A Data-Augmentation Is Worth A Thousand Samples: Analytical Moments And Sampling-Free Training

Randall Balestriero
Meta AI Research, FAIR
NYC, USA
rbalestriero@meta.com

Ishan Misra
Meta AI Research, FAIR
NYC, USA
imisra@meta.com

Yann LeCun
Meta AI Research, FAIR, NYU
NYC, USA
ylecun@meta.com

Abstract

Data-Augmentation (DA) is known to improve performance across tasks and datasets. We propose a method to theoretically analyze the effect of DA and study questions such as: how many augmented samples are needed to correctly estimate the information encoded by that DA? How does the augmentation policy impact the final parameters of a model? We derive several quantities in close-form, such as the expectation and variance of an image, loss, and model's output under a given DA distribution. Up to our knowledge, we obtain the first explicit regularizer that corresponds to using DA during training for non-trivial transformations such as affine transformations, color jittering, or Gaussian blur. Those derivations open new avenues to quantify the benefits and limitations of DA. For example, we show given a loss at hand, common DAs require tens of thousands of samples for the loss to be correctly estimated and for the model training to converge. We then show that for a training loss to have reduced variance under DA sampling, the model's saliency map (gradient of the loss with respect to the model's input) must align with the smallest eigenvector of the sample's covariance matrix under the considered DA augmentation; this is exactly the quantity estimated and regularized by TangentProp. Those findings also hint at a possible explanation on why models tend to shift their focus from edges to textures when specific DAs are employed.

1 Introduction

Data Augmentation (DA) is a prevalent technique in training deep learning models [1, 2, 3]. These Deep Networks (DNs) models f_γ, governed by some parameters $\gamma \in \Gamma$, are trained on the train set and expected to generalize to unseen samples (test set). To combat the tendency of DNs to overfit [4], producing a large train-test performance gap, regularization and in particular DA is heavily employed, amongst other mechanisms such as weight-decay [5, 6]. The benefit of DA over those alternatives is that defining input transformations that preserve the semantics of their inputs is a relatively simple task, at least in computer vision or acoustic processing [7, 8, 9]. Furthermore, if the DA is well designed and rich enough, it can effectively bring the number of non-trivial training samples close to the theoretical limits ensuring that any performance gap between the train and test set vanishes [10] as empirically observed in various scenarios [11, 12]. In fact, DA has proven so useful that novel training methods such as self-supervised learning (SSL) entirely rely on DA [13, 14] to learn meaningful data representations.

Despite its empirical effectiveness, our understanding of DA has many open questions, three of which we propose to study: **(a)** how do different DAs impact the model's parameters during training?; **(b)** how sample-efficient is the DA sampling, i.e., how many DA samples a model must observe to converge?; and **(c)** how sensitive is a loss/model to the DA sampling and how this variance evolves during training as a function of the model's ability to minimize the loss at hand, and as a

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

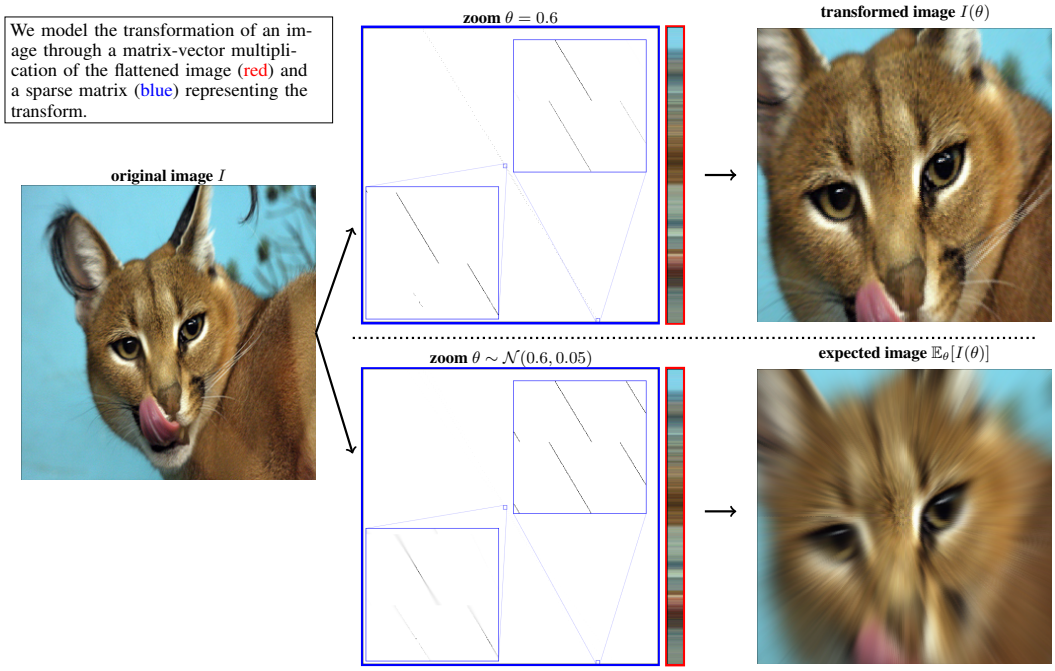

Figure 1: We propose a novel way to express Data Augmentation (DA) analytically that allows us to understand the impact of DA on the learned parameters of a model and quantify DA's sample efficiency. This allows us to compute the analytical expectation and variance of the transformed data (or any function computed with it) in close-form with respect to the transformation parameters.

function of the model's parameters? Our goal is to analytically derive some preliminary answers and insights into those three areas enabled by a novel image transformation operator that we introduce in Section 3.2 coined Data-Space Transform (DST). DST employs a mathematical formulation of input transformations akin to the one of [15] although slightly more general as we do not require the DA to form a group. We study the zoom/translation/shearing/rotation DAs in the main text and Gaussian blur, grayscale, color jitter and random crop in Appendix C. We summarize our contributions below:

(a) we derive the analytical first order moments of augmented samples, and of the losses employing augmented samples (Section 3.3), effectively providing us with the explicit regularizer induced by each DA (Section 3.4)
(b) we quantify the number of DA samples that are required for a model/loss to obtain a correct estimate of the information conveyed by that DA (Section 4.1)
(c) we derive the sensitivity, i.e. variance, of a given loss and model under a DA policy (Section 4.2) leading us to rediscover from first principles a popular deep network regularizer: TangentProp, as being the natural regularization to employ to minimize the loss variance (Section 4.3).

Upshot of gained insights: **(a)** will show us that the explicit DA regularizer corresponds to a generalized Tikhonov regularizer that depends on each sample's covariance matrix largest eigenvectors; the kernel space of the model's Jacobian matrix aligns with the data manifold tangent space as modeled by the DA; **(b)** will show us that the number of augmented samples required for a loss/model to correctly estimate the information provided by a DA of a single sample is on the order of 10^4. Even when combining the information of thousands of samples, we find that the entire train set must be augmented at least $50\times$ for a DA policy to be correctly learned by a model; lastly **(c)** will quantify the loss variance as a simple function of the model's Jacobian matrix, and the eigenvectors of the augmented sample variance matrix. Regardless of the model or task-at-hand, the loss sensitivity to random DA sampling goes down as the kernel of a model's Jacobian matrix aligns with the principal directions of the data manifold tangent space; as a by-product, we recover TangentProp as the natural regularizer minimizing the loss sensitivity to DA sampling.

Limitations: all of our image-space results e.g. expectation and variance of images under DA are exact, no approximation is employed. Results on a model's output/loss are provided both in the case of a linear model and of a nonlinear model. For the former, results are exact, for the latter a

third-order Taylor approximations is employed which is common and has been shown to provide accurate enough approximations to safely rely on the insights/quantities obtained e.g. see Sec. A.2 of [16]. All the proofs and implementation details are provided in the appendix. We also emphasize that our primary goal is to provide a theoretical understanding of DA, to quantify its efficiency and impact onto learned models and to derive and visualize its explicit regularizer; we leave the search for tractable alternatives to DA based on those insights for future work.

2 Background

Existing Explicit Regularizers From Data-Augmentation. It is widely accepted that data-augmentation (DA) regularizes a model towards the transformations that are modeled [5, 17], and that this regularization impacts performances significantly and positively, possibly as much as the regularization offered by the choice of DN architecture [18] and optimizer [19].
To gain precious insights into the impact of DA onto the learned functional f_γ, the most common strategy is to derive the *explicit* regularizer that directly acts upon f_γ in the same manner as if one were to use DA during training. This explicit derivation is however challenging and so far has been limited to DA strategies such as additive white noise or multiplicative binary noise applied identically and independently throughout the image and/or feature maps, as with dropout [20, 21]. In those settings, various works have studied in the linear regime the relation between such DA and its equivalence to using Tikhonov regularization[22] or weight decay [23] as in $\min_{\boldsymbol{W}} \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\sigma)} \left[\|\boldsymbol{y}_n - \boldsymbol{W}(\boldsymbol{x}_n + \boldsymbol{\epsilon})\|_2^2\right] = \min_{\boldsymbol{W}} \sum_{n=1}^{N} \|\boldsymbol{y}_n - \boldsymbol{W}\boldsymbol{x}_n\|_2^2 + \lambda(\sigma)\|\boldsymbol{W}\|_F^2$ [24]. More recently, [25, 26] extended the case of additive white noise to nonlinear models and concluded that additive white noise DA corresponds to adding an explicit Frobenius norm regularization onto the Jacobian of f_θ evaluated at each data sample.
Going to more involved DA strategies e.g. translations or zooms of the input images is challenging and has so far only been studied from an empirical perspective. For example, [27, 28, 29] performed thorough ablation studies on the interplay between DAs and a collection of known explicit regularizers to find correlations between them. It was concluded that weight-decay (the explicit regularizer of additive white noise) does not relate to those more advanced DAs. This also led other studies to suggest that norm-based regularization might be insufficient to describe the implicit regularization of DAs involving advanced image transformations [30]. We debunk this last claim in Sec. 3.3.

Coordinate Space Transformation. Throughout this paper, we will consider a two-dimensional image $I(x, y)$ to be at least square-integrable $I \in L^2(\mathbb{R}^2)$ [31]. Multi-channel images are dealt with by treating each channel as its own single-channel image, as commonly done in practice [1]. As we are interested in practical cases, we will often assume that I has compact support e.g. has nonzero values only within a bounded domain such as $[0, 1]^2$. Visualizing this image thus corresponds to displaying the sampled values of I on a regular grid (pixel positions) of $[0, 1]^2$ [32]. The most common formulation to apply a transformation on the image I to obtain the transformed image T is to transform the image coordinates [33, 34, 35]. That is, a mapping $t : \mathbb{R}^2 \mapsto \mathbb{R}^2$ describes what coordinate $t(u, v)$ of the original image I maps to the coordinate u, v of the transformed image T as in

$$T(u, v) = I(t(u, v)). \tag{1}$$

This mapping t often comes with some parameters θ governing the underlying transformation as in

$$t_\theta(x, y) = [x - \theta_1, y - \theta_2]^T, \;\; t_\theta(x, y) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \;\; t_\theta(x, y) = [\theta_1 x, \theta_2 y]^T, \tag{2}$$

for translation, rotation and zoom respectively. We provide a visual depiction of the zoom transformation applied in coordinate space in Fig. 6, in the appendix.

The formulation of Eq. 1 has two key benefits. First, it allows a simple and intuitive design of t to obtain novel transformations. Second, it is computationally efficient as the coordinate-space of images are $2/3$-dimensional. Those benefits have led to e.g. the Spatial Transformer Network [36]. On the other hand, Eq. 1 has one major drawback for our purpose: the exact moments of the transformed image under random θ parameters are not tractable due to the composition of t with the nonlinear mapping I. And as it will become clear in Section 3.1, those quantities are needed to derive DAs' explicit regularizers.

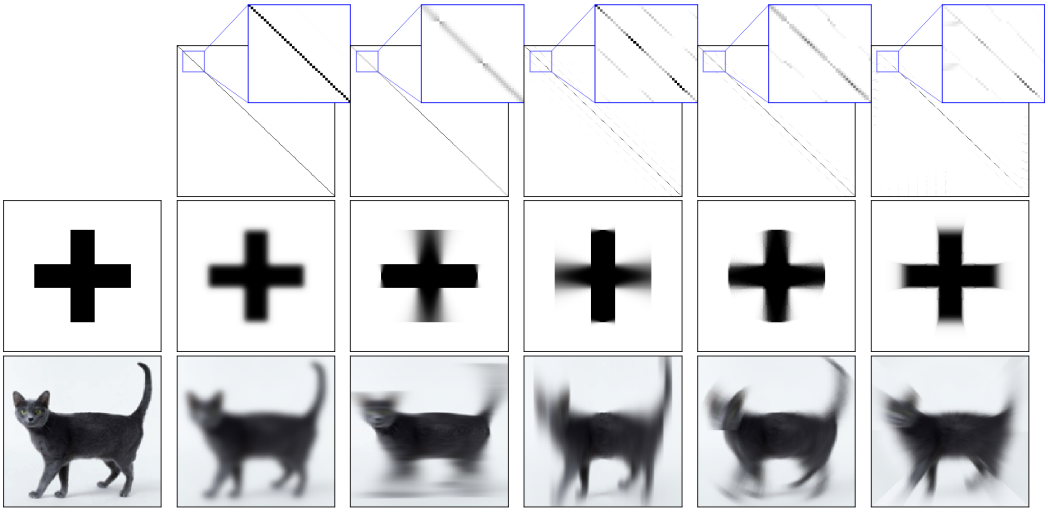

Figure 2: **Top row:** sparse matrix $\mathbb{E}_\theta[\boldsymbol{M}(\theta)]$ producing the expected image via $\mathbb{E}_\theta[\boldsymbol{t}(\theta)] = \mathbb{E}_\theta[\boldsymbol{M}(\theta)]\boldsymbol{x}$ (recall Eq. (7) and the discussion below Theorem 3.3), here with DA distributions of $\mathcal{N}(0, 0.04) \otimes \mathcal{N}(0, 0.04)$, $\mathcal{N}(0, 0.2)$, $\mathcal{N}(0, 0.1)$ and $\mathcal{N}(1, 0.1)$ for each column respectively. This *sparse* matrix is independent of the image and derived in close-form (Theorem 3.3). **Middle & Bottom row:** are two images that depict for each column the "expected image" under each DA. Estimating those images (and any loss employing them) only through sampling the DA augmentations would require tens of thousands of samples (see Fig. 4). *Gaussian blur, grayscale, color jitter and random crop DA are provided in Appendix C along with discussions on how to extend our results to CutOut and MixUp.*

3 Analytical Moments of Transformed Images Enable Infinite Dataset Training

We first motivate this study by formulating the training process under DA sampling as doing a Monte-Carlo estimate of the true (unknown) expected loss under that DA distribution (Section 3.1). Going beyond this sampling/estimation procedures requires knowledge of the average and variance of a transformed sample, i.e. the first two centered-moments. This is our motivation to construct a novel Data-Space Transformation (DST) (Section 3.2) that will enable close-form formula for those moments (Section 3.3). From those, we will be able to remove the need to sample transformed images to train a model, by obtaining the close-form expected loss Section 3.4, under which training can be performed sampling-free.

3.1 Motivation: Current Data-Augmentation Training Performs Monte-Carlo Estimation

Training a model with DA consists in (i) sampling transformed images at each training iteration for each sample $\boldsymbol{x}_n$ as in $\mathcal{T}_{\theta_n}(\boldsymbol{x}_n)$ with $\theta_n \sim \boldsymbol{\theta}$ a randomly sampled DA parameter e.g. the amount of translation to apply, (ii) evaluating the loss $\mathcal{L}$ on the transformed sample/mini-batch/dataset, and (iii) using some flavor of gradient descent to update the parameters γ of the model f_γ.

This training procedure corresponds to a one-sample Monte-Carlo (MC) estimate [37, 38] of the expected loss

$$\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\theta}} \left[\mathcal{L}\left(f_\gamma(\mathcal{T}_{\boldsymbol{\theta}}(\boldsymbol{x}_n))\right)\right] \approx \sum_{n=1}^{N} \mathcal{L}(f_\gamma(\mathcal{T}_{\theta_n}(\boldsymbol{x}_n)), \tag{3}$$

with θ_n i.i.d samples of $\boldsymbol{\theta}$ repeatedly sampled at each epoch. In a supervised setting, the loss would also receive a per-sample target $\boldsymbol{y}_n$ as input. Although a one-sample estimate might be insufficient to apply the central limit theorem [39] and guarantee training convergence, the combination of multiple samples in mini-batch training does provide convergence in most cases. Although, methods such as Self-Supervised Learning that heavily rely on DA, tend to be more sensitive and unstable [40].

To avoid such instabilities, and to incorporate all the DA transformations of $\boldsymbol{x}_n$ into the model's parameter update, one would be tempted to compute the model parameters' gradient on the expected loss (left-hand side of Eq. (3)) (see Section 3.4). Knowledge of the expected loss would also prove useful to measure the quality of the MC estimate (see Section 4.1). Note that even in the least-square setting, such approximation is employed as the expected loss under DAs is not know. Hence, one commonly obtain many DA samples, and then apply the least-square formula on this augmented training set, which is highly inefficient. Alternatively, this study will obtain the optimum least-square parameters under DA sampling (Theorem 3.4) i.e. only using the given training set, we directly obtain the parameters that would correspond to sampling infinitely many DA samples and applying the usual least-square formula on this infinite training set.

The close-form expected loss requires knowledge of the average and variance of the transformed sample $\mathcal{T}_{\boldsymbol{\theta}}(\boldsymbol{x}_n)$ taken with respect to the random variable $\boldsymbol{\theta}$. Hence, we first propose to formulate a novel and tractable augmentation model (Section 3.2) that will allow us to obtain those moments analytically in Section 3.3.

3.2 Proposed Data-Space Transformation (DST)

Instead of altering the coordinate positions of an image, as done in the coordinate-space transformation of Eq. (1), we propose to alter the image basis functions.

Going back to the construction of functions, one easily recalls that any (image) function can be expanded into a basis as in $I(u,v) = \int I(x,y)\delta(u-x, v-y)dxdy$ with δ the usual Dirac distribution. Suppose for now that we consider a horizontal translation by a constant θ. Then, one can obtain that translated image via

$$\mathcal{T}_\theta(I)(u,v) = \int I(x,y)\delta(u-x-\theta, v-y)dxdy, \tag{4}$$

hence, and crucial to our study, *we apply the transformation onto the basis functions onto which the image is evaluated, rather than onto the original image itself.* As the image is now constant with respect to the transformation parameter θ, and as the basis functions have some convenient analytical forms, the derivation of the transformed images moments, under $\boldsymbol{\theta}$, will become straightforward. Because the transformed image T is now obtained by combining its pixel values (recall Eq. (4)) we coin this transform as Data-Space Transform (DST), formally defined below.

Definition 3.1 (Data-Space Transform)**.** We define the DST of an image $I \in L^2(\mathbb{R}^2)$ producing the transformed image $\mathcal{T}_\theta(I) \in L^2(\mathbb{R}^2)$ as

$$\mathcal{T}_\theta(I)(u,v) = \int I(x,y)h_\theta(u,v,x,y)dxdy, \tag{5}$$

with $h_\theta(u,v,.,.) \in \mathbb{C}_0^\infty(\mathbb{R}^2)$ encoding the transformation.

In Definition 3.1, we only impose for $h_\theta(u,v,.,.)$ to be with compact support. In fact, one should interpret $h_\theta(u,v,.,.)$ as a distribution whose purpose is to evaluate I at a desired (coordinate) position on its domain. This evaluation –depending on the form of $h_\theta(u,v,.,.)$– can extract a single pixel-value of the image I at a desired location (as in Eq. (4)), or can combine multiple values e.g. with $h_\theta(u,v,.,.)$ being a bump function. Additionally, we do not impose any restriction on the invertibility of this mapping i.e. $\mathcal{T}_\theta(I)(u,v)$ can disregard parts of the original image.

Coordinate-space transformations as DSTs. The coordinate-space transformation and the proposed DST (Definition 3.1) act in different spaces: the image coordinates and the image pixel values, respectively. Nevertheless, this does not limit the range of transformations that can be applied to an image. The following statement provides a simple recipe to turn any already employed coordinate-space transformation into a DST.

Proposition 3.2. *Any coordinate-space transformation (1) using $t_\theta : \mathbb{R}^2 \mapsto \mathbb{R}^2$ can be expressed as a data-space transformation (5) by setting $h_\theta(u,v,x,y) = \delta(t_\theta(x,y) - [u,v]^T)$.*

We will focus in the main text on zoom/rotation/translation/shearing, and we propose in Appendix C the case of Gaussian blur, color jittering, random crop, along with a discussion on extending our results to CutMix [41] and MixUp [42]. Using Proposition 3.2, we obtain the DST operators $h_\theta(u,v,x,y)$

to be

$$\delta(u-x+\theta_1, v-y+\theta_2),\ \ \delta(u-x-\theta_1 y, v-y-\theta_2 x),\ \ \delta(u-\theta x, v-\theta y),$$
$$\delta(u-\cos(\theta)x+\sin(\theta)y, v-\sin(\theta)x-\cos(\theta)y), \tag{6}$$

for the vertical/horizontal translation, vertical/horizontal shearing, zoom and rotation respectively (compare with Eq. 2). Before focusing on the analytical moments of the DST samples, we describe how those operators are applied in a discrete setting.

Discretized version. We now describe how any DST of a discrete image, flattened as a vector, can be expressed as a matrix-vector product with the matrix entries depending on the employed DA and its parameter θ, as we display in Figs. 1 and 2. The functional form of the DST from Eq. (5) producing the target image at a specific position $\mathcal{T}_\theta(I)(u,v)$ is linear in the original image I. Hence, in the discrete setting, the continuous integral over the image domain is replaced with a summation with indices based on the desired sampling/resolution of I, i.e. the position of the pixel spatial positions. Expressing linear operators as matrix-vector products will greatly ease our development, we will denote by $\boldsymbol{x} \in \mathbb{R}^{hw}$ the flattened $(h \times w)$ discrete images I. Hence, our data-space transformation, given some parameters θ, takes the form of

$$\boldsymbol{t}(\theta) = \boldsymbol{M}(\theta)\boldsymbol{x}, \tag{7}$$

with $\boldsymbol{t}(\theta) \in \mathbb{R}^{hw}$ the flattened transformed image (which can be reshaped as desired) and $\boldsymbol{M}(\theta) \in \mathbb{R}^{hw \times hw}$ the matrix whose rows encode the discrete and flattened $h_\theta(u,v,.,.)$. For example, and employing a uniform grid sampling for illustration, $\boldsymbol{M}(\theta)_{i,j} = h_\theta(i//w, i\%w, j//w, j\%w)$ with $//$ representing the floor division and $\%$ the modulo operation. For the case of multi-channel images, we consider without loss of generality the application of the matrix-vector operation on each channel separately. We depict this operation along with the exact form of $\boldsymbol{M}(\theta)$ for the case of the zoom transformation in Fig. 1. A crucial property of the matrix $\boldsymbol{M}(\theta)$ lies in its sparsity, in fact, for most DAs the transformations rely on displacing pixels rather than combining them. Hence, although $\boldsymbol{M}(\theta)$'s total number of entries grow quadratically with the number of pixels in I, the number of nonzero entries only grows linearly with it.

3.3 Analytical Expectation and Variance of Transformed Images

The above construction (Definition 3.1) turns out to make the analytical form of the first two moments of an augmented sample much simpler to derive. As this derivation is at the core of our main contribution, we propose a step-by-step derivation of $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{T}_{\boldsymbol{\theta}}(I)\right]$ for the case of horizontal translation (recall Eq. (4)).

Let's consider again the continuous model (the discrete version is provided after Theorem 3.3). Using Fubini's theorem to switch the order of integration and recalling Eq. (5), we have that

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{T}_{\boldsymbol{\theta}}(I)(u,v)\right] = \int I(x,y)\mathbb{E}_{\boldsymbol{\theta}}\left[h_{\boldsymbol{\theta}}(u,v,x,y)\right]dxdy. \tag{8}$$

Using the definition for h_θ from Eq. (6) for horizontal translation, $\mathbb{E}_{\boldsymbol{\theta}}\left[h_{\boldsymbol{\theta}}(u,v,x,y)\right]$ becomes

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\delta(u-x-\boldsymbol{\theta}, v-y)\right] = \int \delta(u-x-\theta, v-y)p(\theta)d\theta = p(u-x)\delta(v-y),$$

with p the density function of $\boldsymbol{\theta}$ prescribing how the translation parameter is distributed. As a result, in this univariate translation case, the expected augmented image at coordinate (u,v) is given by

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{T}_{\boldsymbol{\theta}}(I)(u,v)\right] = \int I(x,y)p(u-x)\delta(v-y)dxdy = \int I(x,v)p(u-x)dx, \tag{9}$$

which can be further simplified into $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{T}_{\boldsymbol{\theta}}(I)(.,v)\right] = I(.,v) \star p$. Hence, the expected horizontally-translated image is the convolution (on the x-axis only) between the original image I and the univariate density function p. We formalize this for the transformations of Eq. (6) below.

Theorem 3.3. *The analytical form of* $\mathbb{E}_{\boldsymbol{\theta}}\left[h_{\boldsymbol{\theta}}(u,v,x,y)\right]$*, used to obtain the expected transformed image (recall Eq.* (8)*) is given by*

$$p(u-x, v-y)\ \ \textit{(translation)}, \quad \textit{and} \quad p(u/x)\delta(u/x - v/y)\ \ \textit{(rotation)}, \tag{10}$$

other cases and second-order moment $\mathbb{E}_{\boldsymbol{\theta}}[\mathcal{T}_{\boldsymbol{\theta}}(\boldsymbol{x}_n)\mathcal{T}_{\boldsymbol{\theta}}(\boldsymbol{x}_n)^T]$ *are deferred to the proof. (Proof in Appendix F.)*

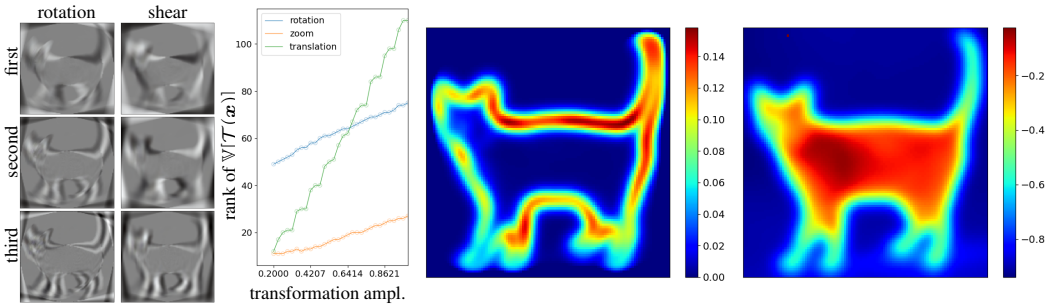

Figure 3: **First column:** top eigenvectors ($\boldsymbol{Q}$ matrix in Theorem 3.4) of the sample variance matrix $\mathbb{V}[\mathcal{T}(\boldsymbol{x})]$ for rotation/shear augmentations with $\mathcal{U}(-15°, 15°)$ and $\mathcal{U}(-15°, 15°) \otimes \mathcal{U}(-15°, 15°)$ respectively. As per Theorem 4.1, training aligns the model's Jacobian matrix kernel to the largest eigenvectors of $\mathbb{V}[\mathcal{T}(\boldsymbol{x})]$ to reduce the loss variance under the respective DA sampling. **Second column:** number of nonzero eigenvalues of $\mathbb{V}[\mathcal{T}(\boldsymbol{x})]$ (nonzero elements in Λ in Theorem 3.4) for increasing transformation amplitude (the datum dimension is $114 \times 114 \times 3$). We observe that the dimension of the subspace spanned by augmented images constraint increases linearly with the amplitude of the transformation. **Third column:** Pixel variance i.e. diagonal of $\mathbb{V}[\mathcal{T}(\boldsymbol{x})]$ reshapes as an image. **Fourth column:** pixel covariance between one background pixel and all other pixels, i.e. 700th row of $\mathbb{V}[\mathcal{T}(\boldsymbol{x})]$ reshaped as an image for the cat of Fig. 2 under translations ($\mathcal{N}(0, 0.1) \otimes \mathcal{N}(0, 0.1)$). As per Theorem 3.4, the variance of the pixel values, seen on the left, being much higher for the edges of the cat than for its interior body/texture pushes the model to focus on the cat texture, a phenomenon empirically observed in deep networks [43, 44].

An interesting observation obtained from the above ties the expected image under random 2-dimensional translations with (2-dimensional) density p to the convolution $I \star p$; this provides a new portal to study and interpret convolutions with nonnegative, sum-to-one filters p. Again, and as per Eq. (7), the discretized version of the expected image takes the form of $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{t}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{M}(\boldsymbol{\theta})]\boldsymbol{x}$ with the entries of $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{M}(\boldsymbol{\theta})]$ given by discretizing Eq. (10). We depict this expected matrix for various transformations as well as their application onto two different discrete images in Fig. 2, and we now proceed on deriving the left-hand side of Eq. (3) i.e. the explicit DA regularizer.

3.4 The Explicit Regularizer of Data-Augmentations

To keep notations as light as possible, we first consider a linear regression model with Mean Squared Error (MSE), the nonlinear case will be studied in Section 4. In that setting, the expected loss under DA sampling (recall Eq. (3)) becomes

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\theta}} \left[\|\boldsymbol{y}_n - \boldsymbol{W}\mathcal{T}_{\boldsymbol{\theta}}(\boldsymbol{x}_n) - \boldsymbol{b}\|_2^2 \right], \tag{11}$$

with $\boldsymbol{x}_n \in \mathbb{R}^D, n = 1, \ldots, N$ the input (flattened) n^{th} image I_n, $\boldsymbol{y}_n \in \mathbb{R}^K$ the nth target vector, and $\boldsymbol{W} \in \mathbb{R}^{K \times D}, \boldsymbol{b} \in \mathbb{R}^K$ the model's parameters. Recall from Section 3.1 that to learn $\boldsymbol{W}$ under DA, the current strategy consists in performing a Monte-Carlo (MC) estimation. Instead, let's derive (detailed derivation in Appendix D) the exact loss of Eq. (11) as a function of the sample mean and variance under the consider DA. We will drop the $\boldsymbol{\theta}$ subscript for clarity to obtain

$$\mathcal{L} = \sum_{n=1}^{N} \|\boldsymbol{y}_n - \boldsymbol{W}\mathbb{E}\left[\mathcal{T}(\boldsymbol{x}_n)\right] - \boldsymbol{b}\|_2^2 + \|\boldsymbol{W}\boldsymbol{Q}(\boldsymbol{x})\Lambda(\boldsymbol{x})^{\frac{1}{2}}\|_F^2, \tag{12}$$

with the spectral decomposition $\boldsymbol{Q}(\boldsymbol{x})\Lambda(\boldsymbol{x})\boldsymbol{Q}(\boldsymbol{x})^T = \mathbb{V}\left[\mathcal{T}(\boldsymbol{x})\right]$. The right term in Eq. (12) is the explicit DA regularizer. It pushes the kernel space of $\boldsymbol{W}$ to align with the largest principal directions of the data manifold tangent space, as modeled by the DA. In fact, the largest eigenvectors in $\boldsymbol{Q}(\boldsymbol{x})$ represent the principal directions of the data manifold tangent space at $\boldsymbol{x}$, as encoded via $\mathbb{V}\left[\mathcal{T}(\boldsymbol{x})\right]$.

We propose in Fig. 3 visualization of $\boldsymbol{Q}$ and Λ for different DAs, illustrating how each DA policy impacts the model's parameter $\boldsymbol{W}$ through the regularization of Eq. (12). The knowledge of $\mathbb{E}\left[\mathcal{T}(\boldsymbol{x})\right]$ and $\mathbb{V}\left[\mathcal{T}(\boldsymbol{x})\right]$ from Theorem 3.3 finally enables to train a (linear) model on the true expected loss (Eq. (12)) as we formalize below.

Theorem 3.4. *Training a linear model with MSE and infinite DA sampling is equivalent to minimizing Eq.* (12) *and produces the optimal* $\boldsymbol{W}^*$ *model's parameter*

$$\boldsymbol{W}^* = \left(\sum_{n=1}^{N} (\boldsymbol{y}_n - \boldsymbol{b}) \mathbb{E}[\mathcal{T}(\boldsymbol{x}_n)]^T \right) \left(\sum_{n=1}^{N} \mathbb{E}[\mathcal{T}(\boldsymbol{x}_n)] \mathbb{E}[\mathcal{T}(\boldsymbol{x}_n)]^T + \sum_{n=1}^{N} \mathbb{V}[\mathcal{T}(\boldsymbol{x}_n)] \right)^{-1}. \tag{13}$$

Not surprisingly, whenever the DA is identity, $\mathbb{E}[\mathcal{T}(\boldsymbol{x}_n)] = \boldsymbol{x}_n, \mathbb{V}[\mathcal{T}(\boldsymbol{x}_n)] = \boldsymbol{0}$ and thus Eq. (13) recovers exactly the standard least-square solution. We visualize $\mathbb{V}[\mathcal{T}(\boldsymbol{x}_n)]$ for the translation DA in Fig. 3. The same line of result can be derived in the nonlinear setting by assuming that the DA is restricted to small transformations. In that case, one leverages a truncated Taylor approximation of the nonlinear model[1] and recovers that (local) DA applies the same regularization as in Eq. (12) but with the model's Jacobian matrix $\boldsymbol{J} f_\gamma(\boldsymbol{x}_n)$ in-place of $\boldsymbol{W}$ (more details in Section 4.3).

Given the above derivation of the exact expected loss, we can now turn to precisely measure how accurate is the MC estimate commonly used to train models under DA sampling.

4 Data-Augmentation Sampling Efficiency and Loss Sensitivity

In this section we study the convergence of the MC estimate (Section 4.1), and provide variance analysis of that estimate as a function of the model's Jacobian matrix and the sample variance eigenvectors (Section 4.2), concluding by the (re-)discovery of TangentProp from first principles (Section 4.3). Those results hold regardless of the type of DA employed. Additionally, the analysis is done on a single sample to ease notation, in the i.i.d. setting, simply sum over all the training samples to analyze the training set.

4.1 Empirical Monte-Carlo Convergence of Transformed Images

Given the close-form *average image* and *average loss*, we empirically measure how efficient is the MC estimation from Eq. (3).
We first propose in Fig. 4 a constructed (64×64) image for which we compute the expected loss (Eq. (12)) and the MC estimate (right-hand side of Eq. (3)). Surprisingly, we obtain that even for such a simple image and translation DA, between 1000 and 10000 samples are required to correctly estimate the MSE loss from the augmented samples. In a more practical scenario, one could rightfully argue that the combination of the DA samples from different images allows to obtain a better estimate with a smaller amount of augmentations per sample. Hence, we provide in Fig. 5 that experiment using a linear model on MNIST [45] with varying train set size. We observe that as the number of samples grows as the required number of augmentation per sample reduces. Nevertheless, even with thousands of samples, at least 50 augmentations per sample are required to provide an accurate estimate (reproduction with nonlinear models given in Fig. 7).
We now propose to specifically quantify the sensitivity of the MC estimate to DA sampling.

4.2 Loss Sensitivity Under Data-Augmentation Sampling in the Linear and Nonlinear Regime

Recalling Section 3.1, current DA training is performed on a MC estimate of the loss. The estimator's variance [46] is proportional to the variance of the quantity being estimated: $\mathbb{V}[(\mathcal{L} \circ f)(\mathcal{T}(\boldsymbol{x}))]$. We now characterize when, and why, would an MC estimator converge for a given model.

By leveraging the delta method [47, 48] i.e. a truncated Taylor expansion of the model and loss function mapping as $\mathcal{L} \circ f$, we have

$$\mathbb{V}[(\mathcal{L} \circ f)(\mathcal{T}(\boldsymbol{x}))] \approx \|\nabla(\mathcal{L} \circ f)(\mathbb{E}[\mathcal{T}(\boldsymbol{x})])\|^2_{\mathbb{V}[\mathcal{T}(\boldsymbol{x})]}, \tag{14}$$

with $\|\boldsymbol{u}\|^2_{\boldsymbol{A}} \triangleq \boldsymbol{u}^T \boldsymbol{A} \boldsymbol{u}$. Noticing that $\mathbb{V}[\mathcal{T}(\boldsymbol{x})] = \mathbb{E}[\mathcal{T}(\boldsymbol{x})\mathcal{T}(\boldsymbol{x})^T] - \mathbb{E}[\mathcal{T}(\boldsymbol{x})]\mathbb{E}[\mathcal{T}(\boldsymbol{x})]^T$ and using the close-form moments of DST samples from Theorem 3.3, it is possible to write out Eq. (14) explicitly for model and DA specific analysis. We visualize $\mathbb{V}[\mathcal{T}(\boldsymbol{x})]$ in Fig. 3.

Linear regression case. To gain some insight into Eq. (14), let's first consider the linear regression case leading to $\mathbb{V}[(\mathcal{L} \circ f)(\mathcal{T}(\boldsymbol{x}))] \approx \|\boldsymbol{y} - \boldsymbol{W}\mathbb{E}[\mathcal{T}(\boldsymbol{x})] - \boldsymbol{b}\|^2_{\boldsymbol{W}\mathbb{V}[\mathcal{T}(\boldsymbol{x})]\boldsymbol{W}^T}$. Hence, the estimated loss variance depends (i) on the model predicting the correct output when observing the expected sample $\mathbb{E}[\mathcal{T}(\boldsymbol{x})]$, and (ii) on the smallest right singular vector of $\boldsymbol{W}$ to align with the largest eigenvectors of $\mathbb{V}[\mathcal{T}(\boldsymbol{x})]$, echoing our observation below Eq. (12).

[1] as commonly done, see e.g. Sec. A.2 from [16] for justification and approximation error analysis

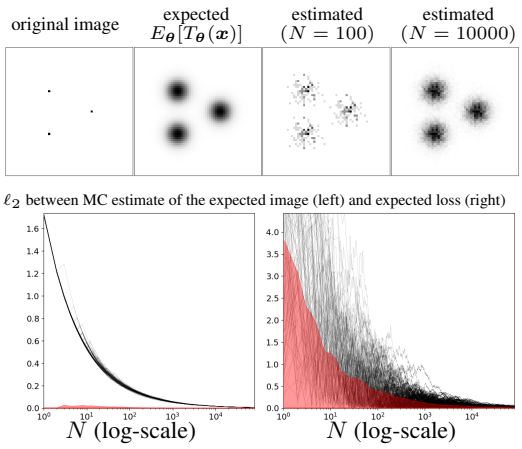

Figure 4: Analytical expected synthetic image under translation ($\mathcal{N}(0, 0.1) \otimes \mathcal{N}(0, 0.1)$) against its N-sample Monte-Carlo estimate (**top row**), ℓ_2 distance between the true and estimated images (**bottom left**) and the true and estimated MSE loss with a random Gaussian $\boldsymbol{y}, \boldsymbol{W}$ (**bottom right**). In red is depicted the standard deviation of the independent Monte-Carlo runs. Clearly we observe that even on a simple (64×64) image and using the translation transformation, thousands of sampled images are necessary to provide an accurate estimate of the loss at hand. That is, sampling based data-augmentations are a rather inefficient medium to employ for injecting prior information into a model as this prior information will only emerge after tens of thousands of images have been sampled.

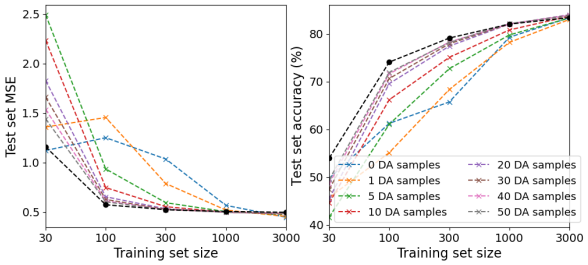

Figure 5: MNIST with linear model and translation DA with varying number of samples (**colored lines**) against the use of the expected loss (**black line**). We depict the MSE loss (**left**) and the accuracy (**right**) on the test set. At first, even a $50\times$ DA-based increase in dataset size fails to provide the performances reached by employing the analytical expected loss. As the training size increases, as the DA samples get redundant with the added samples closing the performance gap (the case of nonlinear model is provided in Fig. 7). See Tables 2 to 5 for additional table of results with standard deviation.

General case. Since the result from Eq. (14) holds in the general setting, we can generalize the above observation and formalize it into the following statement.

Theorem 4.1. *The variance of the loss's MC estimate (recall Eq.* (3)*) for an input* $\boldsymbol{x}_n$ *goes to* 0 *if the loss gradient* $\nabla\mathcal{L}(f(\mathbb{E}[\mathcal{T}(\boldsymbol{x})]))$ *goes to* 0*, or if the kernel of the model's Jacobian matrix* $\boldsymbol{J}f(\mathbb{E}[\mathcal{T}(\boldsymbol{x})])$ *aligns with the largest eigenvectors of the sample variance* $\mathbb{V}[\mathcal{T}(\boldsymbol{x})]$.

The proof of the above statement simply starts from Eq. (14) and applies the Cauchy-Schwarz inequality two times to obtain

$$\mathbb{V}[(\mathcal{L} \circ f)(\mathcal{T}(\boldsymbol{x}))] \leq \|\nabla\mathcal{L}(f(\mathbb{E}[\mathcal{T}(\boldsymbol{x})]))\|_2^4 \times \|\boldsymbol{J}f(\mathbb{E}[\mathcal{T}(\boldsymbol{x})])\boldsymbol{Q}(\boldsymbol{x})\Lambda^{\frac{1}{2}}(\boldsymbol{x})\|_F^4, \tag{15}$$

with decomposition $\mathbb{V}[\mathcal{T}(\boldsymbol{x})] = \boldsymbol{Q}(\boldsymbol{x})\Lambda(\boldsymbol{x})\boldsymbol{Q}(\boldsymbol{x})^T$. From Eq. (15), it is quite realistic to assume that the model's Jacobian does not collapse to 0 (this would not minimize the training loss in general). Hence, the most realistic case of Theorem 4.1 concerns the alignment between the model's Jacobian matrix kernel space and the top eigenvectors of $\mathbb{V}[\mathcal{T}(\boldsymbol{x})]$. However, for models like ResNet DNs [49], the loss variance can only reduce to 0 if the model minimizes the loss at the expected input since the Jacobian matrix is always full-rank. *We recall that the above relied on the delta method to approximate the intractable LHS of Eq.* (14) *(see footnote on Page 7).*

4.3 Explicit Loss Sensitivity Minimization Provably Recovers TangentProp

The expected loss (Eq. (3)) has been derived in the linear regression setting (Eq. (12)). But in a more general scenario, and as discussed below Theorem 3.4, this expectation might not be tractable and thus needs to be approximated e.g. based on a Taylor expansion of the loss and model. Alternatively to using the approximated expectation, one could employ the usual MC estimate of the expectation (right-hand side of Eq. (3)), and leverage the MC estimator variance obtained in Eq. (15) as a regularizer. We show here that both approaches are equivalent, and recover a popular regularizer known as TangentProp [50]. Although various extensions of TangentProp have been introduced [51, 52] no principled derivation of it has yet been proposed. Using the same argument as in Section 4.2, the expectation of a nonlinear transformation of a random variable ($\mathcal{T}(\boldsymbol{x}_n)$) can be

		Train. size (N)	#DA=1	#DA=2	#DA=5	#DA=10	expectation (#DA=∞)
rotation$\in$ [-20,20] degrees translations $\in$[-0.15,0.15]% zoom $\in$[0.9, 1.1]%	MNIST	N=100	10.39$\pm$1.82	11.70$\pm$2.02	10.60 $\pm$1.34	24.30$\pm$1.92	33.44
		N=1000	21.71$\pm$1.61	32.57$\pm$1.95	40.61$\pm$2.20	45.16$\pm$2.38	48.55
		N=10000	47.82$\pm$2.20	51.98$\pm$1.33	54.97$\pm$0.92	56.29$\pm$0.66	57.16
	EMNIST	N=100	4.48$\pm$0.85	4.36$\pm$0.57	3.85$\pm$0.40	8.38$\pm$0.58	13.34
		N=1000	8.45$\pm$0.53	14.79$\pm$0.60	22.51$\pm$0.63	27.49$\pm$0.60	32.18
		N=10000	30.49$\pm$0.75	35.63$\pm$0.50	39.98$\pm$0.49	42.13$\pm$0.40	43.30
	FMNIST	N=100	11.18$\pm$3.31	9.49$\pm$1.40	10.65$\pm$1.66	23.10$\pm$2.17	35.31
		N=1000	24.43$\pm$2.75	38.71$\pm$1.82	47.49$\pm$2.07	52.1$\pm$1.65	54.92
		N=10000	52.60$\pm$2.57	54.9$\pm$1.70	56.84$\pm$1.42	57.29$\pm$1.44	57.63
rotation$\in$ [-5,5] degrees translations $\in$[-0.05,0.05]% zoom $\in$[0.98, 1.02]%	MNIST	N=100	12.72$\pm$1.95	12.42$\pm$2.02	12.45$\pm$1.47	45.82$\pm$1.53	55.83
		N=1000	50.44$\pm$1.07	64.96$\pm$0.95	72.64$\pm$0.67	75.06$\pm$0.38	76.15
		N=10000	78.54$\pm$0.22	79.76$\pm$0.24	80.49$\pm$0.20	80.85$\pm$0.17	80.86
	EMNIST	N=100	4.69$\pm$0.73	4.77$\pm$0.61	4.43$\pm$0.34	12.79$\pm$0.51	19.53
		N=1000	16.74$\pm$0.47	31.10$\pm$0.29	40.59$\pm$0.39	44.45$\pm$0.34	46.53
		N=10000	49.13$\pm$0.30	51.98$\pm$0.13	53.62$\pm$0.14	54.30$\pm$0.16	54.61
	FMNIST	N=100	11.51$\pm$2.12	12.43$\pm$1.99	11.59$\pm$1.40	37.08$\pm$1.23	55.22
		N=1000	46.97$\pm$1.17	66.22$\pm$0.51	72.75$\pm$0.42	74.56$\pm$0.35	75.03
		N=10000	75.2$\pm$0.33	75.89$\pm$0.27	76.49$\pm$0.32	76.63$\pm$0.18	76.71

Table 1: Reprise of Fig. 4 reporting test set classification performances averaged over 10 runs, with standard deviation corresponding to different DA realizations. For each specific run and configuration, all scenarios have access to the exact same training set, #DA represents the number of new samples introduced for each training sample, total training set size is thus N$\times$(1+#DA). We observe that **in the large dataset size regime** $(N = 10000)$ the gap between low number of DA samples and the expected regularizer is marginal which is expected as the introduced DA variations become redundant with the training set samples. This marginal gap can be slightly increased by employing stronger DA (compare top half and bottom half of the table). However, **in the small dataset size regime**, there is an important gain provided by employing the regularizer even when compared to observing many DA samples (#DA=10). This indicates a potentially crucial regime in which introducing the analytical regularizer helps. More surprisingly, this is also the case if part of DA policy is misaligned with the task at hand (e.g. rotation for FashionMNIST). Additional empirical results with more training set sizes ($N \in \{100, 200, 500, 1000, 3000, 10000, 20000\}$) and DA policies (one per table) are available in Tables 2 to 5 for MNIST, in Tables 6 to 9 for EMNIST, and in Tables 10 to 13 for FashionMNIST

approximated from the expectation of the Taylor expansion (detailed derivations in Appendix E) from which an upper-bound is obtained by applying the Cauchy-Schwarz inequality

$$\mathbb{E}[(\mathcal{L} \circ f)(\mathcal{T}(\boldsymbol{x}))] \leq (\mathcal{L} \circ f)(\mathbb{E}[\mathcal{T}(\boldsymbol{x})]) + \kappa(\boldsymbol{x}) \underbrace{\|\boldsymbol{J} f(\mathbb{E}[\mathcal{T}(\boldsymbol{x})]) \boldsymbol{Q}(\boldsymbol{x}) \Lambda(\boldsymbol{x})^{\frac{1}{2}}\|_F^2}_{\text{TangentProp regularization}},$$

with $\kappa(\boldsymbol{x}) \geq 0$ and the notations from Eq. (15). As a result, the TangentProp regularizer naturally appears when using a Taylor approximation of the expected loss, and it corresponds to adding an explicit loss variance regularization term (compare the TangentProp with Eq. (15)). We thus obtained from first principles that TangentProp emerges naturally when considering the second order Taylor approximation of the expected loss given a DA.

5 Conclusions and Limitations

In this paper, we proposed a novel set of mathematical tools (Sections 3.2 and 3.3) under which it is possible to study DA and to provably answer some of the open questions around the efficiency and impact of DA to train a model. We first obtained the explicit regularizer produced by different DAs in Section 3.4. This led to the following observation: the kernel space of the $\boldsymbol{W}$ matrix is pushed to align with the largest eigenvectors of the sample covariance matrix. This was then studied in a more general setting in Section 4.2 for nonlinear models when characterizing the loss variance under DA sampling. We also observed that Monte-Carlo sampling of transformed images is highly inefficient (Section 4.1), even if similar training samples combine their underlying information within a dataset. Lastly, those derivations led us to provably derive a known regularizer —TangentProp— as being the natural minimizer of a model's loss variance. The main limitation of the current analysis concerns nonlinear models for which we rely on a third-order Taylor expansion.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[2] J. Kukačka, V. Golkov, and D. Cremers, "Regularization for deep learning: A taxonomy," *arXiv preprint arXiv:1710.10686*, 2017.

[3] R. Balestriero and Y. LeCun, "Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods," *arXiv preprint arXiv:2205.11508*, 2022.

[4] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[5] B. Neyshabur, R. Tomioka, and N. Srebro, "In search of the real inductive bias: On the role of implicit regularization in deep learning," *arXiv preprint arXiv:1412.6614*, 2014.

[6] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[7] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[8] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.

[9] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[10] M. Unser, "A representer theorem for deep neural networks.," *J. Mach. Learn. Res.*, vol. 20, no. 110, pp. 1–30, 2019.

[11] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[12] A. M. Taqi, A. Awad, F. Al-Azzo, and M. Milanova, "The impact of multi-optimizers and data augmentation on tensorflow convolutional neural network performance," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 140–145, IEEE, 2018.

[13] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

[14] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," *arXiv preprint arXiv:2103.03230*, 2021.

[15] S. Chen, E. Dobriban, and J. Lee, "A group-theoretic framework for data augmentation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21321–21333, 2020.

[16] C. Wei, S. Kakade, and T. Ma, "The implicit and explicit regularization effects of dropout," 2020.

[17] B. Neyshabur, "Implicit regularization in deep learning," *arXiv preprint arXiv:1709.01953*, 2017.

[18] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks," *arXiv preprint arXiv:1806.00468*, 2018.

[19] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] X. Bouthillier, K. Konda, P. Vincent, and R. Memisevic, "Dropout as data augmentation," *arXiv preprint arXiv:1506.08700*, 2015.

[22] A. N. Tikhonov, "On the stability of inverse problems," in *Dokl. Akad. Nauk SSSR*, vol. 39, pp. 195–198, 1943.

[23] G. Zhang, C. Wang, B. Xu, and R. Grosse, "Three mechanisms of weight decay regularization," *arXiv preprint arXiv:1810.12281*, 2018.

[24] P. Baldi and P. J. Sadowski, "Understanding dropout," *Advances in neural information processing systems*, vol. 26, pp. 2814–2822, 2013.

[25] Z. Wang, R. Balestriero, and R. Baraniuk, "A max-affine spline perspective of recurrent neural networks," in *International Conference on Learning Representations*, 2018.

[26] D. LeJeune, R. Balestriero, H. Javadi, and R. G. Baraniuk, "Implicit rugosity regularization via data augmentation," *arXiv preprint arXiv:1905.11639*, 2019.

[27] A. Hernández-García and P. König, "Do deep nets really need weight decay and dropout?," *arXiv preprint arXiv:1802.07042*, 2018.

[28] A. Hernández-García and P. König, "Data augmentation instead of explicit regularization," *arXiv preprint arXiv:1806.03852*, 2018.

[29] A. Hernández-García and P. König, "Further advantages of data augmentation on convolutional neural networks," in *International Conference on Artificial Neural Networks*, pp. 95–103, Springer, 2018.

[30] N. Razin and N. Cohen, "Implicit regularization in deep learning may not be explainable by norms," *arXiv preprint arXiv:2005.06398*, 2020.

[31] S. G. Mallat, "Multifrequency channel decompositions of images and wavelet models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 2091–2110, 1989.

[32] P. Heckbert, "Color image quantization for frame buffer display," *ACM Siggraph Computer Graphics*, vol. 16, no. 3, pp. 297–307, 1982.

[33] A. A. Sawchuk, "Space-variant image restoration by coordinate transformations," *JOSA*, vol. 64, no. 2, pp. 138–144, 1974.

[34] G. Wolberg and S. Zokai, "Robust image registration using log-polar transform," in *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, vol. 1, pp. 493–496, IEEE, 2000.

[35] R. Mukundan, S. Ong, and P. A. Lee, "Image analysis by tchebichef moments," *IEEE Transactions on image Processing*, vol. 10, no. 9, pp. 1357–1364, 2001.

[36] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, pp. 2017–2025, 2015.

[37] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[38] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," 1970.

[39] M. Rosenblatt, "A central limit theorem and a strong mixing condition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 42, no. 1, p. 43, 1956.

[40] Y. Tian, X. Chen, and S. Ganguli, "Understanding self-supervised learning dynamics without contrastive pairs," in *International Conference on Machine Learning*, pp. 10268–10278, PMLR, 2021.

[41] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.

[42] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[43] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.," in *International Conference on Learning Representations*, 2019.

[44] R. Balestriero, L. Bottou, and Y. LeCun, "The effects of regularization and data augmentation are class dependent," *arXiv preprint arXiv:2204.03632*, 2022.

[45] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[46] M. E. Newman and G. T. Barkema, *Monte Carlo methods in statistical physics*. Clarendon Press, 1999.

[47] J. L. Doob, "The limiting distributions of certain statistics," *The Annals of Mathematical Statistics*, vol. 6, no. 3, pp. 160–169, 1935.

[48] G. W. Oehlert, "A note on the delta method," *The American Statistician*, vol. 46, no. 1, pp. 27–29, 1992.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*, pp. 630–645, Springer, 2016.

[50] P. Simard, B. Victorri, Y. LeCun, and J. S. Denker, "Tangent prop-a formalism for specifying selected invariances in an adaptive network," in *NIPS*, vol. 91, pp. 895–903, Citeseer, 1991.

[51] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik, "Vicinal risk minimization," *Advances in neural information processing systems*, pp. 416–422, 2001.

[52] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio, and X. Muller, "The manifold tangent classifier," *Advances in neural information processing systems*, vol. 24, pp. 2294–2302, 2011.

Checklist

1. For all authors...

(a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We have precisely given the scope of our results in the end of the introduction. Additionally, we recall the limitations and assumptions of our results within each section carefully.

(b) Did you describe the limitations of your work? [Yes] The main limitations comes from the approximation done for nonlinear models which we have repeatedly highlight in both the introductions and the corresponding sections. We also highlight those in the conclusion along with future work.

(c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not see any potential negative societal impact for this work

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] we have carefully given the assumptions each result relies on. Especially, at the end of the introduction, we have highlighted a reference paper where it was shown that our assumptions are common in the literature, and that they are good enough to gain insights into nonlinear models.

(b) Did you include complete proofs of all theoretical results? [Yes] Each formal statement and derivation is linked to a proof in the appendix. We also provide details derivations in the appendix of all the equations we have obtained in the main paper, even if they are not part of a formal result.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide samples of the code to reproduce figures. The entire codebase will be released publicly upon completion of the review-process

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We only explore MNIST and provided the training details in the caption of the corresponding figure

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We did not provide error bars but multiple runs with various settings

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The employed experiment is small scale and can be performed without GPU

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] MNIST only

(b) Did you mention the license of the assets? [No]

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]