CONTINUAL INVARIANT RISK MINIMIZATION

Francesco Alesiani & Shujian Yu *
NEC Laboratories Europe GmbH
Kurfuerstenanlage 36,
D-69115 Heidelberg
{francesco.alesiani,shujian.yu}@neclab.eu

ABSTRACT

Empirical risk minimization can lead to poor generalization behaviour on unseen environments if the learned model does not capture invariant feature representations. Invariant risk minimization (IRM) is a recent proposal for discovering environment-invariant representations robust to spurious features. It was introduced by Arjovsky et al. (2019) and extended by Ahuja et al. (2020). The assumption of IRM is that all environments are available to the learning system at the same time. With this work, we generalize the concept of IRM to scenarios where environments are observed sequentially. We show that existing approaches, including those designed for continual learning, fail to identify the invariant features and models across sequentially presented environments. We extend IRM under a variational Bayesian and bilevel framework, creating a general approach to continual invariant risk minimization. We also describe a strategy to solve the optimization problems using a variant of the alternating direction method of multiplier (ADMM). We show empirically using multiple datasets and with multiple sequential environments that the proposed methods outperforms or is competitive with prior approaches.

1 INTRODUCTION AND BACKGROUND KNOWLEDGE

Let us consider a multi-environment setting where, given a set of training environments $E = \{e_1, e_2, \cdots, e_m\}$, the goal is to find parameters θ that generalize well to unseen (test) environments. Each environment e has an associated training data set D_e and a corresponding risk R^e.

$$R^e(w \circ \phi) \doteq E_{(x,y)\sim D_e} \ell_e((w \circ \phi)(x), y), \tag{1}$$

where $f_\theta = w \circ \phi$ is the composition of a feature extraction function ϕ and a classifier w.

As a predominant principle for designing machine learning models, Empirical Risk Minimization (ERM) minimizes the average loss across all training examples, regardless of environment. ERM has strong theoretical foundations in the case of iid data (Vapnik (1992)) but can fail dramatically when test environments differ significantly from training environments.

Improving generalization of learning systems has become a major research topic in recent years. Recently, there have been proposals of approaches that learn environment-invariant representations. The motivating idea is that the behavior of a model being invariant across environments makes it more likely that the model has captured a causal relationship between features and prediction targets. This in turn should lead to a better generalization behavior. Invariant risk minimization (IRM, Arjovsky et al. (2019)), which pioneered this idea, introduces the following bi-level optimization problem:

$$\min_{\phi \in H_\phi, w \in H_w} \sum_{e \in E} R^e(w \circ \phi) \quad \text{s.t. } w \in \arg\min_{w_e \in H_w} R^e(w_e \circ \phi), \forall e \in E, \tag{2}$$

where H_ϕ, H_w are the hypothesis sets for, respectively, feature extractors and classifiers. Unfortunately, solving the IRM bi-level programming problem directly is difficult since solving the outer problem requires solving multiple dependent minimization problems jointly. We can, however, relax

*http://www.neclab.eu

IRM to IRMv1 by fixing a scalar classifier and learning a representation ϕ such that the classifier is "approximately locally optimal" (Arjovsky et al. (2019))

$$\min_{\phi \in H_\phi} \sum_{e \in E} R^e(\phi) + \lambda ||\nabla_{w|w=1.0} R^e(w\phi)||^2, \forall e \in E, \tag{3}$$

where w is a scalar evaluated in 1 and λ controls the strength of the penalty term on gradients on w.

Invariant risk minimization games (IRMG, Ahuja et al. (2020)) expands on IRM from a game-theoretic perspective. However, as already noted by Javed et al. (2020), IRM and IRMG requires sampling data from multiple environments simultaneously for computing a regularization term pertinent to its learning objective, where different environments are defined by intervening on one or more variables of the world.

To address the problem of learning environment-invariant ML models in sequential environements, we make the following contributions:

- We expand both IRM and IRMG under a Bayesian variational framework and develop novel objectives (for the discovery of invariant models) in two scenarios: (1) the standard multi-environment scenario where the learner receives training data from all environments at the same time; and (2) the scenario where data from each environment arrives in a sequential manner.
- We demonstrate that the resulting bilevel problem objectives have an alternative formulation, which allows us to compute a solution efficiently using the alternating direction method of multipliers (ADMM).
- We compare our method to ERM, IRM, IRMG, and various continual learning methods (EWC, GEM, MER, VCL) on a diverse set of tasks, demonstrating comparable or superior performance in most situations.

2 Continual IRM by Approximate Bayesian Inference

Both IRM and IRMG assume the availability of training data from all environments at the same time, which is impractical and unrealistic in numerous applications. Here, we propose to adopt a probabilistic approach, exploiting the propagation of the model distribution over environments using Bayes' rule. We integrate both IRM and IRMG with stochastic models, introducing their variational counterparts that admit a continual extension.

2.1 Equivalent formulation of IRM as a Bilevel Optimization Problem (BIRM)

In order to extend the IRM principle of Equation 2 using the principle of approximate Bayesian inference, by applying Lemma 4 (in supplementary material), we first introduce the following new equivalent definition of IRM (equation 2).

Definition 1 (Bilevel IRM (BIRM)). *Let H_ϕ be a set of feature extractors and let H_w be the set of possible classifiers. An* **invariant predictor** *$w \circ \phi$ on a set of environments E is said to satisfy the Invariant Risk Minimization (IRM) property, if it is the solution to the following bi-level Invariant Risk Minimization (BIRM) problem*

$$\min_{\phi \in H_\phi, w \in H_w} \sum_{e \in E} R^e(w \circ \phi) \tag{4a}$$

$$\textit{s.t.} \quad \nabla_w R^e(w \circ \phi) = 0, \forall e \in E. \tag{4b}$$

This formulation results from substituting the minimization conditions in the constraint set of the original IRM formulation with the Karush–Kuhn–Tucker (KKT) optimality conditions. This new formulation allows us to introduce efficient solution methods and simplifies the conditions of IRM. It also justifies the IRMv1 model; indeed, when the classifier is a scalar value and the equality constraint is included in the optimization cost function, we obtain Equation 3. To solve the BIRM problem, we propose to use the Alternating Direction Method of Multipliers (ADMM) (Boyd et al. (2011)). ADMM is an alternate optimization procedure that improves convergence and exploits the decomposability of the objective function and constraints. Details of the BIRM-ADMM algorithm are presented in the supplementary material.

2.2 Bilevel Variational IRM

In Sec.A.1 we present the Variational definition of the BIRM principle that we call Bilevel Variational IRM (BVIRM). Based on this definition we derive two algorithms: the BVIRM-ADMM and the Continual BVIRM ADMM algorithms.

2.3 The BVIRM ADMM Algorithm

The solution of the variational BVIRM formulation can be obtained by using ADMM (Boyd et al. (2011)). While in general there are no convergence results of ADMM methods for this problem, for local minima, under proper conditions [1], the stochastic version of ADMM converges with rate $O(1/\sqrt{t})$ for convex functions and $O(\log t/t)$ for strongly convex functions (Ouyang et al. (2013)).

The BVIRM-ADMM formulation of the BVIRM problem is presented in Sec.A.2. We provide a pseudo-code implementation leveraging Equation 7 as Algorithm 1. One of the advantages of the ADMM formulation of BVIRM of Eq.7, is that it can be computed in parallel, where only Eq.7b requires synchronization among environments, while the other steps can be computed independently.

Algorithm 1: $w, \phi \leftarrow$ BVIRM-ADMM(E, R^e) ADMM version of the Bilevel Variational IRM Algorithm

Result: $w \circ \phi$: feature extraction and classifier for the environment E

// Randomly initialize the variables
1 $\omega, \omega_e, u_e, v_e, \theta \leftarrow$ Init() ;
// Outer loop (on θ) and Inner loop (on ω)
2 **while** *not converged* **do**
  // Update ϕ using SGD
3   $\theta = \text{SGD}_\theta(\sum_{e \in E} Q^e_\phi(q_w, q_\phi))$;
4   **for** $k = 1, \ldots, K$ **do**
5   **for** $e \in E$ **do**
6       $\omega_e = \text{SGD}_{\omega_e} L_\rho(\omega_e, u_e, \omega, v_e)$;
7       $\omega = 1/|E| \sum_e (\omega_e + u_e)$;
8       $u_e = u_e + (\omega_e - \omega)$;
9       $v_e = v_e + \nabla_\omega Q^e(\omega_e, \theta)$;
10   **end**
11   **end**
12 **end**

Algorithm 2: $w, \phi \leftarrow$ C-BVIRM-ADMM(E, R^e) ADMM version of the Bilevel Variational IRM Algorithm

Result: $w_\omega \circ \phi - \theta$: feature extraction and classifier for the environment E

// Randomly initialize the variables
1 $\omega, \omega_e, u_e, v_e, \theta \leftarrow$ Init() ;
2 $\bar{\omega} = 0$;
3 **for** $e \in E$ **do**
4   **for** $k = 1, \ldots, K$ **do**
5     $\theta = \text{SGD}_\theta(Q^e_\phi(q_w, q_\phi))$;
6   **while** *not converged* **do**
      // Update ω using SGD and ADMM
7       $\omega_e = \text{SGD}_{\omega_e} L_\rho(\omega_e, u_e, \omega, v_e)$;
8       $\omega = 1/2(\omega_e + u_e + \bar{\omega})$;
9       $u_e = u_e + (\omega_e - \omega)$;
10       $v_e = v_e + \nabla_\omega Q^e(\omega_e, \theta)$;
11   **end**
12   **end**
13   $\bar{\omega} = \omega_e$;
14 **end**

2.3.1 The Continual BVIRM ADMM Algorithm

In presence of sequential environments, the priors for the new environment are given by the previous environment's distributions q^-_ϕ and q^-_w, this is obtained by comparing the BVIRM definition in Eqs. (5) with the continual Bayesian learning Equation (10). In Equation 6 we thus now have $Q^e_\phi(q_w, q_\phi) = \mathbb{E}_{w\sim q_w, \phi\sim q_\phi} R^e(w \circ \phi) + \beta D_{\text{KL}}(q_\phi || q^-_\phi) + \beta D_{\text{KL}}(q_w || q^-_w)$ and $Q^e_w(q_w, q_\phi) = \mathbb{E}_{w\sim q_w, \phi\sim q_\phi} R^e(w \circ \phi) + \beta D_{\text{KL}}(q_w || q^-_w)$ Algorithm 2 presents an example implementation of ADMM[2] applied to the continual BVIRM formulation.

[1] These conditions are specific bounds on the magnitude and variance of the (sub-)gradients of the stochastic function (Ouyang et al. (2013)). We used ELU $\in C^\infty$ in the experiments. [2] In Algorithm 1 the ADMM update equation is implemented from line 6 to line 9, while in Algorithm 2, from line 7 to line 10.

Table 1: Mean accuracy ($N = 5$) on train and test environments when training on 2 consecutive environments on MNIST and the b01 color correlation. Baseline methods described in the supplementary material.

	C-BVIRM	C-VIRMG	C-VIRMv1	ERM	EWC	GEM	IRMG	IRMv1	MER	VCL	VCLC
train	71.3	69.1	51.4	86.4	87.4	87.4	86.3	85.3	87.3	**89.3**	**89.3**
	(4.2)	(2.8)	(3.4)	(1.2)	(2.7)	(2.7)	(1.2)	(0.8)	(1.7)	(0.6)	(0.6)
test	29.6	27.9	**46.0**	12.7	15.7	15.7	12.8	9.9	14.8	24.9	24.9
	(3.3)	(8.5)	(2.1)	(2.7)	(4.2)	(4.2)	(2.6)	(0.2)	(3.8)	(1.9)	(1.9)

3 Experimental Evaluation

3.1 Datasets and Experiment Setup

Colored MNIST Figure 1 shows samples from the Colored MNIST dataset. In each training environment, the task is to classify whether the digit is, respectively, even or odd. As in prior work, we add noise to the preliminary label by randomly flipping it with a probability of 0.25. The color of the image is defined by the variable z, which is the noisy label flipped with probability $p_c \in [0.1, 0.2]$. The color of the digit is green if z is even and red if z is odd. Each train environment contains $30,000$ images of size 28×28 pixels, while the test environment contains $10,000$ images where the probability $p_c = 0.9$. The color of the digit (`b01`) is thus generated from the label but depends on the environment. Figure 2 depicts the causal graph (the hammer indicating the effect of the intervention) of the environment. The variable "Color" is inverted when moving from the training to test environment.

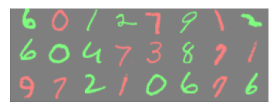


Figure 1: The Colored MNIST dataset with color correction (`b01`).

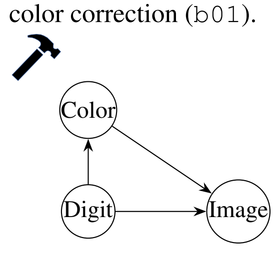


Figure 2: Causal relationships of colored MNIST.

3.2 Results

Table 1 lists the training and test accuracy on the MNIST dataset with the color correction b01 (see Figure 1). Since we introduced label noise by randomly flipping 25 percent of the given labels, a hypothetical optimal classifier would be able to achieve an accuracy of 75% in both training and test environments. ERM, IRMv1, and IRMG perform poorly in the setup where environments are given sequentially. Similarly, reference continual learning methods also fail to learn invariant representation in the new environment. As these models are learning to mainly use spurious features for the classification problems at hand, here: the colors of the digits (red$\sim$odd; green$\sim$even), they perform poorly (much worse than a random baseline) when the spurious feature properties are inverted (green$\sim$odd; red$\sim$even). In contrast, our variational extension IRM achieve a classification accuracy higher than 45% on the test data. This implies that our model is not relying exclusively on spurious correlations present in the color of digits. By comparing the performance between C-VIRMv1 and C-BVIRM, we conclude that (1) our proposed bilevel invariant risk minimization framework (i.e., the BIRM in Definition 1) is an effective alternative to the original formulation Arjovsky et al. (2019); and (2) ADMM is effective in solving the BIRM optimization problem and has the potential to improve the generalization performance. In addition, one can observe that the KL divergence term in VCL and our framework significantly improves the test accuracy with respect to the baseline counterparts. This result further justifies our motivation of using a variational Bayesian framework for the problem of continual invariant risk minimization.

4 Conclusions

We aim to broaden the applicability of IRM to settings where environments are observed sequentially. We show that reference approaches fail in this scenario. We introduce a variational Bayesian approach for the estimation of the invariant models and a solution based on ADMM. We evaluate the proposed approach with reference models, including those from continual learning, and show a significant improvement in generalization capabilities.

REFERENCES

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. *arXiv preprint arXiv:2002.04692*, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

J Andrew Bagnell. Robust supervised learning. In *Proceedings of the national conference on artificial intelligence*, volume 20, pp. 714, 2005.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICML Workshop on Uncertainty and Robustness*, 2020.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. Continual learning: A comparative study on how to defy forgetting in classification tasks. *arXiv preprint arXiv:1909.08383*, 2019.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pp. 2839–2848, 2016.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, pp. 8246–8256, 2018.

Khurram Javed, Martha White, and Yoshua Bengio. Learning causal models online. *arXiv preprint arXiv:2006.07461*, 2020.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. *arXiv preprint arXiv:1903.03448*, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems*, pp. 2687–2696, 2018.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pp. 6467–6476, 2017.

Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.

Hua Ouyang, Niao He, Long Tran, and Alexander Gray. Stochastic alternating direction method of multipliers. In *International Conference on Machine Learning*, pp. 80–88, 2013.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332*, 2015.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Siddharth Swaroop, Cuong V. Nguyen, Thang D. Bui, and Richard E. Turner. Improving and understanding variational continual learning. *arXiv:1905.02099 [cs, stat]*, 2019. URL http://arxiv.org/abs/1905.02099.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pp. 831–838, 1992.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019.

A Supplementary Material

A.1 Bilevel Variational IRM

At this point, we cannot yet directly extend the IRM principle using variational inference. That is because if we observe all environments at the same time, the prior of the single environment is data independent. Therefore, we substitute $q_{t-1}(\theta)$ from Equation 10 with priors $p_\phi(\theta)$ and $p_w(\omega)$, where θ and ω are now the parameters of the two functions ϕ and w. We also substitute $q_t(\theta)$ with the variational distributions $q_\phi(\theta)$ and $q_w(\omega)$.

Definition 2 (Bilevel Variational IRM (BVIRM))**.** *Let P_ϕ be a family of distributions over feature extractors, and let P_w be a family of distributions over classifiers. A* **variational invariant predictor** *on a set of environments E is said to satisfy Bilevel Variational Invariant Risk Minimization (BVIRM) if it is the solution to the following problem:*

$$\min_{\substack{q_\phi \in P_\phi \\ q_w \in P_w}} \sum_{e \in E} Q^e_\phi(q_w, q_\phi) \quad \text{(5a)} \qquad \textit{s.t.} \;\; \nabla_{q_w} Q^e_w(q_w, q_\phi) = 0, \forall e \in E, \quad \text{(5b)}$$

$$\textit{with} \qquad Q^e_\phi(q_w, q_\phi) = \mathbb{E}_{\substack{w \sim q_w \\ \phi \sim q_\phi}} R^e(w \circ \phi) + \beta D_{\mathrm{KL}}(q_\phi || p_\phi) + \beta D_{\mathrm{KL}}(q_w || p_w), \tag{6a}$$

$$\textit{and} \qquad Q^e_w(q_w, q_\phi) = \mathbb{E}_{\substack{w \sim q_w \\ \phi \sim q_\phi}} R^e(w \circ \phi) + \beta D_{\mathrm{KL}}(q_w || p_w), \tag{6b}$$

and where p_ϕ and p_w are the priors of the two distributions. β is a hyper-parameter balancing the ERM and closeness to the prior.

Definition 2 extends Definition 1 with the objective of Eq.10, where the parameters ϕ and w are substituted by their distributions q_ϕ and q_w. The gradient of the cost in the inner problem is taken with respect to the distribution q_w. When we parameterize q_ϕ with θ and q_w with ω, the gradient is evaluated with respect to these parameters[3], since the condition implies that the solution is locally optimal. If $Q(p, q)$ is convex in the first argument, then the solution is globally optimal. This definition extends the IRM principle to the case where we use approximate Bayes inference, shaping the variational distributions q_w and q_ϕ, to be, in expectation, invariant and optimal across environments.

A.2 The BVIRM-ADMM Algorithm

In the following we present th BVIRM-ADMM formulation of the BVIRM problem.

ADMM is defined by the update Eq.7, where we denote with the apexes $^-$ and $^+$ the value of any variable before and after the update. Moreover, we abbreviate as follows $Q(\omega, \theta) = Q(q(\omega), q(\theta))$.

$$\begin{aligned}
\omega_e^+ &= \arg\min_{\omega_e} L_\rho(\omega_e, u_e^-, \omega^-, v_e^-), \forall e \in E, && \text{(7a)} \\
\omega^+ &= 1/|E| \sum_e (\omega_e + u_e) && \text{(7b)} \\
u_e^+ &= u_e^- + (\omega_e^+ - \omega^+) && \text{(7c)} \\
v_e^+ &= v_e^- + \nabla_{q(\omega)} Q^e_w(\omega_e^+, \theta) && \text{(7d)}
\end{aligned}$$

with

$$L_\rho(w_e, u_e, w, v_e) \;\; = \;\; Q^e_\phi(\omega_e, \theta) + \frac{\rho_0}{2} \|\omega_e - \omega + u_e\|^2 + \frac{\rho_1}{2} \|\nabla_{q(\omega)} Q^e_w(\omega_e \circ \phi) + v_e\|^2. \tag{8}$$

Here, ϕ is fixed and θ is updated in an external loop or given (e.g. the identity function). In the experiment we use stochastic Gradient Descent (SGD) to update both the model parameters w_e and the feature extractor parameters ϕ. The result follows by applying Lemma 9 in the supplementary material and substituting $x_i \leftarrow \binom{w_e}{\phi}$, $f_i(x_i) \leftarrow Q^e_\phi(\omega_e, \theta)$ and $g_i(x_i) \leftarrow \nabla_{q(\omega)} Q^e_w(\omega_e^+, \theta)$.

[3] Implementation detail using the mean field parameterization and reparametrization trick is provided in the Supplementary Material

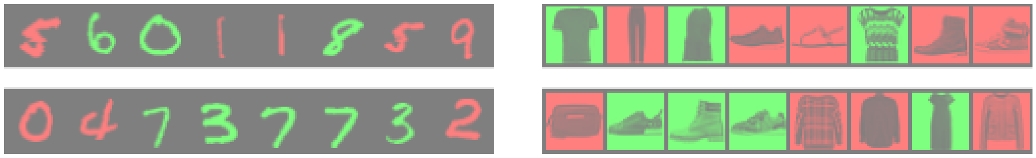


Figure 3: The two color models (on the left `b01`, on the right `b11`) for the train (upper row) and test (lower row) of the MNIST (left) and FashionMNIST (right) datasets.

A.3 Baseline Methods

We compare with a set of popular reference methods in order to show the advantage of the variational Bayesian framework in learning invariant models in the sequential environment setup. For completeness, we also evaluate the performances of four reference continual learning methods. These include Elastic Weight Consolidation (EWC, Kirkpatrick et al. (2017)), Gradient Episodic Memory (GEM, (Lopez-Paz & Ranzato, 2017))[4], Meta-Experience Replay (MER, Riemer et al. (2018))[5], and Variational Continual Learning (VCL, Swaroop et al. (2019); Nguyen et al. (2018))[6]. **ERM** is the classical empirical risk minimization method; we always use the cross-entropy loss. **IRMv1** enforces the gradient of the model with respect to a scalar to be zero. **IRMG** models the problem as a game among environments, where each environment learns a separate model. **EWC** imposes a regularization cost on the parameters that are relevant to the previous task, where the relevance is measured by Fisher Information (FI); **GEM** uses episodic memory and computes the updates such that accuracy on previous tasks is not reduced, using gradients stored from previous tasks; **MER** uses an efficient replay memory and employs the meta-learning gradient update to obtain a smooth adaptation among tasks; **VCL** and Variational Continual Learining with coreset **VCLC** apply variational inference to continual learning. **C-VIRMv1** and **C-VIRMG** refer to, respectively, our proposed variational extensions of **IRMv1** and **IRGM** in sequential environments. **C-BVIRM** is the implementation with ADMM.

All hyper-paramter optimization strategies and simulation configurations are discussed in detail in the supplementary material.

A.4 Extended Experiments

Colored FashionMNIST, KMNIST, and EMNIST Figure 3 (right) shows the Fashion-MNIST dataset, where the variable z defines the background color. Again, we add noise to the preliminary label ($y = 0$ for "t-shirt", "pullover", "coat", "shirt", "bag" and $y = 1$ for "trouser", "dress", "sandal", "sneaker", "ankle boots") by flipping it with 25 percent probability to construct the final label. Besides, we also consider Kuzushiji-MNIST dataset Clanuwat et al. (2018)[7] and the EMNIST Letters dataset Cohen et al. (2017)[8]. The former includes 10 symbols of Hiragana, whereas the latter contains 26 letters in the modern English alphabet. For EMNIST, there are $62,400$ training samples per environment and $20,300$ test samples. We set $y = 0$ for letters 'a', 'c', 'e', 'g', 'i', 'k', 'm', 'o', 'q', 's', 'u', 'v', 'y' and $y = 1$ for remaining ones.

Additional Results Table 2 lists the accuracy on the test environment for: (n) (upper rows) an increasing number of sequential environments (d) (central rows) different datasets, and (c) (lower rows) the two given color correlation schemes. We can observe that there is a general trend in the results. IRMG and IRM, with an accuracy of less than 10%, are not able to learn invariant models. Similarly, the continual learning reference methods (MER, EWC, MER, VCL, VCLC) also fail with a test accuracy of under 25%. The proposed methods on the other hand provide mechanism to learn more robust features and classification models. The higher variance of the accuracy is caused by the stochastic nature of the variational Bayesian formulation.

[4] https://github.com/facebookresearch/GradientEpisodicMemory
[5] https://github.com/mattriemer/mer
[6] https://github.com/nvcuong/variational-continual-learning
[7] https://github.com/rois-codh/kmnist
[8] https://www.nist.gov/itl/products-and-services/emnist-dataset

Table 2: Mean accuracy (over 5 runs) and standard deviation at test time for (n) $2, 6, 10$ environments, (d) across datasets, and (c) for the two color correlations (b01,b11).

		C-BVIRM	C-VIRMG	C-VIRMv1	ERM	EWC	GEM	IRMG	IRMv1	MER	VCL	VCLC
n	2	29.6	27.9	**46.0**	12.7	15.7	15.7	12.8	9.9	14.8	24.9	24.9
		(3.3)	(8.5)	(2.1)	(2.7)	(4.2)	(4.2)	(2.6)	(0.2)	(3.8)	(1.9)	(1.9)
	6	28.8	27.5	**47.1**	11.1	15.6	15.6	12.2	9.7	15.3	15.4	15.4
		(4.1)	(2.6)	(3.0)	(2.9)	(5.7)	(5.7)	(2.8)	(0.3)	(4.4)	(1.1)	(1.1)
	10	21.8	25.2	**31.0**	10.2	17.7	17.7	12.4	10.2	15.5	10.8	10.8
		(2.4)	(4.5)	(7.1)	(0.2)	(5.5)	(5.5)	(2.1)	(0.2)	(4.1)	(0.2)	(0.2)
d	MNIST	29.6	27.1	**46.8**	12.7	15.7	15.7	12.8	9.9	14.8	24.9	24.9
		(3.3)	(7.6)	(2.6)	(2.7)	(4.2)	(4.2)	(2.6)	(0.2)	(3.8)	(1.9)	(1.9)
	Fa-MNIST	36.3	26.7	**48.2**	10.7	15.4	15.3	10.8	9.9	13.2	24.9	24.9
		(4.3)	(8.7)	(3.6)	(1.5)	(5.1)	(5.4)	(1.4)	(0.2)	(2.8)	(2.0)	(2.0)
	KMNIST	32.8	24.2	**46.5**	12.0	14.0	14.0	12.1	9.9	15.6	24.9	24.9
		(4.6)	(6.0)	(1.9)	(2.1)	(3.5)	(3.5)	(2.4)	(0.2)	(4.1)	(2.0)	(2.0)
	EMNIST	32.1	25.0	**45.9**	10.8	15.3	14.8	10.8	10.0	12.6	24.9	24.9
		(4.6)	(7.5)	(2.5)	(1.0)	(3.6)	(3.7)	(1.2)	(0.2)	(2.3)	(2.0)	(2.0)
c	b01	29.6	27.1	**46.8**	14.9	18.8	18.8	14.6	9.8	18.0	24.9	24.9
		(3.3)	(7.6)	(2.6)	(0.8)	(1.3)	(1.3)	(0.8)	(0.1)	(1.0)	(2.0)	(2.0)
	b11	38.8	23.9	**43.3**	9.9	12.5	12.5	9.8	9.9	11.5	24.9	24.9
		(4.2)	(6.9)	(4.5)	(0.2)	(3.6)	(3.6)	(0.1)	(0.2)	(2.0)	(2.0)	(2.0)

A.5 Variational Continual Learning

Following prior work in continual learning (Nguyen et al. (2018)), let D_t be the training data from the t-th environment e^t, let D_1^t be the cumulative data up to the t-th environment, and let θ be the parameters of the feature extractor. When each environment is given in a sequential manner, we can use Bayes' rule and we have (all proofs are provided in the supplementary material)

$$p(\theta|D_1^t) \propto p(\theta|D_1^{t-1})p(D_t|\theta), \tag{9}$$

that is, once we have the posterior distribution $p(\theta|D_1^{t-1})$ at time $t-1$, we can obtain, by applying Bayes rule, the posterior $p(\theta|D_1^t)$ at time t up to a normalization constant. This is achieved by multiplying the previous posterior with the current data likelihood $p(D_t|\theta)$. The posterior distribution is in general not tractable and we use an approximation. With the variational approximation, $p(\theta|D_1^t) \approx q_t(\theta)$, it is thus possible to propagate the variational distribution from one environment to the next. From Corollary 12 (in the supplementary material) we can write the continual variational Bayesian inference objective as

$$q_t(\theta) = \arg\min_{q(\theta)} \mathbb{E}_{(x,y)\sim D_t}\mathbb{E}_{\theta\sim q(\theta)}\{\ell(y, f_\theta(x))\} + D_{\mathrm{KL}}(q(\theta)||q_{t-1}(\theta)), \tag{10}$$

from the variational distribution at step $q_{t-1}(\theta)$, with $f_\theta = w \circ \phi$, a function with parameters θ.

A.6 Related Work

Generalization Domain adaptation (Ben-David et al., 2007; Johansson et al., 2019) aims to learn invariant features or components $\phi(x)$ that have similar $P(\phi(x))$ on different (but related) domains by explicitly minimizing a distribution discrepancy measure, such as the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) or the Correlation Alignment (CORAL) (Sun & Saenko, 2016). The above condition, however, is not sufficient to guarantee successful generalization to unseen domains, even when the class-conditional distributions of all covariates changes between source and target domains (Gong et al., 2016; Zhao et al., 2019). Robust optimization (Hoffman et al., 2018; Lee & Raginsky, 2018), on the other hand, minimizes the worst performance over a set of possible environments E, that is, $\max_{e\in E} R^e(\theta)$. This approach usually poses strong constraint on the closeness between training and test distributions (Bagnell, 2005) which is often violated in practical settings (Arjovsky et al., 2019; Ahuja et al., 2020).

Incorporating the machinery of causality into learning models is a recent trend for improving generalization. (Bengio et al., 2019) argued that causal models can adapt to sparse distributional changes quickly and proposed a meta-learning objective that optimizes for fast adaptation. IRM, on the other hand, presents an optimization-based formulation to find non-spurious actual causal factors to target y. Extensions of IRM include IRMG and the Risk Extrapolation (REx) (Krueger et al., 2020). Our work's motivation is similar to that of online causal learning (Javed et al., 2020), which models the expected value of target y given each feature as a Markov decision process (MDP) and identifies the spurious feature x_i if $\mathbb{E}[y|x_i]$ is not consistent to temporally distant parts of the MDP. The learning is implemented with a gating model and behaves as a feature selection mechanism and, therefore, can be seen as learning the support of the invariant model. The proposed solution, however, is only applicable to binary features and assumes that the aspect of the spurious variables is known (e.g. the color). It also requires careful hyper-parameter tuning. In the cases where data is not divided into environments, Environment Inference for Invariant Learning (EIIL) classification method (Creager et al. (2020)) aims at splitting the samples into environments. This method proves to be effective also when the environment label is present.

Continual Learning Kirkpatrick et al. (2017); De Lange et al. (2019) addresses the problem of learning one classifier that performs well across multiple tasks given in a sequential manner. The focus is on the avoidance of catastrophic forgetting. With our work, we shift the focus of continual learning to the study of a single task that is observed in different environments.

A.7 Variational Invariant Risk Minimization games

The Invariant Risk Minimization Games (IRMG) (Ahuja et al. (2020)) proposes to learn an ensemble of classifiers with each environment controlling one component of the ensemble. Intuitively, the environments play a game where each environment's action is to decide its contribution to the ensemble aiming to minimize its risk. Specifically, IRMG optimizes the following objective:

$$\min_{\phi \in H_\phi} \sum_{e \in E} R^e(\bar{w} \circ \phi) \qquad \text{s.t. } w_e = \arg\min_{w \in H_w} R^e\left(\frac{1}{|E|}(w + w_{-e}) \circ \phi\right), \forall e \in E, \tag{11}$$

where $\bar{w} = \frac{1}{|E|}\sum_{e \in E} w_e$ is the average and $w_{-e} = \sum_{e' \in E, e' \neq e} w_{e'}$ the complement classifier.

We now consider the IRMG objective and extend it with the variational Bayesian inference. If we observe all environment at the same time, the prior of the single environment is data independent. From Equation 10, we thus substitute $q_{t-1}(\theta)$ with a priors $p_\phi(\theta)$ and $q_w(\omega)$, where θ and ω are now the parameters of the two functions ϕ and w. While we substitute $q_t(\theta)$, with the variational distributions $q_\phi(\theta)$ and $q_w(\omega)$. The outer problem is now

$$\min_{q_\phi} \qquad \mathbb{E}_{\phi \sim q(\phi)} R^e(\bar{w} \circ \phi) + \beta D_{\mathrm{KL}}(q_\phi || p_\phi) \tag{12a}$$

$$\text{s.t.} \qquad q_{w_e} = \arg\min_{q_{w_e}} \mathbb{E}_{w \sim q_{w_e}} R^e(\frac{1}{|E|}(w + w_{-e}) \circ \phi) + \beta D_{\mathrm{KL}}(q_{w_e} || p_w) \forall e \in E^{\mathrm{tr}} \tag{12b}$$

where $\bar{w} = \frac{1}{|E|}\sum_{e \in E^{\mathrm{tr}}} w_e,\ w_e \sim q_{w_e}(w)$ is the average classifier and $w_{-e} = \sum_{e' \in E^{\mathrm{tr}}, e' \neq e} w_{e'}, w_{e'} \sim q_{w_{e'}}(w)$ is the complement classifier. In the reformulation of the IRMG model, we weight the distance of the varional distribution to the prior with β. We notice how the difference of the variational formulation of IRMG differs on the presence of the mean on the distribution of the function over the variational distributions and the KL term.

We can now finally extend IRMG when the environments are observed sequentially. Combining the definition of IRMG Eqs. (11) with the continual bayesian learning Equation (10), we obtain the variational objective of IRMG in sequential environment case.

$$\min_{q_\phi} \qquad \mathbb{E}_{\phi \sim q(\phi)}\{\ell(y, \bar{w} \circ \phi)\} + \beta D_{\mathrm{KL}}(q_\phi || q_\phi^{t-1}) \tag{13a}$$

$$\text{s.t.} \qquad \bar{w} = \frac{1}{2}(w + w_{t-1}), w \sim q_w(w), w_{t-1} \sim q_w^{t-1}(w) \tag{13b}$$

$$q_w = \arg\min_{q_w} \mathbb{E}_{w \sim q_{w_e}, \phi \sim q_\phi}\{\ell(y, \frac{1}{2}(w + w_{t-1}) \circ \phi\} + \beta D_{\mathrm{KL}}(q_{w_e} || q_w^{t-1}) \tag{13c}$$

We can similarly extend the definition of IRMv1 when all environments are seen at the same time and sequentially.

A.8 Mean Field Parametrization and reparametrization trick

When we want to implement Equation 7 and Equation 8 and the different variation, we use the mean field approximation and the reparametrization trick Kingma & Welling (2013). In this case the density function of our model is parameterized by θ and ω and constraints becomes $\nabla_{q(\omega)} Q^e_w(\omega^+_e, \theta) = 0 \to \nabla_\omega Q^e_w(\omega, \theta) = 0$. If we then parametrize $\mu(\omega_\mu)$ and $\sigma(\omega_\sigma)$ the mean and standard deviation and model the distribution as $q_\omega(w) = \mu(\omega_\mu) + \epsilon\sigma(\omega_\sigma)$, with $\epsilon \sim N(0,1)$ We now want to compute the gradient (in the following we ignore the dependence on the ϕ and its parameters)

$$\nabla_\omega Q(\omega) = \nabla_\omega \mathbb{E}_{w\sim q(\omega)} R(w \circ \phi) + \beta \nabla_\omega D_{\mathrm{KL}}(q(\omega)||p)$$

The second term is

$$\nabla_\omega D_{\mathrm{KL}}(q||p) = \nabla_\mu D_{\mathrm{KL}}(q||p) \nabla_\omega \mu + \nabla_\sigma D_{\mathrm{KL}}(q||p) \nabla_\omega \sigma$$

with

$$\nabla_\omega \mu = 1, \nabla_\omega \sigma = \frac{1}{\epsilon}$$

$$\nabla_\mu D_{\mathrm{KL}}(q||p) = -\sigma_p^{-1}(\mu_p - \mu_q)$$

$$\nabla_\sigma D_{\mathrm{KL}}(q||p) = -\operatorname{diag}(\sigma_q)^{-1} + \operatorname{diag}(\sigma_p)^{-1}$$

where we assume σ_p, σ_q to be diagonal, in this way the previous equation can be evaluated element-wise and where the $D_{\mathrm{KL}}(q||p)$ is defined as

$$D_{\mathrm{KL}}(q||p) = \ln \frac{|\Sigma_p|}{|\Sigma_q|} - n + \operatorname{tr}\left\{\Sigma_p^{-1}\Sigma_q\right\} + (\mu_p - \mu_q)^T \Sigma_p^{-1}(\mu_p - \mu_q)$$

The first term is evaluated by Monte Carlo sampling

$$\nabla_\omega \mathbb{E}_{w\sim q(\omega)} R(w) \approx \nabla_\omega \frac{1}{N}\sum_{i=1}^N R(w_i)$$

with

$$w_i = \mu(\omega) + \epsilon_i \odot \sigma(\omega)$$

and $w_i \sim N(0,1)$. Also in this case

$$\nabla_\omega \frac{1}{N}\sum_{i=1}^N R(w_i) = \nabla_\mu \frac{1}{N}\sum_{i=1}^N R(w_i)\nabla_\omega \mu + \nabla_\sigma \frac{1}{N}\sum_{i=1}^N R(w_i)\nabla_\omega \sigma$$

A.9 The BIRM-ADMM Algorithm

We observe that to solve BIRM we can use Lemma 9 and write the following algorithm

$$w_e^+ = \arg\min_{w_e} L_\rho(w_e, u_e^-, w^-, v_e^-), \forall e \in E \tag{14a}$$

$$w^+ = 1/|E| \sum_i (w_e + u_e) \tag{14b}$$

$$u_e^+ = u_e^- + (w_e^+ - w^+) \tag{14c}$$

$$v_e^+ = v_e^- + \nabla_w R^e(w_e^+ \circ \phi) \tag{14d}$$

where

$$L_\rho(w_e, u_e, w, v_e) = R^e(w_e \circ \phi) + \frac{\rho_0}{2}\|w_e - w + u_e\|^2 + \frac{\rho_1}{2}\|\nabla_w R^e(w_e \circ \phi) + v_e\|^2 \tag{15}$$

We denote $.^+, .^-$ the values of the variable after and before the update. In order to implement the method we use the SGD to update the model w_e and in a outer loop updating for ϕ.

Algorithm 3: $w, \phi \leftarrow$ BIRM-ADMM(E, R^e) ADMM version of the Bilevel IRM Algorithm

Result: $w \circ \phi$: feature extraction and classifier for the environment E

`// Randomly initialize the variables`

$w, w_e, u_e, v_e, \phi \leftarrow$ Init() ;

`// Outer (on` ϕ`) and Inner loop (on` w`)`

while *not converged* **do**

  `// Update` ϕ `using stochastic gradient descent(SGD)`

  $\phi = \text{SGD}_\phi(\sum_e R^e(w \circ \phi))$;

  for $k = 1, \dots, K$ **do**

    for $e \in E$ **do**

      $w_e = \text{SGD}_{w_e} L_\rho(w_e, u_e, w, v_e)$;

      $w = 1/|E| \sum_e (w_e + u_e)$;

      $u_e = u_e + (w_e - w)$;

      $v_e = v_e + \nabla_w R^e(w_e \circ \phi)$;

    end

  end

end

A.10 Variational Invariant Risk Minimization

Definition 3 (VIRM). *Give a set of distribution over the mapping P_ϕ and a distribution over the set of classifier P_w, a* **variational invariant predictor** *on a set of environments E is said to satisfy the Variational Invariant Risk Minimization (VIRM) if it is the solution of the following problem*

$$\min_{\substack{q_\phi \in P_\phi \\ q_w \in P_w}} \quad \sum_{e \in E} Q^e_\phi(q_w, q_\phi) \tag{16a}$$

$$\textit{s.t.} \quad q_w \in \arg\min_{q^e_w \in P_w} Q^e_w(q_w, q_\phi), \forall e \in E \tag{16b}$$

$$\textit{where} \quad Q^e_\phi(q_w, q_\phi) = \mathbb{E}_{\substack{w \sim q_w \\ \phi \sim q_\phi}} R^e(w \circ \phi) + \beta D_{\mathrm{KL}}(q_\phi || p_\phi) + \beta D_{\mathrm{KL}}(q_w || p_w), \tag{16c}$$

$$Q^e_w(q_w, q_\phi) = \mathbb{E}_{\substack{w \sim q_w \\ \phi \sim q_\phi}} R^e(w \circ \phi) + \beta D_{\mathrm{KL}}(q_w || p_w) \tag{16d}$$

and p_ϕ, p_w are the priors of the two distributions.

A.11 Bilevel Alternative Formulation

We state here a general result on solving Bilevel Optimization Problems

Lemma 4 (Bilevel Reformulation)**.**

$$\min_x \quad F(x,y) | G(x, y(x)) \leq 0 \tag{17a}$$

$$\textit{s.t.} \quad y(x) \in \arg\min_y f(x,y) | g(x,y) \leq 0 \tag{17b}$$

then we can solve the equivalent problem

$$\min_{x,y,u} \quad F(x,y) | G(x, y(x)) \leq 0, \tag{18a}$$

$$\nabla_y L(x,y,u) = 0, \tag{18b}$$

$$u \geq 0, \tag{18c}$$

$$g(x,y) \leq 0, \tag{18d}$$

$$u^T g(x,y) = 0 \tag{18e}$$

$$L(x,y,u) \quad = \quad f(x,y) + u^T g(x,y) \tag{18f}$$

Proof of Lemma 4 . Lemma 4 follows by applying the Karush-Kuhn-Tucker conditions (Chapter 5 Boyd et al. (2004)) to Eq.17, where the Lagrangian function is $L(x,y,u) = f(x,y) + u^T g(x,y)$. □

Lemma 5 (Equivalence of Definition 1)**.** *Definition 1 is equivalent to Eq. 2, the Invariant Risk Minimization.*

Proof of Lemma 5 . The result follows by apply Lemma 4 to Eq.2. ☐

Lemma 6 (Definition 2)**.** *Definition 2 is the extension of Eq. 4, the Bilevel Invariant Risk Minimization, when the function is described by the distributions of their variable ϕ and w.*

Proof of Lemma 6 . The result follows by inspecting Eq. 4. The equation requires the minimisation of the aggregated loss function, which is now, from Eq.10:

$$Q^e_\phi(q_w, q_\phi) = \mathbb{E}_{\substack{w\sim q_w \\ \phi\sim q_\phi}} R^e(w \circ \phi) + \beta D_{\mathrm{KL}}(q_\phi||p_\phi) + \beta D_{\mathrm{KL}}(q_w||p_w), \tag{19}$$

where we have separated the two contributions in ϕ and w, and used genetic prior distributions p_ϕ and p_w. This is by the additive property of KL divergence:

$$D_{\mathrm{KL}}(q_\phi q_w||p_\phi p_w) = D_{\mathrm{KL}}(q_\phi||p_\phi) + D_{\mathrm{KL}}(q_w||p_w), \tag{20}$$

since we model the two distributions independently, i.e. $q_{\phi,w} = q_\phi q_w$ and $q_{\phi,w} = p_\phi p_w$. Since the classifiers' losses shall be minimal for all environments, this condition is substituted by requiring the gradient with respect to q_w to be zero, $\forall e$. The gradient w.r.t. q_w of the second term of Eq.19 is zero. ☐

A.12 Out of distribution Generalization

The question arises if the property of generalization to out of distributions given by Theorem 9 in Arjovsky et al. (2019) also holds for BIRM and BVIRM.

Lemma 7. *If ϕ and w are linear functions and $w \circ \phi = \Phi^T w$ is a solution of Eq.4 it then satisfies*

$$\Phi\mathbb{E}_{X^e}\left[X^e X^{eT}\right]\Phi^T w = \Phi\mathbb{E}_{X^e,Y^e}\left[X^e Y^{eT}\right]$$

Proof. Lemma 7 follows from the fact that

$$\begin{aligned} \nabla_w R^e(w\circ\phi) &= \Phi\mathbb{E}_{X^e}\left[X^e X^{eT}\right]\Phi^T w - \Phi\mathbb{E}_{X^e,Y^e}\left[X^e Y^{eT}\right] && (21)\\ &= 0 && (22)\end{aligned}$$

☐

For BIRM thus Theorem 9 of Arjovsky et al. (2019) applies directly. A similar results holds for the the BVIRM model

Lemma 8. *If $\phi \sim p_\phi$ and $w \sim p_\phi$ are linear functions and $w \circ \phi = \Phi^T w$ is a solution of Eq.5, with $\beta = 0$, it then satisfies*

$$\mathbb{E}_{\phi\sim q_\phi}\left\{\Phi\mathbb{E}_{X^e}\left[X^e X^{eT}\right]\Phi^T\right\}\bar{w} = \bar{\Phi}\mathbb{E}_{X^e,Y^e}\left[X^e Y^{eT}\right]$$

where $\bar{\Phi} = \mathbb{E}_{\Phi\sim p_\phi}[\Phi]$ and $\bar{w} = \mathbb{E}_{w\sim p_w}[w]$ are the mean values.

Proof. Lemma 8 follows from the fact that

$$\begin{aligned} \nabla_{q_w} Q^e_w(q_w, q_\phi)|_{\beta=0} &= \nabla_{q_w} \mathbb{E}_{\substack{w\sim q_w \\ \phi\sim q_\phi}} R^e(w\circ\phi) && (23)\\ &= 0 && (24)\end{aligned}$$

We now take the Fréchet directional derivative in the η direction that is the limit of

$$\delta_{q_w,\eta}\,\mathbb{E}_{\substack{w\sim q_w \\ \phi\sim q_\phi}} R^e(w\circ\phi) = \lim_{\epsilon\to 0}\frac{1}{\epsilon}(\mathbb{E}_{\substack{w\sim q_w \\ \phi\sim q_\phi}} R^e(((w+\epsilon\eta)\circ\phi) - \mathbb{E}_{\substack{w\sim q_w \\ \phi\sim q_\phi}} R^e(w\circ\phi))$$

which is obtained when we differentiate the distribution $q_w \rightarrow q_w + \epsilon\eta$. Since $\delta_{q_w,\eta}\,\mathbb{E}_{w\sim q_w\,\phi\sim q_\phi} R^e(w\circ\phi) = \mathbb{E}_{w\sim q_w\,\phi\sim q_\phi} 2\eta^T\Phi\mathbb{E}_{X^e}\left[X^e X^{eT}\right]\Phi^T w - 2\eta^T\Phi\mathbb{E}_{X^e,Y^e}\left[X^e Y^{eT}\right]$ we can factorize for the direction η and obtain

$$\delta_{q_w} R^e(w\circ\phi) = 2\,\mathbb{E}_{\substack{w\sim q_w \\ \phi\sim q_\phi}} \Phi\mathbb{E}_{X^e}\left[X^e X^{eT}\right]\Phi^T w - \Phi\mathbb{E}_{X^e,Y^e}\left[X^e Y^{eT}\right]$$

We can now derive the Lemma by requiring $\delta_{q_w} R^e(w\circ\phi) = 0$ ☐

Theorem 9 of Arjovsky et al. (2019) now holds when ϕ has $\operatorname{rank} r > 0$ in expectation with respect to the invariant distribution q_ϕ, i.e. $\mathbb{E}_{\phi \sim q_\phi} \operatorname{rank}(\Phi) = r$.

A.13 Generalized ADMM

The following generalization of ADMM holds:

Lemma 9 (GADMM)**.** *Suppose we want to minimized*

$$\min_x \qquad \sum_i f_i(x) | g_i(x) = 0, \forall i \in I \tag{25}$$

we can equivalently solve the following problem

$$\min_{x_i, z} \qquad \sum_i f_i(x_i) | x_i = z, g_i(x_i) = 0, \forall i \in I \tag{26}$$

using the following update role (scaled ADMM)

$$x_i^+ = \arg\min_{x_i} L_\rho(x_i, x_{-i}^-, u_i^-, z^-, v_i^-), \forall i \in I \tag{27a}$$

$$z^+ = 1/N \sum_i (x_i + u_i) \tag{27b}$$

$$u_i^+ = u_i^- + (x_i^+ - z^+) \tag{27c}$$

$$v_i^+ = v_i^- + g_i(x_i^+) \tag{27d}$$

where the augmented Lagrangian

$$L_\rho(x_i, u_i, z, v_i) = \sum_i f_i(x_i) + \frac{\rho_0}{2} \sum_i \|x_i - z + u_i\|^2 + \frac{\rho_1}{2} \sum_i \|g_i(x_i) + v_i\|^2 \tag{28}$$

and $x_{-i} = \{x_j, j \neq i\}$ is the set of all other variable, expect the i-th.

A.14 Continual Variational Inference

Following Nguyen et al. (2018) we can state the following lemma.

Lemma 10 (Variational Continual Learning)**.** *Suppose we have a sequence of datasets $D_i, i = 1, \ldots, t$ drown i.i.d, then the variational estimation of the distribution q_t at step t is given as projection on KL divergence*

$$q_t(\theta) = \arg\min_{q(\theta)} D_{\mathrm{KL}}\left(q(\theta) || \frac{1}{Z_t} q_{t-1}(\theta) p(D_t|\theta) \right)$$

with $Z_t = \int q_{t-1}(\theta) p(D_t|\theta) d\theta$ the normalization factor, which does not depends on q.

Proof of Lemma 10. Let denote $D_1^t = \bigcup_{i=1}^t D_i$, from i.i.d. $p(D_1^t) = \prod_{i=1}^t p(D_i)$. We are interested to maximase the a posteriori probability of the paramters give the data $p(\theta|D_1^t)$

$$p(\theta|D_1^t) = \frac{1}{p(D_t)} p(\theta|D_1^{t-1}) p(D_t|\theta)$$

since

$$\begin{aligned} p(\theta, D_1^t) &= p(\theta|D_1^t) p(D_1^t) \\ &= p(\theta) p(D_1^t|\theta) \\ &= p(\theta) \prod_{i=1}^t p(D_i|\theta) \\ &= p(\theta) p(D_1^{t-1}|\theta) p(D_t|\theta) \\ &= p(\theta, D_1^{t-1}) p(D_t|\theta) \\ &= p(\theta|D_1^{t-1}) p(D_1^{t-1}) p(D_t|\theta) \end{aligned}$$

thus

$$\begin{aligned} p(\theta|D_1^t) &= \frac{1}{p(D_1^t)} p(\theta|D_1^{t-1}) p(D_1^{t-1}) p(D_t|\theta) \\ &= \frac{p(D_1^{t-1})}{p(D_1^t)} p(\theta|D_1^{t-1}) p(D_t|\theta) \\ &= \frac{1}{p(D_t)} p(\theta|D_1^{t-1}) p(D_t|\theta) \end{aligned}$$

We now use a probability distribution which approximates the distribution at step $t-1$

$$q_{t-1}(\theta) \approx p(\theta|D_1^{t-1})$$

when then want to approximate at time t

$$q_t(\theta) \approx \frac{1}{p(D_t)} q_{t-1}(\theta) p(D_t|\theta)$$

This can be obtain by minimizing the KL divergence of the variational distribution q_t and the distribution induced by the previous step approximation, thus

$$q_t(\theta) = \arg\min_{q(\theta)} D_{\mathrm{KL}}\left(q(\theta) || \frac{1}{Z_t} q_{t-1}(\theta) p(D_t|\theta) \right)$$

□

Lemma 11 (VCLv2). *The minimization of the VCL defined in Lemma 10, is equivalent to solve the following minimization*

$$q_t(\theta) = \arg\max_{q(\theta)} \mathbb{E}_{\theta\sim q(\theta)}\{\log p(D_t|\theta)\} - D_{\mathrm{KL}}(q(\theta)||q_{t-1}(\theta))$$

with N_t i.i.d. samples

$$\mathbb{E}_{\theta\sim q(\theta)}\{\log p(D_t|\theta)\} = \frac{1}{N_t}\sum_{i=1}^{N_t} E_{\theta\sim q(\theta)}\{\log p(y_i^t|\theta, x_t^t)\}$$

Where the second term can be computed in closed form for known distribution as for example with the Gaussian distributions, whereas the expectation can be approximated by Monte Carlo sampling. For a general loss function we can substitute the reconstruction probability with the loss function associated with a neural network parametrized by θ

$$\log p(y_i^t|\theta, x_t^t) \leftarrow \ell(y_i^t, (w\circ\phi)_\theta(x_t^t))$$

$$\mathbb{E}_{\theta\sim q(\theta)}\{\log p(D_t|\theta)\} \leftarrow \frac{1}{N_t}\sum_{i=1}^{N_t} E_{\theta\sim q(\theta)}\ell(y_i^t, (w\circ\phi)_\theta(x_t^t))\}$$

Proof of Lemma 11. The Lamma follows from the definition of the KL diveregnce

$$\begin{aligned} D_{\mathrm{KL}}\left(q(\theta) || \frac{1}{Z_t} q_{t-1}(\theta) p(D_t|\theta) \right) &= \mathbb{E}_q(\ln q(\theta) - \ln q_{t-1}(\theta) - \ln p(D_t|\theta) + \ln Z_t) \\ &= \mathbb{E}_q(\ln q(\theta) - \ln q_{t-1}(\theta)) - \mathbb{E}_q \ln p(D_t|\theta) + \mathbb{E}_q \ln Z_t \\ &= D_{\mathrm{KL}}(q(\theta)||q_{t-1}(\theta)) - \mathbb{E}_q \ln p(D_t|\theta) + \ln Z_t \end{aligned}$$

The last term does not depend on q. Thus the result follows. □

If we substitute the log of the posterior probability with a specific loss function we obtain the following Corollary.

Corollary 12 (Continual Variational Bayesian Inference). *Given a loss function $\ell(y,\hat{y})$, the variational continual learning is formulated as*

$$q_t(\theta) = \arg\min_{q(\theta)} \mathbb{E}_{(x,y)\sim D_t}\mathbb{E}_{\theta\sim q(\theta)}\{\ell(y, f_\theta(x))\} + D_{\mathrm{KL}}(q(\theta)||q_{t-1}(\theta)), \tag{29}$$

with $f_\theta = (w\circ\phi)_\theta$

A.15 DATASETS AND COLOR CORRECTION

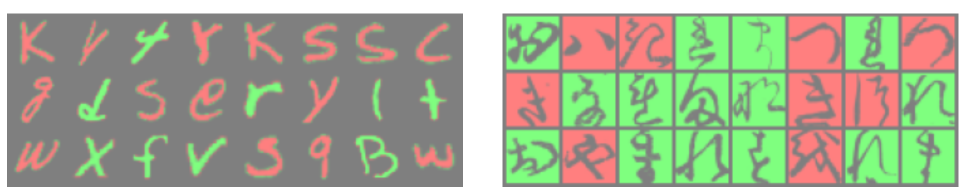
We here visualize few of the dataset and color correlations. Figure 4 shows Fashion-MNIST and the b11 color correlation. In the test environment the background color of each class is inverted. In Figure 5 we show the dataset as generated from Ahuja et al. (2020). In Figure 6 we show the EMINST (letter) and KMNIST dataset.

(a) (b)

Figure 4: Fashion MNIST dataset training (a) and testing (b) environments; the color is inverted based on the b11 color correlation scheme, where the background color depends on the class of the image. In the test environment the dependency is inverted.

(a) (b)

Figure 5: MNIST dataset (a) and Fashion MNIST (b) environments as defined in Ahuja et al. (2020)



(a) (b)

Figure 6: Examples of the EMNIST dataset (a) and of the KMNIST (b).

A.16 HYPER-PARAMETER SEARCH AND EXPERIMENTAL SETUP

We performed hyper-parameter search around the suggested values from the original works and the values selected based on the best performance on the test environment. To implement a complete comparison we used for training $1'000$ samples randomly drawn from each environment. All methods were trained on the same data, using random seed reset. We trained all method with 100 epochs on a batch size of 256.

- **IRM**: $\gamma = 91257$, threshold = 1/2 epochs, learning rate $2.5e^{-4}$
- **IRMG**: warm start=300, termination accuracy 0.6, learning rate $2.5e^{-4}$, dropout probability 75%, weight decays = .00125
- **ERM**: learning rate $1e^{-3}$, dropout probability 75%, weight decays = .00125
- **MER**: memory size 100 (10% of the samples), learning rate $1e^{-3}$, replay batch size =5, $\beta = .03$, $\gamma = 1.0$

Table 3: Mean accuracy (over 5 runs) on Synthetic Dataset (Arjovsky et al. (2019),Creager et al. (2020)). BIRM refers to our bilevel objective Eq. (8) optimized with ADMM.

	Causal MSE	Noncausal MSE
ERM	0.827 ± 0.016	0.824 ± 0.015
ICP	1.000 ± 0.000	0.756 ± 0.423
IRM	0.280 ± 0.006	0.290 ± 0.009
BIRM	0.183 ± 0.005	$\mathbf{0.184} \pm 0.002$
EIIL(IRM)	$\mathbf{0.180} \pm 0.026$	0.188 ± 0.033

- **GEM**: memory size 100 (10% of the samples), learning rate $1e^{-3}$
- **EWC**: memory size 100 (10% of the samples), learning rate $1e^{-3}$, regularization 0.1
- **VCL,VCLC**: learning rate = $5e^{-3}$, corset size 100 (10% of the samples),
- **C-BVIRM, C-VIRMG, C-VIRMv1**: weight decays = .00125, $\beta = 1.$, number evaluations 5, $\rho_0 = \rho_1 = 10$, step threshold =1/2 epochs, $\delta\rho = 100$, learning rate $1e^{-3}$

The neural network architecture is composed of 2 non-linear Exponential Linear Unit (ELU) activated Full Connected layers of size 100, followed by a linear full connected layer. Each layer with dropout. Dropout is not present in VCL/VCLC since not implemented in the original work. Training loss is the Cross Entropy. We tested also with the feature extraction layer separated, but with no advantage, since the test set-up only consist of one task.

The IRMv1, IRMG and ERM methods, similarly to the other methods, are trained sequentially as data from each new environment arrives. The Continual Learning methods are allowed to have a limited memory of samples from previous environments.

A.17 Hyper-parameters of environment inference for continual invariant learning

We list below the values of hyper-parameters in EIIL for continual invariant learning:
2 layers,
390 hidden neurons,
501 epochs,
l2 regularizer weight: 0.00110794568,
learning rate: 0.0004898536566546834,
numbber of runs: 10,
number of EIIL iterations: $10'000$,
number Monte Carlo evaluations: 3,
penalty anneal iterations: 190,
penalty weight: 191257.18613115902,
prior weight: $1e^{-6}$

A.18 Synthetic Dataset

The Synthetic Dataset is described in (Arjovsky et al. (2019)) for testing IRM and it is defined by a Structural Causal Model (Pearl (2009)), where a variable $y \in \mathbb{R}^N$ is generated by $x_1 \in \mathbb{R}^N$, while $x_2 \in \mathbb{R}^N$ is generated by y. The observed variable is $x = (x_1, x_2)$. The structural equations are

$$x_1 = \epsilon_1, \ \epsilon_1 \sim N(0, \sigma_1^2) \quad (30)$$

$$y = x_1 + \epsilon_y, \ \epsilon_y \sim N(0, \sigma_e^2) \quad (31)$$

$$x_1 = y + \epsilon_2, \ \epsilon_2 \sim N(0, 1) \quad (32)$$

with σ_1 fixed and σ_e dependent on the environment. We compared with ERM, IRM (Arjovsky et al. (2019)), IPC (Invariant Prediction), which is the method proposed in Peters et al. (2015), and EIIL (Creager et al. (2020)). We use a similar set up of Creager et al. (2020), with $N = 4$. The invariant model is given by $w = (1, 0)$.