Appendices

A Further Experimental Details

We used the same hyperparameters as [1] and terminated training after the same number of examples specifid in [1]. We used a temperature $\mathbf{c}$ of 1, which was recommended in [2]. We initialized $\mathbf{T}$ to be an identity matrix and all $\boldsymbol{\zeta}$ to zero.

Compated to MAML [1], training a convolutional MT-net takes roughly 0.4 times longer (omniglot 40k steps took 7h 19m for MT-net and 5h 14m for MAML). This gap is fairly small because 1×1 convolutions require little compute compared to regular convolutions. This gap is larger (roughly 1.1 times) for fully connected MT-nets. We additionally observed that MT-nets take less training steps to converge compared to MAML.

We provide our official implementation of MT-nets at `https://github.com/yoonholee/MT-net`.

References

[1] C. Finn, P. Abbeel., and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017.

[2] E. Jang, S. Gu, and B. Poole. Categorical Reparameterization with Gumbel-Softmax. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

B Proofs for Section 4

B.1 MT-nets Learn a Subspace

Proposition 1. *Fix* $\mathbf{x}$ *and* $\mathbf{A}$. *Let* $\mathbf{U}$ *be a d-dimensional subspace of* $\mathbb{R}^n$ *(*$d \leq n$*). There exist configurations of* $\mathbf{T}, \mathbf{W}$, *and* $\boldsymbol{\zeta}$ *such that the span of* $\mathbf{y}^{new} - \mathbf{y}$ *is* $\mathbf{U}$ *while satisfying* $\mathbf{A} = \mathbf{TW}$.

Proof. We show by construction that Proposition 1 is true.

Suppose that $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a basis of $\mathbb{R}^n$ such that $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d\}$ is a basis of $\mathbf{U}$. Let $\mathbf{T}$ be the $n \times n$ matrix $[\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n]$. $\mathbf{T}$ is invertible since it consists of linearly independent columns. Let $\mathbf{W} = \mathbf{T}^{-1}\mathbf{A}$ and let $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \dots, \boldsymbol{\zeta}_d \to \infty$ and $\boldsymbol{\zeta}_{d+1}, \dots, \boldsymbol{\zeta}_n \to -\infty$. The resulting mask $\mathbf{M}$ that $\boldsymbol{\zeta}$ generates is a matrix with only ones in the first d rows and zeroes elsewhere.

$$\begin{aligned} \mathbf{y}^{\text{new}} - \mathbf{y} &= \mathbf{T}(\mathbf{W}^{\text{new}} - \mathbf{W})\mathbf{x} \\ &= \mathbf{T}(\mathbf{M} \odot \nabla_{\mathbf{W}} \mathcal{L}_{\mathcal{T}})\mathbf{x} \end{aligned} \tag{1}$$

Since all but the first d rows of $\mathbf{M}$ are $\mathbf{0}$, $(\mathbf{M} \odot \nabla_{\mathbf{W}} \mathcal{L}_{\mathcal{T}})\mathbf{x}$ is an n-dimensional vector in which nonzero elements can only appear in the first d dimensions. Therefore, the vector $\mathbf{T}(\mathbf{M} \odot \nabla_{\mathbf{W}} \mathcal{L}_{\mathcal{T}})\mathbf{x}$ is a linear combination of $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d\}$. Thus the span of $\mathbf{y}^{\text{new}} - \mathbf{y}$ is $\mathbf{U}$. $\square$

B.2 MT-nets Learn a Metric in their Subspace

Proposition 2. *Fix* $\mathbf{x}$*,* $\mathbf{A}$*, and a loss function* $\mathcal{L}_\mathcal{T}$*. Let* $\mathbf{U}$ *be a d-dimensional subspace of* $\mathbb{R}^n$*, and* $g(\cdot,\cdot)$ *a metric tensor on* $\mathbf{U}$*. There exist configurations of* $\mathbf{T}, \mathbf{W}$*, and* $\boldsymbol{\zeta}$ *such that the vector* $\mathbf{y}^{new} - \mathbf{y}$ *is in the steepest direction of descent on* $\mathcal{L}_\mathcal{T}$ *with respect to the metric* du*.*

Proof. We show Proposition 2 is true by construction as well.

We begin by constructing a representation for the arbitrary metric tensor $g(\cdot,\cdot)$. Let $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be a basis of $\mathbb{R}^n$ such that $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d\}$ is a basis of $\mathbf{U}$. Vectors $\mathbf{u}_1, \mathbf{u}_2 \in \mathbf{U}$ can be expressed as $\mathbf{u}_1 = \sum_{i=0}^d c_{1i}\mathbf{v}_i$ and $\mathbf{u}_2 = \sum_{i=0}^d c_{2i}\mathbf{v}_i$. We can express any metric tensor $g(\cdot,\cdot)$ using such coefficients c:

$$g(\mathbf{u}_1, \mathbf{u}_2) = \underbrace{\begin{bmatrix} c_{11} & \ldots & c_{1d} \end{bmatrix}}_{\mathbf{c}_1^\top} \underbrace{\begin{bmatrix} g_{11} & \cdots & g_{1d} \\ \vdots & \ddots & \vdots \\ g_{d1} & \cdots & g_{dd} \end{bmatrix}}_{\mathbf{G}} \underbrace{\begin{bmatrix} c_{21} \\ \vdots \\ c_{2d} \end{bmatrix}}_{\mathbf{c}_2}, \tag{2}$$

where $\mathbf{G}$ is a positive definite matrix. Because of this, there exists an invertible $d \times d$ matrix $\mathbf{H}$ such that $\mathbf{G} = \mathbf{H}^\top\mathbf{H}$. Note that $g(\mathbf{u}_1, \mathbf{u}_2) = (\mathbf{H}\mathbf{c}_1)^\top(\mathbf{H}\mathbf{c}_2)$: the metric $g(\cdot,\cdot)$ is equal to the inner product after multiplying $\mathbf{H}$ to given vectors $\mathbf{c}$.

Using $\mathbf{H}$, we can alternatively parameterize vectors in $\mathbf{U}$ as

$$\mathbf{u}_1 = \underbrace{\begin{bmatrix} \mathbf{v}_1 & \ldots & \mathbf{v}_d \end{bmatrix}}_{\mathbf{V}} \mathbf{c}_1 \tag{3}$$

$$= \mathbf{V}\mathbf{H}^{-1}\left(\mathbf{H}\mathbf{c}_1\right). \tag{4}$$

Here, we are using $\mathbf{H}\mathbf{c}_1$ as a d-dimensional parameterization and the columns of the $n \times d$ matrix $\mathbf{V}\mathbf{H}^{-1}$ as an alternative basis of $\mathbf{U}$.

Let $\mathbf{v}_1^H, \ldots, \mathbf{v}_d^H$ be the columns of $\mathbf{V}\mathbf{H}^{-1}$, and set $\mathbf{T} = [\mathbf{v}_1^\mathbf{H}, \ldots, \mathbf{v}_d^\mathbf{H}, \mathbf{v}_{d+1}, \ldots, \mathbf{v}_n]$. Since $\mathbf{H}$ is invertible, $\{\mathbf{v}_1^\mathbf{H}, \ldots, \mathbf{v}_d^\mathbf{H}\}$ is a basis of $\mathbf{U}$ and thus $\mathbf{T}$ is an invertible matrix. As in Proposition 1, set $\mathbf{W} = \mathbf{T}^{-1}\mathbf{A}$, $\zeta_1, \zeta_2, \ldots, \zeta_d \to \infty$, and $\zeta_{d+1}, \ldots, \zeta_n \to -\infty$. Note that this configuration of $\boldsymbol{\zeta}$ generates a mask $\mathbf{M}$ that projects gradients onto the first d rows, which will later be multiplied by the vectors $\{\mathbf{v}_1^\mathbf{H}, \ldots, \mathbf{v}_d^\mathbf{H}\}$.

We can express $\mathbf{y}$ as $\mathbf{y} = V\mathbf{c}_\mathbf{y} = \mathbf{V}\mathbf{H}^{-1}(\mathbf{H}\mathbf{c}_\mathbf{y})$, where $\mathbf{c}_\mathbf{y}$ is again a d-dimensional vector. Note that $\mathbf{V}\mathbf{H}^{-1}$ is constant in the network and change in $\mathbf{W}$ only affects $\mathbf{H}\mathbf{c}_\mathbf{y}$. Since $\nabla_\mathbf{W}\mathcal{L}_\mathcal{T} = (\nabla_{\mathbf{W}\mathbf{x}}\mathcal{L}_\mathcal{T})\mathbf{x}^\top$, the task-specific update is in the direction of steepest descent of $\mathcal{L}_\mathcal{T}$ in the space of $\mathbf{H}\mathbf{c}_y$ (with the Euclidean metric). This is exactly the direction of steepest descent of $\mathcal{L}_\mathcal{T}$ in $\mathbf{U}$ with respect to the metric $g(\cdot,\cdot)$. $\square$

C Additional Experiments

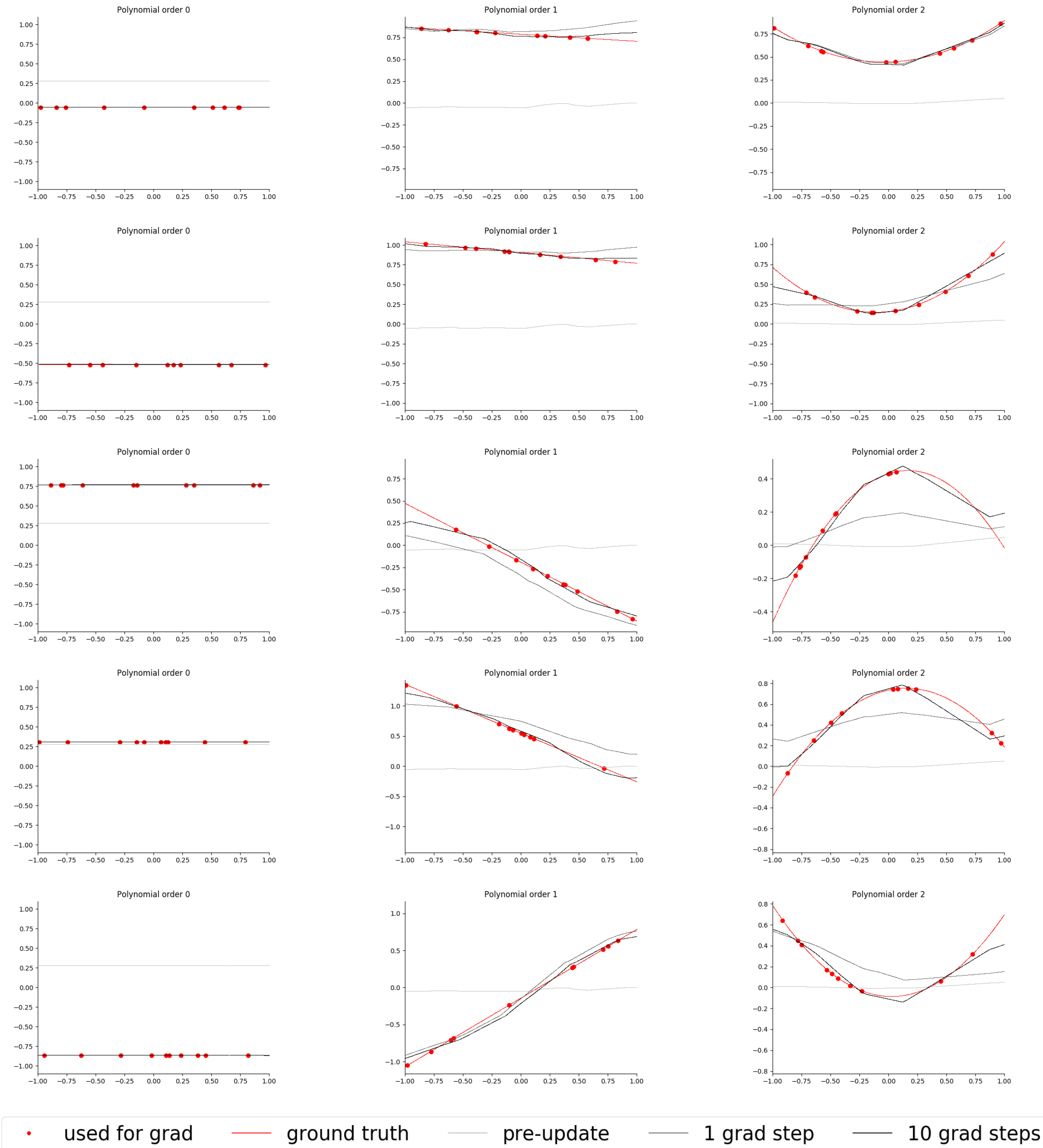

Figure 1: Additional qualitative results from the polynomial regression task