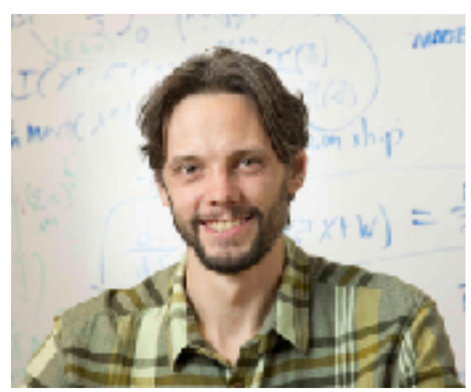
The Gaussian equivalence of generative models for learning with shallow neural networks

Sebastian Goldt, Bruno Loureiro, Galen Reeves,
Florent Krzakala, Marc Mézard, and Lenka Zdeborová

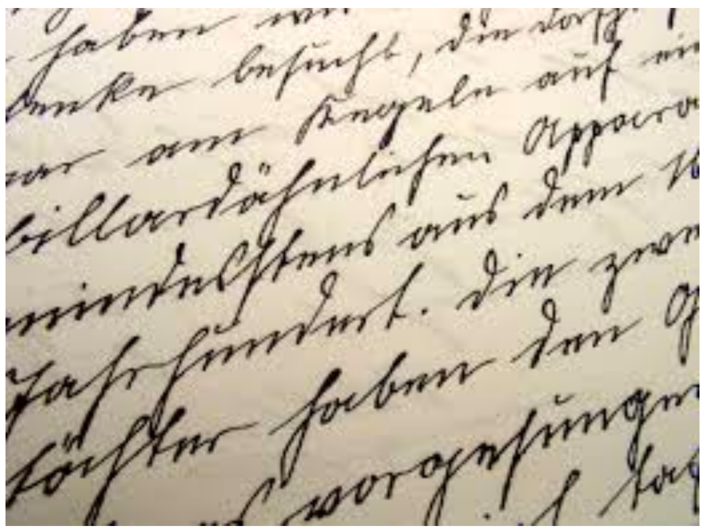
MSML 2021






The impact of data structure on learning

The data sets we care about in machine learning contain a lot of structure.



Written text (NLP)

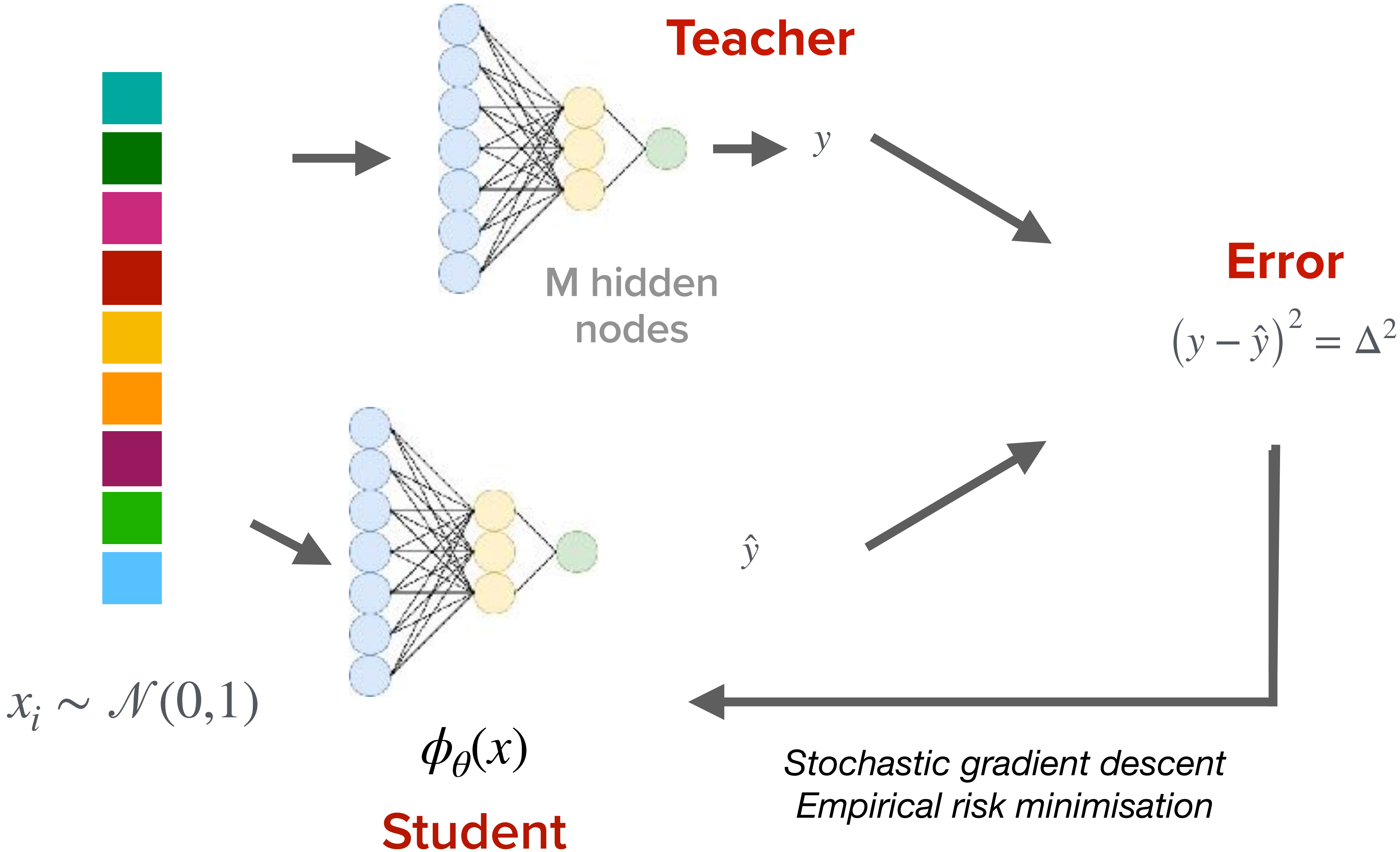
Images

Games of Go

How does data structure impact learning in neural networks?

The teacher-student setup

Gardner & Derrida (1989)
Seung, Sompolinsky, Tishby (1993)



Goal: $$\mathrm{pmse}(\theta, \tilde{\theta}) = \mathbb{E}_{q(x)} \left[\sum_k^K v^k g\left(w^k x\right) - \sum_m^M \tilde{v}^m g\left(\tilde{w}^m x\right) \right]^2$$

The Gaussian Equivalence Property

Goal: compute the prediction mean-squared error at all times.

For the **vanilla-teacher student** with i.i.d. inputs *x:*

Saad & Solla, (1995)
Biehl & Schwarze (1995)

$$\text{pmse}\left(\theta,\tilde{\theta}\right)=\mathbb{E}_x\left(\sum_{k=1}^{K}v^k g\left(w^k x\right)-\sum_{m=1}^{M}\tilde{v}^m\tilde{g}\left(\tilde{w}^m x\right)\right)^2$$

Student network
(trying to learn)

Teacher network
(creates the data)

The Gaussian Equivalence Property

Goal: compute the prediction mean-squared error at all times.

For the **vanilla-teacher student** with i.i.d. inputs *x:*

Saad & Solla, (1995)
Biehl & Schwarze (1995)

$$\mathrm{pmse}\left(\theta, \tilde{\theta}\right) = \mathop{\mathbb{E}}_{x} \left(\sum_{k=1}^{K} v^k g\left(w^k x\right) - \sum_{m=1}^{M} \tilde{v}^m \tilde{g}\left(\tilde{w}^m x\right) \right)^2$$

Average over the inputs x

$\lambda^k \sim w^k x$

$\nu^m \sim \tilde{w}^m x$

The Gaussian Equivalence Property

Goal: compute the prediction mean-squared error at all times.

For the **vanilla-teacher student** with i.i.d. inputs *x:*

Saad & Solla, (1995)
Biehl & Schwarze (1995)

$$\mathrm{pmse}\left(\theta, \tilde{\theta}\right) = \underset{\lambda,\nu}{\mathbb{E}} \left(\sum_{k=1}^{K} v^k g\left(\lambda^k\right) - \sum_{m=1}^{M} \tilde{v}^m \tilde{g}\left(\nu^m\right) \right)^2$$

Average over the local fields (λ, ν)

$\lambda^k \sim w^k x$

$\nu^m \sim \tilde{w}^m x$

Key random variables
for online learning
and replicas (batch)

The Gaussian Equivalence Property

Goal: compute the prediction mean-squared error at all times.

Saad & Solla, (1995)
Biehl & Schwarze (1995)

For the **vanilla-teacher student** with i.i.d. inputs *x:*

$$\mathrm{pmse}\left(\theta, \tilde{\theta}\right) = \mathbb{E}_{\lambda,\nu}\left(\sum_{k=1}^{K} v^k g\left(\lambda^k\right) - \sum_{m=1}^{M} \tilde{v}^m \tilde{g}\left(\nu^m\right)\right)^2$$

$$\mathbb{E}\, x_i x_j = \delta_{ij}$$

$$\lambda^k \sim \sum_i w_i^k x_i$$

$$\nu^m \sim \sum_i \tilde{w}_i^m x_i$$

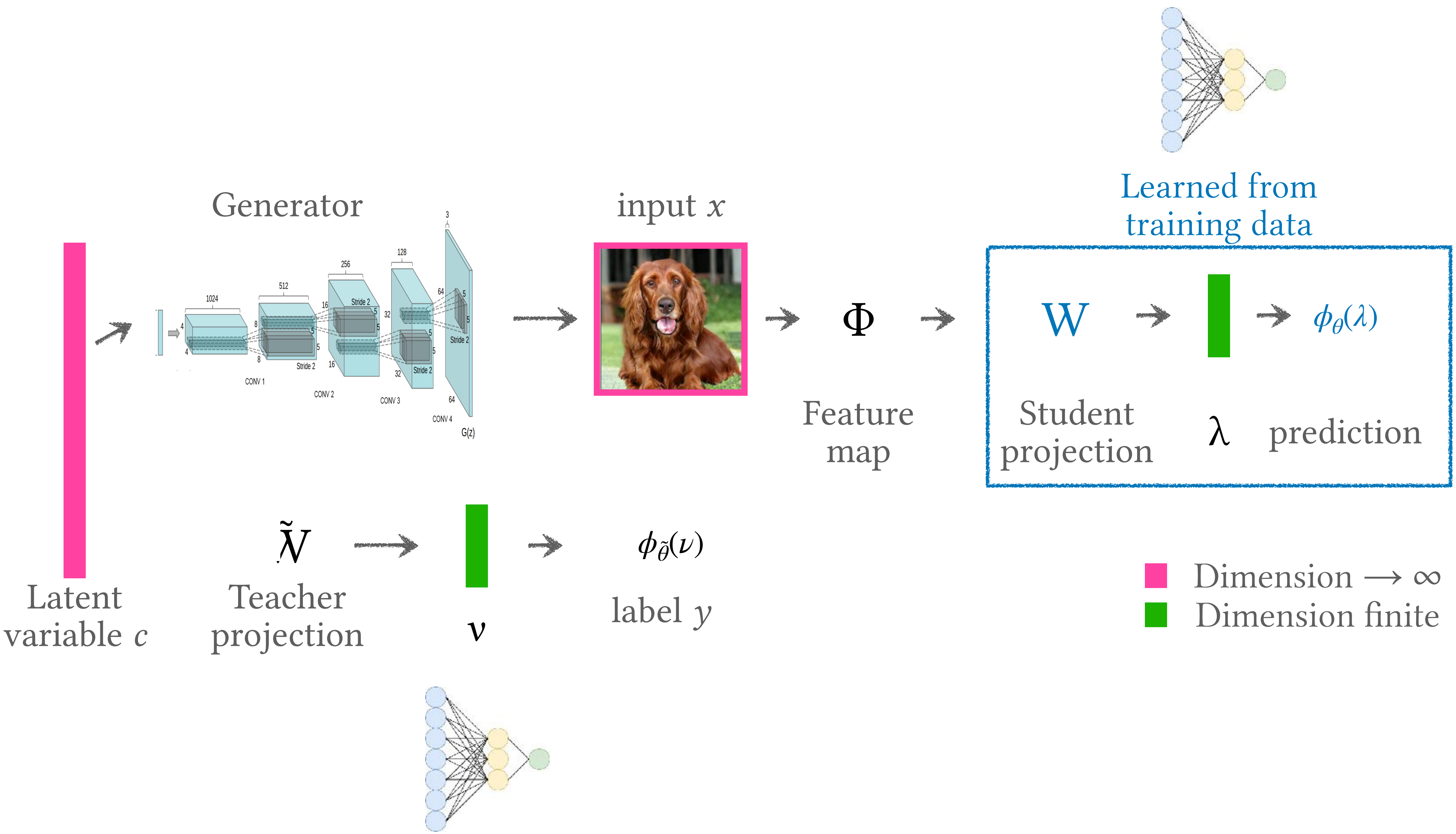
Gaussian Equivalence Property:
(λ, ν) are jointly Gaussian

Hence, the *pmse* is a function of only the second moments of (λ, ν):

$$Q^{k\ell} \equiv \mathbb{E}\, \lambda^k \lambda^\ell, \quad R^{km} \equiv \mathbb{E}\, \lambda^k \nu^m, \quad T^{mn} \equiv \mathbb{E}\, \nu^m \nu^n$$

The hidden manifold model

SG, M. Mézard, F. Krzakala, L. Zdeborová
Phys. Rev. X, **10** (4), 041044



Our contributions

Gaussian Equivalence Theorem

We give rigorous conditions under which we can analyse learning from data coming from single-layer generators.

Dynamical equations for two-layer students

The equations track the test error of two-layer students trained on deep generative models.

Replica analysis for random feature regression

Closed set of fixed point equations that characterise the performance after full-batch training.

The Gaussian Equivalence Theorem

Setup: Fully connected, single layer generator $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^N$

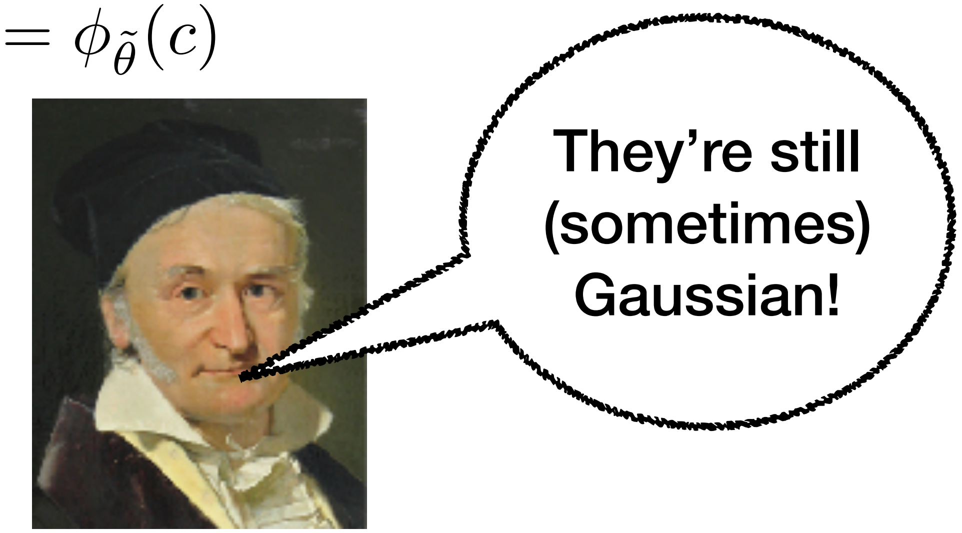
$$x_n = \mathcal{G}_n(c) = \sigma(a_n^\top c)$$

with the teacher acting on the latent variable *c:* $y = \phi_{\tilde{\theta}}(c)$

$$\boxed{\mathbb{E}\, x_i x_j = \Omega_{ij}}$$

$$\lambda^k \sim \sum_i w_i^k x_i$$

$$\nu^m \sim \sum_r \boxed{\tilde{w}_r^m c_r}$$



Theorem: Let *P* be the distribution of the pair (λ, ν) and let $\hat{P}$ be the Gaussian distribution with the same first and second moments. Then…

$$d_{\mathrm{MS}}(P, \hat{P}) = O\left(\left\| \frac{1}{\sqrt{N}} W M_1^{1/2} \right\|^2 + \left\| \frac{1}{\sqrt{N}} W M_2^{1/2} \right\| + \frac{1}{\sqrt{N}} \left\| \frac{1}{\sqrt{D}} \tilde{W} A^\top \right\|^2 + \frac{1}{\sqrt{N}} \right)$$

The Gaussian Equivalence Theorem

Theorem: Let P be the distribution of the pair (λ, ν) and let $\hat{P}$ be the Gaussian distribution with the same first and second moments. Then…

$$\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^N$$

$$x_n = \mathcal{G}_n(c) = \sigma(a_n^\top c)$$

$$y = \phi_{\tilde{\theta}}(c)$$

Generator weights

Student weights *Teacher weights*

$$d_{\mathrm{MS}}(P, \hat{P}) = O\left(\left\| \tfrac{1}{\sqrt{N}} W M_1^{1/2} \right\|^2 + \left\| \tfrac{1}{\sqrt{N}} W M_2^{1/2} \right\| + \tfrac{1}{\sqrt{N}} \left\| \tfrac{1}{\sqrt{D}} \tilde{W} A^\top \right\|^2 + \frac{1}{\sqrt{N}} \right)$$

Related to input correlations

Related work

- Works in wide network limit rely on RMT and thus random weights
- Mei & Montanari; Couillet et al. introduce related equivalent Gaussian models for integrals w.r.t. spectral densities.
- Large body of work on low-dim projections of high-dim data being Gaussian - we quantify how Gaussian they look like.

Dynamical equations for two-layer students

Setup: Fully connected, single layer generator $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^N$

$$x_n = \mathcal{G}_n(c) = \sigma(a_n^\top c)$$

with the teacher acting on the latent variable *c:* $y = \phi_{\tilde{\theta}}(c)$

- Train the student using online SGD:

$$\theta_{\mu+1} = \theta_\mu - \eta \nabla_\theta \mathcal{L}(\theta)|_{\theta_\mu, x_\mu, y^*_\mu}$$

Goal: Derive a closed set of equations for the order parameters

$$Q^{k\ell} \equiv \mathbb{E}\,\lambda^k \lambda^\ell, \quad R^{km} \equiv \mathbb{E}\,\lambda^k \nu^m$$

that track the dynamics of a two-layer student
trained using online SGD on the deep hidden manifold.

Dynamical equations for two-layer students

Train the student using online SGD: $\theta_{\mu+1} = \theta_\mu - \eta \nabla_\theta \mathcal{L}(\theta)|_{\theta_\mu, x_\mu, y^*_\mu}$

Goal: Derive a closed set of equations for the order parameters

Saad & Solla (1995)
Biehl & Riegler (1995)

$$Q^{k\ell} \equiv \mathbb{E}\, \lambda^k \lambda^\ell, \quad R^{km} \equiv \mathbb{E}\, \lambda^k \nu^m$$

$$Q^{k\ell} = \int \mathrm{d}\mu_\Omega(\rho)\ \rho\ q^{k\ell}(\rho)$$

$$\frac{\partial q^{k\ell}(\rho)}{\partial t} = -\eta \Bigg(\rho \sum_{j \neq k}^{K} \left[v^k v^j q^{k\ell}(\rho) h^{kj}_{(1)}(Q) + v^k v^j q^{j\ell}(\rho) h^{kj}_{(2)}(Q) \right] + \rho v^k v^k q^{k\ell}(\rho) h^k_{(3)}(Q)$$
$$- v^k \sum_n^M \left[\rho \tilde{v}^n q^{k\ell}(\rho) h^{kn}_{(4)}(Q,R,T) + \frac{1}{\sqrt{\delta}} \tilde{v}^n r^{\ell n}(\rho) h^{kn}_{(5)}(Q,R,T) \right]$$
$$+ \text{all of the above with } \ell \to k, k \to \ell \Bigg) + \eta^2 \gamma v^k v^\ell h^{k\ell}_{(6)}(Q,R,T,v,\tilde{v}).$$

Spectral density of input-input covariance

$$R^{km} = \frac{1}{\sqrt{\delta}} \int \mathrm{d}\mu_\Omega(\rho)\ r^{km}(\rho)$$

$$\frac{\partial r^{km}(\rho)}{\partial t} = -\eta v^k \Bigg(\rho \sum_{j \neq k}^{K} \left[v^j r^{km}(\rho) h^{kj}_{(1)}(Q) + v^j \rho r^{jm}(\rho) h^{kj}_{(2)}(Q) \right] + v^k \rho r^{km}(\rho) h^k_{(3)}(Q)$$
$$- \sum_n^M \left[\rho \tilde{v}^n r^{km}(\rho) h^{kn}_{(4)}(Q,R,T) + \frac{1}{\sqrt{\delta}} \tilde{v}^n h^{kn}_{(5)}(Q,R,T) \right] \Bigg).$$

Dynamical equations for two-layer students

Statement:

$$Q^{k\ell} \equiv \mathbb{E}\,\lambda^k\lambda^\ell, \quad R^{km} \equiv \mathbb{E}\,\lambda^k\nu^m$$

$$Q^{k\ell} = \int \mathrm{d}\mu_\Omega(\rho)\;\rho\;q^{k\ell}(\rho) \qquad\qquad R^{km} = \frac{1}{\sqrt{\delta}}\int \mathrm{d}\mu_\Omega(\rho)\;r^{km}(\rho)$$

Remarkably, the generator only appears via two covariance matrices:

$$\Omega_{ij} = \mathbb{E}\;x_i x_j \qquad\qquad \Phi_{ir} = \mathbb{E}\;x_i c_r$$

Input-input correlations

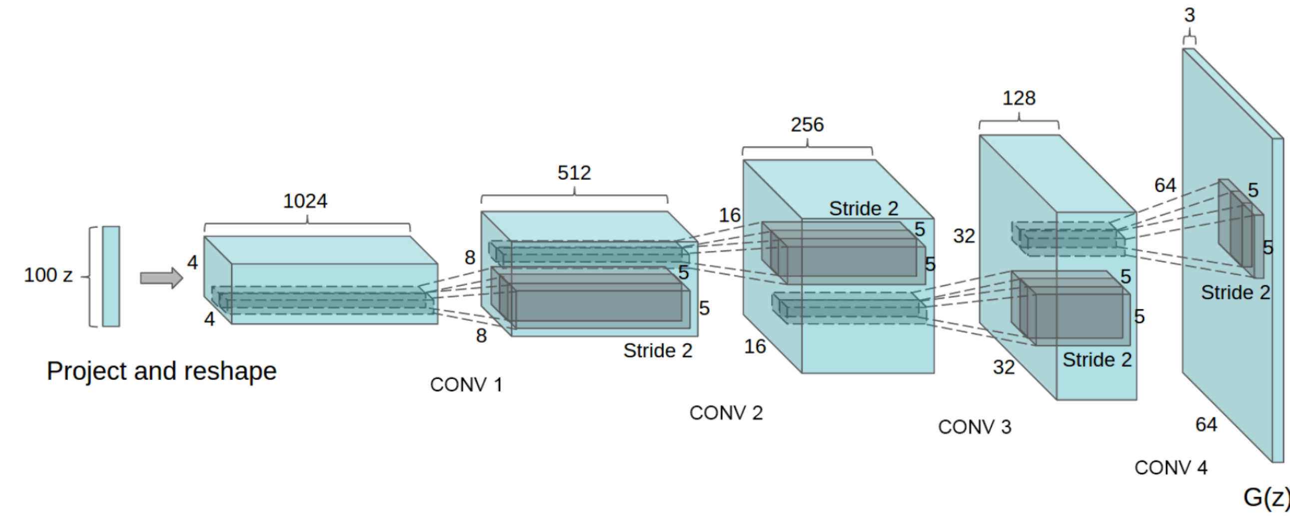
Input-latent correlations

Testing the equations with deep generators

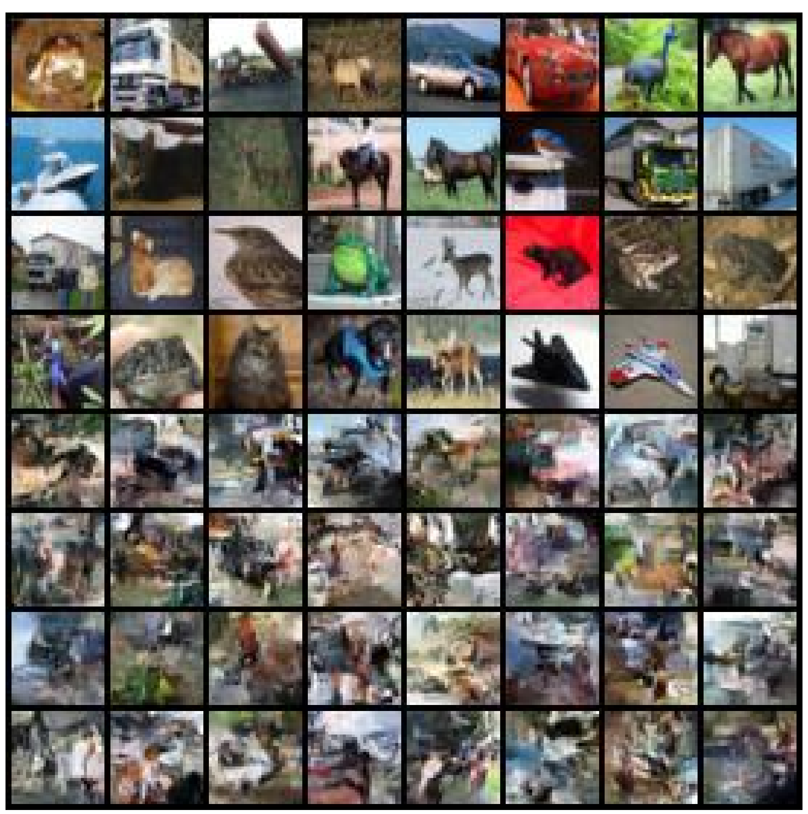
Used pre-trained dcGAN (Radford ’15) and normalising flows (Dinh ’17) to generate inputs

$$x = \mathcal{G}(c) = \mathcal{G}^L \cdots \mathcal{G}^3 \circ \mathcal{G}^2 \circ \mathcal{G}^1(c)$$

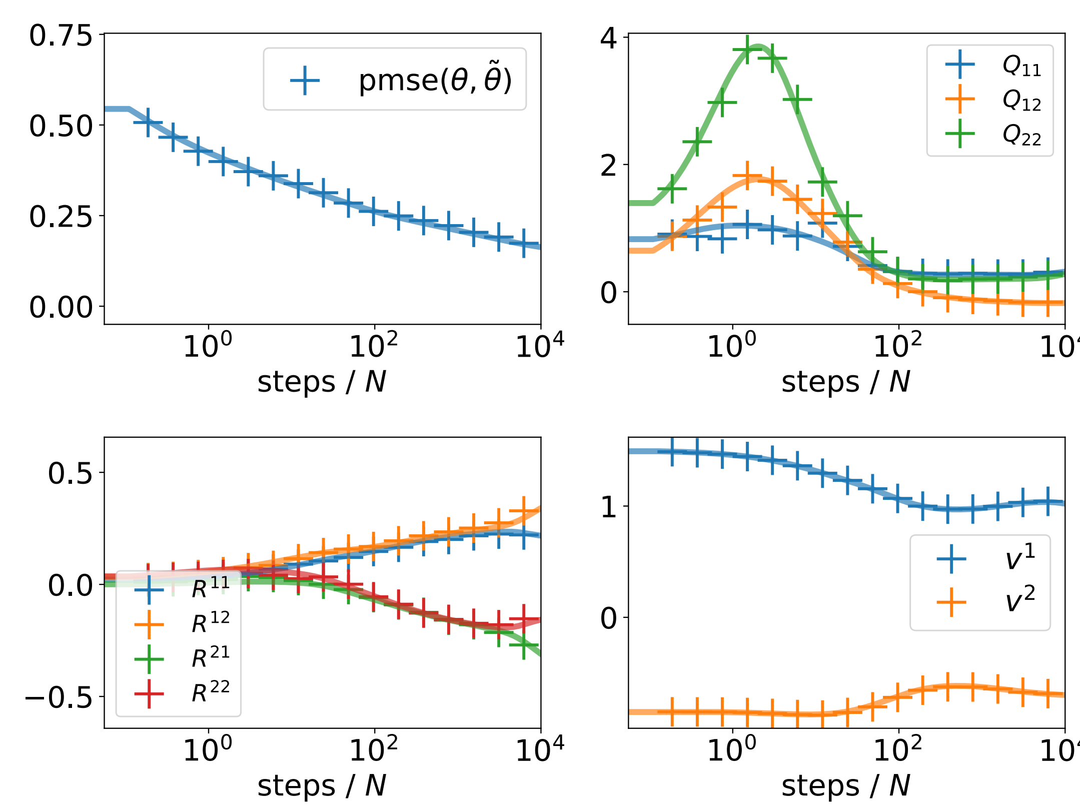
$$c \sim \mathcal{N}(0, I_D) \qquad y = \phi_{\tilde{\theta}}(c)$$



Deep Convolutional GAN (Radford et al., ICLR 2016)



Top half: CIFAR10 images
Bottom half: realNVP samples

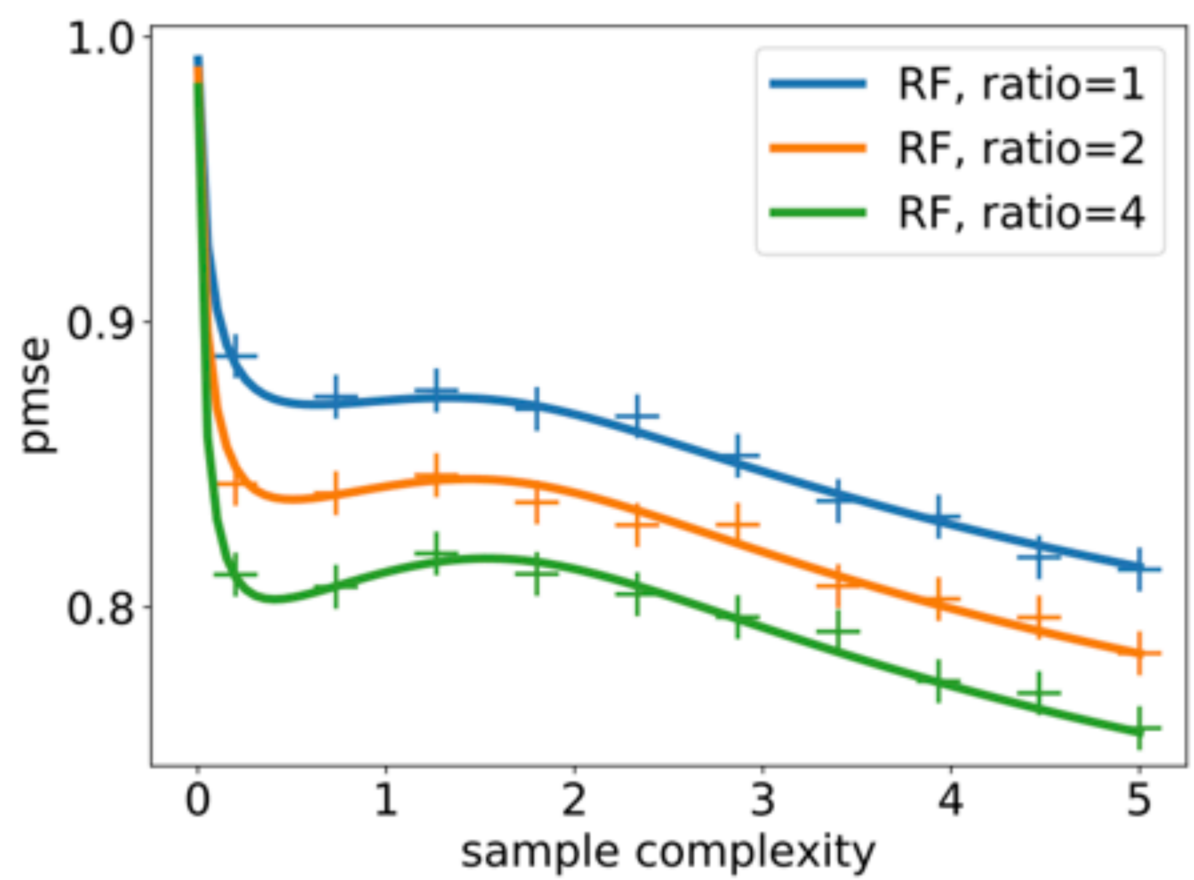


M=K=2, $\eta = 0.2$, D=3072, N=3072

The batch case: random-features logistic regression

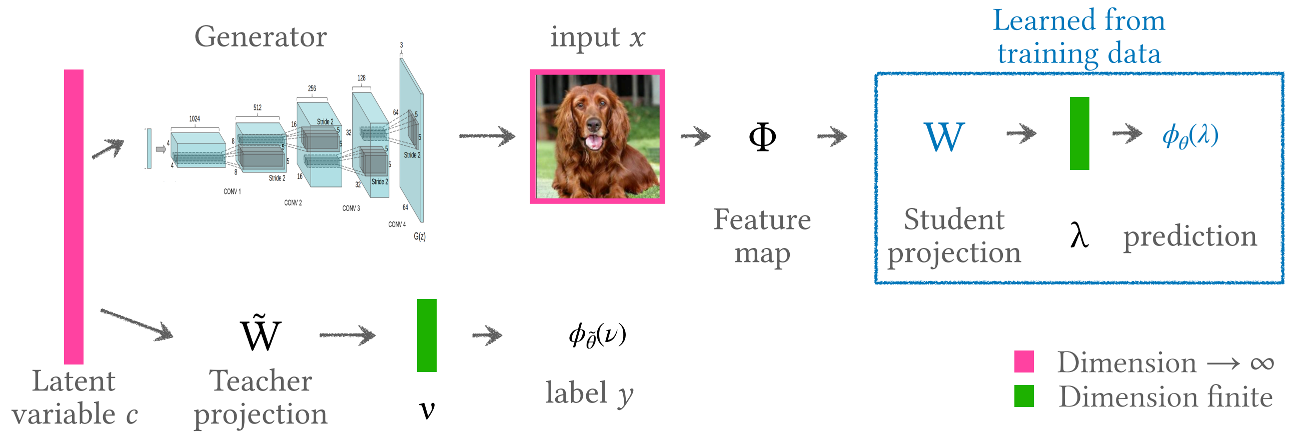
- Replica calculation provides generalisation error of full-batch logistic regression with random features.

Top half: Graysacle CIFAR10 images
Bottom half: Samples from dcGAN
(Radford et al. ’15)



Fixed weight decay λ = 10⁻².

Concluding perspectives



$$\begin{bmatrix} \nu \\ \lambda \end{bmatrix} \in \mathbb{R}^{K+M} \sim \mathcal{N}\left(0, \begin{bmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{bmatrix}\right)$$

- Proof of convergence for empirical risk

B. Loureiro, C. Gerbelot, H. Cui
SG, M. Mézard, F. Krzakala, L. Zdeborová,
arXiv:**2102.08127**

  - Complementary proof of risk convergence: Hu & Lu (arXiv:2009.07669)

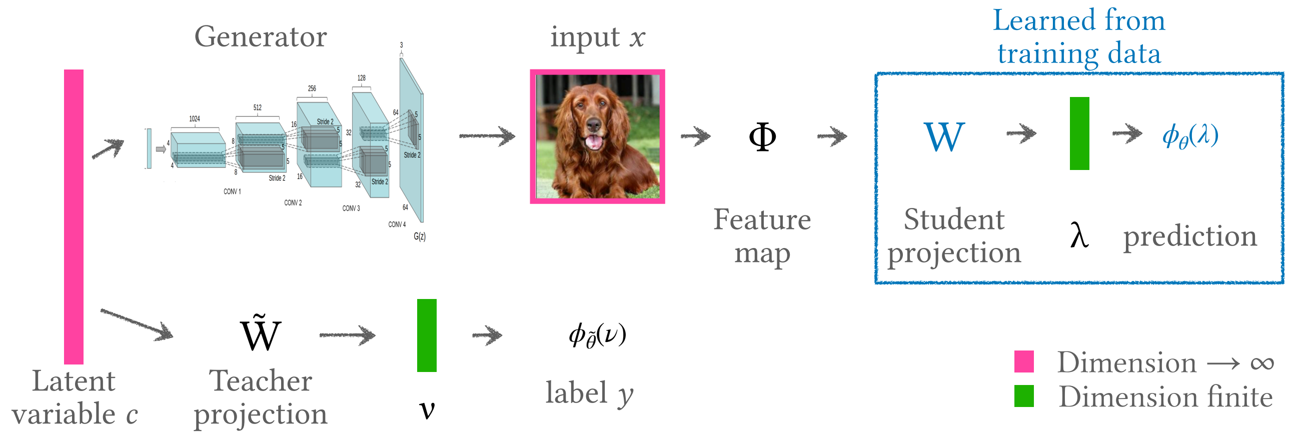
Theorem 1. *(Training loss and generalisation error) Under Assumption (C.1), there exist constants $C, c, c' > 0$ such that, for any optimal solution $\hat{\mathbf{w}}$ to (1.3), the training loss and generalisation error respectively defined by equations (2.2) and (2.3) verify, for any $0 < \epsilon < c'$:*

$$\mathbb{P}\left(\left|\mathcal{E}_{\mathrm{train}}(\hat{\mathbf{w}}) - \mathcal{E}^*_{\mathrm{train}}\right| \geqslant \epsilon\right) \leqslant \frac{C}{\epsilon} e^{-cn\epsilon^2}, \tag{2.10}$$

$$\mathbb{P}\left(\left|\mathcal{E}_{\mathrm{gen}}(\hat{\mathbf{w}}) - \mathbb{E}_{\omega,\xi}\left[\hat{g}(f_0(\omega), \hat{f}(\xi))\right]\right| \geqslant \epsilon\right) \leqslant \frac{C}{\epsilon} e^{-cn\epsilon^2},$$

Concluding perspectives

B. Loureiro, C. Gerbelot, H. Cui
SG, M. Mézard, F. Krzakala, L. Zdeborová,
arXiv:**2102.08127**



$$\begin{bmatrix}\nu \\ \lambda\end{bmatrix} \in \mathbb{R}^{K+M} \sim \mathcal{N}\left(0, \begin{bmatrix}\Psi & \Phi \\ \Phi^\top & \Omega\end{bmatrix}\right)$$

- Proof of convergence for empirical risk
  - Complementary proof of risk convergence: Hu & Lu (arXiv:2009.07669)
- Pre-trained teacher with static feature map for more realistic learning curves.

Goals: Establish the limits of Gaussian equivalence, go beyond Gaussian models of data!