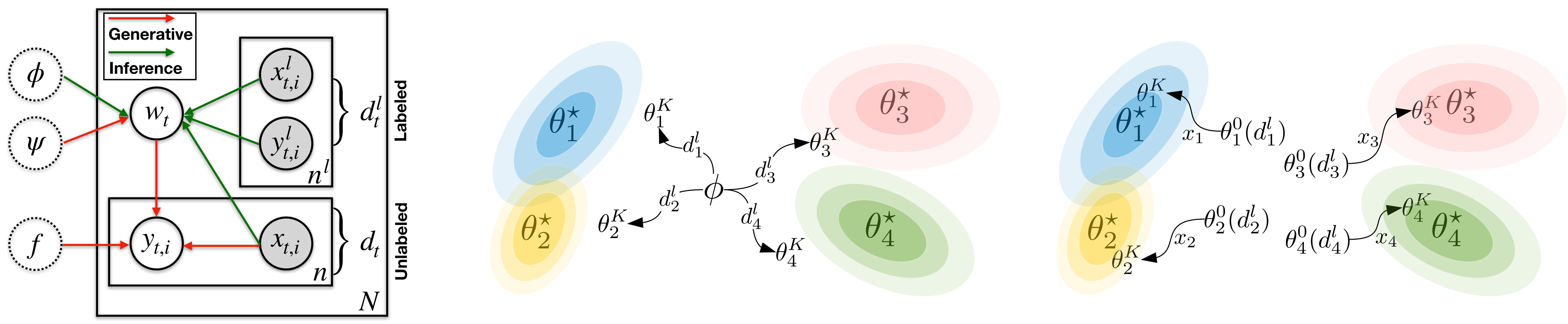
Empirical Bayes Transductive Meta-Learning with Synthetic Gradients

Shell X. Hu[1], Pablo G. Moreno[2], Xi Shen[1], Yang Xiao[1],
Guillaume Obozinski[3], Neil D. Lawrence[2,4] and Andreas Damianou[2]

[1]École des Ponts ParisTech, [2]Amazon, [3]Swiss Data Science Center, [4]University of Cambridge

How can we make use of the unlabeled data (aka., the query set) in meta-learning?



(a) Graphical model of EB (b) MAML (c) Our method (SIB)

Summary

- Formulate transductive meta-learning with empirical Bayes model.
- Implement transductive amortized inference using synthetic gradient descent.

Figure 1. **A comparison between MAML and our method (SIB)** is shown in (b) and (c). MAML is an inductive method since, for a task t, it first constructs a variational posterior $q_{\theta_t^K}$ (a Dirac delta distribution) as a function of the labeled set d_t^l, and then apply $q_{\theta_t^K}$ on the unlabeled set x_t; while SIB constructs a better variational posterior as a function of both d_t^l and x_t: it starts with an initialization $\theta_t^0(d_t^l)$ generated using the labeled set d_t^l, and then yields θ_t^K by running K synthetic gradient steps on the unlabeled set x_t.

From hierarchical Bayes to empirical Bayes

Consider N tasks and the associated data $\mathcal{D} := \{d_t := (x_t, y_t)\}_{t=1}^N$:

$$\text{HB} \to \text{EB}: \quad p_f(\mathcal{D}) \to p_{\psi,f}(\mathcal{D}) = \int_\psi \Big[\prod_{t=1}^N \int_{w_t} p_f(d_t|w_t)p(w_t|\psi)\Big]p(\psi),$$

where $\log p_f(d_t|w_t) = -\sum_{i=1}^n \ell_t\Big(\hat{y}_{t,i}(f(x_{t,i}), w_t), y_{t,i}\Big) + C$. The ELBO is

$$\log p_{\psi,f}(\mathcal{D}) \geq \sum_{t=1}^N \Big[\mathbb{E}_{w_t \sim q_{\theta_t}}\big[\log p_f(d_t|w_t)\big] - D_{\mathsf{KL}}\big(q_{\theta_t}(w_t)\|p_\psi(w_t)\big)\Big].$$

Learning with variational inference in EB

$$\text{Exact}: \quad \min_{\psi,f}\min_{\theta_1,\dots,\theta_N} \sum_{t=1}^N D_{\mathsf{KL}}\Big(q_{\theta_t}(w_t) \,\Big\|\, p_{\psi,f}(w_t|d_t)\Big)$$

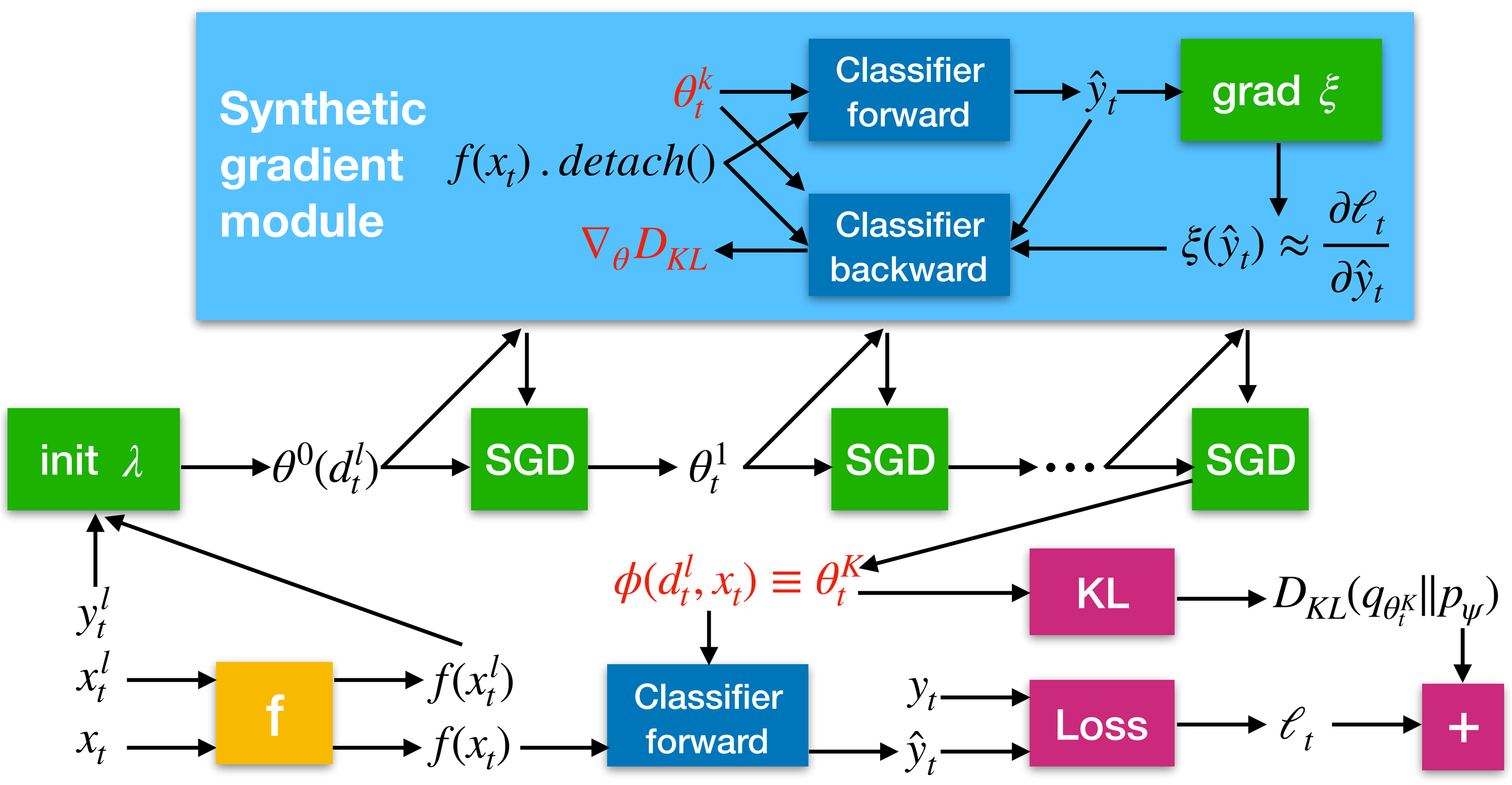
$$\text{Inductive}: \quad \min_{\psi,f}\min_{\phi} \sum_{t=1}^N D_{\mathsf{KL}}\Big(q_{\phi(d_t^l)}(w_t) \,\Big\|\, p_{\psi,f}(w_t|d_t)\Big)$$

$$\text{Transductive}: \quad \min_{\psi,f}\min_{\phi} \sum_{t=1}^N D_{\mathsf{KL}}\Big(q_{\phi(d_t^l,x_t)}(w_t) \,\Big\|\, p_{\psi,f}(w_t|d_t)\Big)$$

Unrolling exact inference with synthetic gradient [2]

How do we implement the amortization network $\phi(d_t^l, x_t)$? The best is through the exact inference $\phi(d_t^l, x_t) = \arg\min_{\theta_t} D_{\mathsf{KL}}\Big(q_{\theta_t}(w_t) \,\Big\|\, p_{\psi,f}(w_t|d_t)\Big)$. However, we don't have access to y_t at test time. Instead, we unroll the optimization by parameterizing (a) the **initialization** θ_t^0 and (b) the **gradient**

$$\nabla_{\theta_t} D_{\mathsf{KL}}\Big(q_{\theta_t}\|p_{\psi,f}\Big) = \mathbb{E}_\epsilon\Big[\sum_{i=1}^n \frac{\partial \ell_t(\hat{y}_{t,i}, y_{t,i})}{\partial \hat{y}_{t,i}} \frac{\partial \hat{y}_{t,i}}{\partial w_t} \frac{\partial w_t(\theta_t, \epsilon)}{\partial \theta_t}\Big] + \nabla_{\theta_t} D_{\mathsf{KL}}\Big(q_{\theta_t}\|p_\psi\Big)$$



Link to information bottleneck [3]

Consider an abstract variational posterior $q(w|d,t)$ with inference & generative processes:

Inference : $q(w, d, t) = q(t)q(d|t)q(w\,|\,d, t)$

Generative : $p(w, d, t) = p(d\,|\,w, t)p(w)q(t)$

Theorem (generalization analysis of EB via IB)

If ℓ_t is σ-subgaussian under $q(w|t)q(z|t)$, then

$$\min_{p(w)} \mathbb{E}_{q(t)}\mathbb{E}_{q(d|t)}\Big[D_{\mathsf{KL}}\big(q(w\,|\,d,t) \,\|\, p(w\,|\,d,t)\big)\Big]$$
$$\geq I_q(w; d\,|\,t) - \beta\, I_{q,p}(w; d\,|\,t) \text{ with } \beta = 1$$
$$\geq \frac{n}{2\sigma^2}\mathsf{gen}(q)^2 - \beta\, I_{q,p}(w; d\,|\,t),$$

where I_q and $I_{q,p}$ are mutual information and cross mutual information respectively and

$$\mathsf{gen}(q) = \mathbb{E}_{q(t)q(d|t)q(w|d,t)}\Big[\underbrace{\mathbb{E}_{b\sim q(\cdot|t)}\log\frac{p(d\,|\,w,t)}{p(b\,|\,w,t)}}_{\text{gen-error wrt } w}\Big]$$

Few-shot classification on Mini-ImageNet

Method	FeatNet f	Mini-ImageNet, 5-way 1-shot	Mini-ImageNet, 5-way 5-shot	CIFAR-FS, 5-way 1-shot	CIFAR-FS, 5-way 5-shot
MAML [1]	Conv-4-64	48.7±1.8%	63.1±0.9%	58.9±1.9%	71.5±1.0%
cc+rot [4]	Conv-4-64	54.8±0.4%	**71.9±0.3%**	63.5±0.3%	**79.8±0.2%**
SIB K=0	Conv-4-64	50.0±0.4%	67.0±0.4%	59.2±0.5%	75.4±0.4%
SIB K=3	Conv-4-64	**58.0±0.6%**	70.7±0.4%	**68.7±0.6%**	77.1±0.4%
cc+rot [4]	WRN-28-10	62.9±0.5%	**79.9±0.3%**	73.6±0.3%	**86.1±0.2%**
SIB K=0	WRN-28-10	60.6±0.4%	77.5±0.3%	70.0±0.5%	83.5±0.4%
SIB K=1	WRN-28-10	67.3±0.5%	78.8±0.4%	76.8±0.5%	84.9±0.4%
SIB K=3	WRN-28-10	69.6±0.6 %	78.9±0.4%	78.4±0.6%	85.3±0.4%
SIB K=5	WRN-28-10	**70.0±0.6%**	79.2±0.4%	**80.0±0.6%**	85.3±0.4%

Bibliography

[1] Finn et al. Model-agnostic meta-learning for fast adaptation of deep networks. ICML 2017.

[2] Jaderberg et al. Decoupled neural interfaces using synthetic gradients. ICML 2017.

[3] Achille and Soatto. Emergence of invariance and disentangling in deep representations. JMLR 2018.

[4] Gidaris et al. Boosting Few-Shot Visual Learning with Self-Supervision. ICCV 2019.

Paper and code