Analysis of Trainability of Gradient-based Multi-environment Learning from Gradient Norm Regularization Perspective

Shiro Takagi
Department of Complexity Science and Engineering
The University of Tokyo
Chiba, Japan
shiro-takagi@g.ecc.u-tokyo.ac.jp

Yoshihiro Nagano
Department of Complexity Science and Engineering
The University of Tokyo
Chiba, Japan
nagano@k.u-tokyo.ac.jp

Yuki Yoshida
KARAKURI Inc.
Tokyo, Japan
y.yoshida@karakuri.ai

Masato Okada
Department of Complexity Science and Engineering
The University of Tokyo
Chiba, Japan
okada@edu.k.u-tokyo.ac.jp

***Abstract*—Adaptation and invariance to multiple environments are both crucial abilities for intelligent systems. Model-agnostic meta-learning (MAML) is a meta-learning algorithm to enable such adaptability, and invariant risk minimization (IRM) is a problem setting to achieve the invariant representation across multiple environments. We can formulate both methods as optimization problems with the environment-dependent constraint and this constraint is known to hamper optimization. Therefore, understanding the effect of the constraint on the optimization is important. In this paper, we provide a conceptual insight on how the constraint affects the optimization of MAML and IRM by analyzing the trainability of the gradient descent on the loss with the gradient norm penalty, which is easier to study but is related to both MAML and IRM. We conduct numerical experiments with practical datasets and architectures for MAML and IRM and validate that the analysis of the gradient norm penalty loss captures well the empirical relationship between the constraint and the trainability of MAML and IRM.**

***Index Terms*—model-agnostic meta-learning, invariant risk minimization, gradient norm penalty, regularization, trainability**

I. Introduction

The capability to cope with various environments is essential for intelligent systems. On the one hand, the intelligent systems should rapidly adapt to a new environment from the environment in which they have been trained. On the other hand, they also have to be robust against the noise in each environment to assure that they learn the essential patterns useful to multiple environments. These adaptation and invariance with respect to environment are the crucial capabilities for the multi-environment learning.

For the adaptability for the new environment, model-agnostic meta-learning (MAML), an optimization-based meta-learning algorithm, is known to be a successful algorithm for finding the solutions that are adaptive to unseen environments[1] [2]. For the invariance to the environment-specific noise, invariant risk minimization (IRM) is attracting increasing attention as a problem setting that induces the optimal classifier invariant to the environmental noise [1]. Both MAML and IRM, and other multi-environment learning methods, aim to minimize some energy function for all environments, while satisfying some condition in each environment. Hence, they are formulated as the optimization problems with environment-dependent constraints. Because the constraint for MAML determines how adaptive the obtained solution is and that for IRM controls the invariance to the variations among the environments, this constraint is essential for achieving the optimization problem and should be as strict as possible.

Meanwhile, since the constraint affects the shape of the overall energy landscape to be minimized, a too strict constraint is likely to increase the difficulty of the optimization [3], [4]. This trade-off highlights the importance of comprehending the relationship between the magnitude of the constraint and the trainability of the optimization. The goal of this paper is to elucidate this relationship for MAML and IRM.

However, the constraint makes the study of the optimization problem complicated and challenges our conceptual understanding of its trainability. This motivates us to investigate the trainability of an alternative problem that is related to MAML and IRM but is easier to analyze. A key observation is that both MAML and IRM are associated with loss minimization with

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research(A) (No. 18H04106).

[1] To discuss MAML and IRM in a unified notation, we use the term *environment*, instead of the term *task*, which has been conventionally used in previous studies. We use the term *environment* as a general term that refers to anything underlying the learning process, following the definition in [1]. Although the definition of a task varies among the different reports in the literature, one reasonable interpretation in the context of supervised learning is that *a task is an index variable on which the data are conditioned*. This definition of a task is included in the above definition of environment.

the gradient norm penalty as we explain in Sections III and IV. In this paper, therefore, we will examine the trainability of the optimization on the loss with a gradient penalty. Specifically, we will focus on gradient descent for this regularized loss because gradient descent is the most commonly currently used optimization method. In gradient descent, learning rate is a crucial parameter that determines the trainability and a too large learning rate results in loss divergence. Thus, we will consider the maximum possible learning rate for a given regularization coefficient such that gradient descent can find local minima. Analysis of this alternative problem provides a useful explanation on how the training of MAML and IRM is determined by their constraints.

The rest of this paper is organized as follows. In Section II, we introduce the gradient penalty loss. In Sections III and IV, we show that MAML and IRM are associated with loss minimization with the gradient penalty of negative and positive signs, respectively. In Section V, we explain what the flip in sign of the penalty between MAML and IRM may indicate. In Sections VII and VIII, we discuss the acceptable regularization coefficient and learning rate of negative and positive gradient penalty loss for gradient descent to locally reach a local minimum. For negative gradient penalty loss, we find that the maximum possible learning rate increases when the regularization coefficient is close to its upper bound. By contrast, for the positive penalty, we show that the possible learning rate decays quickly with increasing regularization. In Section IX, we explain the result of the experiment to verify the statements in the previous sections. The results show that the relationship between the constraint and trainability of MAML and IRM in Sections VII and VIII is qualitatively consistent with the relationship observed in the experiment. This result supports that the analysis of the gradient penalty loss is useful for the elucidation of how the constraint influences the successful training of MAML and IRM.

II. Gradient Penalty Loss

Suppose that we have a dataset $\mathcal{D}$ and a model $f_{\boldsymbol{\theta}}$ parameterized by p-dimensional parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. By using the dataset and the model, we calculate the empirical risk $L(\boldsymbol{\theta})$. Here, we define *gradient penalty loss* as follows:

$$L_{-}(\boldsymbol{\theta}) \coloneqq L(\boldsymbol{\theta}) - \alpha \left\| \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \right\|^2, \tag{1}$$

$$L_{+}(\boldsymbol{\theta}) \coloneqq L(\boldsymbol{\theta}) + \alpha \left\| \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \right\|^2, \tag{2}$$

where α is the regularization coefficient and $\|\cdot\|$ is the $L2$ norm. Gradient descent is performed on these loss functions with the learning rate β. We refer to (1) as *negative gradient penalty loss*, and (2) as *positive gradient penalty loss*.

Instead of using α above, in Sections VII and VIII, we use α scaled by $1/2$. This scaling is just for notational simplicity and does not affect the qualitative relationship between α and the upper bound of β. We note that in the following sections, we use the notation of α and β a several times to mean different concepts so that we can highlight the correspondence of MAML and IRM to the gradient penalty loss.

III. Relationship between Model-agnostic Meta-learning and Negative Gradient Penalty

A. Background

MAML finds the solution from which it can quickly reach the optimal solution $\boldsymbol{\theta}^*_\tau$ for environment τ with few data. For this purpose, MAML employs the bi-level gradient-based optimization that consists of the inner-loop where the parameter is updated in an environment-specific manner and the outer-loop where the environment-invariant representation is learned.

Suppose that we sample environment τ from an environment distribution $P(\tau)$. We split the data conditioned on each environment into the data for the inner-loop and those for the outer-loop. Usually, the former is called the training dataset and the later is called the test dataset. The training and the test dataset of environment τ are denoted by $\mathcal{D}^{tr}_\tau$ and $\mathcal{D}^{te}_\tau$.

In the inner-loop, we sample a batch of data $\{D^{tr}_{\tau\nu}\}^K_{\nu=1} \subset \mathcal{D}^{tr}_\tau$ for each τ, and update the parameter by gradient descent:

$$\boldsymbol{\theta}_\tau \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \frac{1}{K} \sum_{\nu}^{K} \ell\left(D^{tr}_{\tau\nu}, \boldsymbol{\theta}\right), \tag{3}$$

where α is a step size known as the inner learning rate and $\ell(\cdot,\cdot)$ is the loss function for each environment.

In the outer-loop, we compute the loss for the parameter updated in the inner-loop with a batch of test data $\{D^{te}_{\tau\mu}\}^M_{\mu=1} \subset \mathcal{D}^{te}_\tau$. Taking the average over the loss for each sample, MAML minimizes the following loss

$$\tilde{L}(\boldsymbol{\theta}) \coloneqq \frac{1}{E} \sum_{\tau}^{E} \frac{1}{M} \sum_{\mu}^{M} \ell\left(D^{te}_{\tau\mu}, \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \frac{1}{K} \sum_{\nu}^{K} \ell\left(D^{tr}_{\tau\nu}, \boldsymbol{\theta}\right)\right), \tag{4}$$

by gradient descent

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta \nabla_{\boldsymbol{\theta}} \tilde{L}(\boldsymbol{\theta}), \tag{5}$$

where E is the number of environments, $\tilde{L}(\boldsymbol{\theta})$ is the loss that MAML minimizes in the outer-loop, and β is the learning rate called the meta-learning rate. This whole process is called the meta-training. After the meta-training, the model is checked if it can fine-tune to the new environment with few steps; this process is called the meta-test.

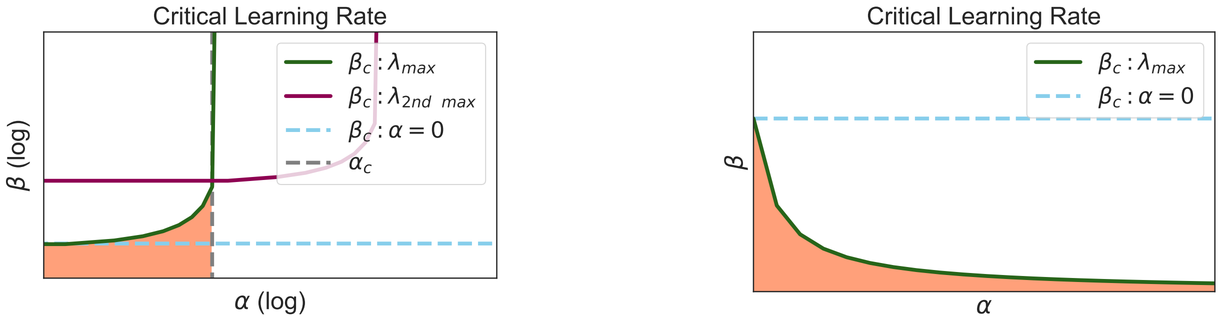
B. Model-agnostic Meta-learning Virtually Minimizes Negative Gradient Penalty Loss

To observe the correspondence between MAML and negative gradient penalty loss minimization, we take the Taylor series of the MAML loss (4) for the first-order term of inner learning rate α. Then, we obtain

$$\tilde{L}(\boldsymbol{\theta}) = \frac{1}{E} \sum_{\tau}^{E} L^{te}_\tau(\boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} L^{tr}_\tau(\boldsymbol{\theta})) \tag{6}$$

$$\approx \frac{1}{E} \sum_{\tau}^{E} L^{te}_\tau(\boldsymbol{\theta}) - \alpha \nabla_{\boldsymbol{\theta}} L^{te}_\tau(\boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} L^{tr}_\tau(\boldsymbol{\theta}) \tag{7}$$

$$= \frac{1}{E} \sum_{\tau}^{E} L^{te}_\tau(\boldsymbol{\theta}) - \alpha \left\| \nabla_{\boldsymbol{\theta}} L^{te}_\tau(\boldsymbol{\theta}) \right\| \left\| \nabla_{\boldsymbol{\theta}} L^{tr}_\tau(\boldsymbol{\theta}) \right\| \cos(a_\tau), \tag{8}$$



(a) Negative Gradient Penalty

(b) Positive Gradient Penalty

Fig. 1: Diagram of critical regularization coefficient α_c of α and critical learning rate β_c of β with respect to gradient descent of (a) negative gradient penalty loss $L(\boldsymbol{\theta}) - \frac{\alpha}{2} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ and (b) positive gradient penalty loss $L(\boldsymbol{\theta}) + \frac{\alpha}{2} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. The solid curves are β_c for the maximum and second maximum eigenvalues λ_{max}, $\lambda_{\text{2nd max}}$ of $L(\boldsymbol{\theta})$'s Hessian, respectively. The dashed vertical and horizontal lines indicate α_c and β_c when $\alpha = 0$. Parameters α and β should be in the colored area, where α is smaller than α_c and β is below both β_c curves. For (a) negative gradient penalty loss, we find that β_c of α close to α_c is larger than β_c of $\alpha = 0$. This suggests that α close to α_c allows larger β_c (Section VII). On the other hand, for (b) positive gradient penalty loss, we find that critical learning rate β_c quickly decays as α increases (Section VIII).

where $L_\tau^{tr}(\boldsymbol{\theta}) := \frac{1}{K} \sum_\nu^K \ell(D_{\tau\nu}^{tr}, \boldsymbol{\theta})$, $L_\tau^{te}(\boldsymbol{\theta}) := \frac{1}{M} \sum_\mu^M \ell\left(D_{\tau\mu}^{te}, \boldsymbol{\theta}\right)$, and a_τ is the angle between $\nabla_{\boldsymbol{\theta}} L_\tau^{te}(\boldsymbol{\theta})$ and $\nabla_{\boldsymbol{\theta}} L_\tau^{tr}(\boldsymbol{\theta})$. That is, MAML has a bias to increase the inner product of the gradient vectors. Once $\cos(a_\tau)$ becomes negative, the penalty in (8) will be positive, drastically increasing the total loss. Thus, it is fair to say that MAML has a bias to make the cosine similarity positive. In Section IX-B, we will explain that our simulation confirms that MAML keeps cosine similarity positive in practice. Therefore, we can conclude that MAML in effect minimizes the negative gradient penalty loss. Specifically, if $L(\boldsymbol{\theta}) = \frac{1}{E} \sum_\tau^E L_\tau^{te}(\boldsymbol{\theta})$ and $s \cdot \|\nabla_{\boldsymbol{\theta}} L_\tau^{tr}(\boldsymbol{\theta})\| \cos(a_\tau) = \|\nabla_{\boldsymbol{\theta}} L_\tau^{te}(\boldsymbol{\theta})\|$ are taken with positive scalar s, (8) corresponds to (15) up to scale. When you consider the full-batch training, (8) and (15) are identical.

The above conclusion suggests that MAML prefers the larger gradient norm. In Section IX-B, we will explain that MAML does in fact keep the gradient norm larger during meta-training in our simulation. Since the solution with the large gradient is favorable for quick loss minimization by gradient descent, this bias may explain why MAML is successful for few-shot adaptation. In fact, another simulation shows that MAML finds the solution with the large gradient even in meta-test, as will be explained in Section IX-B.

IV. Relationship between Invariant Risk Minimization and Positive Gradient Penalty

A. Background

Empirical Risk Minimization lets machines exploit spurious correlation and fails to learn the pattern that they truly want to learn. Based on the hypothesis that the spurious correlation is unique to each environment, while the important correlations are environment-independent, IRM aims to learn the representation that induces the environment-invariant optimal predictor. Slightly abusing the notation of $L_\tau(\boldsymbol{\theta})$ to also mean $L_\tau(f_{\boldsymbol{\theta}})$, this problem is formulated as follows:

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H} \\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{\tau\in\mathcal{E}} L_\tau(w \circ \Phi) \tag{9}$$

$$\text{subject to } w \in \underset{\bar{w}:\mathcal{H}\to\mathcal{Y}}{\arg\min} L_\tau(\bar{w} \circ \Phi), \text{ for all } \tau \in \mathcal{E}, \tag{10}$$

where $\mathcal{X}$ is the input domain, $\mathcal{H}$ is the hidden space, $\mathcal{Y}$ is the output domain, and $\mathcal{E}$ is a set of environments. We note that Φ is the feature extractor that maps the data to the hidden representation, w is the predictor on top of the representation, and $\bar{w}$ is the optimal predictor.

B. Relation between Invariant Risk Minimization and Positive Gradient Penalty

The optimization problem of (9) and (10) is difficult to solve because of its bi-level structure. Therefore, IRMv1 was proposed as an alternative heuristic problem formulation of IRM [1]. IRMv1 rewrites the bi-level optimization into a single-level optimization of a regularized loss:

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{\tau\in\mathcal{E}} L_\tau(\Phi) + \alpha \left\|\nabla_{w|w=1.0} L_\tau(w \circ \Phi)\right\|^2. \tag{11}$$

In this paper, we will refer to the loss to be minimized in (11) as the IRMv1 loss.

Although there are some differences, it is observed that IRMv1 loss is roughly associated with the positive gradient penalty loss if you take $L(\boldsymbol{\theta}) = \sum_{\tau\in\mathcal{E}} L_\tau(\Phi)$. Specifically, in (11), if the gradient is taken with respect to $\boldsymbol{\theta}$, and minimization and loss computation is performed for not only Φ but also for $\boldsymbol{\theta}$, IRMv1 loss corresponds to (2). Thus, we conjecture that the analysis of the trainability of the positive gradient penalty is helpful for the study of the trainability of IRMv1.

V. Relation between Model-agnostic Meta-learning and Invariant Risk Minimization from Gradient Penalty Perspective

In Section III and IV, we explained that MAML and IRM are related to the gradient penalty losses of different signs. Prior to the analysis of trainability, we investigate the connection between MAML and IRM in greater depth.

The problem that MAML solves is formulated as follows:

$$\min_{\boldsymbol{\theta}} \min_{\|d_{\tau \in \mathcal{E}}\| \leq \epsilon} L_\tau(\boldsymbol{\theta} + d_\tau), \tag{12}$$

where d_τ is the p-dimensional vector and ϵ is an infinitesimal value. Usually, few-step gradient descent is used to execute the inner minimization, as explained in Section III. On the other hand, IRM solves the out-of-distribution (OOD) problem defined as follows [5], [6]:

$$\min_{\boldsymbol{\theta}} \max_{\tau \in \mathcal{E}} L_\tau(\boldsymbol{\theta}). \tag{13}$$

When the environment-dependent adjustment d_τ is allowed for (13) and the inner maximization is approximated by the maximization with respect to d_τ, (13) becomes the following problem:

$$\min_{\boldsymbol{\theta}} \max_{\|d_{\tau \in \mathcal{E}}\| \leq \epsilon} L_\tau(\boldsymbol{\theta} + d_\tau). \tag{14}$$

Thus, if we perform the inner maximization by the single-step gradient ascent, we can rewrite (14) as loss minimization with the positive gradient penalty, in the similar manner to Section III. That is, we can interpret the flip in the sign of the penalty between MAML and IRM as the difference in the minimization and maximization in the inner optimization.

This formal similarity is not only a coincidence. Because the goal of MAML is adaptation to each environment, MAML requires the minimization of the loss for each environment. By contrast, IRM seeks to exclude the environment-specific information as much as possible, and it considers the environmentwise loss maximization. That is, MAML and IRM differ in the sign of the penalty term because of the difference in their requirements for each environment. Although this paper focuses on the trainability of MAML and IRM, this formal similarity is likely to highlight the hidden relationships between the various multi-environment learning algorithms. We would like to address these questions in future work.

VI. Related Works

A. Model-agnostic Meta-learning

Meta-learning algorithms can be categorized into three subcategories: black box/model-based [7], [8], metric learning based [9], [10], and optimization-based [2], [11], [12]. MAML is a well-known optimization-based algorithm and has recently been extensively studied. For example, MAML is used for continual learning [13], [14], reinforcement learning [2], [15], [16] and probabilistic inference [17], [18].

Since MAML is known for being difficult to train [3], several papers studied the trainability of MAML. Some proposed heuristics to perform better training [3], [4], while the others provide the mathematical proofs of the convergence condition of MAML and its variants [13], [19]–[23]. In contrast to these studies, the goal of our paper is to provide a conceptual insight on how the inner learning rate influences the maximum possible meta-learning rate in practice. Specifically, we describe this critical meta-learning rate as a function of the inner learning rate, which has not been done in previous studies.

B. Invariant Risk Minimization

IRM is known to deal with the out-of-distribution generalization problem [5], [6] and can be formulated as game [24]. There are several seminal subsequent works that proposed the new algorithms or studied the property of IRM [5], [6], [24]–[29]. Among them, a few works studied a topic related to the trainability of IRM. Some works discussed the condition for IRM to successfully work [5], [29] and a work studied the convergence of the IRM game [25]. Our study differs from these studies in that we focus on the trainability of solving IRMv1 by gradient descent and elucidate how the regularization coefficient affects the acceptable learning rate for the gradient descent to work. Therefore, our paper contributes to the first step for the study of this research direction.

VII. Trainability on Negative Gradient Penalty Loss and Model-Agnositc Meta-Learning

In this section and the next section, we will consider the condition to satisfy such that the gradient descent of the gradient penalty loss find local minima. Specifically, we will discuss the conditions for the learning rate and regularization coefficient that should be satisfied for steepest gradient descent to locally converge to a local minimum from any point in the vicinity of it. In this section, we will examine the gradient descent for the negative gradient penalty loss to study the trainability of MAML.

A. Condition That α Should Satisfy for a Fixed Point to be a Local Minimum

To analyze the behavior around a local minimum, we first consider when a fixed point of negative gradient penalty loss will be a local minimum. Taking the Taylor series for the second-order term at a fixed point $\boldsymbol{\theta}^*$, the negative gradient penalty loss around the fixed point is given by

$$L_-(\boldsymbol{\theta}) \approx L_-(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top H_-(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \tag{15}$$

where $H_- = H - \alpha\left(T\boldsymbol{g} + H^2\right)$ is the Hessian matrix of $L_-(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$. Note that $\boldsymbol{g} = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*) \in \mathbb{R}^d$, $H = \nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*) \in \mathbb{R}^{d \times d}$, and $T = \nabla^3_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*) \in \mathbb{R}^{d \times d \times d}$. Here, we presume that $T\boldsymbol{g}$ is negligible compared to H^2 and will thus assume that $H_- = H - \alpha H^2$. Considering the purpose of studying the trainability of MAML, this is justified because we confirm that the empirical $T\boldsymbol{g}$ is much smaller than the empirical H^2 of MAML, which we will explain in Section IX-C.

With a diagonal matrix Λ_{H_-} for which the entries are the eigenvalues of H_- and a matrix P the rows of which the are eigenvectors of H_-, $P\Lambda_{H_-}P^\top = P[\Lambda_H - \alpha\Lambda_H^2]P^\top$. The necessary condition for $\boldsymbol{\theta}^*$ to be a local minimum is that

H_- is positive semi-definite, or all of its eigenvalues are non-negative. Therefore, the condition of α for $\boldsymbol{\theta}^*$ to be a local minimum is

$$\forall i, \quad \lambda(H_-)_i = \lambda(H)_i - \alpha\lambda(H)_i^2 \geq 0 \tag{16}$$

$$\Rightarrow \forall i, \quad \alpha \leq \frac{1}{\lambda(H)_i} \quad , (\lambda(H)_i \neq 0) \quad \text{or} \quad \lambda(H)_i = 0, \tag{17}$$

where $\lambda(A)_i$ is the ith eigenvalue of a matrix A. Defining $1/0$ to be ∞, the condition for $\boldsymbol{\theta}^*$ to be a local minimum is

$$\forall i, \quad \alpha \leq \frac{1}{\lambda(H)_i}. \tag{18}$$

Hence, the maximum bound, or critical regularization coefficient α_c of α is the inverse of the largest eigenvalue of H.

B. Condition That β Should Satisfy for Gradient Descent to Locally Converge to the Local Minimum

Next, we will consider how large we can set the learning rate β for the optimizer to reach the local minimum from the vicinity of it, which is an extension of [30]. Since $PP^\top = I$, the negative gradient penalty loss can be written as

$$L_-(\boldsymbol{\theta}) \approx L_-(\boldsymbol{\theta}^*) + \frac{1}{2}((\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top P)P^\top H_- P(P^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)). \tag{19}$$

Therefore, the update equation of the parameter $\boldsymbol{\theta}$ with gradient descent is

$$\boldsymbol{\theta}(t+1) - \boldsymbol{\theta}(t) = -\beta\nabla_{\boldsymbol{\theta}} L_-(\boldsymbol{\theta}) = -\beta H_-(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \tag{20}$$

where t is the number of iterations. Hence, $\boldsymbol{\theta}(t+1) - \boldsymbol{\theta}^* = (I - \beta H_-)(\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*)$ and the negative gradient penalty loss is $L_-(\boldsymbol{v}) \approx L_-(0) + \frac{1}{2}\boldsymbol{v}^\top \Lambda_{H_-}\boldsymbol{v}$ by denoting $P^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ by $\boldsymbol{v}$. Because the gradient of $L_-(\boldsymbol{v})$ for $\boldsymbol{v}$ is $\nabla_{\boldsymbol{v}} L_-(\boldsymbol{v}) = \Lambda_{H_-}\boldsymbol{v}$, the equation for the updating of $\boldsymbol{v}$ is

$$\boldsymbol{v}(t+1) = \boldsymbol{v}(t) - \beta\Lambda_{H_-}\boldsymbol{v}(t) = (I - \beta\Lambda_{H_-})\boldsymbol{v}(t), \tag{21}$$

where $\boldsymbol{v}(t)$ is the value of $\boldsymbol{v}$ at iteration t. Assuming that (28) holds, the condition of β is as follows: for all i,

$$|1 - \beta\lambda(H - \alpha H^2)_i| = |1 - \beta(\lambda(H)_i - \alpha\lambda(H)_i^2)| < 1 \tag{22}$$

$$\Rightarrow -1 + \beta(\lambda(H)_i - \alpha\lambda(H)_i^2) < 1 \tag{23}$$

$$(\because \lambda(H)_i - \alpha\lambda(H)_i^2 \geq 0 \text{ holds because of (28)}) \tag{24}$$

$$\Rightarrow \beta < \frac{2}{\lambda(H)_i - \alpha\lambda(H)_i^2}. \tag{25}$$

C. β_c Becomes Large When α is Close to α_c

In Sections VII-A and VII-B, we discussed the conditions of α and β. Consequently, the condition for locally converging to local minima is given by

$$\forall i, \quad \alpha \leq \frac{1}{\lambda(H)_i} \quad \wedge \quad \beta \leq \frac{2}{\lambda(H)_i - \alpha\lambda(H)_i^2}. \tag{26}$$

If $\alpha = 0$, $\beta \leq \frac{2}{\lambda_{\text{max}}}$ is the condition that β must satisfy, where λ_{max} is the largest eigenvalue of H. However, when α is close to its upper bound α_c, the upper bound β_c of β becomes larger than that with $\alpha = 0$. This is because $\frac{2}{\lambda(H)_i - \alpha\lambda(H)_i^2}$ goes to infinity as α approaches $\frac{1}{\lambda(H)_i}$ for each i, and (26) should hold for all i. That is, β_c of the negative gradient penalty loss is larger than that of $\alpha = 0$ when α is close to α_c, and the bound is determined by both the maximum and second-maximum eigenvalues of the Hessian. The diagram of this relation between α and β_c is shown in Fig. 1 (a). The x-axis and y-axis indicate α and β, and the curves correspond to $\frac{2}{\lambda(H)_i - \alpha\lambda(H)_i^2}$ of the largest and the second largest $\lambda(H)_i$, respectively. Considering that α and β correspond to the inner learning rate and the meta-learning rate for MAML, we conjecture that MAML can take larger critical meta-learning rate when the inner learning rate is close to its upper bound.

VIII. Trainability on Positive Gradient Penalty Loss and Invariant Risk Minimization

In this section, we will derive the condition for the gradient descent of the positive gradient penalty loss to find local minima for studying the trainability of IRM. The procedure for the analysis is basically the same as that in Section VII.

A. Condition That α Should Satisfy for a Fixed Point to be a Local Minimum

The positive gradient penalty loss at a fixed point $\boldsymbol{\theta}^*$ is

$$L_+(\boldsymbol{\theta}) \approx L_+(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top H_+(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \tag{27}$$

where $H_+ = H + \alpha\left(H^2 + T\boldsymbol{g}\right)$. Similar to Section VII-A, we will ignore $T\boldsymbol{g}$. Then, similar calculus reveals the condition that α should satisfy:

$$\forall i, \quad \lambda(H_+)_i = \lambda(H)_i + \alpha\lambda(H)_i^2 \geq 0. \tag{28}$$

If $\lambda(H)_i$ is positive, (28) always holds since α is positive by definition. If $\lambda(H)_i$ is negative, (28) will be $\alpha \leq -\frac{1}{\lambda(H)_i}$ for all i. Yet, it is known that negative eigenvalues of Hessian in neural network training are scarce and very small [31], [32]. Thus, this is not likely to affect the optimization in practice.

B. Condition That β Should Satisfy for Gradient Descent to Locally Converge to the Local Minimum

We follow the similar procedure to that in Section VII-B again. Then, the condition for β is as follows: for all i,

$$|1 - \beta\lambda(H + \alpha H^2)_i| = |1 - \beta(\lambda(H)_i + \alpha\lambda(H)_i^2)| < 1 \tag{29}$$

$$\Rightarrow \beta < \frac{2}{\lambda(H)_i + \alpha\lambda(H)_i^2}. \tag{30}$$

C. β_c Decreases Quickly as α Increases

Summarizing the condition of α and β, the condition to locally converge to local minima is given by

$$\forall\lambda(H)_i \leq 0, \alpha \leq -\frac{1}{\lambda(H)_i} \wedge \forall i, \beta \leq \frac{2}{\lambda(H)_i + \alpha\lambda(H)_i^2}. \tag{31}$$

As explained above, negative eigenvalues are not likely to affect the trainability in practice and thus we only consider the positive eigenvalues. The diagram of the critical learning rate is shown in Fig. 1 (b). The x-axis and y-axis indicate α and β, and the curve corresponds to $\frac{2}{\lambda(H)_i - \alpha\lambda(H)_i^2}$ of the largest $\lambda(H)_i$. It is observed from Fig. 1 (b) that the critical learning rate β_c decreases with increasing regularization coefficient α, in contrast to the case for negative gradient penalty. Therefore, we expect that also for IRMv1, the critical learning rate decreases rapidly as regularization coefficient increases.

IX. Numerical Simulation

A. *Setup for Model-agnositc Meta-learning*

To empirically validate our findings in previous sections, we conducted simulations with benchmark tasks and practical model architectures. Specifically, we performed Omniglot and Mini-ImageNet classification, and sinusoid regression for MAML [2], [8], [33]. The setup of each experiment is described below. Unless otherwise noted, all of the experiments about MAML in the following sections follow this setup.

1) Sinusoid Regression: The task in each environment is to regress a sine wave with an amplitude in the range of $[0.1, 5.0]$ and phase in the range of $[0, \pi]$ based on the data points in the range of $[-5.0, 5.0]$. A ReLU multilayer perceptron with two hidden layers of size 40 was trained with SGD. The batch size of data is 10, the number of environments is 100, and one step is taken for the update in the inner-loop.

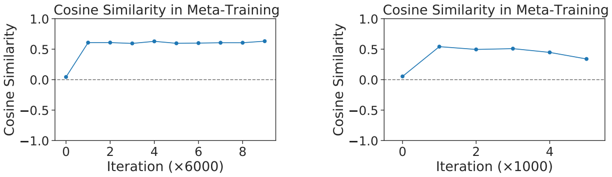
2) Omniglot and Mini-ImageNet Classification: The Omniglot and Mini-ImageNet datasets are benchmark datasets for few-shot classification. The model used is the same as that in [2], and hence, [9] used. The task is a five-way one-shot classification, where the query size is 15, the number of update steps is one, and the batch size of the environments is 32 for Omniglot and four for Mini-ImageNet. In this setup, we trained 60000 iterations for the Omniglot and 12 epochs (10000 iterations per epoch) for Mini-ImageNet. SGD is used as the optimizer for both the inner-loop and the outer-loop.

B. *Cosine Similarity and Gradient Norm of Model-agnostic Meta-learning*

We will check if three statements in Section III hold: cosine similarity remains positive in the meta-training, gradient norm is kept large in the meta-training, and MAML finds the solution at which the gradient in the meta-test is large. To that end, we conducted Omniglot and Mini-ImageNet classification.

1) Meta-training: We computed the gradient norm with the test data and the cosine similarity between the gradient with the training and the test data every 6000 iteration for Omniglot and 10000 iteration for Mini-ImageNet. Specifically, we computed the cosine similarity and the gradient norm per environment, and subsequently the average is computed, respectively. The parameters α and β are the following values; for Omniglot, $\alpha = 4e - 1$ and $\beta = 1e - 3$, and for Mini-ImageNet, $\alpha = 1e - 2$ and $\beta = 1e - 3$. We used the Adam optimizer for the outer-loop training with these β [34].

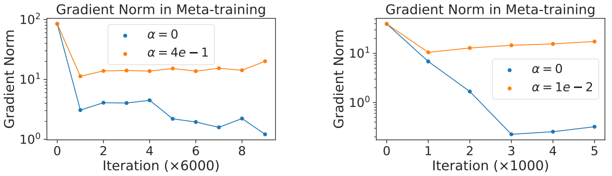
The results for the cosine similarity of (a) Omniglot and (b) Mini-ImageNet are shown in Fig. 2. Here, the x-axis is the number of iterations and the y-axis is the cosine similarity. The dashed line indicates the cosine similarity equal to zero. It is clear that cosine similarity remains positive throughout the meta-training, validating our conjecture in Section II.



(a) Omniglot

(b) Mini-ImageNet

Fig. 2: Cosine similarity between the gradient of the training loss and test loss for (a) Omniglot and (b) Mini-ImageNet. Cosine similarity remains positive during meta-training, supporting the statement in III.



(a) Omniglot

(b) Mini-ImageNet

Fig. 3: Gradient norm of MAML ($\alpha > 0$) compared to that of $\alpha = 0$ for (a) Omniglot and (b) Mini-ImageNet. Gradient norm is larger for MAML throughout meta-training, confirming that MAML has a bias toward large gradient.

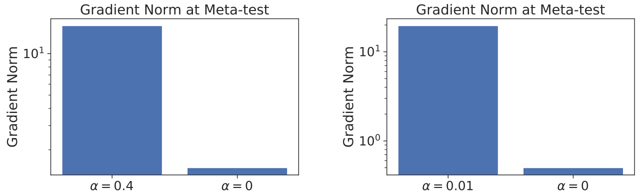
Fig. 3 shows the gradient norm for (a) Omniglot and (b) Mini-ImageNet in the meta-training. We compared the gradient norm of MAML ($\alpha > 0$) with that of $\alpha = 0$. Here, the x-axis is the number of iterations and the y-axis is the gradient norm. The blue and orange lines are the gradient norms of $\alpha = 0$ and MAML ($\alpha = 4e - 1$). We confirm that MAML retains a greater gradient norm than the case of $\alpha = 0$.

2) Meta-test: After the meta-training, we computed the gradient norm with meta-test data for Omniglot and Mini-ImageNet. The batch size of the environment is one and we computed the averaged gradient norm over 992 runs for Omniglot and 100 runs for Mini-ImageNet. The results are shown in Fig. 4 for (a) Omniglot and (b) Mini-ImageNet, where the y-axis is the gradient norm. It is observed that the gradient norm of MAML ($\alpha > 0$) is larger than that of $\alpha = 0$ even during the meta-test.

C. *Magnitude of $T\boldsymbol{g}$ and H^2 of Model-agnostic Meta-learning*

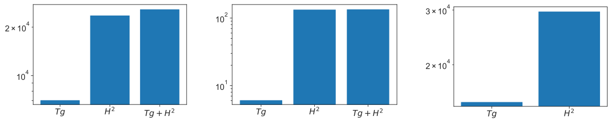
To validate ignoring $T\boldsymbol{g}$ in Section VII-A, we empirically calculated $T\boldsymbol{g}$ and H^2 of MAML and compared them. To that end, we conducted sinusoid regression in the setup of Section IX-A because the calculation of $T\boldsymbol{g}$ and H^2 is computationally expensive for architectures used in Omniglot and MiniImageNet classification. Concretely, we used the training error after 50000 iterations to compute them.

To compare $T\boldsymbol{g}$ and H^2, we calculated the top and the second top eigenvalues and Frobenius norm of $T\boldsymbol{g}$, H^2 and



(a) Omniglot (b) Mini-ImageNet

Fig. 4: Comparison of gradient norm between MAML ($\alpha > 0$) and that of $\alpha = 0$ at meta-test. Both the results of (a) Omniglot and (b) Mini-ImageNet show that gradient norm is large for MAML, suggesting that the solution MAML finds has larger gradient even for meta-test loss.



(a) Top eigenvalue (b) 2nd top eigenvalue (c) Frobenius norm

Fig. 5: Top (a) and second top (b) eigenvalue of $T\boldsymbol{g}$, H^2 and $T\boldsymbol{g}+H^2$, and Frobenius norm of $T\boldsymbol{g}$ and H^2 (c). The top and the second top eigenvalue of $T\boldsymbol{g} + H^2$ is almost the same as that of H^2, while that of $T\boldsymbol{g}$ is much smaller. Frobenius norm of $T\boldsymbol{g}$ is much smaller than that of H^2.

$T\boldsymbol{g} + H^2$. Frobenius norm is a common measure to compare two matrices. The reason why we calculated eigenvalues is that large eigenvalues are important for upper bound, as explained in Section VII-C. The results of the top eigenvalue, the second top eigenvalue, and Frobenius norm are shown in Figs. 5 (a), (b), and (c), respectively. Here, the y-axis indicates the magnitude of $T\boldsymbol{g}$, H^2, and $T\boldsymbol{g} + H^2$ that are indicated in x-axis. In all cases, magnitude for $T\boldsymbol{g}$ is much smaller than that of H^2, supporting the assumption that we can ignore $T\boldsymbol{g}$.

D. Large inner learning Rate Allows Large Critical Meta-Learning Rate for Model-agnostic Meta-learning

In this section, we will present the results of experiments to confirm that MAML allows a larger meta-learning rate β if the inner learning rate α is close to its upper bound. For that purpose, we computed the training loss after fixed iterations for various set of α and β. This experiment finds the critical learning rates above which the training loss diverges.

1) Sinusoid Regression: We computed the training loss after 500 iterations with α in the range of $[1e-4, 9e-1]$ and β in the range of $[1e-2, 9e-0]$. Fig. 6 (a) shows the training losses with various values of α and β. The dashed line indicates β of $\alpha = 0$ over which the loss diverges. According to Fig. 6 (a), if α is close to the value above which the losses diverge, a larger β can be used. This result confirms that MAML allows larger β if α is close to its critical value α_c.

2) Classification: We computed the training losses after 100 iterations for Omniglot and one epoch for Mini-ImageNet with various values of α and β; for Omniglot, α is in the range of $[1e-3, 9e-0]$ and β is in the range of $[1e-1, 9e+1]$, and for Mini-ImageNet, α is in the range of $[1e-4, 9e-1]$ and β is in the range of $[1e-2, 9e-0]$. Figs. 6 (b) and (c) show the training losses at various values of α and β. The dashed line indicates β of $\alpha = 0$ above which the training loss diverges. As shown in Figs. 6 (b) and (c), the maximum β is larger at large α. Even though the architecture is a convolutional neural network with batch normalization [34] and practical dataset is used, our expectation in Sections VII and VIII is confirmed.

E. Invariant Risk Minimization's Large Regularization Coefficient Does Not Allow Large Critical Learning Rate

We will explain the result of the experiment to validate that the critical learning rate β_c decays quickly as the regularization coefficient α of IRMv1 increases. We performed colored MNIST classification, a bench mark task of IRM, by IRMv1 [1]. We employed a ReLU fully-connected neural network with one hidden layer as the feature extractor and fixed the classifier to 1: this is the architecture used in [1]. For the feature extractor, the width of hidden layers is 256, the input dimension is 14 × 14, and the output dimension is one. We computed the averaged training loss for 10 runs for various sets of α and β: α was in the range of $[1e-2, 9e+2]$ and β was in the range of $[1e-2, 9e+1]$. Each run is a full-batch gradient descent training of 500 iterations. To focus on the pure effect of IRMv1, we discarded all of the tricks used in [1], namely penalty annealing, penalty weight scaling, and weight decay. Fig. 6 (d) shows the training losses at the various values of α and β. The dashed line indicates β of $\alpha = 0$ above which the training loss diverges. As shown in Fig. 6 (d) the maximum β quickly decreases with increasing α. This result supports our conjecture that critical learning rate decreases quickly as the regularization coefficient of IRMv1 gets larger.

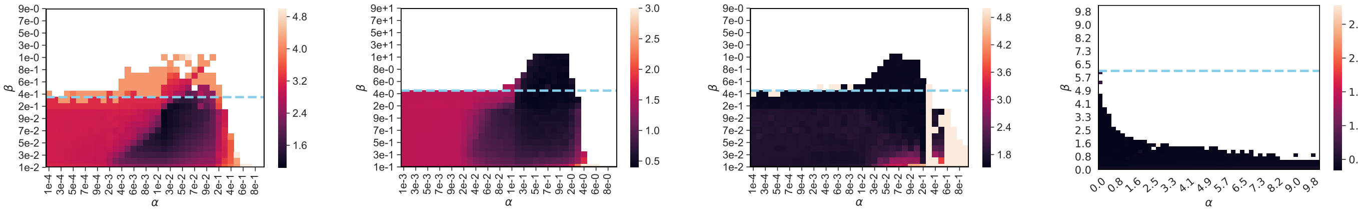
X. Conclusion

We studied the trainability of MAML and IRM by analyzing the trainability of the gradient descent of the loss with the gradient penalty. We explained that MAML and IRM are related to loss minimization with negative and positive gradient penalties, respectively, and that this relationship reflects the requirement of MAML and IRM for each environment.

For negative and positive penalty losses, we investigated the acceptable learning rate and regularization coefficient for gradient descent to locally reach a local minimum. We found that the critical learning rate for the negative gradient penalty becomes larger when the regularization coefficient is close to its critical value, while that for its positive counterpart decreases rapidly as the coefficient increases. Our experiment supports all of the findings empirically with the practical dataset and architecture for both MAML and IRM.

References

[1] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.



(a) MAML / Sinusoid

(b) MAML / Omniglot

(c) MAML / Mini-ImageNet

(d) IRMv1 / Colored MNIST

Fig. 6: Training losses of MAML for (a) sinusoid regression, (b) Omniglot classification, and (c) Mini-ImageNet classification, and those of IRMv1 for (d) colored MNIST classification at various values of α and β after a fixed number of iterations. The parameter α is the inner learning rate for MAML and regularization coefficient for IRMv1, and β is the meta-learning rate for MAML and the learning rate for IRMv1. The area with no color represents the diverged losses, and the dashed line indicates the values of β above which the loss diverges for $\alpha = 0$. For MAML ((a), (b), and (c)), the maximum possible β is larger than that at $\alpha = 0$ if α is close to the value above which the losses diverge, which is similar to Fig. 1 (a). On the other hand, for IRMv1 (d), the maximum possible β decays quickly as α increases, resembling Fig. 1 (b). These results are consistent with the relationship between α and critical β predicted in Sections VII and VIII

[2] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International Conference on Machine Learning (ICML)*, 2017.

[3] A. Antoniou, H. Edwards, and A. Storkey, “How to train your maml,” *International Conference on Learning Representations (ICLR)*, 2019.

[4] H. S. Behl, A. G. Baydin, and P. H. Torr, “Alpha maml: Adaptive model-agnostic meta-learning,” *International Conference on Machine Learning (ICML)*, 2019.

[5] K. Ahuja, J. Wang, A. Dhurandhar, K. Shanmugam, and K. R. Varshney, “Empirical or invariant risk minimization? a sample complexity perspective,” *arXiv preprint arXiv:2010.16412*, 2020.

[6] M. Koyama and S. Yamaguchi, “Out-of-distribution generalization with maximal invariant predictor,” *arXiv preprint arXiv:2008.01883*, 2020.

[7] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, “Meta-Learning with Memory-Augmented Neural Networks,” *arXiv preprint arXiv:1902.10644*, 2019.

[8] S. Ravi and H. Larochelle, “Optimization as a Model for Few-shot Learning,” in *International Conference on Learning Representations (ICLR)*, 2017.

[9] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, “Matching networks for one shot learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[10] J. Snell, K. Swersky, and R. Zemel, “Prototypical Networks for Few-shot Learning,” in *Neural Information Processing Systems (NeurIPS)*, 2017.

[11] Z. Li, F. Zhou, S. Chen, and H. Li, “Meta-SGD: Learning to Learn Quickly for Few-Shot Learning,” *arXiv preprint arXiv:1707.09835*, 2017.

[12] K. Li and J. Malik, “Learning To Optimize,” in *International Conference on Learning Representations (ICLR)*, 2017.

[13] C. Finn, A. Rajeswaran, S. Kakade, and S. Levine, “Online Meta-Learning,” in *International Conference on Machine Learning (ICML)*, 2019.

[14] G. Jerfel, E. Grant, T. L. Griffiths, and K. Heller, “Reconciling Meta-learning and Continual Learning with Online Mixtures of Tasks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[15] M. Al-Shedivat, T. Bansal, Y. Burda, I. Sutskever, and I. Mordatch, “Continuous Adaptation Via Meta-Learning In Nonstationary And Competetive Environments,” in *International Conference on Learning Representations*, 2018.

[16] A. Gupta, B. Eysenbach, C. Finn, and S. Levine, “Unsupervised Meta-Learning for Reinforcement Learning,” *arXiv:1806.04640*, 2018.

[17] C. Finn, K. Xu, and S. Levine, “Probabilistic Model-Agnostic Meta-Learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[18] E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths, “Recasting Gradient-Based Meta-Learning as Hierarchical Bayes,” in *International Conference on Learning Representations (ICLR)*, 2018.

[19] A. Fallah, A. Mokhtariy, and A. Ozdaglar, “On the Convergence Theory of Gradient-Based Model-Agnostic Meta-Learning Algorithms,” in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[20] K. Ji, J. D. Lee, Y. Liang, and H. V. Poor, “Convergence of meta-learning with task-specific adaptation over partial parameters,” *arXiv preprint arXiv:2006.09486*, 2020.

[21] K. Ji, J. Yang, and Y. Liang, “Theoretical convergence of multi-step model-agnostic meta-learning,” *arXiv preprint arXiv:2002.07836*, 2020.

[22] H. Wang, R. Sun, , and B. Li, “Global convergence and induced kernels of gradientbased meta-learning with neural nets,” *arXiv preprint arXiv:2006.14606*, 2020.

[23] L. Wang, Q. Cai, Z. Yang, and Z. Wang, “On the global optimality of model-agnostic meta-learning,” in *International conference on machine learning*, 2020.

[24] K. Ahuja, K. Shanmugam, K. R. Varshney, and A. Dhurandhar, “Invariant risk minimization games,” *arXiv preprint arXiv:2002.04692*, 2020.

[25] K. Ahuja, K. Shanmugam, and A. Dhurandhar, “Linear regression games: Convergence guarantees to approximate out-of-distribution solutions,” *arXiv preprint arXiv:2010.15234*, 2020.

[26] S. Schneider, S. Krishna, L. Eck, W. Brendel, M. W. Mathis, and M. Bethge, “Generalized invariant risk minimization: Relating adaptation and invariant representation learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[27] Y. J. Choe, J. Ham, and K. Park, “An empirical study of invariant risk minimization,” in *International Conference on Machine Learning (ICLR)*, 2020.

[28] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. Courville, “Out-of-distribution generalization via risk extrapolation,” *arXiv preprint arXiv:2003.00688*, 2020.

[29] D. Rosenfeld, P. Ravikumar, and A. Risteski, “The risks of invariant risk minimization,” *arXiv preprint arXiv:2010.05761*, 2020.

[30] Y. LeCun, L. Bottou, B. G. Orr, and K.-R. Müller, “Efficient backprop,” *Neural networks: Tricks of the trade*, pp. 9 – 50, 1998.

[31] Y. L. Levent Sagun, Leon Bottou, “Eigenvalues of the hessian in deep learning: Singularity and beyond,” *arXiv:1611.07476*, 2016.

[32] B. Ghorbani, S. Krishnan, and Y. Xiao, “An investigation into neural net optimization via hessian eigenvalue density,” in *International Conference on Machine Learning (ICML)*, 2019.

[33] B. M. Lake, R. Salakhutdinov, J. Gross, and J. B. Tenenbaum, “One shot learning of simple visual concepts,” in *Proceedings of the 33th Annual Meeting of the Cognitive Science Society, CogSci 2011, Boston, Massachusetts, USA, July 20-23, 2011*, 2011.

[34] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in *International Conference on Machine Learning (ICML)*, 2015.