Toward Understanding Generative Data Augmentation

Chenyu Zheng[1] Guoqiang Wu[2] Chongxuan Li[1]

[1]Gaoling Schoold of AI, Renmin University of China [2]School of Software, Shandong University

Abstract

Generative data augmentation, which scales datasets by obtaining fake labeled examples from a trained conditional generative model, boosts classification performance in various learning tasks including (semi-)supervised learning, few-shot learning, and adversarially robust learning. However, little work has theoretically investigated the effect of generative data augmentation. To fill this gap, we establish a general stability bound in this not independently and identically distributed (non-i.i.d.) setting, where the learned distribution is dependent on the original train set and generally not the same as the true distribution. Our theoretical result includes the divergence between the learned distribution and the true distribution. It shows that *generative data augmentation can enjoy a faster learning rate when the order of divergence term is* $o(\max\left(\log(m)\beta_m, 1/\sqrt{m}\right))$, where m is the train set size and β_m is the corresponding stability constant. We further specify the learning setup to the Gaussian mixture model and generative adversarial nets. We prove that *in both cases, though generative data augmentation does not enjoy a faster learning rate, it can improve the learning guarantees at a constant level when the train set is small, which is significant when the awful overfitting occurs*. Simulation results on the Gaussian mixture model and empirical results on generative adversarial nets support our theoretical conclusions. Our code is available at *https://github.com/ML-GSAI/Understanding-GDA*.

Highlights

- We establish a **general theoretical framework** for the GDA in the supervised classification setting.
- We particularize the general results to the **binary Gaussian mixture model (bGMM)** and **generative adversarial nets (GANs)**.
- We conduct simulation and empirical experiments in both cases to **validate our theoretical findings**.

Notations and definitions

We introduce some basic notations and definitions in learning theory.

- Data:
  - Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the input space and $\mathcal{Y}$ be the label space.
  - We denote by $\mathcal{D}$ the population distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.
  - Given a set $S = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}$, we define $S^{\setminus i}$ as the set after removing the i-th data point in the set S, and S^i as the set after replacing the i-th data point with $\mathbf{z}_i'$ in the set S.
- Learning algorithm:
  - Let $\mathcal{A}$ be a learning algorithm.
  - Let $\mathcal{A}(S) \in (\mathcal{Y})^{\mathcal{X}}$ be the hypothesis learned on the dataset S.
- Evaluation:
  - Loss function $\ell : (\mathcal{Y})^{\mathcal{X}} \times \mathcal{Z} \to \mathbb{R}_+$.
  - True error $\mathcal{R}_{\mathcal{D}}(\mathcal{A}(S))$ with respect to the data distribution $\mathcal{D}$ is defined as $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(\mathcal{A}(S), \mathbf{z})]$.
  - Empirical error $\widehat{\mathcal{R}}_S(\mathcal{A}(S))$ is defined as $\frac{1}{m}\sum_{i=1}^{m} \ell(\mathcal{A}(S), \mathbf{z}_i)$.

We then introduce some notations and definitions for the GDA.

- **Training generative model:** given a dataset S with m_S i.i.d. examples from $\mathcal{D}$, we can train a conditional generative model G with the model distribution $\mathcal{D}_G(S)$.
- **GDA:** we then obtain a new dataset S_G with m_G i.i.d. samples from $\mathcal{D}_G(S)$, where m_G is a hyperparameter.
- We denote the total number of the data in augmented set $\widetilde{S} = S \cup S_G$ by m_T.
- We define the mixed distribution after augmentation as $\widetilde{\mathcal{D}}(S) = \frac{m_S}{m_T}\mathcal{D} + \frac{m_G}{m_T}\mathcal{D}_G(S)$

Our goal

We interested in the generalization error *Gen-error* $= |\mathcal{R}_{\mathcal{D}}(\mathcal{A}(\widetilde{S})) - \widehat{\mathcal{R}}_{\widetilde{S}}(\mathcal{A}(\widetilde{S}))|$. We will derive a high probability bound for it by using the algorithmic stability technique.

General Generalization bound for GDA

Theorem 1(Generalization bound for GDA). Assume that $\mathcal{A}$ is a β_m-stable learning algorithm and the loss function ℓ is bounded by M. Given an augmented set $\widetilde{S}$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that

$$|\text{Gen-error}| \lesssim \underbrace{\frac{m_G}{m_T} M \mathcal{D}_{\mathrm{TV}}\left(\mathcal{D}, \mathcal{D}_G(S)\right)}_{\text{Distributions' divergence}} + \frac{M(\sqrt{m_S}+\sqrt{m_G}) + m_S\sqrt{m_G}\beta_{m_T}}{m_T}\sqrt{\log\left(\frac{1}{\delta}\right)}$$
$$+ \frac{\beta_{m_T}\left(m_S \log m_S + m_G \log m_G\right) + m_S \log m_S M \mathcal{T}(m_S, m_G)}{m_T} \log\left(\frac{1}{\delta}\right),$$

where $\mathcal{T}(m_S, m_G) = \sup_i \mathcal{D}_{\mathrm{TV}}\left(\mathcal{D}_G^{m_G}(S), \mathcal{D}_G^{m_G}(S^i)\right)$.

Remark (Selection of augmentation size). An efficient augmentation size $m^*_{G,\text{order}}$ with regard to the order of m_S can be defined as:

$$m^*_{G,\text{order}} = \inf_{m_G} \left\{\text{generalization error w.r.t. mixed distribution} \lesssim \text{distributions' divergence}\right\}.$$

Corollary 2(Sufficient conditions for GDA with (no) faster learning rate). Assume the assumptions in Theorem 1 hold, then

- if $\mathcal{D}_{\mathrm{TV}}\left(\mathcal{D}, \mathcal{D}_G(S)\right) = o\left(\max\left(\log(m)\beta_m, 1/\sqrt{m}\right)\right)$, then GDA enjoys a faster learning rate.
- if $\mathcal{D}_{\mathrm{TV}}\left(\mathcal{D}, \mathcal{D}_G(S)\right) = \Omega\left(\max\left(\log(m)\beta_m, 1/\sqrt{m}\right)\right)$, then GDA can not enjoy a faster learning rate.

Theoretical results for binary Gaussian mixture model

Theorem 3 (Generalization bound for bGMM). Given a set S with m_S i.i.d. samples from the bGMM distribution $\mathcal{D}$ and an augmented set S_G with m_G i.i.d. samples drawn from the learned Gaussian mixture distribution, then with high probability at least $1-\delta$, it holds that

$$|\textit{Gen-error}| \lesssim \begin{cases} \frac{\log(m_S)}{\sqrt{m_S}} & \text{if fix } d \text{ and } m_G = 0, \\ \frac{\log^2(m_S)}{\sqrt{m_S}} & \text{if fix } d \text{ and } m_G = \Theta(m_S), \\ \frac{\log(m_S)}{\sqrt{m_S}} & \text{if fix } d \text{ and } m_G = m^*_{G,\text{order}}, \quad \textbf{(No faster learning rate)} \\ d & \text{if fix } m_S. \quad \textbf{(Improvement at a constant level)} \end{cases}$$

Theoretical results for binary Gaussian mixture model

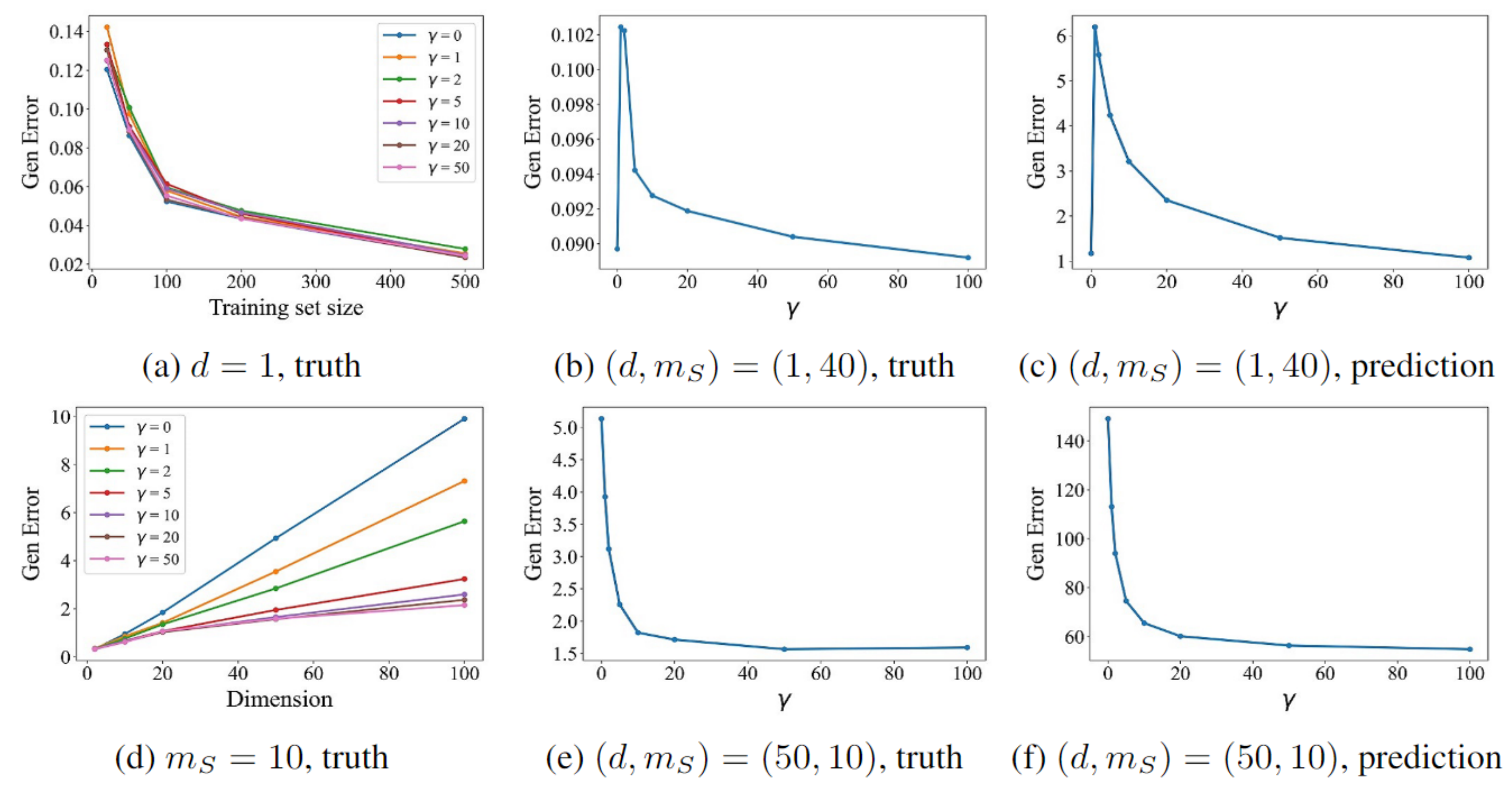
Theorem 4(Generalization bound for GAN). Given a set S with m_S i.i.d. samples from any distribution $\mathcal{D}$ and an augmented set S_G with m_G i.i.d. examples sampled from the distribution $\mathcal{D}_G(S)$ learned by GANs, then for any fixed $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that

$$\mathbb{E}|\textit{Gen-error}| \lesssim \begin{cases} \frac{1}{\sqrt{m_S}} & \text{if fix } W, L, d, \text{ let } m_G = 0, \\ \left(\frac{\log(m_S)}{m_S}\right)^{\frac{1}{4}} & \text{if fix } W, L, d, \text{ let } m_G = m^*_{G,\text{order}}, \quad \textbf{(Worse rate)} \\ dL^2\left(\prod_{l=1}^{L} \|W_l\|_2\right)^2 & \text{if fix } m_S. \quad \textbf{(Improvement at a constant level)} \end{cases}$$

Simulation results on binary Gaussian mixture model

We validate Theorem 3 on a **binary mixture of Gaussian distribution**.

1. We investigate the case that data dimension d is fixed ($d=1$),
2. We conduct simulations in the case that m_S is fixed as a small constant ($m_S = 10$),
3. We design experiments to validate whether the explicit upper bound can predict the trend of generalization error.



(a) $d = 1$, truth (b) $(d, m_S) = (1, 40)$, truth (c) $(d, m_S) = (1, 40)$, prediction

(d) $m_S = 10$, truth (e) $(d, m_S) = (50, 10)$, truth (f) $(d, m_S) = (50, 10)$, prediction

Figure 1. Simulations results with $\boldsymbol{\mu} = (1/\sqrt{d}, \ldots, 1/\sqrt{d})^\top$ and $\sigma^2 = 0.6^2$

Empirical results on GANs

GANs are chosen to empirically validate Theorem 4.

1. We choose a **"good" GAN (StyleGAN2-ADA)** to verify that GANs can not improve the test performance obviously when the m_S is approximately large (with S.A.).
2. We choose a **"bad" GAN (DCGAN)** to empirically verify that GANs can improve the test performance when m_S is small and awful overfitting happens (without S.A.).

Table 1. Accuracy on the CIFAR-10 test set, where S.A. denotes standard augmentation.

Generator	Classifier	S.A.	GDA (m_G)					
			0	100k	300k	500k	700k	1M
cDCGAN	ResNet18	×	85.76	86.8	87.83	87.59	87.52	86.47
		√	94.4	93.92	93.41	93.81	93.01	92.6
	ResNet34	×	85	86.9	87.93	87.56	87.17	86.28
		√	94.59	94.83	94.21	93.64	93.69	93.18
	ResNet50	×	82.85	87.49	88.59	86.67	86.3	85.2
		√	94.69	94.43	93.86	93.74	93.12	92.63
StyleGAN2-ADA	ResNet18	×	85.76	90.22	91.33	91.37	91.25	91.38
		√	94.4	94.68	94.46	94.4	94.11	94.12
	ResNet34	×	85	90.24	91.23	91.45	91.56	90.91
		√	94.59	95.05	94.9	94.4	94.43	94.21
	ResNet50	×	82.85	90.85	92.29	92.29	92.29	91.61
		√	94.69	94.74	95.04	94.56	94.76	94.28
EDM	ResNet18	×	85.76	92.8	94.87	95.43	96.24	96.28
		√	94.4	96.15	96.74	97.09	97.28	97.5
	ResNet34	×	85	93.42	94.93	95.59	96.14	96.44
		√	94.59	96.47	96.96	97.36	97.53	97.51
	ResNet50	×	82.85	93.29	95.29	95.95	96.1	96.64
		√	94.69	96.09	96.87	97.28	97.6	97.74

Project page: https://github.com/ML-GSAI/Understanding-GDA

NeurIPS 2023, Poster

Emails: chongxuanli@ruc.edu.cn; chenyu.zheng666@gmail.com