1. Introduction

In this supplementary material, we show the concrete mathematical derivation of the Nearest Class Mean (NCM) classifier in iCaRL [15] from a Bayesian perspective (Sec. 3.1), details of Group ImageNet (Sec. 4.1), implementation details of all class incremental learning methods (Sec. 4.1), and more qualitative results of the geometric perspective (Sec. 4.3).

2. Mathematical Derivation

The discriminant function of class $c$ based on minimum-error-rate classification is [4]:

$$g_c(z) = \log p(z|c) + \log p(c)$$

(1)

Assuming that the class conditioned probability $p(z|c)$ is multivariate Gaussian, Eq. 1 can be derived as:

$$g_c(z) = -\frac{1}{2}(z - \mu_c)^T \Sigma_c^{-1}(z - \mu_c) - \frac{d}{2} \log(2\pi)$$

$$-\frac{1}{2} \log |\Sigma_c| + \log p(c)$$

(2)

In Eq. 2, $z$ is the feature and $d$ is its dimension. $\mu_c$ and $\Sigma_c$ are the mean and the covariance matrix of the features for class $c$. Apparently, $\frac{d}{2} \log(2\pi)$ is constant for all classes and can be dropped. Assuming that the prior probability $p(c)$ is equal, the term $\log p(c)$ can be ignored. If the covariance matrix is identical for all classes (i.e. $\Sigma_c = \Sigma$), the term $\frac{1}{2} \log |\Sigma_c|$ can also be ignored. Therefore, Eq. 2 can be simplified as:

$$g_c(z) = -\frac{1}{2}(z - \mu_c)^T \Sigma^{-1}(z - \mu_c)$$

(3)

Eq. 3 is based on the squared Mahalanobis distance. It can be easily found that the quadratic term $z^T \Sigma^{-1}z$ is equal for all classes. Thus, the decision boundary between two classes $i$ and $j$ is linear, which is:

$$(\mu_i - \mu_j)^T \Sigma^{-1}(z - \frac{1}{2}(\mu_i + \mu_j)) = 0$$

(4)

From Eq. 4, it can be observed that the midpoint between two class means $\mu_i$ and $\mu_j$ (i.e. $\frac{1}{2}(\mu_i + \mu_j)$) is on the decision boundary, and it can be easily proved that the distances between the two class means and the decision boundary is equal. Thus, we argue that the assumption of the identical covariance matrix may alleviate the class imbalance since the decision boundary does not bias towards either class.

However, calculating the matrix inverse is time-consuming in Eq. 3, and the overhead can be reduced by assuming the covariance matrix $\Sigma$ to be isotropic (i.e. $\Sigma = \sigma^2 I$). Then, Eq. 3 can be simplified as:

$$g_c(z) = -\frac{1}{2\sigma^2} \|z - \mu_c\|^2_2$$

(5)

By ignoring the constant term $\frac{1}{2\sigma^2}$, the corresponding prediction function is:

$$\arg \max_c g_c(z) = \arg \min_c \|z - \mu_c\|^2_2$$

(6)

Eq. 6 indicates that we can find the closest class mean via Euclidean distance and predict to the corresponding class. The class mean can be estimated by the average feature of exemplars. The more exemplars are, the more accurate the estimated class mean is. The procedure above is exactly what the NCM classifier does. Note that in iCaRL [15], the authors normalize $z$ and the feature mean before predicting the label via Eq. 6. Since the authors attach less importance to this trick, we do not revisit it in this literature.

If the reader is not familiar with the derivations above, he or she could refer to [4] for more information. One open question is that it is more intuitive to define the biasing phenomenon when the decision boundary is linear. However, if the covariance matrix $\Sigma_c$ is arbitrary, the decision boundary

---

1Text in blue indicates sections or graphs in the main paper.
is quadratic and it might be difficult to judge if the classifier is biased or not.

3. Implementation Details

3.1. Dataset

**Group ImageNet.** The 10 superclasses we select are: *fish, dog, monkey, bird, lizard, fruit, vegetable, vessel, vehicle*, and *clothes*. The subclasses in these superclasses are:

- Fish: anemone fish, sturgeon, coho, barracouta, electric ray, puffer, goldfish, tench, lionfish, rock beauty.
- Dog: borzoi, welsh springer spaniel, weimaraner, mexican hairless, blenheim spaniel, saint bernard, samoyed, dalmatian, african hunting dog, leonberg.
- Monkey: siamang, gibbon, gorilla, guenon, chimpanzee, howler monkey, colobus, baboon, orangutan, proboscis monkey.
- Bird: robin, jacamar, african grey, chickadee, bald eagle, macaw, junco, great grey owl, lorikeet, bee eater.
- Lizard: alligator lizard, american chameleon, agama, gila monster, frilled lizard, african chameleon, banded gecko, common iguana, whiptail, komodo dragon.
- Fruit: strawberry, orange, lemon, pomegranate, banana, fig, pineapple, granny smith, custard apple, jackfruit.
- Vegetable: butternut squash, cucumber, artichoke, cauliflower, acorn squash, spaghetti squash, bell pepper, head cabbage, broccoli, cardoon.
- Vessel: submarine, schooner, yawl, container ship, speedboat, liner, lifeboat, fireboat, gondola, canoe.
- Vehicle: streetcar, snowplow, snowmobile, garbage truck, jinrikisha, steam locomotive, ambulance, trolleybus, model T, bullet train.
- Clothes: cardigan, poncho, hoopskirt, pajama, apron, jean, jersey, kimono, abaya, lab coat.

The class list above will be released together with the codes. The corresponding images are down-sampled to $64 \times 64$ according to [3]. Such a down-sampling procedure is also used in [5].

3.2. Method

For CIFAR-100 [10], we adopt LeNet and a modified ResNet-34 [7] that is suitable for a $32 \times 32$ input. For Group ImageNet, we adopt MobileNetV2 [16] and a modified ResNet-18 [7] that accepts an input size of $64 \times 64$. An Adam optimizer [9] is used and the base learning rate is 0.001. The learning rate is divided by 10 at epoch 49 and 63 inspired by the implementation of iCaRL [15]. The weight decay is $10^{-4}$. As for the exemplar selection strategy, we simply use random selection since it does not exhibit appreciable difference of performance with other carefully designed selection strategies as noted by [8, 18, 2].

**Learning without Forgetting (LwF)** [12]. The implementation is similar to the original Learning without Forgetting (LwF) [12] except that an exemplar memory which is similar to iCaRL is used. As for training, there are two losses—softmax cross entropy loss and distillation loss which are calculated on a combination of new samples and old exemplars. The weight of the distillation loss is simply set to 1 and the temperature is set to 2 inspired by [1].

**iCaRL** [15]. The implementation is similar to the original iCaRL [15], and the difference with iCaRL is that a softmax cross entropy loss is employed instead of a binary cross entropy loss, because empirically we find that softmax cross entropy yields better performance.

**End-to-End Incremental Learning (EEIL)** [1]. The difference between EEIL and LwF is that EEIL adds a balanced fine-tuning stage, which is to under-sample equal numbers of samples from each class and fine-tune the network to remove the classifier bias. The number of epochs in the balanced fine-tuning stage is set to 30 which is the same as the setting in the original paper of EEIL.

**Large Scale Incremental Learning (LSIL)** [17]. LSIL differs from LwF in adding a bias correction stage, which is to prevent the classifier from biasing towards new classes. Inspired by the parameters used by the authors, we set the number of epochs in the 2nd stage (i.e. bias correction stage) to be twice the number of epochs in the 1st stage. The weighting factor of the L2 regularization on $\beta$ in the 2nd stage is 0.1, and the ratio of train/validation split on the exemplars is 9:1 as the authors recommend in the paper.

**Maintaining Discrimination and Fairness in Class Incremental Learning (MDFCIL)** [20]. We only implement weight aligning which is the core technique of MDFCIL. Other techniques mentioned in [20] such as weight clipping are not used, since they are general deep learning techniques and are not the main focus of this paper.

**GDumb** [14]. We do not use the techniques mentioned in GDumb [14] such as SGDR [13], cutmix [19] or early stopping with patience, since they are not much related to class imbalance. Also, the greedy balancing sampler is simply implemented by a random under-sampling, because theoretically there is little difference between them.

**ADASYN** [6]. On Group ImageNet, we find that setting sampling strategy to auto gives errors in imblearn [11]. Therefore, we set it to minority and report the corresponding results on Group ImageNet.

**Lowerbound.** No exemplar memory is leveraged. The network is trained on new samples with the vanilla softmax
cross-entropy loss.

**Upperbound.** The exemplar memory has a boundless capacity. The network is trained on all samples (i.e., old and new) with the vanilla softmax cross-entropy loss.

4. More Qualitative Results

In this main paper, we only show the polar chart that depicts the sample distribution of the weight aligning technique on Group ImageNet (Fig. 3). Here, we show the polar chart on CIFAR-100 (Fig. 1). The interpretation of the graph and the conclusion from it is the same as the main paper.

References


Figure 1. Polar charts of the *weight aligning* method in the final class increment on CIFAR-100. The interpretation of this chart is similar to Fig. 3 in the main paper.