Online Hyperparameter Meta-Learning with Hypergradient Distillation

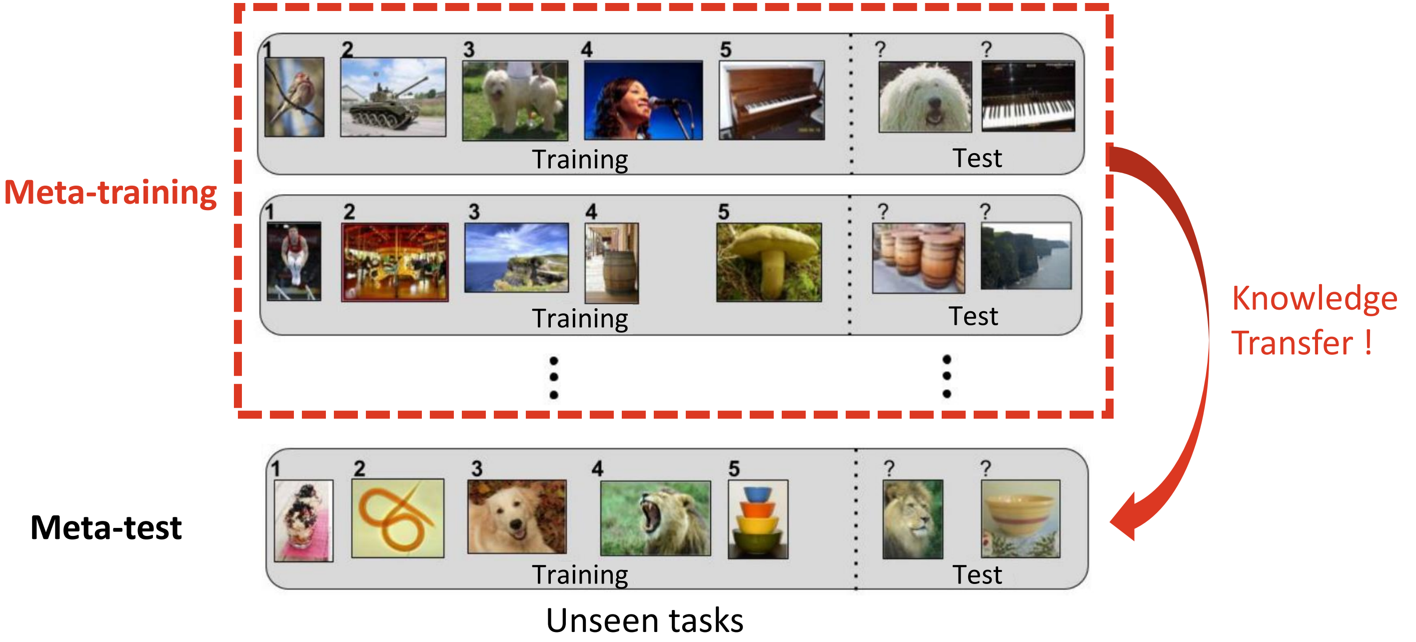
Hae Beom Lee[1], Hayeon Lee[1], Jaewoong Shin[3],
Eunho Yang[1,2], Timothy Hospedales[4,5], Sung Ju Hwang[1,2]

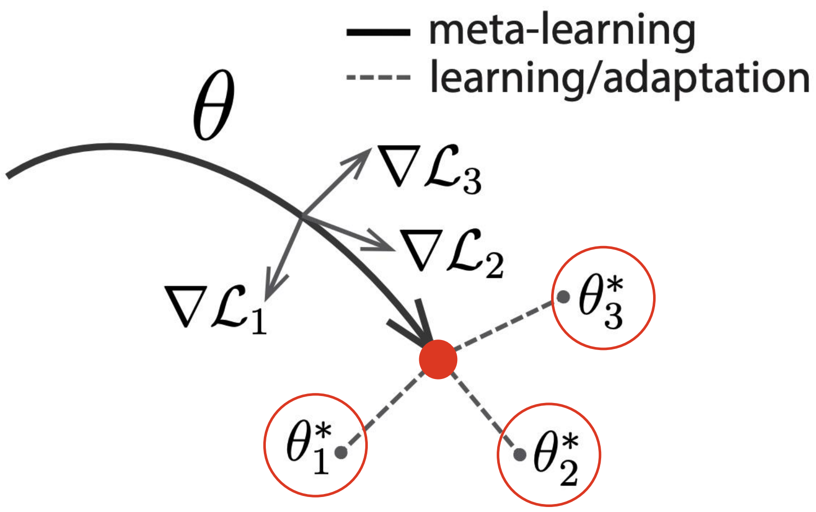
KAIST[1], AITRICS[2], Lunit[3], University of Edinburgh[4], Samsung AI Centre Cambridge[5]

ICLR 2022 spotlight

Meta-Learning

- Humans generalize well because we never learn from scratch.
- Learn a model that can generalize over **a distribution of tasks.**





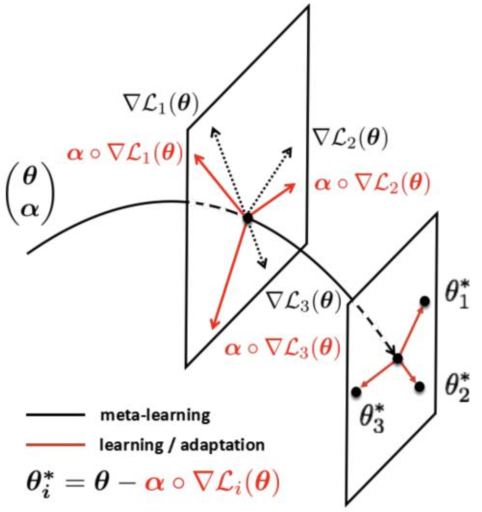
MAML (Finn et al., '17)

S. Ravi, H. Larochelle, Optimization as a Model for Few-shot Learning, ICLR 2017

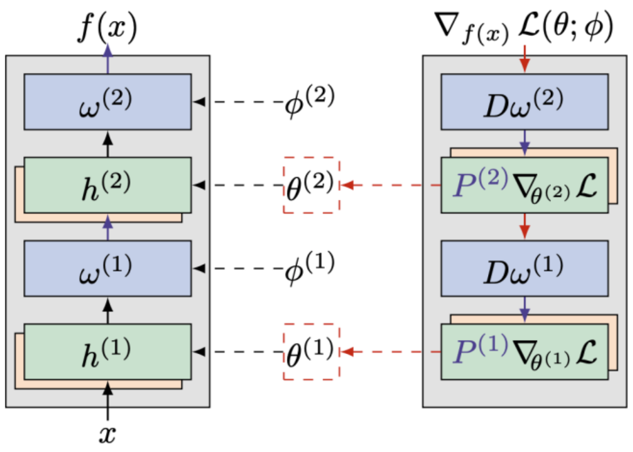
C. Finn, P, Abbeel, S. Levine, Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, ICML 2017

Hyperparameters in Meta-Learning

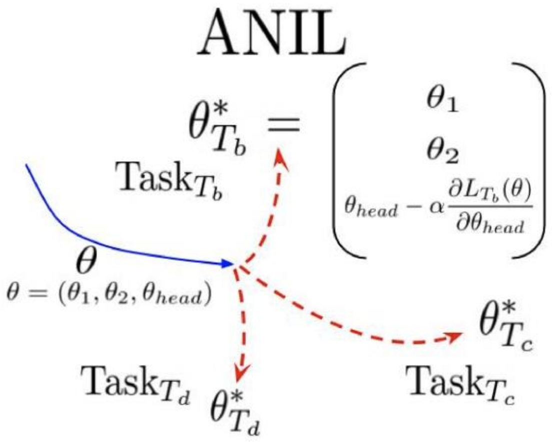
- The parameters that do not participate in inner-optimization → **Hyperparameters** in meta-learning.
- They are usually **high-dimensional.**



Element-wise learning rates



Interleaved (e.g. "Warp") layers



Whole feature extractor

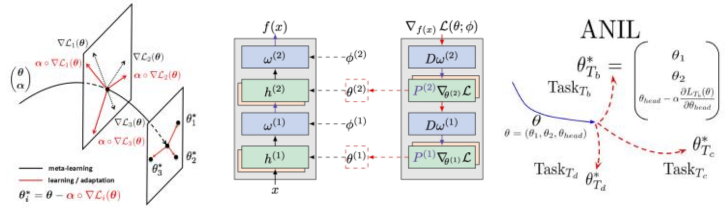
Z. Li, F. Zhou, F. Chen, H. Li, Meta-SGD: Learning to Learn Quickly for Few-Shot Learning, 2017

S. Flennerhag, A. A. Rusu, R. Pascanu, F. Visin, H. Yin, R. Hadsell, Meta-Learning with Warped Gradient Descent, ICLR 2020

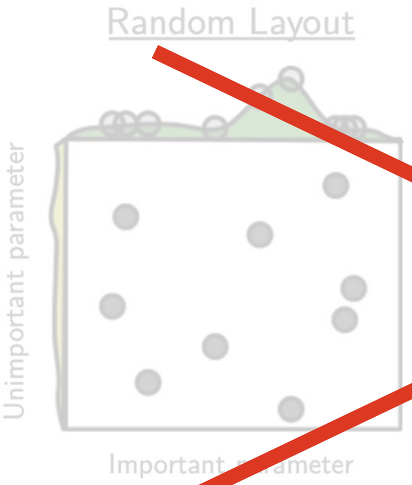
A. Raghu*, M. Raghu*, S. Bengio, O. Vinyals, Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML, ICLR 2020

Hyperparameter Optimization (HO)

- **Hyperparameter optimization (HO):** a problem of choosing a set of optimal hyperparameters for a learning algorithm.
- Which method should we use for such high-dimensional hyperparams?

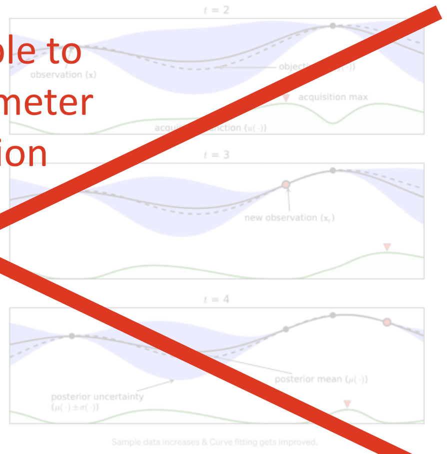


Element-wise learning rates Interleaved (e.g. “Warp”) layers Whole feature extractor

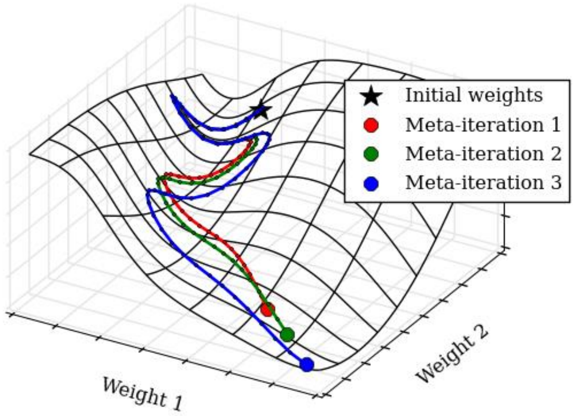


Random Search

Not scalable to hyperparameter dimension



Bayesian Optimization



Gradient-based HO

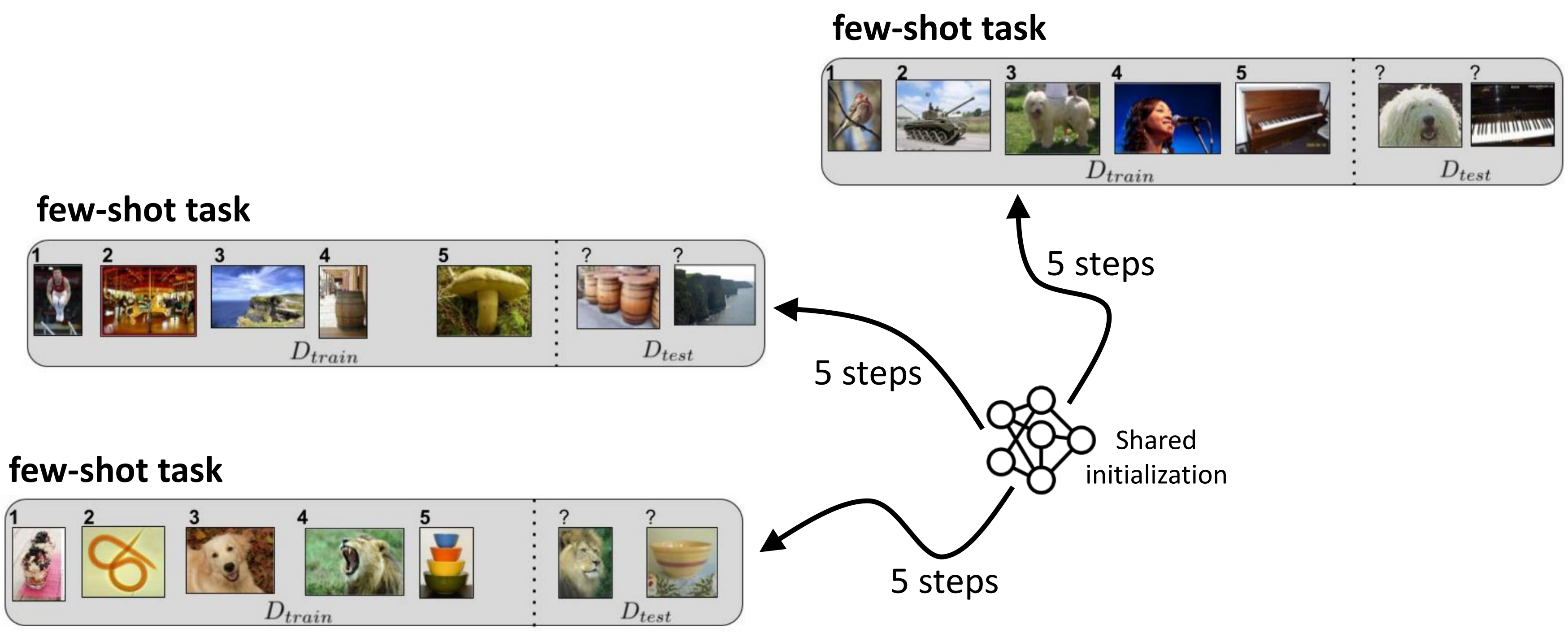
J. Bergstra, Y. Bengio. Random search for hyper-parameter optimization, 2012

https://towardsdatascience.com/shallow-understanding-on-bayesian-optimization-324b6c1f7083

D. Maclaurin, D. Duvenaud, R. P. Adams, Gradient-based Hyperparameter Optimization through Reversible Learning, ICML 2015

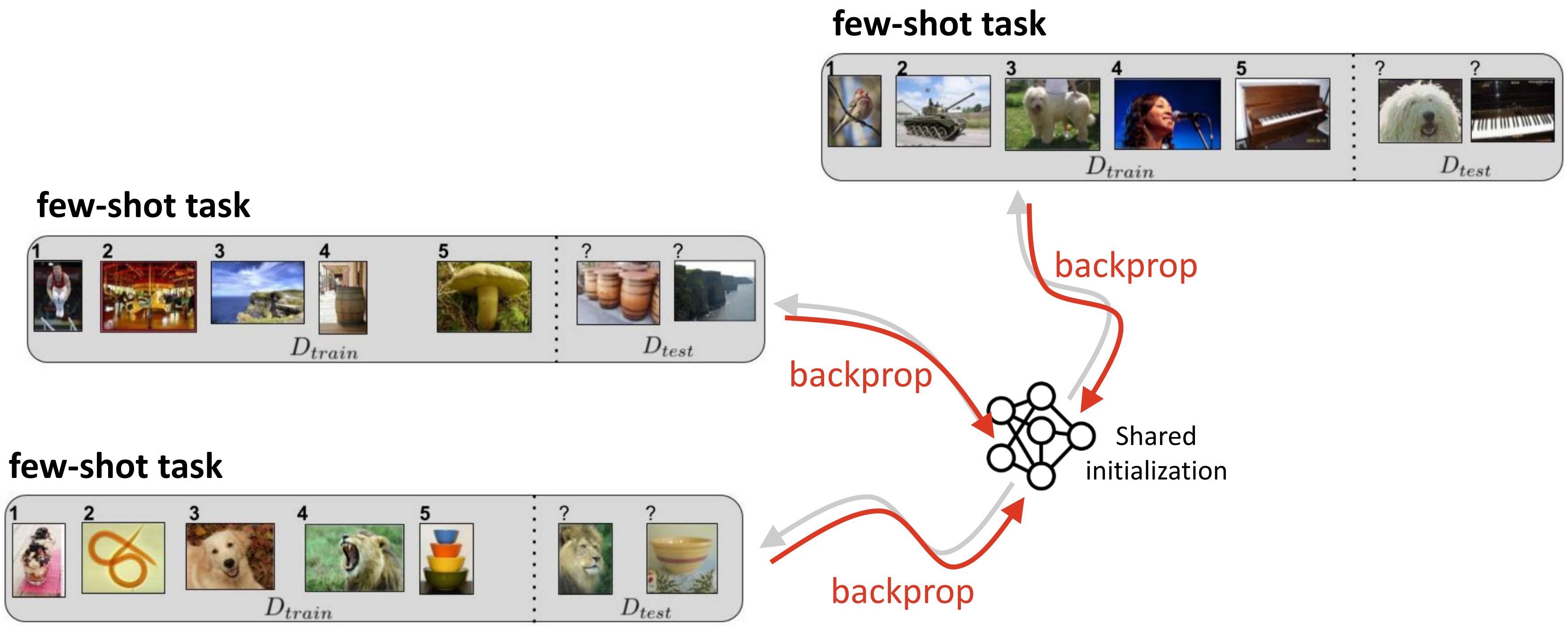
In Case of Few-shot Learning

- In case of few-shot learning, computing the exact gradient w.r.t. the hyperparameter (i.e. hypergradient) is not too expensive.
- **A few-gradient steps** are sufficient for each task.



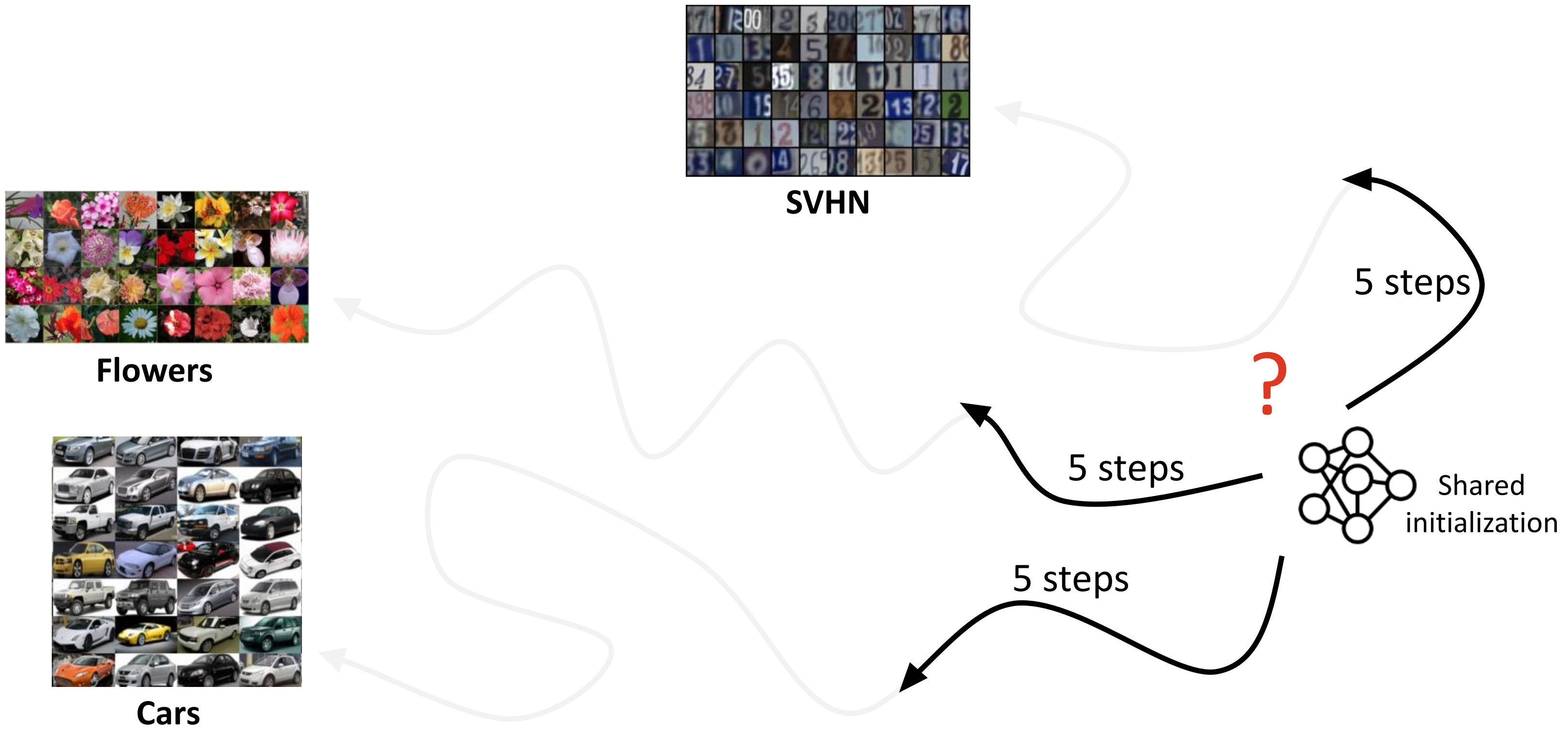
In Case of Few-shot Learning

- In case of few-shot learning, computing the exact gradient w.r.t. the hyperparameter (i.e. hypergradient) is not too expensive.
- **A few-gradient steps** are sufficient for each task.



HO Does Matter when Horizon Gets Longer

- Many-shot learning → Only **a few** gradient steps?
  → Meta-learner may suffer from the **short-horizon bias** (Wu et al. ‘18).

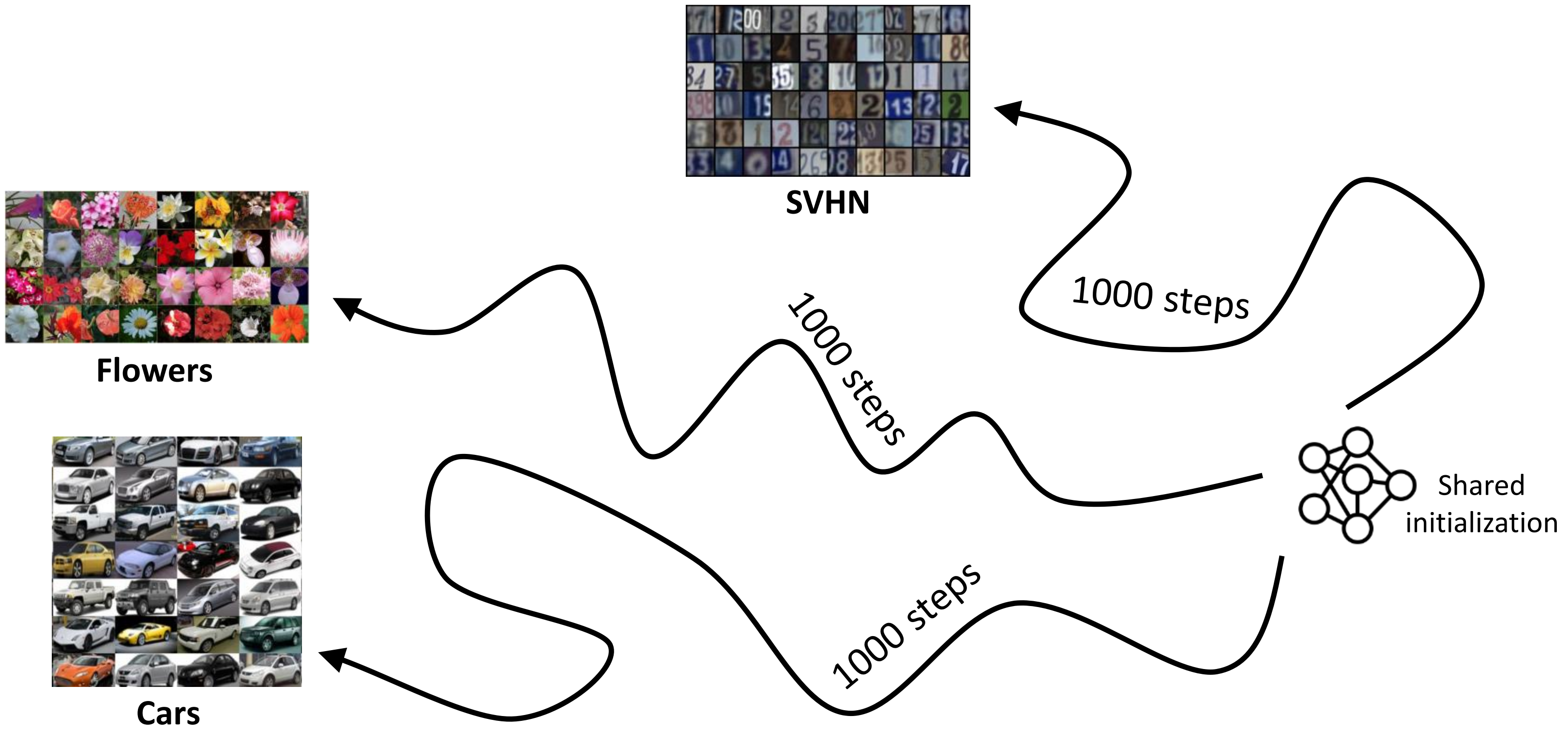


Y. Wu*, M. Ren*, R. Liao, R. Grosse, Understanding Short-Horizon Bias in Stochastic Meta-Optimization, ICLR 2018

J. Shin*, H. B. Lee*, B. Gong, S. J. Hwang, Large-Scale Meta-Learning with Continual Trajectory Shifting, ICML 2021

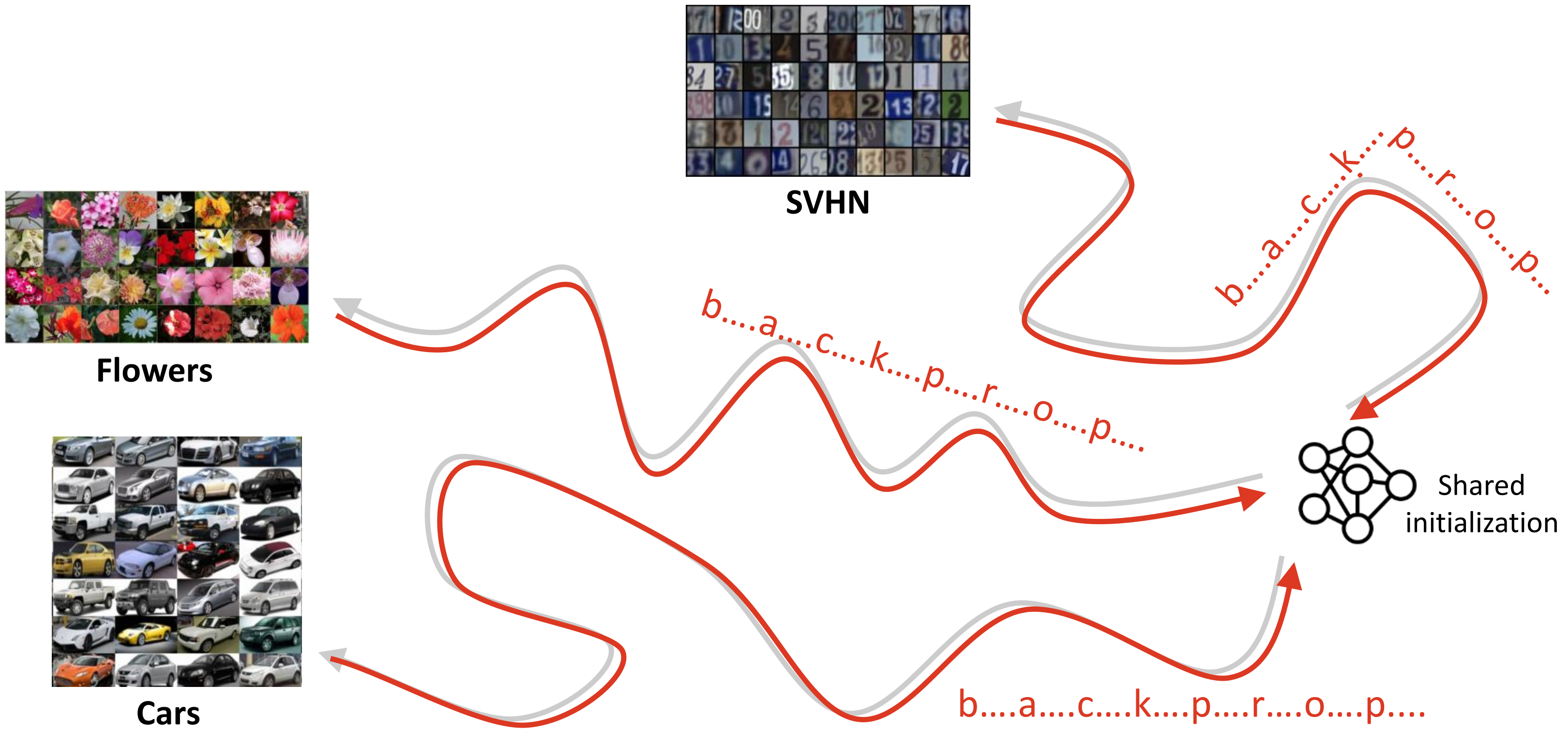
HO Does Matter when Horizon Gets Longer

- Many-shot learning → requires longer inner-learning trajectory



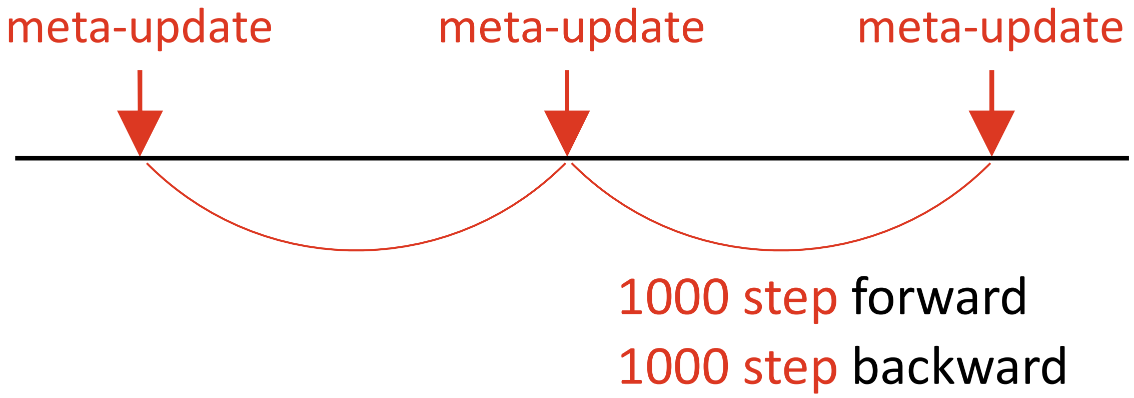
HO Does Matter when Horizon Gets Longer

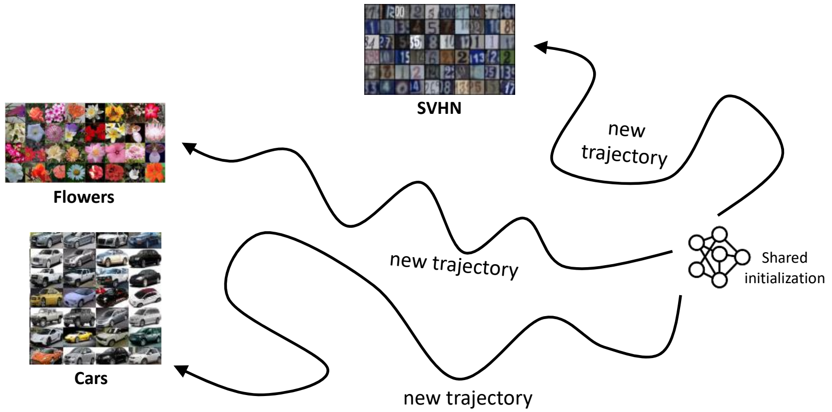
- Many-shot learning → requires longer inner-learning trajectory
  → **Computing a single hypergradient becomes too expensive!**



HO Does Matter when Horizon Gets Longer

- Many-shot learning → requires longer inner-learning trajectory
  → **Offline** method: **interval between two adjacent meta-updates is too long…**
  → Meta-convergence is poor.



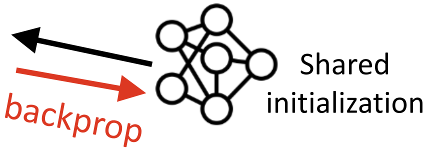


HO Does Matter when Horizon Gets Longer

- Many-shot learning → requires longer inner-learning trajectory

  → Online method: update hyperparamer every inner-grad step!

Flowers

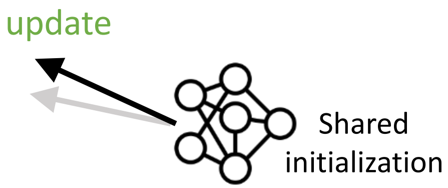


HO Does Matter when Horizon Gets Longer

- Many-shot learning → requires longer inner-learning trajectory

  → Online method: update hyperparamer every inner-grad step!

Flowers

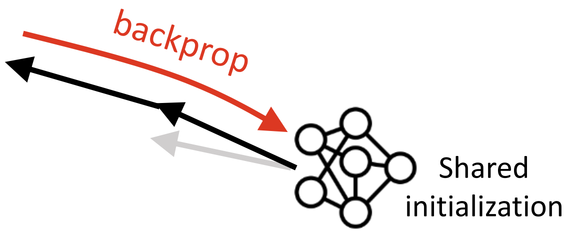


HO Does Matter when Horizon Gets Longer

- **Many-shot** learning → requires **longer** inner-learning trajectory

  → **Online** method: update hyperparamer **every** inner-grad step!

Flowers

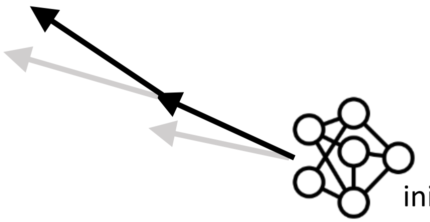


HO Does Matter when Horizon Gets Longer

- Many-shot learning → requires longer inner-learning trajectory

  → Online method: update hyperparamer every inner-grad step!

Flowers

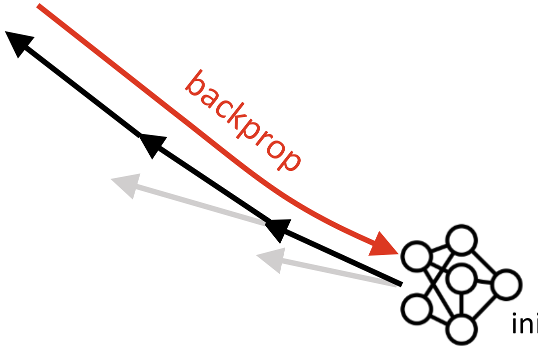


HO Does Matter when Horizon Gets Longer

- Many-shot learning → requires longer inner-learning trajectory

  → Online method: update hyperparamer every inner-grad step!

Flowers

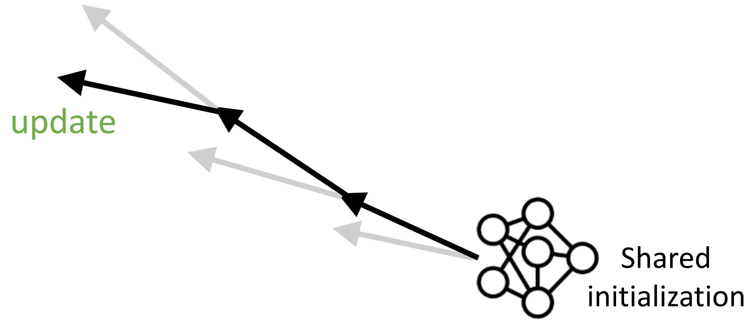


HO Does Matter when Horizon Gets Longer

- Many-shot learning → requires longer inner-learning trajectory

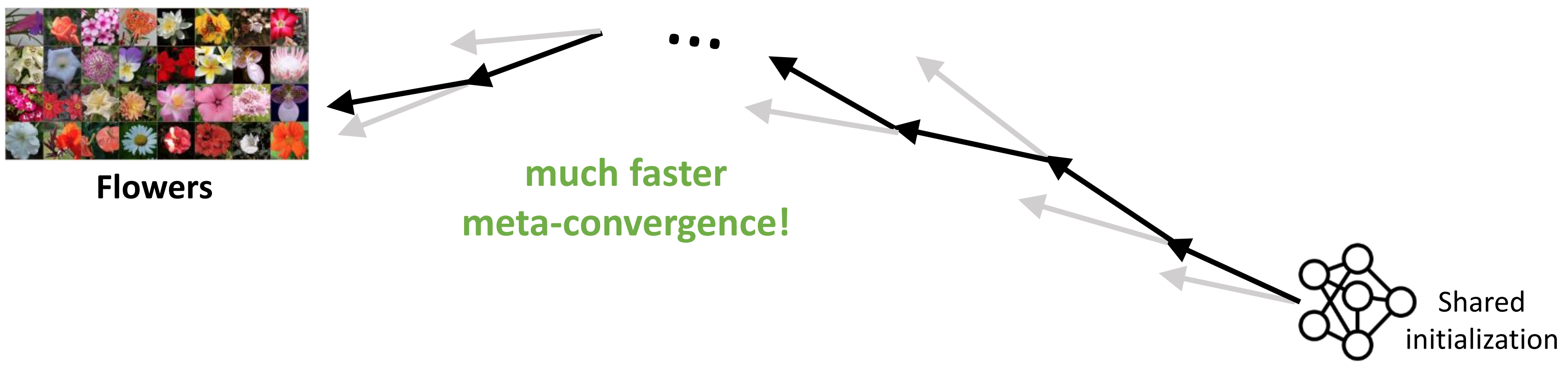
  → Online method: update hyperparamer every inner-grad step!

Flowers

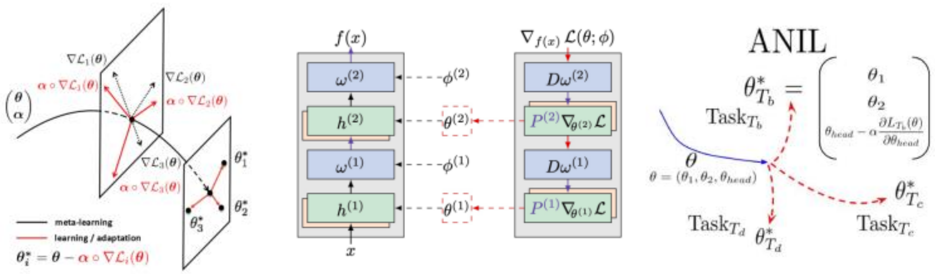


HO Does Matter when Horizon Gets Longer

- **Many-shot** learning → requires **longer** inner-learning trajectory
  → **Online** method: update hyperparamer **every** inner-grad step!

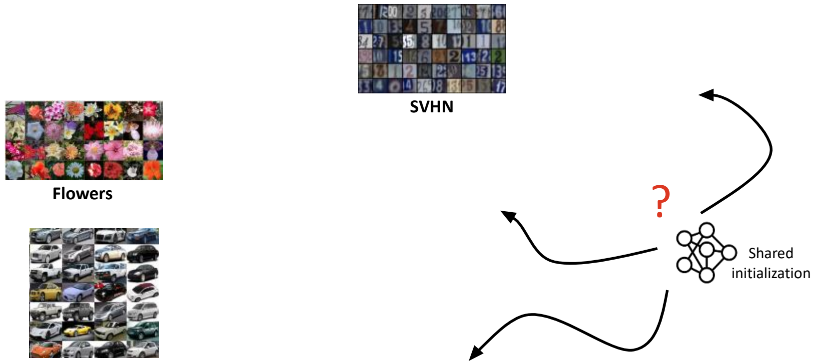


Criteria of Good HO Algorithm for Meta-Learning

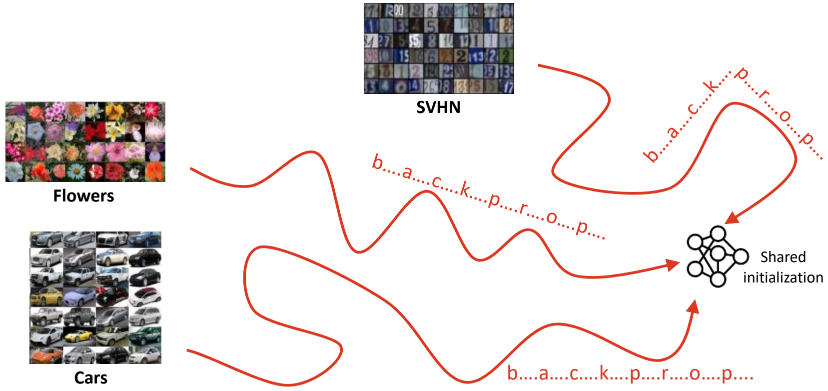


Element-wise learning rates | Interleaved (e.g. “Warp”) layers | Whole feature extractor

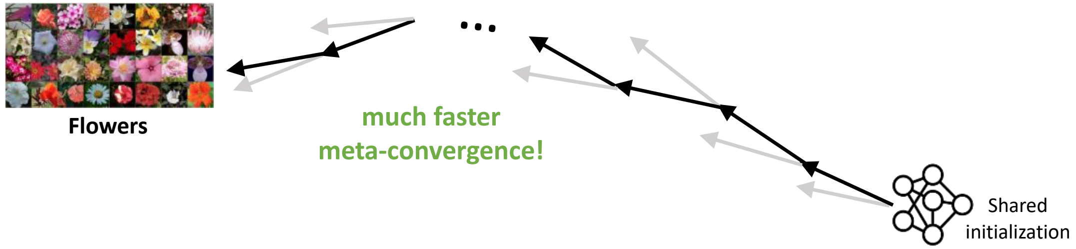
1. Scalable to hyperparameter dimension



2. Less or no short-horizon bias



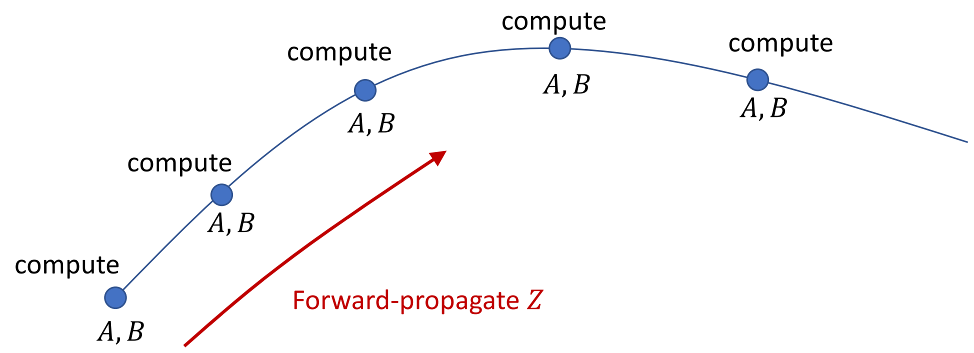
3. Computing a single hypergradient should not be too expensive



4. Update hyperparam every inner-grad step i.e. online optimization

Limitations of Existing Grad-based HO Algs

Unfortunately, the existing gradient-based HO algorithms do not satisfy all the criteria simultaneously.

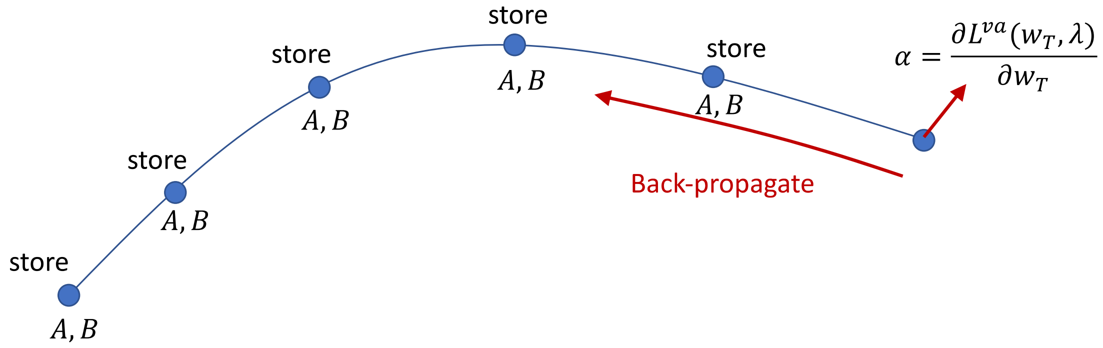


Criteria	FMD
1. Scalable to hyperparam dim	X
2. Less or no short horizon bias	O
3. Constant memory cost	O
4. Online optimization	O

L. Franceschi, M. Donini, P. Frasconi, M. Pontil, Forward and Reverse Gradient-Based Hyperparameter Optimization, ICML 2017

Limitations of Existing Grad-based HO Algs

Unfortunately, the existing gradient-based HO algorithms do not satisfy all the criteria simultaneously.



Criteria	FMD	RMD
1. Scalable to hyperparam dim	X	O
2. Less or no short horizon bias	O	O
3. Constant memory cost	O	X
4. Online optimization	O	X

L. Franceschi, M. Donini, P. Frasconi, M. Pontil, Forward and Reverse Gradient-Based Hyperparameter Optimization, ICML 2017

Limitations of Existing Grad-based HO Algs

Unfortunately, the existing gradient-based HO algorithms do not satisfy all the criteria simultaneously.

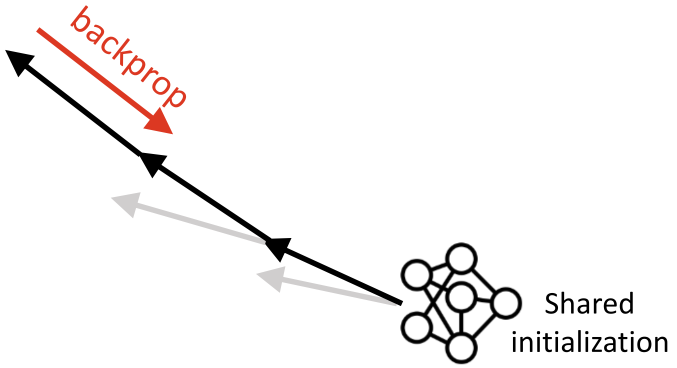
$$\frac{\partial \mathbf{w}^*}{\partial \boldsymbol{\lambda}}\bigg|_{\boldsymbol{\lambda}'} = -\underbrace{\left[\frac{\partial^2 \mathcal{L}_\mathrm{T}}{\partial \mathbf{w} \partial \mathbf{w}^T}\right]^{-1}}_{\text{training Hessian}} \times \underbrace{\frac{\partial^2 \mathcal{L}_\mathrm{T}}{\partial \mathbf{w} \partial \boldsymbol{\lambda}^T}}_{\text{training mixed partials}}\Bigg|_{\boldsymbol{\lambda}', \mathbf{w}^*(\boldsymbol{\lambda}')} \quad \text{(IFT)}$$

Criteria	FMD	RMD	IFT
1. Scalable to hyperparam dim	X	O	O
2. Less or no short horizon bias	O	O	O
3. Constant memory cost	O	X	O
4. Online optimization	O	X	△

J. Lorraine, P. Vicol, D. Duvenaud, Optimizing Millions of Hyperparameters by Implicit Differentiation, AISTATS 2020

Limitations of Existing Grad-based HO Algs

Unfortunately, the existing gradient-based HO algorithms do not satisfy all the criteria simultaneously.



Criteria	FMD	RMD	IFT	1-step
1. Scalable to hyperparam dim	X	O	O	O
2. Less or no short horizon bias	O	O	O	X
3. Constant memory cost	O	X	O	O
4. Online optimization	O	X	△	O

J. Luketina, M. Berglund, K. Greff, T. Raiko, Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters, ICML 2016

Goal of This Paper

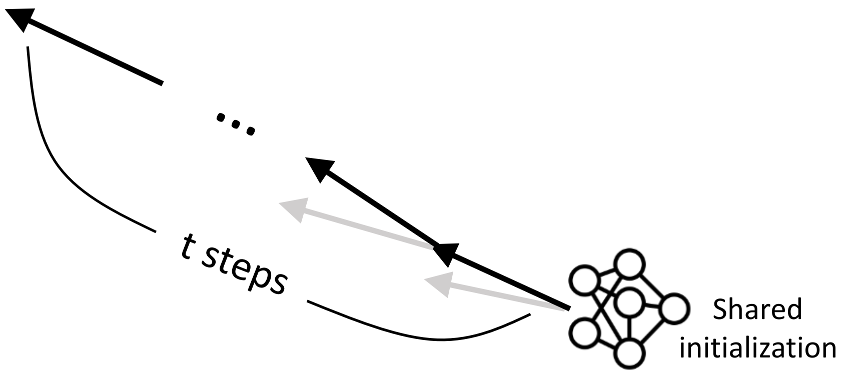
This paper aims to overcome all the aforementioned limitations at the same time.

Hypergradient distillation

$$\pi_t^*, w_t^*, D_t^* = \underset{\pi, w, D}{\arg\min} \left\| \pi f_t(w, D) - g_t^{\mathrm{SO}} \right\|_2$$

Criteria	FMD	RMD	IFT	1-step	Ours
1. Scalable to hyperparam dim	X	O	O	O	O
2. Less or no short horizon bias	O	O	O	X	O
3. Constant memory cost	O	X	O	O	O
4. Online optimization	O	X	△	O	O

Hypergradient Distillation



Hypergradient Distillation

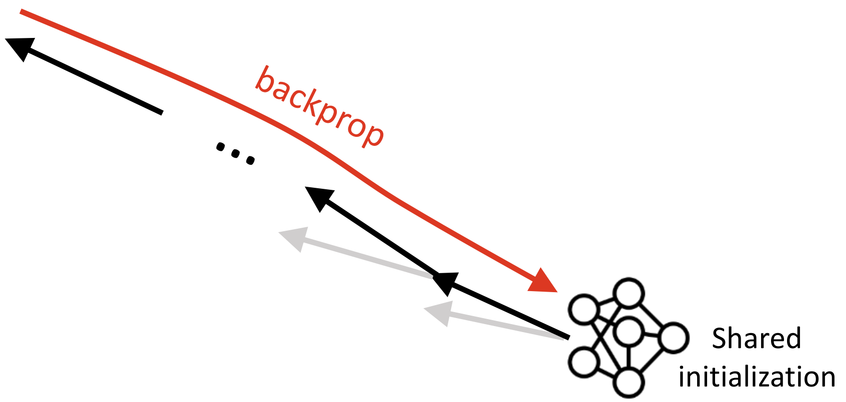
hypergradient

respose Jacobian

$$\frac{d\mathcal{L}^{\text{val}}(w_t,\lambda)}{d\lambda} = \underbrace{\frac{\partial\mathcal{L}^{\text{val}}(w_t,\lambda)}{\partial\lambda}}_{g_t^{\text{FO}}:\ \text{First-order term}} + \underbrace{\frac{\partial\mathcal{L}^{\text{val}}(w_t,\lambda)}{\partial w_t}\boxed{\frac{dw_t}{d\lambda}}}_{g_t^{\text{SO}}:\ \text{Second-order term}}$$

$$\boxed{\frac{dw_t}{d\lambda}} = \sum_{i=1}^{t}\left(\prod_{j=i+1}^{t} A_j\right) B_i$$

requires 2t – 1 **JVP** computations (e.g. RMD)



Hypergradient Distillation

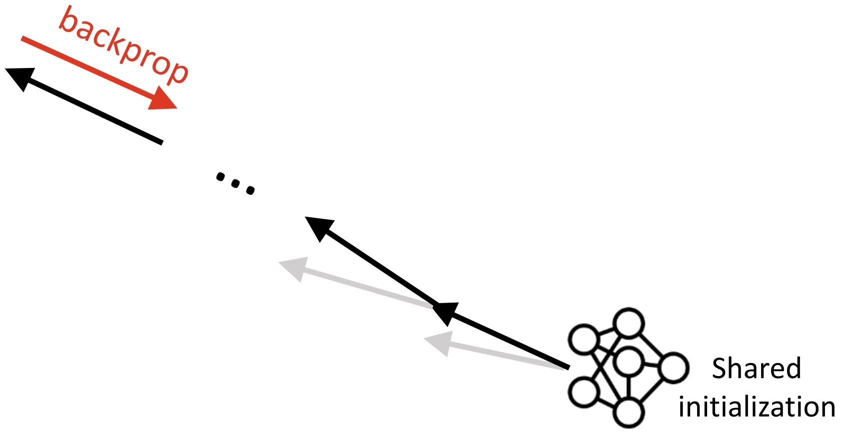
hypergradient

respose Jacobian

$$\frac{d\mathcal{L}^{\text{val}}(w_t,\lambda)}{d\lambda} = \underbrace{\frac{\partial\mathcal{L}^{\text{val}}(w_t,\lambda)}{\partial\lambda}}_{g_t^{\text{FO}}:\text{ First-order term}} + \underbrace{\frac{\partial\mathcal{L}^{\text{val}}(w_t,\lambda)}{\partial w_t}\boxed{\frac{dw_t}{d\lambda}}}_{g_t^{\text{SO}}:\text{ Second-order term}}$$

$$\boxed{\frac{dw_t}{d\lambda}} \approx \frac{\partial w_t}{\partial\lambda}\bigg|_{w_{t-1}} = B_t$$

Requires only 1 JVP computation
But it suffers from **short horizon bias**



Hypergradient Distillation

hypergradient

$$\frac{d\mathcal{L}^{\text{val}}(w_t,\lambda)}{d\lambda} = \underbrace{\frac{\partial\mathcal{L}^{\text{val}}(w_t,\lambda)}{\partial\lambda}}_{g_t^{\text{FO}}:\ \text{First-order term}} + \underbrace{\frac{\partial\mathcal{L}^{\text{val}}(w_t,\lambda)}{\partial w_t}\frac{dw_t}{d\lambda}}_{g_t^{\text{SO}}:\ \text{Second-order term}}$$

2t – 1 **JVP**

distill

$$\pi f_t(w, D)$$

a single JVP

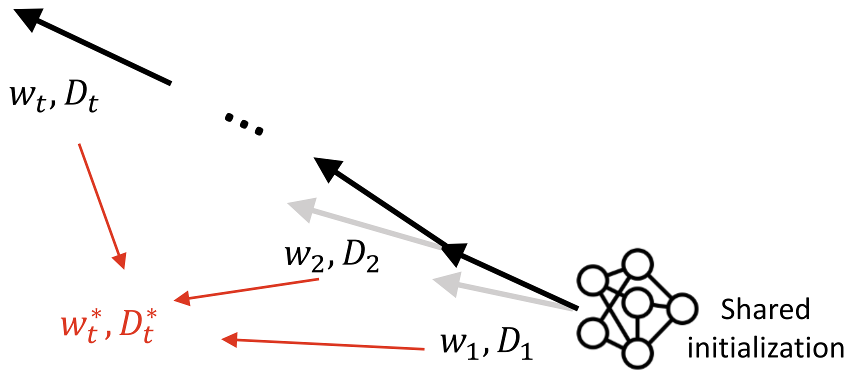
$$\pi_t^*, w_t^*, D_t^* = \underset{\pi,w,D}{\arg\min} \left\| \pi f_t(w,D) - g_t^{\text{SO}} \right\|_2$$

scaling factor
→ hypergrad size

distilled weight and dataset
→ hypergrad direction

For each online HO step t,

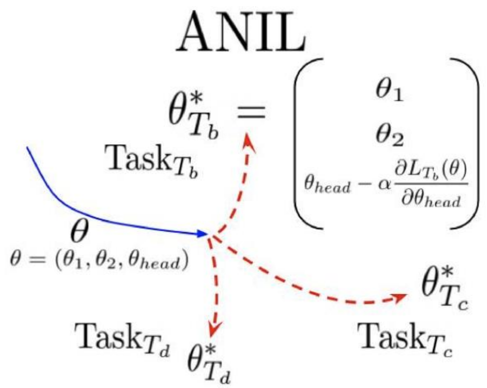
- it does not require computing the actual g_t^{SO}.
- we only need to keep updating a moving average of w_t^* and D_t^*.
- the scaling factor π^* is also efficiently estimated with a function approximator.
- we can approximately solve the distillation problem efficiently.



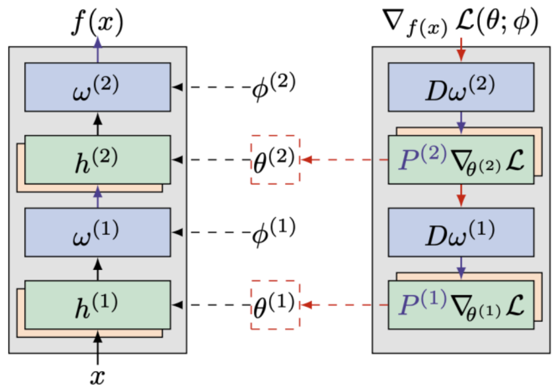
Please read the main paper for the technical details !

Experimental Setup

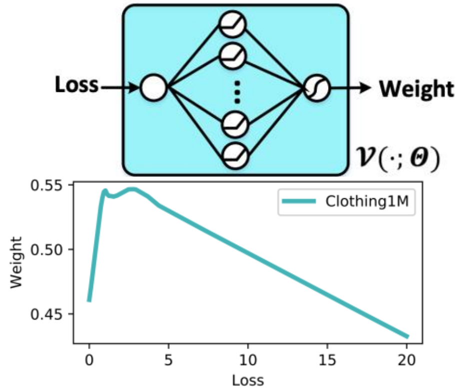
Meta-learning models



ANIL (Raghu et al. '20)

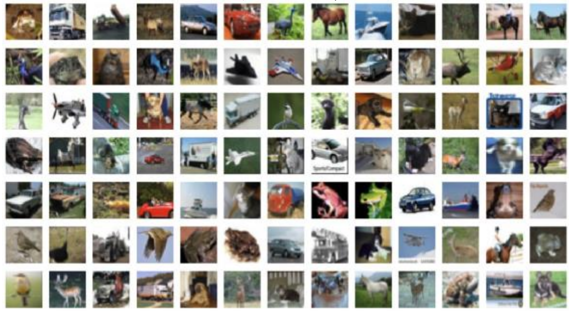


WarpGrad (Flennerhag et al. '20)



Meta-Weight-Net (Shu et al. '20)

tinyImageNet



CIFAR100

Task distribution

- 10-way 250-shot

Other details

- Inner-grad step = 100
- Use Reptile for learning shared initialization

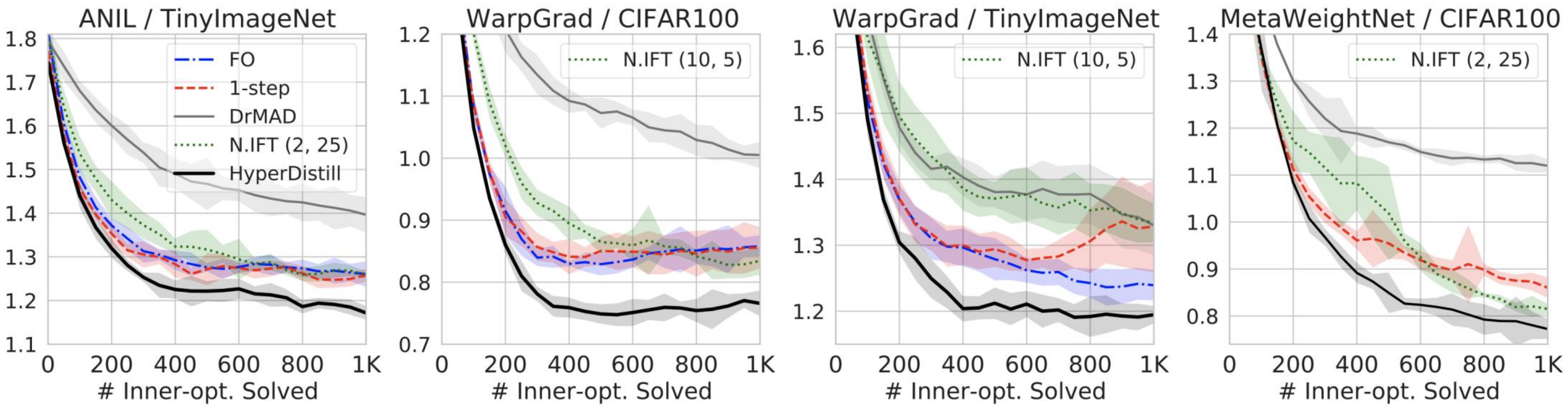
J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, D. Meng, Meta-Weight-Net: Learning an Explicit Mapping For Sample Weighting, NeurIPS 2019

A. Nichol, J. Achiam, J. Schulman, On First-Order Meta-Learning Algorithms, 2018

Experimental Results

Q1. Does HyperDistill provide **faster convergence?**

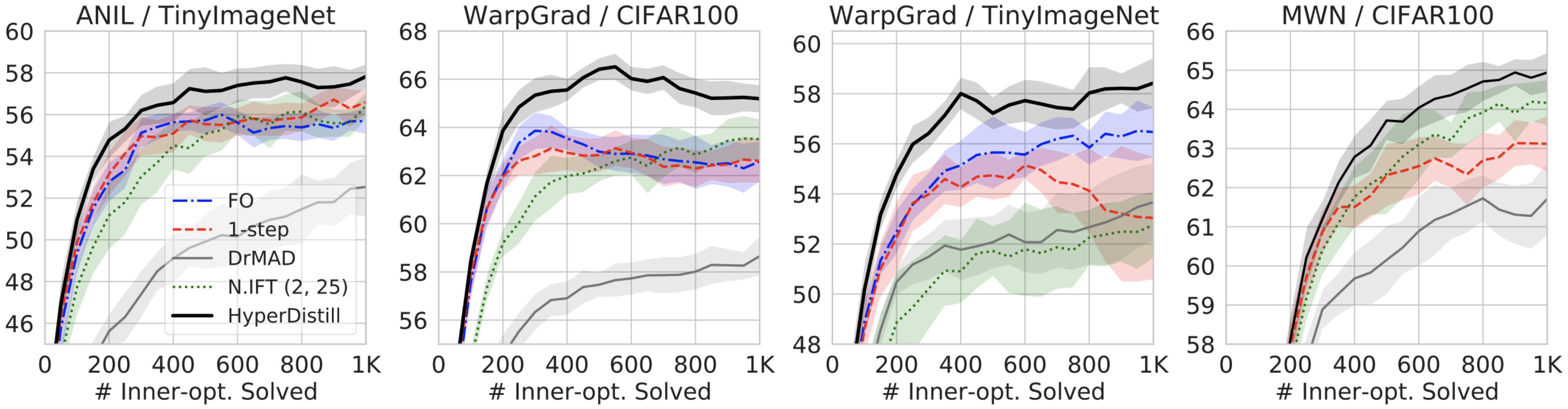
Meta-training convergence (Test Loss)



Experimental Results

Q2. Does HyperDistill provide **better generalization performance?**

Meta-validation performance (Test Acc)



Meta-test performance (Test Acc)

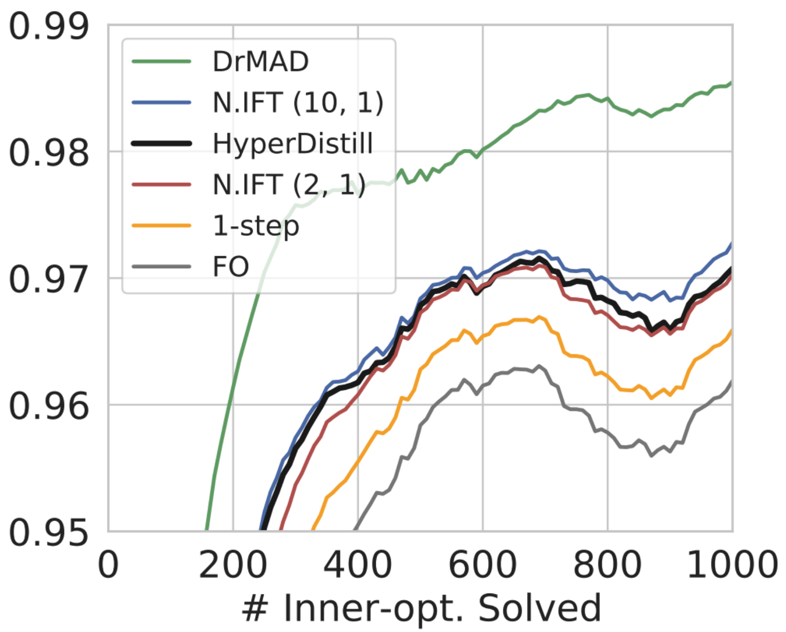
	Online optim.	# JVPs / inner-opt.	ANIL tinyImageNet	WarpGrad CIFAR100	WarpGrad tinyImageNet	MetaWeightNet CIFAR100
FO	O	0	53.62±0.06	58.16±0.52	53.54±0.74	N/A
1-step	O	50	53.90±0.43	58.18±0.52	49.97±2.46	58.45±0.40
DrMAD	X	199	49.84±1.35	55.13±0.64	50.71±1.16	57.03±0.42
Neumann IFT	△	{55, 60, 75}	53.76±0.31	58.88±0.65	50.15±0.98	59.34±0.27
HyperDistill	O	≈ 58	**56.37**±0.27	**60.91**±0.27	**55.04**±0.52	**60.82**±0.33

Experimental Results

Q3. Is HyperDistill **a reasonable approximation** to the true hypergradient?

Q4. Is HyperDistill **computationally efficient?**

Cosine similarity to the true hypergradient



GPU memory consumption and wall-clock runtime

	ANIL tinyImageNet (Mb) / (s / inner-opt.)
FO	1430 / 6.23
1-step	1584 / 6.80
DrMAD	1442 / 20.88
Neumann IFT	1392 / 7.98
HyperDistill	1638 / 6.92

Conclusion

- The existing gradient-based HO algorithms do not satisfy the four criteria that should be met for their practical use in meta-learning.

- In this paper, we showed that for each online HO step, it is possible to efficiently distill the whole hypergradient indirect term into a single JVP, satisfying the four criteria simultaneously.

- Thank to the accurate hypergradient approximation, HyperDistill could improve meta-training convergence and meta-testing performance, in a computationally efficient manner.

github.com/haebeom-lee/hyperdistill