Supplementary materials for the paper
"Meta-Learning with a Geometry-Adaptive Preconditioner"

A. Toy example

To build an intuition for the effect of Riemannian metric, we construct a 2-D toy example over the parameter space. A learner minimizes an objective function of the form below.

$$\begin{aligned} f(x_1, x_2) =& x_1^2 + x_2^2 + x_1 x_2 \\ &+ \frac{1}{2}(\sin^2 5x_1 + \sin^2 5x_2) \\ &- \frac{1}{2}(\cos^2 3x_1 + \cos^2 3x_2) \end{aligned} \tag{13}$$

We set the initial point to $(x_1, x_2) = (-4, -2)$ and the learning rate to 0.1. In Figure 2 (a), we train the learner for 50 iterations. In Figure 2 (b), we define a preconditioner $\mathbf{P}_1$ as follows:

$$\mathbf{P}_1 = \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & -0.3 \end{bmatrix} \tag{14}$$

and train the learner with $\mathbf{P}_1$ for 13 iterations. In Figure 2 (c), we derive a preconditioner $\mathbf{P}_2$, which is the Riemannian metric corresponding to the parameter space (Eq. 13) as follows [32]:

$$\mathbf{P}_2 = \begin{bmatrix} 1 + u^2 & uv \\ uv & 1 + v^2 \end{bmatrix} \tag{15}$$

where $u = 2x_1 + x_2 + 3\sin(3x_1)\cos(3x_1) + 5\cos(5x_1)\sin(5x_1)$ and $v = 2x_2 + x_1 + 3\sin(3x_2)\cos(3x_2) + 5\cos(5x_2)\sin(5x_2)$. We train the learner with $\mathbf{P}_2$ for 50 iterations.

B. Proofs of Theorems

Definition 1. *Two $n \times n$ matrices A and B are similar if there exists an invertible $n \times n$ matrix P such that*

$$B = P^{-1}AP \tag{16}$$

Lemma 1. *Let $A = blkdiag(A_1, \cdots, A_n)$ be a block diagonal matrix such that the main-diagonal blocks A_i are $k \times k$ positive definite matrices. Then A is a positive definite matrix.*

Proof. First, we show that A is a positive definite matrix. For all non-zero $x = (x_1, \cdots, x_n) \in \mathbb{R}^{nk}$ where $x_i \in \mathbb{R}^k$, we can derive the following.

$$\begin{aligned} x^T A x &= x^T \text{blkdiag}(A_1, \cdots, A_n) x \\ &= x_1^T A_1 x_1 + \cdots x_n^T A_n x_n \\ &> 0 \ (\because A_i \text{ is a positive definite}) \end{aligned} \tag{17}$$

Next, we show that A is a symmetric matrix. Since A_i is a symmetric matrix (i.e., $A_i = A_i^T$), we find that the following is satisfied.

$$\begin{aligned} A^T &= \text{blkdiag}(A_1, \cdots, A_n)^T \\ &= \text{blkdiag}(A_1^T, \cdots, A_n^T) \\ &= \text{blkdiag}(A_1, \cdots, A_n) \\ &= A \end{aligned} \tag{18}$$

Hence, A is a symmetric matrix. Therefore, A is a positive definite matrix. □

Theorem 1. *Let $\tilde{\mathbf{G}}_{\tau,k}^l \in \mathbb{R}^{m \times n}$ be the 'l-layer k-th inner-step' gradient matrix transformed by meta parameter $\mathbf{M}^l$ for task τ. Then preconditioner $\mathbf{P}_{GAP}$ induced by $\tilde{\mathbf{G}}_{\tau,k}^l$ is a Riemannian metric and depends on the task-specific parameters $\theta_{\tau,k}$.*

Proof. We can rewrite the $\tilde{\mathbf{G}}_{\tau,k}^l$ as follows:

$$\begin{aligned} \tilde{\mathbf{G}}_{\tau,k}^l &= \mathbf{U}_{\tau,k}^l (\mathbf{M}^l \cdot \mathbf{\Sigma}_{\tau,k}^l) {\mathbf{V}_{\tau,k}^l}^T \\ &= (\mathbf{U}_{\tau,k}^l \mathbf{M}^l {\mathbf{U}_{\tau,k}^l}^T) \mathbf{U}_{\tau,k}^l \mathbf{\Sigma}_{\tau,k}^l {\mathbf{V}_{\tau,k}^l}^T \\ &= \mathbf{D}_{\tau,k}^l \mathbf{G}_{\tau,k}^l, \end{aligned} \tag{19}$$

where $\mathbf{D}_{\tau,k}^l = \mathbf{U}_{\tau,k}^l \mathbf{M}^l {\mathbf{U}_{\tau,k}^l}^T$. To induce preconditioner in Eq. (19), we reformulate Eq. (19) as the general gradient descent form (i.e., matrix-vector product):

$$\begin{aligned} \text{vec}(\tilde{\mathbf{G}}_{\tau,k}^l) &= \text{blkdiag}(\underbrace{\mathbf{D}_{\tau,k}^l, \cdots, \mathbf{D}_{\tau,k}^l}_{n \text{ times}}) \cdot \text{vec}(\mathbf{G}_{\tau,k}^l) \\ &= \mathbf{P}_{\text{GAP}} \cdot \text{vec}(\mathbf{G}_{\tau,k}^l) \end{aligned} \tag{20}$$

where $\mathbf{P}_{\text{GAP}}$ is a block diagonal matrix such that the main-diagonal blocks are $\mathbf{D}_{\tau,k}^l$'s. Now, we prove that block $\mathbf{D}_{\tau,k}^l$ is a positive definite matrix. Since $\mathbf{D}_{\tau,k}^l$ is similar to $\mathbf{M}^l$ by Definition 1, they have the same eigenvalues. In addition, all eigenvalues of $\mathbf{D}_{\tau,k}^l$ are positive because all eigenvalues of $\mathbf{M}^l$ are positive. Next, we show that $\mathbf{D}_{\tau,k}^l$ is a symmetric matrix as below.

$$\begin{aligned} (\mathbf{D}_{\tau,k}^l)^T &= (\mathbf{U}_{\tau,k}^l \mathbf{M}^l {\mathbf{U}_{\tau,k}^l}^T)^T \\ &= \mathbf{U}_{\tau,k}^l \mathbf{M}^l {\mathbf{U}_{\tau,k}^l}^T \\ &= \mathbf{D}_{\tau,k}^l \end{aligned} \tag{21}$$

Therefore, $\mathbf{D}_{\tau,k}^l$ is a positive definite matrix. By Lemma 1, $\mathbf{P}_{\text{GAP}}$ is a positive definite matrix.

Since the unitary matrix $\mathbf{U}_{\tau,k}^l$ depends on the gradient matrix $\tilde{\mathbf{G}}_{\tau,k}^l$, it depends on the task-wise parameters $\theta_{\tau,k}$.

Hence, $\mathbf{P}_{\text{GAP}}$ depends on the task-wise parameters $\theta_{\tau,k}$ because it depends on the unitary matrix $\mathbf{U}^l_{\tau,k}$.

Since $\mathbf{P}_{\text{GAP}}$ depends on the task-wise parameters $\theta_{\tau,k}$, it can be expressed as a function which is a smooth function mapping from the given $\theta_{\tau,k}$ to a positive definite matrix $\text{blkdiag}(\mathbf{D}^l_{\tau,k}, \cdots, \mathbf{D}^l_{\tau,k})$. Hence, $\mathbf{P}_{\text{GAP}}$ is a Riemannian metric.

Therefore, $\mathbf{P}_{\text{GAP}}$ is a Riemannian metric and depends on the task-specific parameters $\theta_{\tau,k}$. □

Lemma 2. *If a random vector $\boldsymbol{x} = (X_1, \cdots, X_n) \in \mathbb{R}^n$ follows an uniform distribution on the $(n-1)$-dimensional unit sphere, the variance of the random variable X_i satisfies the following.*

$$\mathbb{V}(X_i) = \frac{1}{n} \tag{22}$$

Proof. Since $X_1, \cdots, X_n$ follow an identical distribution, $\mathbb{V}(X_i) = \mathbb{V}(X_j)$ holds for all i, j. Thus,

$$n\mathbb{V}(X_i) = \sum_{i=1}^{n} \mathbb{V}(X_i). \tag{23}$$

Then, we derive the sum of variance as follows:

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{V}(X_i) &= \sum_{i=1}^{n} \mathbb{E}(X_i^2) \ (\because \mathbb{E}(X) = 0) \\
&= \mathbb{E}(\sum_{i=1}^{n} X_i^2) \\
&= \mathbb{E}(||X||_2^2) \\
&= 1.
\end{aligned} \tag{24}$$

By Eq. (23) and (24), we have

$$\mathbb{V}(X_i) = \frac{1}{n}. \tag{25}$$

□

Lemma 3. *If two independent random vectors $\boldsymbol{x} = (X_1, \cdots, X_n)$, $\boldsymbol{y} = (Y_1, \cdots, Y_n) \in \mathbb{R}^n$ follow a uniform distribution on the $(n-1)$-dimensional unit sphere, then*

$$P(|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| > \epsilon) \leq \frac{1}{n\epsilon^2}. \tag{26}$$

Proof. Since we can rotate coordinate so that $\boldsymbol{y} = (1, 0, \cdots, 0) \in \mathbb{R}^n$, we have

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = X_1. \tag{27}$$

Following Eq. (27), we show that its expectation is equal to:

$$\begin{aligned}
\mathbb{E}[\langle \boldsymbol{x}, \boldsymbol{y} \rangle] &= \mathbb{E}[X_1], \\
&= 0
\end{aligned} \tag{28}$$

and its variance is equal to:

$$\begin{aligned}
\mathbb{V}[\langle \boldsymbol{x}, \boldsymbol{y} \rangle] &= \mathbb{V}[X_1], \\
&= \frac{1}{n} \ \text{(by Lemma 2)}.
\end{aligned} \tag{29}$$

By applying Chebyshev's inequality [12] on $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, we have

$$P(|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \geq \frac{k}{\sqrt{n}}) \leq \frac{1}{k^2}, \tag{30}$$

for any real number $k > 0$. Let $\frac{k}{\sqrt{n}}$ be a ϵ. Then we rewrite the in Eq. (30) as follows:

$$P(|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \geq \epsilon) \leq \frac{1}{n\epsilon^2}. \tag{31}$$

This result indicates that the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ become asymptotically orthogonal as n increases. □

Assumption 2. *The elements of gradient matrix follows an i.i.d. normal distribution with zero mean.*

Theorem 2. *Let $\mathbf{G} \in \mathbb{R}^{m \times n}$ be a gradient matrix and $\tilde{\mathbf{G}}$ be the gradient matrix transformed by meta parameter $\mathbf{M}$. Under the Assumption 2, as n becomes large, $\tilde{\mathbf{G}}$ asymptotically becomes equivalent to $\mathbf{MG}$ as follows:*

$$\tilde{\mathbf{G}} \cong \mathbf{MG} \tag{32}$$

Proof. Let $\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_m$ are the row vectors of $\mathbf{G}$. Then,

$$\mathbf{G} = \begin{bmatrix} ||\boldsymbol{g}_1|| & & \\ & \ddots & \\ & & ||\boldsymbol{g}_m|| \end{bmatrix} \begin{bmatrix} \frac{\boldsymbol{g}_1}{||\boldsymbol{g}_1||} \\ \vdots \\ \frac{\boldsymbol{g}_m}{||\boldsymbol{g}_m||} \end{bmatrix}. \tag{33}$$

Under the Assumption 2, $\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_m$ follow an i.i.d multivariate normal distribution. Then, we have

$$\frac{\boldsymbol{g}_i}{||\boldsymbol{g}_i||} \perp\!\!\!\perp \frac{\boldsymbol{g}_j}{||\boldsymbol{g}_j||} \ \ (\forall i \neq j), \tag{34}$$

and $\frac{\boldsymbol{g}_i}{||\boldsymbol{g}_i||}, \frac{\boldsymbol{g}_j}{||\boldsymbol{g}_j||}$ are located on the $(n-1)$-dimensional unit sphere [41]. Since independent vectors $\frac{\boldsymbol{g}_i}{||\boldsymbol{g}_i||}, \frac{\boldsymbol{g}_j}{||\boldsymbol{g}_j||}$ are located on the $(n-1)$-dimensional unit sphere, the vectors are asymptotically orthogonal as n increases by Lemma 2. Now, we rewrite $\mathbf{G}$ as follows.

$$\mathbf{G} = \mathbf{I} \begin{bmatrix} ||\boldsymbol{g}_1|| & & \\ & \ddots & \\ & & ||\boldsymbol{g}_m|| \end{bmatrix} \begin{bmatrix} \frac{\boldsymbol{g}_1}{||\boldsymbol{g}_1||} \\ \vdots \\ \frac{\boldsymbol{g}_m}{||\boldsymbol{g}_m||} \end{bmatrix} \tag{35}$$

Since $\mathbf{I}$ is a unitary matrix and $(\frac{\boldsymbol{g}_1}{||\boldsymbol{g}_1||}, \cdots, \frac{\boldsymbol{g}_m}{||\boldsymbol{g}_m||})^T$ approximately becomes semi-unitary matrices as n increases, the singular values of $\mathbf{G}$ asymptotically become $||\boldsymbol{g}_1||, \cdots, ||\boldsymbol{g}_m||$.

By Eq. (35), the following holds under the Assumption 2 as n becomes sufficiently large.

$$\tilde{\mathbf{G}} \cong \mathbf{MG} \tag{36}$$

□

C. Implementation Details

For the reproducibility, we provide the details of implementation. Our implementations are based on Torchmeta [15] library. Our implementation code is available at: https://github.com/Suhyun777/CVPR23-GAP.

C.1. Hyper-parameters

For all the experiments, we use the hyper-parameters in Table 9.

Hyper-parameter	Sinusoid			mini-ImageNet		tiered-ImageNet		Cross-domain	
	5 shot	10 shot	20 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
Bathc size	4	4	4	4	2	4	2	4	2
Total training iteration	70000	70000	70000	80000	80000	130000	200000	80000	80000
inner learning rate α	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
outer learning rate β_1	0.001	0.001	0.001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
outer learning rate β_2	0.001	0.001	0.0001	0.003	0.0001	0.003	0.0001	0.003	0.0001
The number of training inner steps	5	5	5	5	5	5	5	5	5
The number of testing inner steps	10	10	10	10	10	10	10	10	10
Data augmentation	None			random flip		random flip		random flip	

Table 9. Hyper-parameters used for training GAP on various few-shot learning experiments.

C.2. Backbone Architecture

C.2.1 2-layer MLP network.

For the few-shot regression experiment, we use a simple Multi-Layer Perceptron (MLP) with 1-dimensional input/output and 40-dimensional hidden layers as in [20].

C.2.2 4-Conv network.

For the few-shot classification and cross-domain few-shot classification experiments, we use the standard Conv-4 backbone used in [56], comprising 4 modules with 3×3 convolutions, with 128 filters followed by batch normalization [26], ReLU non-linearity, and 2×2 max-pooling.

C.3. Optimization

We use ADAM optimizer [28]. For tiered-ImageNet experiment, the learning rate (LR) is scheduled by the cosine learning rate decay [38] for every 500 iterations. In all the experiments except for the tiered-ImageNet, the learning rate is unscheduled.

C.4. Preconditioning

In the few-shot regression experiment, we apply preconditioner only to the hidden layer. In both few-shot classification and cross-domain few-shot classification, we only apply preconditioner to 4 convolutional layers.