{\mathbf{x}}^2$. Both $\boldsymbol{\psi}_t$ in MAML (with one inner gradient step) and $\boldsymbol{\theta}_t$ in CommonMean converge to $\bar{\mathbf{w}} = \mathbb{E}_\tau \mathbf{w}_\tau^*$.*

Proposition 2. *Assume that $\gamma < 1/\sigma_{\mathbf{x}}^2$. We have $\bar{\mathbf{w}} = \operatorname{argmin}_{\boldsymbol{\theta}} \mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\mathbf{x}^\top \mathbf{w}_\tau^{(prox)} - y)^2 = \operatorname{argmin}_{\boldsymbol{\psi}} \mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\mathbf{x}^\top \mathbf{w}_\tau^{(gd)} - y)^2$.*

We first prove Proposition 2 that the mean regressor is the unique minimizer. Then, we prove Proposition 1 by showing that MAML (with one inner gradient step) and CommonMean algorithms achieve global convergence.

1.1 Proof of Proposition 2

Proof. For each task τ, let $\mathbf{v}_\tau = \mathbf{w}_\tau^* - \bar{\mathbf{w}}$, then $\{\mathbf{v}_\tau\}$ are i.i.d. random variables with zero mean. Denote $\mathbf{C}_\tau = \left(\lambda \mathbf{I} + \mathbf{X}_\tau^\top \mathbf{X}_\tau\right)^{-1}$. As $\mathbf{w}_\tau^{(\text{prox})} = \mathbf{C}_\tau \left(\lambda \boldsymbol{\theta} + \mathbf{X}_\tau^\top \mathbf{y}_\tau\right)$ and $\mathbf{y}_\tau = \mathbf{X}_\tau \mathbf{w}_\tau^* + \boldsymbol{\xi}_\tau$, it follows that

$$
\begin{aligned}
&\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\mathbf{x}^\top \mathbf{w}_\tau^{(\text{prox})} - y)^2 \\
=&\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\lambda \mathbf{x}^\top \mathbf{C}_\tau \boldsymbol{\theta} + \mathbf{x}^\top \mathbf{C}_\tau \mathbf{X}_\tau^\top (\mathbf{X}_\tau \mathbf{w}_\tau^* + \boldsymbol{\xi}_\tau) - \mathbf{x}^\top \mathbf{w}_\tau^* - \xi)^2 \\
=&\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\lambda \mathbf{x}^\top \mathbf{C}_\tau \boldsymbol{\theta} + \mathbf{x}^\top \mathbf{C}_\tau \mathbf{X}_\tau^\top (\mathbf{X}_\tau \bar{\mathbf{w}} + \mathbf{X}_\tau \mathbf{v}_\tau + \boldsymbol{\xi}_\tau) - \mathbf{x}^\top \bar{\mathbf{w}} - \mathbf{x}^\top \mathbf{v}_\tau - \xi)^2 \\
=&\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\lambda \mathbf{x}^\top \mathbf{C}_\tau \boldsymbol{\theta} + \mathbf{x}^\top \mathbf{C}_\tau \mathbf{X}_\tau^\top \mathbf{X}_\tau \bar{\mathbf{w}} - \mathbf{x}^\top \bar{\mathbf{w}})^2 + \text{constant} \qquad (1) \\
=&\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\lambda \mathbf{x}^\top \mathbf{C}_\tau (\boldsymbol{\theta} - \bar{\mathbf{w}}))^2 + \text{constant} \\
=&\lambda^2 \sigma_{\mathbf{x}}^2 n_q \mathbb{E}_\tau \mathbb{E}_{S_\tau} (\boldsymbol{\theta} - \bar{\mathbf{w}})^\top \mathbf{C}_\tau^2 (\boldsymbol{\theta} - \bar{\mathbf{w}}) + \text{constant},
\end{aligned}
$$

where we have used the setting that $\mathbf{x}, \xi, \mathbf{X}_\tau, \boldsymbol{\xi}_\tau$, and $\mathbf{v}_\tau$ are independent to obtain (1). Since $\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbf{C}_\tau^2 \succeq \lambda^{-2} \mathbf{I}$, we conclude that $\boldsymbol{\theta} = \bar{\mathbf{w}}$ is the unique optima.

For MAML with one gradient step $\mathbf{w}_\tau^{(\text{gd})} = \boldsymbol{\psi} - \gamma \mathbf{X}_\tau^\top (\mathbf{X}_\tau \boldsymbol{\psi} - \mathbf{y}_\tau)$, it follows that

$$
\begin{aligned}
&\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\mathbf{x}^\top \mathbf{w}_\tau^{(\text{gd})} - y)^2 \\
=&\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\mathbf{x}^\top (\mathbf{I} - \gamma \mathbf{X}_\tau^\top \mathbf{X}_\tau) \boldsymbol{\psi} + \gamma \mathbf{x}^\top \mathbf{X}_\tau^\top \mathbf{y}_\tau - y)^2 \\
=&\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\mathbf{x}^\top (\mathbf{I} - \gamma \mathbf{X}_\tau^\top \mathbf{X}_\tau) \boldsymbol{\psi} + \gamma \mathbf{x}^\top \mathbf{X}_\tau^\top (\mathbf{X}_\tau \bar{\mathbf{w}} + \mathbf{X}_\tau \mathbf{v}_\tau + \boldsymbol{\xi}_\tau) - \mathbf{x}^\top \bar{\mathbf{w}} - \mathbf{x}^\top \mathbf{v}_\tau - \xi)^2 \\
=&\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} \left(\mathbf{x}^\top (\mathbf{I} - \gamma \mathbf{X}_\tau^\top \mathbf{X}_\tau)(\boldsymbol{\psi} - \bar{\mathbf{w}})\right)^2 + \text{constant} \\
=&n_q \sigma_{\mathbf{x}}^2 \mathbb{E}_\tau \mathbb{E}_{S_\tau} \|(\mathbf{I} - \gamma \mathbf{X}_\tau^\top \mathbf{X}_\tau)(\boldsymbol{\psi} - \bar{\mathbf{w}})\|^2 + \text{constant}.
\end{aligned}
$$

As $\gamma < {}^1/_{\sigma_{\mathbf{x}}^2}$, we conclude that $\boldsymbol{\psi} = \bar{\mathbf{w}}$ is the unique optima. □

1.2 Proof of Proposition 1

Proof. (i) Notice that $\mathbf{w}_\tau^{(\text{prox})}$ is affine in $\boldsymbol{\theta}$, thus, $\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\mathbf{x}^\top \mathbf{w}_\tau^{(\text{prox})} - y)^2$ is convex in $\boldsymbol{\theta}$. The CommonMean algorithm is using stochastic gradient descent to minimize the population risk, and the global convergence of $\boldsymbol{\theta}_t$ follows from the stochastic convex optimization [1].

(ii) Similarly, $\mathbf{w}_\tau^{(\text{gd})}$ is affine in $\boldsymbol{\psi}$, thus, the loss $\mathbb{E}_\tau \mathbb{E}_{S_\tau} \mathbb{E}_{Q_\tau} \sum_{(\mathbf{x},y)\in Q_\tau} (\mathbf{x}^\top \mathbf{w}_\tau^{(\text{gd})} - y)^2$ is convex in $\boldsymbol{\psi}$. Using stochastic gradient descent, $\boldsymbol{\psi}_t$ achieves global convergence [1]. By Proposition 2, $\bar{\mathbf{w}}$ is the unique optima, and we finish the proof. □

1.3 Proof of Proposition 4

The task index τ' will be omitted for simplifying notations in Proposition 4.

Proposition 4. $\mathbb{E}_{\boldsymbol{\xi}}\|\mathbf{w}^{(prox)} - \mathbf{w}^*\|^2 = \|\tilde{\mathbf{b}}\|^2 + \sum_{j=1}^{n_s}\left(\frac{\lambda \tilde{\mathrm{a}}_j}{\lambda+\nu_j^2}\right)^2 + \sum_{j=1}^{n_s}\left(\frac{\sigma_\xi}{(\lambda/\nu_j)+\nu_j}\right)^2$, *where the expectation is over the label noise vector* $\boldsymbol{\xi}$*.*

Proof. The ridge regression has a closed-form solution $\mathbf{w}^{(\text{prox})} = \left(\lambda\mathbf{I} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\lambda\boldsymbol{\theta} + \mathbf{X}^\top\mathbf{y}\right)$. Using the SVD decomposition of $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ and $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}$, we obtain

$$\begin{aligned}\mathbf{w}^{(\text{prox})} &= \left(\mathbf{I} + \lambda^{-1}\mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^\top\right)^{-1}\left(\mathbf{V}\mathbf{a}_0 + \mathbf{V}^\perp\mathbf{b}_0 + \lambda^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y}\right) \\ &= \left(\mathbf{I} + \lambda^{-1}\mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^\top\right)^{-1}\left(\mathbf{V}\mathbf{a}_0 + \mathbf{V}^\perp\mathbf{b}_0 + \lambda^{-1}\mathbf{V}\boldsymbol{\Sigma}^2\mathbf{a}^* + \lambda^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}\boldsymbol{\xi}\right) \qquad (2)\\ &= \mathbf{V}^\perp\mathbf{b}_0 + \mathbf{V}(\mathbf{I} + \lambda^{-1}\boldsymbol{\Sigma}^2)^{-1}\left(\mathbf{a}_0 + \lambda^{-1}\boldsymbol{\Sigma}^2\mathbf{a}^*\right) + \mathbf{V}\left(\lambda\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{U}^\top\boldsymbol{\xi}, \qquad (3)\end{aligned}$$

where we have used $\mathbf{U}^\top\mathbf{y} = \mathbf{U}^\top(\mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}) = \mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top(\mathbf{V}\mathbf{a}^* + \mathbf{V}^\perp\mathbf{b}^*) + \mathbf{U}^\top\boldsymbol{\xi} = \boldsymbol{\Sigma}\mathbf{a}^* + \mathbf{U}^\top\boldsymbol{\xi}$ in (2) and the Woodbury identity in (3). Then the estimation error is

$$\mathbf{w}^{(\text{prox})} - \mathbf{w}^* = \mathbf{V}^\perp(\mathbf{b}_0 - \mathbf{b}^*) + \mathbf{V}(\mathbf{I} + \lambda^{-1}\boldsymbol{\Sigma}^2)^{-1}\left(\mathbf{a}_0 - \mathbf{a}^*\right) + \mathbf{V}\left(\lambda\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{U}^\top\boldsymbol{\xi}.$$

Taking the square ℓ_2-norm and then expectation over $\boldsymbol{\xi}$ on both sides, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\xi}}\|\mathbf{w}^{(\text{prox})} - \mathbf{w}^*\|^2 \\ =&\|\mathbf{V}^\perp(\mathbf{b}_0 - \mathbf{b}^*)\|^2 + \|\mathbf{V}(\mathbf{I} + \lambda^{-1}\boldsymbol{\Sigma}^2)^{-1}\left(\mathbf{a}_0 - \mathbf{a}^*\right)\|^2 + \mathbb{E}_{\boldsymbol{\xi}}\|\mathbf{V}\left(\lambda\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{U}^\top\boldsymbol{\xi}\|^2 \qquad (4)\\ =&\|\mathbf{b}_0 - \mathbf{b}^*\|^2 + \|(\mathbf{I} + \lambda^{-1}\boldsymbol{\Sigma}^2)^{-1}\left(\mathbf{a}_0 - \mathbf{a}^*\right)\|^2 + \mathbb{E}_{\boldsymbol{\xi}}\|\left(\lambda\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{U}^\top\boldsymbol{\xi}\|^2 \\ =&\|\tilde{\mathbf{b}}\|^2 + \sum_{j=1}^{n_s}\left(\frac{\lambda\tilde{\mathrm{a}}_j}{\lambda + \nu_j^2}\right)^2 + \sum_{j=1}^{n_s}\left(\frac{\nu_j\sigma_\xi}{\lambda + \nu_j^2}\right)^2,\end{aligned}$$

where (4) follows from the fact that $\mathbf{V}^\perp$ is $\mathbf{V}$'s orthogonal complement and $\boldsymbol{\xi}$ is independent with $\mathbf{X}$ (also the $\boldsymbol{\Sigma}$, $\mathbf{U}$ and $\mathbf{V}$). $\square$

1.4 Proof of Theorem 1

Lemma 1. $\mathcal{L}_{meta}(\boldsymbol{\theta}, \boldsymbol{\phi})$ *is Lipschitz-smooth w.r.t.* $(\boldsymbol{\theta}, \boldsymbol{\phi})$ *with a Lipschitz constant* β_{meta}*.*

Lipschitz-smoothness is a basic assumption to establish convergence of gradient descent algorithms in stochastic non-convex optimization [4, 8] and meta-learning in non-convex settings [2, 11].

Proof of Lemma 1. As $\mathcal{L}_{\text{meta}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \equiv \sum_{\tau\in\mathcal{T}}\sum_{(\mathbf{x},y)\in Q_\tau}\ell(\hat{y}, y)$, it suffices to show that $\ell(\hat{y}, y)$ is Lipschitz-smooth in $(\boldsymbol{\theta}, \boldsymbol{\phi})$.

Using the chain rule, we have

$$\begin{aligned}\nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\ell(\hat{y}, y) &= \nabla_1\ell(\hat{y}, y)\nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\hat{y}, \qquad (5)\\ \nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\hat{y} &= \nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}f_{\boldsymbol{\theta}}(\mathbf{z}) + (\nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\mathcal{K}(\mathbf{Z}_\tau, \mathbf{z}))^\top\boldsymbol{\alpha}_\tau + (\nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\boldsymbol{\alpha}_\tau)^\top\mathcal{K}(\mathbf{Z}_\tau, \mathbf{z}). \qquad (6)\end{aligned}$$

The Lipschitz properties of direct derivatives $\nabla_1\ell(\hat{y}, y), \nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}f_{\boldsymbol{\theta}}(\mathbf{z}), \nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\mathcal{K}(\mathbf{Z}_\tau, \mathbf{z})$, and $\mathcal{K}(\mathbf{Z}_\tau\mathbf{z})$ follow from the Assumption 1. It remains to claim $\boldsymbol{\alpha}_\tau$ and $\nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\boldsymbol{\alpha}_\tau$ are Lipschitz. Let $\mathbf{p} = [f_{\boldsymbol{\theta}}(\mathbf{z}_1); \ldots; f_{\boldsymbol{\theta}}(\mathbf{z}_{n_s}); \mathcal{K}(\mathbf{Z}_\tau, \mathbf{z}_1); \ldots; \mathcal{K}(\mathbf{Z}_\tau, \mathbf{z}_{n_s})] \in \mathbb{R}^{n_s + n_s^2}$ be the input of the dual problem.

(i) Claim: $\underline{\boldsymbol{\alpha}_\tau \text{ is Lipschitz w.r.t. } (\boldsymbol{\theta},\boldsymbol{\phi}) \text{ and } \boldsymbol{\alpha}_\tau(\mathbf{p}) \text{ is Lipschitz-smooth w.r.t. } \mathbf{p}}$. To show $\boldsymbol{\alpha}_\tau$ is Lipschitz w.r.t. $(\boldsymbol{\theta},\boldsymbol{\phi})$, it suffices to show that $\|\nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\boldsymbol{\alpha}_\tau\|$ is bounded. By the chain rule, $\nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\boldsymbol{\alpha}_\tau = \nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau\nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\mathbf{p}$. Denote the dual objective by $g(\mathbf{p}, \boldsymbol{\alpha})$. By the implicit function theorem [9], $\nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau = -\left(\nabla^2_{\boldsymbol{\alpha}}g(\mathbf{p}, \boldsymbol{\alpha}_\tau)\right)^{-1}\frac{\partial^2}{\partial\mathbf{p}\partial\boldsymbol{\alpha}}g(\mathbf{p}, \boldsymbol{\alpha}_\tau)$, where $\nabla^2_{\boldsymbol{\alpha}}g(\mathbf{p}, \boldsymbol{\alpha}_\tau) = \sum_{(\mathbf{x}_i,y_i)\in S_\tau}\nabla^2_1\ell(f_\tau(\mathbf{z}_i), y_i)\mathcal{K}(\mathbf{Z}_\tau, \mathbf{z}_i)\mathcal{K}(\mathbf{Z}_\tau, \mathbf{z}_i)^\top + \mathcal{K}(\mathbf{Z}_\tau, \mathbf{Z}_\tau)$, $\frac{\partial^2}{\partial\mathbf{p}\partial\boldsymbol{\alpha}}g(\mathbf{p}, \boldsymbol{\alpha}_\tau) = \left[\mathcal{K}(\mathbf{Z}_\tau, \mathbf{Z}_\tau)\mathbf{D} \mid (\mathcal{K}(\mathbf{Z}_\tau, \mathbf{Z}_\tau)\mathbf{D}) \otimes \boldsymbol{\alpha}_\tau^\top + \mathbf{v}^\top \otimes \mathbf{I} + \mathbf{I} \otimes \boldsymbol{\alpha}_\tau^\top\right]$, $\mathbf{D} =$

$\mathrm{diag}([\nabla_1^2\ell(f_\tau(\mathbf{z}_1),y_1);\ldots;\nabla_1^2\ell(f_\tau(\mathbf{z}_{n_s}),y_{n_s})])$, $\mathbf{v} = [\nabla_1\ell(f_\tau(\mathbf{z}_1),y_1);\ldots;\nabla_1\ell(f_\tau(\mathbf{z}_{n_s}),y_{n_s})]$, where $\otimes$ is the Kronecker product. It follows from the Assumption 1 that both $\nabla^2_{\boldsymbol{\alpha}} g(\mathbf{p},\boldsymbol{\alpha}_\tau)$ and $\frac{\partial^2}{\partial\mathbf{p}\partial\boldsymbol{\alpha}} g(\mathbf{p},\boldsymbol{\alpha}_\tau)$ are Lipschitz w.r.t. $\mathbf{p}$. Hence, we conclude that $\nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau(\mathbf{p})$ is Lipschitz, $\boldsymbol{\alpha}_\tau(\mathbf{p})$ is Lipschitz-smooth w.r.t. $\mathbf{p}$, and $\|\nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau(\mathbf{p})\|$ is bounded. Again, the boundedness of $\nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\mathbf{p}$ follows from the Lipschitz-smoothness of $\mathbf{p}$ w.r.t. $(\boldsymbol{\theta},\boldsymbol{\phi})$. We conclude that $\boldsymbol{\alpha}_\tau$ is Lipschitz w.r.t. $(\boldsymbol{\theta},\boldsymbol{\phi})$.

(ii) Claim: $\underline{\nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\boldsymbol{\alpha}_\tau \text{ is Lipschitz w.r.t. } (\boldsymbol{\theta},\boldsymbol{\phi})}$. Given $(\boldsymbol{\theta},\boldsymbol{\phi})$ and $(\boldsymbol{\theta}',\boldsymbol{\phi}')$, we show $\|\nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\boldsymbol{\alpha}_\tau(\boldsymbol{\theta},\boldsymbol{\phi}) - \nabla_{(\boldsymbol{\theta},\boldsymbol{\phi})}\boldsymbol{\alpha}_\tau(\boldsymbol{\theta}',\boldsymbol{\phi}')\| \leq \beta\|(\boldsymbol{\theta},\boldsymbol{\phi}) - (\boldsymbol{\theta}',\boldsymbol{\phi}')\|$ for some $\beta > 0$. For notation simplicity, let $\boldsymbol{\varphi} = (\boldsymbol{\theta},\boldsymbol{\phi})$ and $\boldsymbol{\varphi}' = (\boldsymbol{\theta}',\boldsymbol{\phi}')$, then we have

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{\varphi}}\boldsymbol{\alpha}_\tau(\boldsymbol{\varphi}) - \nabla_{\boldsymbol{\varphi}}\boldsymbol{\alpha}_\tau(\boldsymbol{\varphi}')\| \\
=&\|\nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau(\mathbf{p}(\boldsymbol{\varphi}))\nabla_{\boldsymbol{\varphi}}\mathbf{p}(\boldsymbol{\varphi}) - \nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau(\mathbf{p}(\boldsymbol{\varphi}'))\nabla_{\boldsymbol{\varphi}}\mathbf{p}(\boldsymbol{\varphi}')\| \\
=&\|\nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau(\mathbf{p}(\boldsymbol{\varphi}))\nabla_{\boldsymbol{\varphi}}\mathbf{p}(\boldsymbol{\varphi}) - \nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau(\mathbf{p}(\boldsymbol{\varphi}'))\nabla_{\boldsymbol{\varphi}}\mathbf{p}(\boldsymbol{\varphi}') \pm \nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau(\mathbf{p}(\boldsymbol{\varphi}))\nabla_{\boldsymbol{\varphi}}\mathbf{p}(\boldsymbol{\varphi}')\| \\
\leq&\|\nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau(\mathbf{p}(\boldsymbol{\varphi}))\|\|\nabla_{\boldsymbol{\varphi}}\mathbf{p}(\boldsymbol{\varphi}) - \nabla_{\boldsymbol{\varphi}}\mathbf{p}(\boldsymbol{\varphi}')\| + \|\nabla_{\boldsymbol{\varphi}}\mathbf{p}(\boldsymbol{\varphi}')\|\|\nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau(\mathbf{p}(\boldsymbol{\varphi})) - \nabla_{\mathbf{p}}\boldsymbol{\alpha}_\tau(\mathbf{p}(\boldsymbol{\varphi}'))\|.
\end{aligned}
$$

As $\mathbf{p}(\boldsymbol{\varphi})$ and $\boldsymbol{\alpha}_\tau(\mathbf{p})$ are Lipschitz-smooth, there exists $\beta > 0$ such that

$$
\begin{aligned}
\|\nabla_{\boldsymbol{\varphi}}\boldsymbol{\alpha}_\tau(\boldsymbol{\varphi}) - \nabla_{\boldsymbol{\varphi}}\boldsymbol{\alpha}_\tau(\boldsymbol{\varphi}')\| &\leq \beta\|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\| + \beta\|\mathbf{p}(\boldsymbol{\varphi}) - \mathbf{p}(\boldsymbol{\varphi}')\| \\
&\leq \beta\|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\| + \beta\|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\| \\
&= 2\beta\|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\|.
\end{aligned}
$$

We conclude that $\nabla_{\boldsymbol{\varphi}}\boldsymbol{\alpha}_\tau$ is Lipschitz.

By (i) and (ii), ℓ is Lipschitz-smooth w.r.t. the meta-parameters $\boldsymbol{\varphi}$. Therefore, $\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi})$ is Lipschitz-smooth w.r.t. $\boldsymbol{\varphi}$ with a Lipschitz constant $\beta_{\text{meta}} > 0$. □

Theorem 1. *Let the step size be* $\eta_t = \min(1/\sqrt{T}, 1/\beta_{meta})$. *Algorithm 3 satisfies* $\min_{1\leq t\leq T}\mathbb{E}\|\nabla_{(\boldsymbol{\theta}_t,\boldsymbol{\phi}_t)}\mathcal{L}_{meta}(\boldsymbol{\theta}_t,\boldsymbol{\phi}_t)\|^2 = \mathcal{O}\left(\sigma^2_{\mathbf{g}}/\sqrt{T}\right)$*, where the expectation is taken over the random training samples.*

The proof is similar to non-convex stochastic programming [4].

Proof of Theorem 1. Let $\boldsymbol{\varphi} = (\boldsymbol{\theta},\boldsymbol{\phi})$. Let $\boldsymbol{\zeta}_t = \nabla_{\boldsymbol{\varphi}_t}\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t) - \frac{1}{b}\sum_{\tau\in\mathcal{B}_t}\mathbf{g}_\tau$, where $\frac{1}{b}\sum_{\tau\in\mathcal{B}_t}\mathbf{g}_\tau$ is an unbiased estimation of $\nabla_{\boldsymbol{\varphi}_t}\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t)$, Using the Taylor expansion, we have

$$
\begin{aligned}
&\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_{t+1}) \\
&\leq \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t) + \nabla_{\boldsymbol{\varphi}_t}\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t)^\top(\boldsymbol{\varphi}_{t+1} - \boldsymbol{\varphi}_t) + \frac{1}{2}\beta_{\text{meta}}\|\boldsymbol{\varphi}_{t+1} - \boldsymbol{\varphi}_t\|^2 \\
&\leq \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t) - \eta_t\Big(1 - \frac{\beta_{\text{meta}}\eta_t}{2}\Big)\|\nabla_{\boldsymbol{\varphi}_t}\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t)\|^2 + \eta_t\nabla^\top_{\boldsymbol{\varphi}_t}\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t)\boldsymbol{\zeta}_t + \frac{1}{2}\beta_{\text{meta}}\eta_t^2\sigma^2_{\mathbf{g}}.
\end{aligned}
$$

Taking conditional expectation over $\boldsymbol{\zeta}_{t-1}$ on both sides and then take the expectation over the random training samples, we have

$$
\mathbb{E}\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_{t+1}) \leq \mathbb{E}\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t) - \frac{\eta_t}{2}\mathbb{E}\|\nabla_{\boldsymbol{\varphi}_t}\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t)\|^2 + \frac{1}{2}\beta_{\text{meta}}\eta_t^2\sigma^2_{\mathbf{g}}, \tag{7}
$$

where we have used $1 - \frac{\beta_{\text{meta}}\eta_t}{2} \geq \frac{1}{2}$. Rearranging the above inequality and summing over t, we have

$$
\sum_{t=1}^T \frac{\eta_t}{2}\mathbb{E}\|\nabla_{\boldsymbol{\varphi}_t}\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t)\|^2 \leq \mathbb{E}\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_1) + \beta_{\text{meta}}\sigma^2_{\mathbf{g}}\sum_{t=1}^T\eta_t^2. \tag{8}
$$

Since $\eta_t = \min(1/\sqrt{T}, 1/2\beta_{\text{meta}})$, we have $\sum_{t=1}^T\eta_t^2 \leq 1$. Diving both sides by $1/\sqrt{T}$, we conclude that $\min_{1\leq t\leq T}\mathbb{E}\|\nabla_{\boldsymbol{\varphi}_t}\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t)\|^2 = \mathcal{O}\left(\sigma^2_{\mathbf{g}}/\sqrt{T}\right)$. □

1.5 Proof of Theorem 2

Theorem 2. *Assume that* $\mathcal{M}(\boldsymbol{\theta},\boldsymbol{\phi})$ *is uniform conditioning. (i) Let* $\eta_t = \min(1/\sqrt{T}, 1/2\beta_{meta})$*. Algorithm 3 satisfies* $\min_{1\leq t\leq T}\mathbb{E}\mathcal{L}_{meta}(\boldsymbol{\theta}_t,\boldsymbol{\phi}_t) - \min_{(\boldsymbol{\theta},\boldsymbol{\phi})}\mathcal{L}_{meta}(\boldsymbol{\theta},\boldsymbol{\phi}) = \mathcal{O}\left(\sigma^2_{\mathbf{g}}/\sqrt{T}\right)$*, where the expectation is taken over the random training samples. (ii) Let* $\eta_t = \eta < \min(1/2\beta_{meta}, 4|\mathcal{T}|/\rho\mu)$ *and* $\mathcal{B}_t = \mathcal{T}$*. Algorithm 3 satisfies* $\mathcal{L}_{meta}(\boldsymbol{\theta}_t,\boldsymbol{\phi}_t) - \min_{(\boldsymbol{\theta},\boldsymbol{\phi})}\mathcal{L}_{meta}(\boldsymbol{\theta},\boldsymbol{\phi}) = \mathcal{O}((1 - \eta\rho\mu/4|\mathcal{T}|)^t)$.

Proof of Theorem 2. Let $\boldsymbol{\varphi} = (\boldsymbol{\theta}, \boldsymbol{\phi})$. By the chain rule, we have

$$\nabla_{\boldsymbol{\varphi}} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \sum_{(\mathbf{x}, y) \in Q_\tau} \nabla_1 \ell(\hat{y}, y) \nabla_{\boldsymbol{\varphi}} \hat{y} \tag{9}$$

$$= \frac{1}{|\mathcal{T}|} \mathcal{G}(\boldsymbol{\varphi})^\top \nabla_{\boldsymbol{\varphi}} \mathcal{M}(\boldsymbol{\varphi}), \tag{10}$$

where $\mathcal{G}(\boldsymbol{\varphi}) \equiv [\cdots \quad \nabla_1 \ell(\hat{y}, y) \quad \cdots] \in \mathbb{R}^{n_q |\mathcal{T}|}$ stacks all gradients of the losses on query examples as a vector. Hence, we establish the Polyak-Lojasiewicz (PL) inequality [7] as follows

$$\begin{aligned}
\|\nabla_{\boldsymbol{\varphi}} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi})\|^2 &= \frac{1}{|\mathcal{T}|^2} \left\| \mathcal{G}(\boldsymbol{\varphi})^\top \nabla_{\boldsymbol{\varphi}} \mathcal{M}(\boldsymbol{\varphi}) \right\|^2 \\
&= \frac{1}{|\mathcal{T}|^2} \mathcal{G}(\boldsymbol{\varphi})^\top \nabla_{\boldsymbol{\varphi}} \mathcal{M}(\boldsymbol{\varphi}) \nabla_{\boldsymbol{\varphi}}^\top \mathcal{M}(\boldsymbol{\varphi}) \mathcal{G}(\boldsymbol{\varphi}) \\
&\geq \frac{\mu}{|\mathcal{T}|^2} \|\mathcal{G}(\boldsymbol{\varphi})\|^2 && \text{(uniform conditioning)} \\
&= \frac{\mu}{|\mathcal{T}|^2} \sum_{\tau \in \mathcal{T}} \sum_{(\mathbf{x}, y) \in Q_\tau} (\nabla_1 \ell(\hat{y}, y))^2 \\
&\geq \frac{\mu\rho}{2|\mathcal{T}|^2} \sum_{\tau \in \mathcal{T}} \sum_{(\mathbf{x}, y) \in Q_\tau} (\ell(\hat{y}, y) - \min_{y'} \ell(y', y)) && \text{(strongly convex)} \\
&\geq \frac{\mu\rho}{2|\mathcal{T}|} \left(\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}) - \min_{\boldsymbol{\varphi}} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}) \right).
\end{aligned}$$

The PL inequality is commonly used in proving the global convergence of nonconvex optimization [5, 6]. Then, $\min_{1 \leq t \leq T} \mathbb{E} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t) - \min_{\boldsymbol{\varphi}} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}) = \mathcal{O}\left(\sigma_{\mathbf{g}}^2 / \sqrt{T}\right)$ follows directly from Theorem 1.

For full gradient descent, the gradient noise $\boldsymbol{\zeta}_t = \nabla_{\boldsymbol{\varphi}_t} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t) - \frac{1}{b} \sum_{\tau \in \mathcal{B}_t} \mathbf{g}_\tau = \mathbf{0}$, thus, the noisy gradient will be the true gradient. By the Taylor expansion, it follows that

$$\begin{aligned}
&\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_{t+1}) - \min_{\boldsymbol{\varphi}} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}) \\
\leq& \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t) + \nabla_{\boldsymbol{\varphi}_t}^\top \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t)(\boldsymbol{\varphi}_{t+1} - \boldsymbol{\varphi}_t) + \frac{\beta_{\text{meta}}}{2} \|\boldsymbol{\varphi}_{t+1} - \boldsymbol{\varphi}_t\|^2 - \min_{\boldsymbol{\varphi}} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}) \\
=& \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t) - \eta \|\nabla_{\boldsymbol{\varphi}_t} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t)\|^2 + \frac{\eta^2 \beta_{\text{meta}}}{2} \|\nabla_{\boldsymbol{\varphi}} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t)\|^2 - \min_{\boldsymbol{\varphi}} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}) \\
\leq& \left(1 - \frac{\eta\mu\rho}{4|\mathcal{T}|}\right) (\mathcal{L}_{\text{meta}}(\boldsymbol{\varphi}_t) - \min_{\boldsymbol{\varphi}} \mathcal{L}_{\text{meta}}(\boldsymbol{\varphi})),
\end{aligned}$$

and we obtain the exponential convergence. □

2 Additional Experiments

2.1 Compared with MAML using a wide network on *Sine*

As the network width is critical to MAML, we perform few-shot regression experiments on *Sine* using the setting in [10]. We compare MetaProx with MAML that uses a larger (denoted by LargeMAML) and wider (denoted by VeryWideMAML) network. As can be seen from Table 1, MetaProx achieves the best performance.

2.2 MetaProx with RBF kernel on *Sine*

In this section, we evaluate the performance of MetaProx with the radial basis function (RBF) kernel on *Sine*. The RBF kernel is $\mathcal{K}(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2}\right)$, where $\sigma > 0$. Table 2 reports the results when σ varies from $\{0.01, 0.05, 0.1, 0.5, 1.0, 5.0\}$. As can be seen, a simple linear kernel is better.

Table 1: Average MSE (with 95% confidence intervals) of few-shot regression on *Sine* using the settings in [10]. Results of baselines are from [10].

method	5-shot	10-shot
OriginalMAML [3]	0.390 ± 0.156	0.114 ± 0.010
LargeMAML	0.208 ± 0.009	0.061 ± 0.004
VeryWideMAML	0.205 ± 0.013	0.059 ± 0.010
MetaFun [10]	0.040 ± 0.008	0.017 ± 0.005
MetaProx (proposed)	$\mathbf{0.010 \pm 0.001}$	$\mathbf{0.002 \pm 0.001}$

Table 2: Average MSE (with 95% confidence intervals) of MetaProx with different base kernels on *Sine* (noise-free).

kernel	2-shot	5-shot
RBF (0.01)	2.92 ± 0.19	2.78 ± 0.18
RBF (0.05)	2.72 ± 0.18	2.36 ± 0.17
RBF (0.1)	2.50 ± 0.17	2.25 ± 0.14
RBF (0.5)	2.38 ± 0.16	1.71 ± 0.13
RBF (1.0)	2.36 ± 0.16	1.68 ± 0.12
RBF (5.0)	2.38 ± 0.15	1.72 ± 0.13
linear	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.01 \pm 0.00}$

References

[1] J. C. Duchi. Introductory lectures on stochastic optimization. *The mathematics of data*, 25:99, 2018.

[2] A. Fallah, A. Mokhtari, and A. Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1082–1092, 2020.

[3] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135, 2017.

[4] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[5] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[6] C. Liu, L. Zhu, and M. Belkin. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. Preprint arXiv:2003.00307, 2020.

[7] B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[8] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pages 314–323, 2016.

[9] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 1976.

[10] J. Xu, J.-F. Ton, H. Kim, A. Kosiorek, and Y. W. Teh. MetaFun: Meta-learning with iterative functional updates. In *International Conference on Machine Learning*, pages 10617–10627, 2020.

[11] P. Zhou, X. Yuan, H. Xu, S. Yan, and J. Feng. Efficient meta learning via minibatch proximal update. In *Neural Information Processing Systems*, pages 1534–1544, 2019.