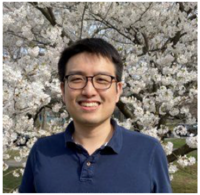
# When is Task Vector *Provably* Effective for Model Editing? A Generalization Analysis of Nonlinear Transformers

Hongkang Li[1], Yihua Zhang[2], Shuai Zhang[3], Pin-Yu Chen[4], Sijia Liu[2,4], Meng Wang[1]

[1]Rensselaer Polytechnic Institute, [2]Michigan State University, [3]New Jersey Institute of Technology, [4]IBM Research



# Task Vectors and Task Arithmetic

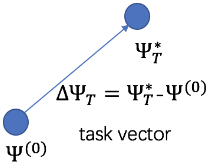



*Figure 1:* *Task vector.*

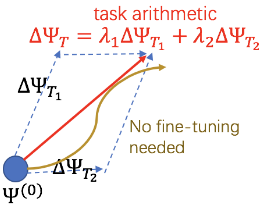



*Figure 2:* *Task arithmetic by adding up two task vectors for inference. No fine-tuning on the two tasks are needed.*

**Task vector** is the difference between the fine-tuned model and the pre-trained model.

$$\Delta\Psi_{\mathcal{T}} = \Psi^*_{\mathcal{T}} - \Psi^{(0)}, \quad (1)$$

where $\Psi^*_{\mathcal{T}}$ is the model fine-tuned on $(\boldsymbol{X}, y) \sim \mathcal{D}_{\mathcal{T}}$ for task $\mathcal{T}$, and $\Psi^{(0)}$ is the pre-trained model.

**Task arithmetic** refers to adding a linear combination of task vectors of different tasks.
Given $\Psi^{(0)}$ and a set of task vectors $\{\Delta\Psi_{\mathcal{T}_i}\}_{i\in\mathcal{V}}$,

$$\Psi = \Psi^{(0)} + \sum_{i\in\mathcal{V}} \lambda_i \Delta\Psi_{\mathcal{T}_i}, \quad (2)$$

for the inference on the downstream task.

# Task Vectors and Task Arithmetic

Applications: multi-task learning, unlearning, and out-of-domain generalization in vision and language generation tasks.

Advantage: No need of fine-tuning for new tasks.

- Linear coefficient selection: Simple averaging [Ilharco et al.22, Wortsman et al.2022], Fisher-weighted averaging [Metena & Raffel, 2022] for multi-task learning; negation for unlearning [Ilharco et al.22].
- Task vector construction: sparsification [Yadav et al.2023, Yu et al.24], linearization [Ortiz-Jimenez et al.23].

# Task Correlations Affect Task Arithmetic

**Experiments on Colored-MNIST dataset**:

- Classify the parity of digits.
- Control the fraction of red/green digit colors for different task correlations/distributions.

| | "Irrelevant" Tasks | | "Aligned" Tasks | | "Contradictory" Tasks | |
|---|---|---|---|---|---|---|
| | Multi-Task | Unlearning | Multi-Task | Unlearning | Multi-Task | Unlearning |
| Best $\lambda$ | 1.4 | -0.6 | 0.2 | 0.0 | 0.6 | -1.0 |
| $\mathcal{T}_1$ Acc | 91.83 (-3.06) | 95.02 (-0.56) | 95.62 (0.00) | 95.20 (-0.42) | 79.54 (-16.70) | 94.21 (-0.61) |
| $\mathcal{T}_2$ Acc | 88.40 (-5.65) | 50.34 (-45.24) | 92.46 (-3.23) | 90.51 (-5.18) | 62.52 (-33.72) | 4.97 (-89.85) |

*Figure 3: Test accuracy (%) of $\Psi = \Psi^{(0)} + \Delta\Psi_{\mathcal{T}_1} + \lambda\Delta\Psi_{\mathcal{T}_2}$ on task $\mathcal{T}_1$ and $\mathcal{T}_2$. Different task correlations $\Rightarrow$ Different arithmetic coefficients.*

| | Fine-Tuning | $\Psi^*_{\mathcal{T}_1}$ | $\Psi^*_{\mathcal{T}_2}$ | Searching $\lambda_1, \lambda_2$ in $[-2, 3]$ |
|---|---|---|---|---|
| $(\lambda_1, \lambda_2)$ | N/A | $(1, 0)$ | $(0, 1)$ | $(1.2, -0.6)$ |
| $\mathcal{T}'$ Acc | 92.21 | 88.10 | 45.06 | **91.74** |

*Figure 4: Test $\Psi = \Psi^{(0)} + \lambda_1\Delta\Psi_{\mathcal{T}_1} + \lambda_2\Delta\Psi_{\mathcal{T}_1}$ on task $\mathcal{T}'$. $\mathcal{T}'$ shares a different distribution from $\mathcal{T}_1$ or $\mathcal{T}_2$. The optimal $\lambda_1$ and $\lambda_2$ generates a model that outperforms any separately trained model $\Psi^*_{\mathcal{T}_1}$ and $\Psi^*_{\mathcal{T}_2}$. $\mathcal{T}'$ and $\mathcal{T}_1$ are positively correlated; $\mathcal{T}'$ and $\mathcal{T}_2$ are negatively correlated.*

# Problems to Solve

**Q1**: Can we provide generalization guarantees for task arithmetic?

**Q2**: How does task correlation quantitatively affect the performance of task arithmetic?

**Q3**: Why do the arithmetic operations of task vectors perform well for out-of-domain generalization?

# Related Theoretical Works

- Some works [Ginart et al.2019, Guo et al.2020, Neel et al.2021, Mu & Klabjan, 2024] theoretically analyze the performance of machine unlearning from an optimization perspective.

- [Izmailov et al.2018, Frankle et al.2020] propose linear mode connectivity, concluding the existence of a small-loss connected region in the loss landscape.

- [Ortiz-Jimenez et al.23] study task arithmetic in model editing with the Neural Tangent Kernel (NTK) framework to linearize the models.

# Problem Formulation

We study binary classification tasks that map each $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_P)$ to $y \in \{+1, -1\}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$, $i \in [P]$.

The **learner model** is considered as a one-layer nonlinear Transformer with $\Psi$ as the set of parameters, where $\boldsymbol{W}, \boldsymbol{V} \in \Psi$ are trainable,

$$f(\boldsymbol{X}; \Psi) = \frac{1}{P} \sum_{l=1}^{P} \boldsymbol{a}_{(l)}^\top \text{Relu}\Big(\sum_{s=1}^{P} \boldsymbol{V}\boldsymbol{x}_s \text{softmax}_l({\boldsymbol{x}_s}^\top \boldsymbol{W}\boldsymbol{x}_l)\Big). \tag{3}$$

**Data formulation**: Let $\boldsymbol{\mu}_\mathcal{T}$ be the discriminative pattern of $\mathcal{T}$. Each token is chosen from $\{\boldsymbol{\mu}_\mathcal{T}, -\boldsymbol{\mu}_\mathcal{T}\}$ or other irrelevant patterns. If $y = 1$ ($y = -1$), the number of tokens equal to $\boldsymbol{\mu}_\mathcal{T}$ (or $-\boldsymbol{\mu}_\mathcal{T}$) is larger than that of $-\boldsymbol{\mu}_\mathcal{T}$ (or $\boldsymbol{\mu}_\mathcal{T}$).

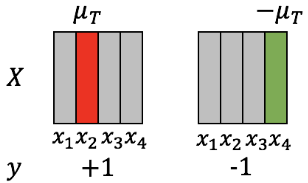



Figure 5: Data formulation.

# Theoretical Results (Multi-Task learning and Unlearning)

Let $\Psi = \Psi^{(0)} + \Delta\Psi_{\mathcal{T}_1} + \lambda\Delta\Psi_{\mathcal{T}_2}$. $\beta = \Theta(1/d)$. Loss function $\ell(\cdot)$: Hinge loss.

- Define $\alpha = \boldsymbol{\mu}_{\mathcal{T}_1}^\top \boldsymbol{\mu}_{\mathcal{T}_2}$ as the correlation between $\mathcal{T}_1$ and $\mathcal{T}_2$.
- $\alpha > 0$, $< 0$, or $= 0$, corresponds to "aligned", "contradictory", or"irrelevant" relationship.
- $\Psi^*_{\mathcal{T}_1}$ and $\Psi^*_{\mathcal{T}_2}$ are trained to achieve an $\epsilon$ generalization error on $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively.

**Theorem 1 (Success of Multi-Task Learning on Irrelevant and Aligned Tasks)**

*Then, as long as $\alpha \geq 0$ and $\lambda \geq 1 - \alpha + \beta$, we have a desired multi-task learning performance with $\Psi$, i.e., $\mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}_{\mathcal{T}_1}} \ell(\boldsymbol{X}, y; \Psi) \leq \Theta(\epsilon) + |\lambda| \cdot \beta$, and $\mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}_{\mathcal{T}_2}} \ell(\boldsymbol{X}, y; \Psi) \leq \Theta(\epsilon)$.*

**Theorem 2 (Success of Unlearning on Irrelevant and Contradictory Tasks)**

*As long as $\alpha \leq 0$ and $-\Theta(\alpha^{-2}) \leq \lambda \leq 0$,we have a desired unlearning performance with $\Psi$, i.e., $\mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}_{\mathcal{T}_1}} \ell(\boldsymbol{X}, y; \Psi) \leq \Theta(\epsilon) + |\lambda| \cdot \beta$, and $\mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}_{\mathcal{T}_2}} \ell(\boldsymbol{X}, y; \Psi) \geq \Theta(1)$.*

# Theoretical Results (Out-of-Domain Generalization)

**Out-of-domain generalization** on the task $\mathcal{T}'$, given task vectors of tasks $\{\mathcal{T}_i\}_{i\in\mathcal{V}_\Psi}$. Suppose

- all $\boldsymbol{\mu}_{\mathcal{T}_i}$ are orthogonal to each other,
- the discriminative pattern of $\mathcal{T}'$ is $\boldsymbol{\mu}_{\mathcal{T}'} = \sum_{i\in\mathcal{V}_\Psi}\gamma_i\boldsymbol{\mu}_{\mathcal{T}_i} + \kappa\cdot\boldsymbol{\mu}'_\perp$ with $\boldsymbol{\mu}'_\perp \perp \{\boldsymbol{\mu}_{\mathcal{T}_i}\}_{i\in\mathcal{V}_\Psi}$,
- not all $\gamma_i$ are zero.

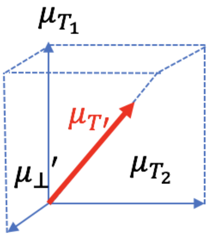



*Figure 6:* ***An illustration of*** $\boldsymbol{\mu}_{\mathcal{T}'}$.

**Theorem 3 (Out-of-domain generalization using task arithmetic)**

*Let* $\Psi = \Psi^{(0)} + \sum_{i\in\mathcal{V}_\Psi}\lambda_i\Delta\Psi_{\mathcal{T}_i}, \lambda_i \neq 0$. *Then, for some* $c\in(0,1)$ *and all* $i\in\mathcal{V}_\Psi$, *and a non-empty region of* $\lambda_i$, $i\in\mathcal{V}_\Psi$, *where*

$$\begin{cases}\sum_{i\in\mathcal{V}_\Psi}\lambda_i\gamma_i \geq 1+c,\\ \sum_{i\in\mathcal{V}_\Psi}\lambda_i\gamma_i^2 \geq 1+c,\\ |\lambda_i|\cdot\beta \leq c,\end{cases} \tag{4}$$

*we have* $\mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}_{\mathcal{T}'}}\ell(\boldsymbol{X},y;\Psi) \leq \Theta(\epsilon)$.

# Theoretical Results (Efficiency)

Recall that $\boldsymbol{W}, \boldsymbol{V} \in \Psi$. $\Delta \boldsymbol{W}_{\mathcal{T}} = \boldsymbol{W}^*_{\mathcal{T}} - \boldsymbol{W}^{(0)}$, $\Delta \boldsymbol{V}_{\mathcal{T}} = \boldsymbol{V}^*_{\mathcal{T}} - \boldsymbol{V}^{(0)}$.

**Corollary 1 (Low-rank Approximation)**

*For any task $\mathcal{T}$ defined above, there exists rank-1 $\Delta \boldsymbol{W}_{LR}$ and $\Delta \boldsymbol{V}_{LR}$, such that*

$$\|\Delta \boldsymbol{W}_{\mathcal{T}} - \Delta \boldsymbol{W}_{LR}\|_F \leq M \cdot \epsilon + \frac{1}{\log M}, \quad \textit{and} \quad \|\Delta \boldsymbol{V}_{\mathcal{T}} - \Delta \boldsymbol{V}_{LR}\|_F \leq \Theta(\epsilon), \tag{5}$$

**Corollary 2 (Sparsification)**

*Let $\boldsymbol{u}_i$ be the $i$-th row of $\Delta \boldsymbol{V}_{\mathcal{T}}$. Then, for a constant fraction of $\boldsymbol{u}_i$, we have $\|\boldsymbol{u}_i\| \geq \Omega(m^{-1/2})$; for the remaining neurons, we have $\|\boldsymbol{u}_i\| \leq O(m^{-1/2}\epsilon)$ (pruning these neurons still ensures Theorems 1-3 to hold.)*

## Experiments

### Image classification on Colored-MNIST with ViT-Small/16

- Consider a merged model $\Psi = \Psi^{(0)} + \lambda_1 \Delta\Psi_{\mathcal{T}_1} + \lambda_2 \Delta\Psi_{\mathcal{T}_2}$ constructed by two task vectors for the targeted task $\mathcal{T}'$. We estimate $\gamma_1 \approx 0.792$, $\gamma_2 \approx -0.637$.
- The result justifies the sufficient conditions in Theorem 3.

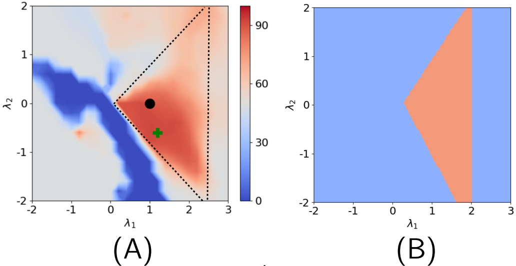



*Figure 7: (A) The heatmap of the testing accuracy on $\mathcal{T}'$ using the merged model $\Psi$. The black dot is the baseline, while the green cross is the best $\lambda_1$, $\lambda_2$. (B) The red region satisfies (4), while the blue region does not.*

## Language generation with Phi-3-small (7B)

- Given "Harry Potter 1" (HP1), "Harry Potter 2" (HP2) by J.K. Rowling, and "Pride and Prejudice" (PP) by Jane Austen.
- Estimate task correlations $\hat{\alpha}(\Psi^*_{\mathcal{T}_1}, \Psi^*_{\mathcal{T}_2}) = \mathbb{E}_{\boldsymbol{X}}[Sim(f(\boldsymbol{X}; \Psi^*_{\mathcal{T}_1}), f(\boldsymbol{X}; \Psi^*_{\mathcal{T}_1}))]$. HP1 and HP2 are semantically similar, while PP is less aligned with HP1 or HP2.
- Unlearning $\mathcal{T}_{\mathrm{HP1}}$ can effectively degrade the performance of the aligned ($\mathcal{T}_{\mathrm{HP2}}$) as well, while the degradation on the less aligned ($\mathcal{T}_{\mathrm{PP}}$) is relatively smaller.

| $\lambda$ | 0 (baseline) | $-0.2$ | $-0.4$ | $-0.6$ | $-0.8$ | $-1$ |
|---|---|---|---|---|---|---|
| $\mathcal{T}_{\mathrm{HP1}}$ | 0.2573 | 0.1989 | 0.1933 | 0.1888 | 0.1572 | **0.1142** (55.61% ↓) |
| $\mathcal{T}_{\mathrm{HP2}}$ | 0.2688 | 0.2113 | 0.1993 | 0.1938 | 0.1622 | **0.1563** (52.29% ↓) |
| $\mathcal{T}_{\mathrm{PP}}$ | 0.1942 | 0.1825 | 0.1644 | 0.1687 | 0.1592 | **0.1541** (20.65% ↓) |

*Figure 8:* *Rouge-L scores of $\mathcal{T}_{HP1}$ $\mathcal{T}_{HP2}$, and $\mathcal{T}_{PP}$ by $\Psi = \Psi^{(\mathbf{0})'} + \lambda \cdot \Delta\Psi^{LR}_{HP1}$ using low-rank task vector $\Delta\Psi^{LR}_{HP1}$ (Phi-3-small).*

# Summary

- We quantitatively characterize the selection of arithmetic hyper-parameters and their dependence on task correlations so that the resulting task vectors achieve desired multi-task learning, unlearning, and out-of-domain generalization.
- We also demonstrate the validity of using sparse or low-rank task vectors.
- Theoretical results are justified on vision models and large language models.
- Future work: analyzing task vectors in more complex models and designing more robust task vector selection methods.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, et al.
Editing models with task arithmetic.
In *International Conference on Learning Representations 2022.*

Guillermo Ortiz-Jimenez, Alessandro Favero, Pascal Frossard.
Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models.
In *Conference on Neural Information Processing Systems 2023.*

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, Yongbin Li.
Language models are super mario: Absorbing abilities from homologous models as a free lunch.
In *Conference on Machine Learning 2024.*

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al.
Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time.

In *Conference on Machine Learning 2022.*

Matena, Michael S and Raffel, Colin A
Merging models with fisher-weighted averaging.
In *Conference on Neural Information Processing Systems 2022.*

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal.
Ties-merging: Resolving interference when merging models.
In *Conference on Neural Information Processing Systems 2023.*

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin.
Linear mode connectivity and the lottery ticket hypothesis.
In *Conference on Machine Learning 2020.*

P. Izmailov, A.G. Wilson, D. Podoprikhin, D. Vetrov, and T. Garipov.
Averaging Weights Leads to Wider Optima and Better Generalization.
In *Conference on Uncertainty in Artificial Intelligence 2018.*

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou.
Making ai forget you: Data deletion in machine learning.

In *Conference on Neural Information Processing Systems 2019.*

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten.
Certified data removal from machine learning models.
In *Conference on Machine Learning 2020.*

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi.
Descent-to-delete: Gradient-based methods for machine unlearning.
In *Algorithmic Learning Theory 2021.*

Siqiao Mu and Diego Klabjan.
Rewind-to-delete: Certified machine unlearning for nonconvex functions.
*arXiv preprint arXiv:2409.09778, 2024.*