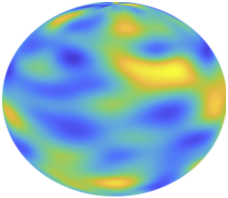
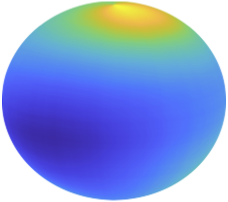
On Energy-Based Models with Overparametrized Shallow Neural Networks

Carles Domingo-Enrich[a], Alberto Bietti[b],
Eric Vanden-Eijnden[a], Joan Bruna[a,b]
[a]Courant Institute (NYU), [b]Center for Data Science (NYU)

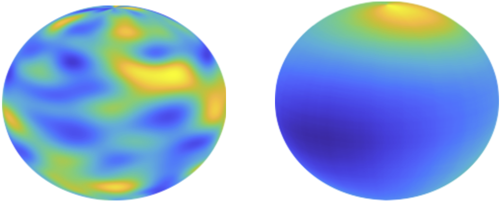
ICML 2021
July, 2021

Introduction

– In regression, learning a generic Lipschitz function in dimension d up to error ϵ requires a number of samples of order $\epsilon^{-\Omega(d)}$: *curse of dimensionality!*

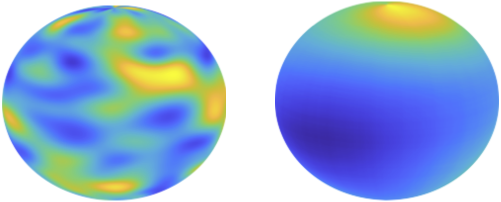
Introduction



– In regression, learning a generic Lipschitz function in dimension d up to error ϵ requires a number of samples of order $\epsilon^{-\Omega(d)}$: *curse of dimensionality!*

– If the target function has certain k-dimensional structure, then only $\epsilon^{-O(k)}$ are needed [Bach, 2017].

Introduction



– In regression, learning a generic Lipschitz function in dimension d up to error ϵ requires a number of samples of order $\epsilon^{-\Omega(d)}$: *curse of dimensionality!*

– If the target function has certain k-dimensional structure, then only $\epsilon^{-O(k)}$ are needed [Bach, 2017].

– In our work, we develop similar adaptivity results for the task of learning distributions via energy-based models.

Background: generative modeling

– Central problem in ML: to learn generative models of a distribution through its samples.

Background: generative modeling

– Central problem in ML: to learn generative models of a distribution through its samples.

– One approach: implicit generative modeling.
Black-box, no meaningful estimates computed, e.g. GANs, normalizing flows.

Background: generative modeling

– Central problem in ML: to learn generative models of a distribution through its samples.

– One approach: implicit generative modeling. Black-box, no meaningful estimates computed, e.g. GANs, normalizing flows.

– Another approach: explicit generative modeling. Estimates of the density/energy computed and used to generate samples, e.g **energy-based models (EBMs)**.

Background: EBMs (1)

– Let $K \subseteq \mathbb{R}^{d+1}$ with base probability measure τ.

– EBMs: learned models are Gibbs measures $\nu_f \in \mathcal{P}(K)$ defined through an *energy function* $f : K \to \mathbb{R}$, with a density proportional to $\exp(-f(x))$:

$$\frac{d\nu_f}{d\tau}(x) := \frac{e^{-f(x)}}{Z_f}, \text{ with } Z_f := \int_K e^{-f(y)} d\tau(y) \ .$$

Background: EBMs (1)

– Let $K \subseteq \mathbb{R}^{d+1}$ with base probability measure τ.

– EBMs: learned models are Gibbs measures $\nu_f \in \mathcal{P}(K)$ defined through an *energy function* $f : K \to \mathbb{R}$, with a density proportional to $\exp(-f(x))$:

$$\frac{d\nu_f}{d\tau}(x) := \frac{e^{-f(x)}}{Z_f}, \text{ with } Z_f := \int_K e^{-f(y)} d\tau(y) \ .$$

– Given samples $\{x_i\}_{i=1}^n$ from a target measure ν, training an EBM consists in selecting the best ν_f with energy f within a certain function class $\mathcal{F}$, according to a given criterion.

Background: EBMs (2)

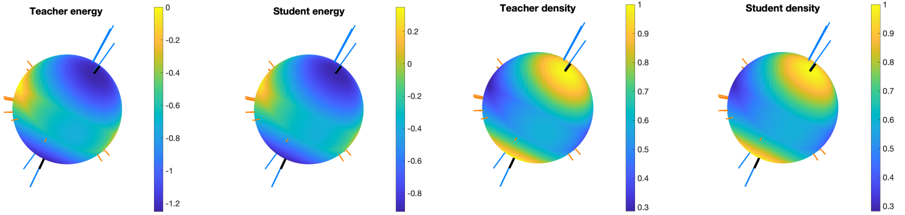


Figure: 3D synthetic EBM experiments.

Background: EBMs (2)

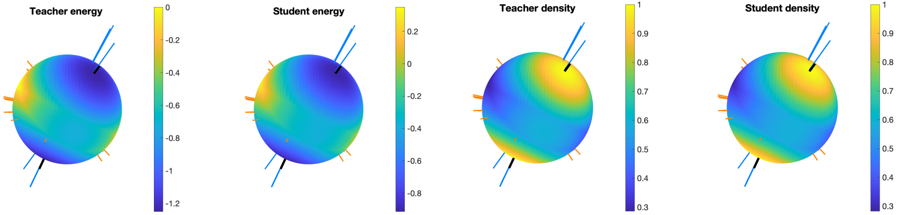


Figure: 3D synthetic EBM experiments.

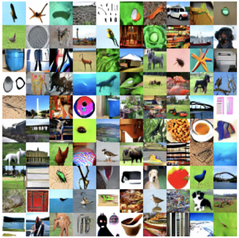


Figure: ImageNet 32x32 EBM samples from [Du and Mordatch, 2019].

Background: Overparametrized shallow NN

Two-layer overparametrized NN spaces provide a simplified and manageable framework to study neural networks.

Background: Overparametrized shallow NN

Two-layer overparametrized NN spaces provide a simplified and manageable framework to study neural networks. They come in two flavors [Bach, 2017]:

– Feature learning regime: $\mathcal{F}_1$ or Barron space [Barron, 1993]. Features are learned, weak theoretical optimization guarantees (works well in practice).

Background: Overparametrized shallow NN

Two-layer overparametrized NN spaces provide a simplified and manageable framework to study neural networks. They come in two flavors [Bach, 2017]:

– Feature learning regime: $\mathcal{F}_1$ or Barron space [Barron, 1993]. Features are learned, weak theoretical optimization guarantees (works well in practice).

– Kernel regime: $\mathcal{F}_2$ space. Features are fixed, it is an RKHS, smaller than $\mathcal{F}_1$, optimization has guarantees.

Contributions of the paper

We study the statistics of learning EBMs with overparametrized shallow NN energies ($\mathcal{F}_1$ and $\mathcal{F}_2$):

Contributions of the paper

We study the statistics of learning EBMs with overparametrized shallow NN energies ($\mathcal{F}_1$ and $\mathcal{F}_2$):

1. Generalization bounds for the learned measures in terms of training metrics (maximum likelihood, Stein discrepancies).

Contributions of the paper

We study the statistics of learning EBMs with overparametrized shallow NN energies ($\mathcal{F}_1$ and $\mathcal{F}_2$):

1. Generalization bounds for the learned measures in terms of training metrics (maximum likelihood, Stein discrepancies).
2. **Adaptivity to low-dimensional structure:** for energies in $\mathcal{F}_1$, target measures with low-dimensional structure can be learnt at a rate controlled by the intrinsic dimension, not the ambient dimension.

Contributions of the paper

We study the statistics of learning EBMs with overparametrized shallow NN energies ($\mathcal{F}_1$ and $\mathcal{F}_2$):

1. Generalization bounds for the learned measures in terms of training metrics (maximum likelihood, Stein discrepancies).
2. **Adaptivity to low-dimensional structure:** for energies in $\mathcal{F}_1$, target measures with low-dimensional structure can be learnt at a rate controlled by the intrinsic dimension, not the ambient dimension.
3. **Separation between $\mathcal{F}_1$ and $\mathcal{F}_2$:** experimentally, $\mathcal{F}_1$ energies can learn simple synthetic distributions with planted NN energies, $\mathcal{F}_2$ energies fail.

Framework: Maximum likelihood

– A natural estimator $\hat{f}$ for the energy is the **maximum likelihood** estimator (MLE), i.e.,
$\hat{f} = \operatorname{argmax}_{f \in \mathcal{F}} \prod_{i=1}^{n} \frac{d\nu_f}{d\tau}(x_i)$.

Framework: Maximum likelihood

– A natural estimator $\hat{f}$ for the energy is the **maximum likelihood** estimator (MLE), i.e., $\hat{f} = \mathrm{argmax}_{f \in \mathcal{F}} \prod_{i=1}^{n} \frac{d\nu_f}{d\tau}(x_i)$.

– Equivalently, $\hat{f}$ minimizes the cross-entropy with the samples:

$$\hat{f} = \underset{f \in \mathcal{F}}{\mathrm{argmin}}\, H(\nu_n, \nu_f) = \underset{f \in \mathcal{F}}{\mathrm{argmin}} -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{d\nu_f}{d\tau}(x_i) \right)$$

$$= \underset{f \in \mathcal{F}}{\mathrm{argmin}}\, \frac{1}{n} \sum_{i=1}^{n} f(x_i) + \log Z_f .$$

Framework: Maximum likelihood

– A natural estimator $\hat{f}$ for the energy is the **maximum likelihood** estimator (MLE), i.e.,
$\hat{f} = \mathrm{argmax}_{f \in \mathcal{F}} \prod_{i=1}^{n} \frac{d\nu_f}{d\tau}(x_i)$.
– Equivalently, $\hat{f}$ minimizes the cross-entropy with the samples:

$$\hat{f} = \underset{f \in \mathcal{F}}{\mathrm{argmin}}\, H(\nu_n, \nu_f) = \underset{f \in \mathcal{F}}{\mathrm{argmin}} -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{d\nu_f}{d\tau}(x_i) \right)$$

$$= \underset{f \in \mathcal{F}}{\mathrm{argmin}}\, \frac{1}{n} \sum_{i=1}^{n} f(x_i) + \log Z_f .$$

– The estimated distribution is simply $\nu_{\hat{f}}$, and samples can be obtained by the MCMC algorithm of choice.

Framework: NN energy classes (1)

We focus on two different energy classes $\mathcal{F}$. From now on, $\sigma(x) = \max\{0, x\}$ is the ReLu activation.

Framework: NN energy classes (1)

We focus on two different energy classes $\mathcal{F}$. From now on, $\sigma(x) = \max\{0, x\}$ is the ReLu activation.

Feature learning regime:

– Define $\mathcal{F}_1$ as the Banach space of functions $f : K \to \mathbb{R}$ such that for all $x \in K$ we have $f(x) = \int_{\mathbb{S}^d} \sigma(\langle \theta, x \rangle)\, d\gamma(\theta)$, for some signed Radon measure $\gamma \in \mathcal{M}(\mathbb{S}^d)$.

Framework: NN energy classes (1)

We focus on two different energy classes $\mathcal{F}$. From now on, $\sigma(x) = \max\{0, x\}$ is the ReLu activation.

Feature learning regime:

– Define $\mathcal{F}_1$ as the Banach space of functions $f : K \to \mathbb{R}$ such that for all $x \in K$ we have $f(x) = \int_{\mathbb{S}^d} \sigma(\langle \theta, x \rangle)\, d\gamma(\theta)$, for some signed Radon measure $\gamma \in \mathcal{M}(\mathbb{S}^d)$.

– The norm of $\mathcal{F}_1$ is defined as $\|f\|_{\mathcal{F}_1} = \inf \left\{ |\gamma|_{\mathrm{TV}} \mid f(\cdot) = \int_{\mathbb{S}^d} \sigma(\langle \theta, \cdot \rangle)\, d\gamma(\theta) \right\}$. $|\cdot|_{\mathrm{TV}}$ is the total variation norm.

– $\mathcal{F}$ is the ball $\mathcal{B}_{\mathcal{F}_1}(\beta)$ of radius $\beta > 0$ of $\mathcal{F}_1$.

Framework: NN energy classes (2)

Kernel regime:

– Define $\mathcal{F}_2$ as the RKHS of functions $f : K \to \mathbb{R}$ such that for some $h \in L^2(\mathbb{S}^d, \tau)$, we have that for all $x \in K$, $f(x) = \int_{\mathbb{S}^d} \sigma(\langle \theta, x \rangle) h(\theta) \; d\tau(\theta)$.

Framework: NN energy classes (2)

Kernel regime:

– Define $\mathcal{F}_2$ as the RKHS of functions $f : K \to \mathbb{R}$ such that for some $h \in L^2(\mathbb{S}^d, \tau)$, we have that for all $x \in K$, $f(x) = \int_{\mathbb{S}^d} \sigma(\langle \theta, x \rangle) h(\theta) \, d\tau(\theta)$.

– The RKHS norm of $\mathcal{F}_2$ is defined as
$\|f\|_{\mathcal{F}_2} = \inf \left\{ \|h\|_{L^2(\mathbb{S}^d,\tau)} \mid f(\cdot) = \int_{\mathbb{S}^d} \sigma(\langle \theta, \cdot \rangle) h(\theta) \, d\tau(\theta) \right\}$
where $\|h\|_{L^2(\mathbb{S}^d,\tau)}^2 := \int_{\mathbb{S}^d} |h(\theta)|^2 \, d\tau(\theta)$.

– $\mathcal{F}$ is the ball $\mathcal{B}_{\mathcal{F}_2}(\beta)$ of radius $\beta > 0$ of $\mathcal{F}_2$.

Framework: NN energy classes (3)

Quick facts:

– Cauchy-Schwarz inequality $\implies \mathcal{F}_2 \subset \mathcal{F}_1$ and $\mathcal{B}_{\mathcal{F}_2}(\beta) \subset \mathcal{B}_{\mathcal{F}_1}(\beta)$.

Framework: NN energy classes (3)

Quick facts:

– Cauchy-Schwarz inequality $\implies \mathcal{F}_2 \subset \mathcal{F}_1$ and $\mathcal{B}_{\mathcal{F}_2}(\beta) \subset \mathcal{B}_{\mathcal{F}_1}(\beta)$.

– [Bach, 2017] shows that single ReLU units belong to $\mathcal{F}_1$ but not to $\mathcal{F}_2$, and their L^2 approximations in $\mathcal{F}_2$ have exponentially high norm in the dimension.

Framework: NN energy classes (3)

Quick facts:

– Cauchy-Schwarz inequality $\implies \mathcal{F}_2 \subset \mathcal{F}_1$ and $\mathcal{B}_{\mathcal{F}_2}(\beta) \subset \mathcal{B}_{\mathcal{F}_1}(\beta)$.

– [Bach, 2017] shows that single ReLU units belong to $\mathcal{F}_1$ but not to $\mathcal{F}_2$, and their L^2 approximations in $\mathcal{F}_2$ have exponentially high norm in the dimension.

– The ball radius β determines the expressiveness. $\beta >> 1 \implies$ expressive models with lower approximation error but higher statistical error.

Statistical guarantees for MLE EBMs

Theorem

– Assume that the class $\mathcal{F}_\beta$ has a (distribution-free) Rademacher complexity bound $\mathcal{R}_n(\mathcal{F}_\beta) \leq \frac{\beta C}{\sqrt{n}}$ and L^∞ norm unif. bounded by β.

– Given samples $\{x_i\}_{i=1}^n$ from the target measure ν, consider the MLE $\hat{\nu} := \nu_{\hat{f}}$.

Statistical guarantees for MLE EBMs

Theorem

– Assume that the class $\mathcal{F}_\beta$ has a (distribution-free) Rademacher complexity bound $\mathcal{R}_n(\mathcal{F}_\beta) \leq \frac{\beta C}{\sqrt{n}}$ and L^∞ norm unif. bounded by β.
– Given samples $\{x_i\}_{i=1}^n$ from the target measure ν, consider the MLE $\hat{\nu} := \nu_{\hat{f}}$.
– If $\frac{d\nu}{d\tau}(x) \propto e^{-g(x)}$ for some $g : K \to \mathbb{R}$, i.e. $-g$ is the log-density of ν up to a constant term, then with probability at least $1-\delta$,

$$D_{KL}(\nu||\hat{\nu}) \leq \underbrace{\frac{4\beta C}{\sqrt{n}} + \beta\sqrt{\frac{8\log(1/\delta)}{n}}}_{\text{statistical error}} + \underbrace{2\inf_{f\in\mathcal{F}_\beta} \|g-f\|_\infty}_{\text{approximation error}}.$$

Adaptivity of MLE to low-dimensional structure (1)

Assumption (Low-dimensional structure)

– Let $K = K_0 \times \{R\}$*, where* $K_0 \subseteq \{x \in \mathbb{R}^d \,|\, \|x\|_2 \leq R\}$*.*

Adaptivity of MLE to low-dimensional structure (1)

Assumption (Low-dimensional structure)

– *Let* $K = K_0 \times \{R\}$, *where* $K_0 \subseteq \{x \in \mathbb{R}^d \,|\, \|x\|_2 \leq R\}$.

– *Suppose the target probability measure* ν *is absolutely continuous w.r.t.* τ, *with energy* $-\log\left(\frac{d\nu}{d\tau}(x, R)\right) = \sum_{j=1}^{J} \phi_j(U_j x)$, *where*

- ϕ_j *are* (ηR^{-1})*-Lipschitz continuous functions on the* R*-ball of* $\mathbb{R}^k$ *such that* $\|\phi_j\|_\infty \leq \eta$,
- *and* $U_j \in \mathbb{R}^{k \times d}$ *with orthonormal rows.*

Adaptivity of MLE to low-dimensional structure (1)

Assumption (Low-dimensional structure)

– *Let* $K = K_0 \times \{R\}$, *where* $K_0 \subseteq \{x \in \mathbb{R}^d \mid \|x\|_2 \leq R\}$.

– *Suppose the target probability measure* ν *is absolutely continuous w.r.t.* τ, *with energy* $-\log\left(\frac{d\nu}{d\tau}(x,R)\right) = \sum_{j=1}^{J} \phi_j(U_j x)$, *where*

- ϕ_j *are* (ηR^{-1})*-Lipschitz continuous functions on the* R*-ball of* $\mathbb{R}^k$ *such that* $\|\phi_j\|_\infty \leq \eta$,
- *and* $U_j \in \mathbb{R}^{k \times d}$ *with orthonormal rows.*

Shallow NN models with Lipschitz activation satisfy the assumption with $k = 1$!

Adaptivity of MLE to low-dimensional structure (2)

Corollary

Let $\mathcal{F}_\beta = B_{\mathcal{F}_1}(\beta)$. Assume that the low-dimensional structure assumption holds. Then, we can choose $\beta > 0$ such that with probability at least $1-\delta$, the MLE $\hat{\nu} := \nu_{\hat{f}}$ fulfills

$$\boxed{D_{KL}(\nu || \hat{\nu}) \leq \tilde{O}\left(\left(1 + \sqrt{\log(1/\delta)}\right) J\eta R^{-\frac{2}{k+3}} n^{-\frac{1}{k+3}}\right)}$$

where the notation $\tilde{O}$ indicates that we overlook logarithmic factors and constants depending only on the dimension k.

Adaptivity of MLE to low-dimensional structure (3)

$$D_{KL}(\nu||\hat{\nu}) \leq \tilde{O}\left(\left(1+\sqrt{\log(1/\delta)}\right) J\eta R^{-\frac{2}{k+3}} n^{-\frac{1}{k+3}}\right)$$

– **Idea of the proof:** Leverage low-dimensional structure to show $\inf_{f \in B_{\mathcal{F}_1}(\beta)} \|g - f\|_\infty$ is $\mathcal{O}\left(C(k) J\eta \left(R\beta/\eta J\right)^{-2/(k+1)}\right)$ using spherical harmonics arguments from [Bach, 2017]. Find β with the optimal tradeoff between statistical and approximation error.

Adaptivity of MLE to low-dimensional structure (3)

$$D_{KL}(\nu||\hat{\nu}) \leq \tilde{O}\left(\left(1+\sqrt{\log(1/\delta)}\right) J\eta R^{-\frac{2}{k+3}} n^{-\frac{1}{k+3}}\right)$$

– **Idea of the proof:** Leverage low-dimensional structure to show $\inf_{f \in B_{\mathcal{F}_1}(\beta)} \|g - f\|_\infty$ is $\mathcal{O}\left(C(k)J\eta\left(R\beta/\eta J\right)^{-2/(k+1)}\right)$ using spherical harmonics arguments from [Bach, 2017]. Find β with the optimal tradeoff between statistical and approximation error.

– **Why is this result relevant?** Without additional structure, the approximation error $\inf_{f \in B_{\mathcal{F}_1}(\beta)} \|g - f\|_\infty$ goes as $n^{-O(1/d)} \implies D_{KL}(\nu||\hat{\nu})$ would go as $n^{-O(1/d)}$. *Curse of dimensionality*! We would need $n = \epsilon^{-\Omega(d)}$ samples to get test error ϵ.

Algorithms

Algorithms for $\mathcal{F} = \mathcal{B}_{\mathcal{F}_1}(\beta)$:
We switch from a constrained problem to a lifted, penalized problem:

$$\inf_{\mu \in \mathcal{P}(\mathbb{R}^{d+2})} F(\mu) := R\left(\int \Phi(w,\theta) d\mu\right) + \lambda \int (|w|^2 + \|\theta\|_2^2)\ d\mu,$$

where R is the cross-entropy or SD loss. We discretize μ and train by gradient descent:
$G((w^{(i)}, \theta^{(i)})_{i=1}^m) := F\left(\frac{1}{m}\sum_{i=1}^m \delta_{(w^{(i)},\theta^{(i)})}\right) =$
$R\left(\frac{1}{m}\sum_{i=1}^m \Phi(w^{(i)}, \theta^{(i)})\right) + \frac{\lambda}{m}\sum_{i=1}^m(|w^{(i)}|^2 + \|\theta^{(i)}\|_2^2)$.

Algorithms

Algorithms for $\mathcal{F} = \mathcal{B}_{\mathcal{F}_1}(\beta)$:
We switch from a constrained problem to a lifted, penalized problem:

$$\inf_{\mu \in \mathcal{P}(\mathbb{R}^{d+2})} F(\mu) := R\left(\int \Phi(w,\theta) d\mu\right) + \lambda \int (|w|^2 + \|\theta\|_2^2)\ d\mu,$$

where R is the cross-entropy or SD loss. We discretize μ and train by gradient descent:
$G((w^{(i)}, \theta^{(i)})_{i=1}^m) := F\left(\frac{1}{m}\sum_{i=1}^m \delta_{(w^{(i)},\theta^{(i)})}\right) =$
$R\left(\frac{1}{m}\sum_{i=1}^m \Phi(w^{(i)}, \theta^{(i)})\right) + \frac{\lambda}{m}\sum_{i=1}^m (|w^{(i)}|^2 + \|\theta^{(i)}\|_2^2).$

Algorithms for $\mathcal{F} = \mathcal{B}_{\mathcal{F}_2}(\beta)$: Same discretization, but training only $w^{(i)}$ and keeping $\theta^{(i)}$ (random features kernel discretization).

Experimental setup

– We illustrate our theory on simple synthetic datasets generated by teacher models with energies $f^*(x) = \frac{1}{J}\sum_{j=1}^{J} w_j^* \sigma(\langle \theta_j^*, x\rangle)$, with $\theta_j^* \in \mathbb{S}^{d-1}$ for all j.

Experimental setup

– We illustrate our theory on simple synthetic datasets generated by teacher models with energies $f^*(x) = \frac{1}{J}\sum_{j=1}^{J} w_j^* \sigma(\langle \theta_j^*, x \rangle)$, with $\theta_j^* \in \mathbb{S}^{d-1}$ for all j.

– We train models with (i) maximum likelihood, (ii) $\mathcal{F}_1$-SD, (iii) KSD.

– We evaluate test error in KL divergence and the corresponding training metric (if different from maximum likelihood).

Experiments in $d = 15$ with one planted neuron

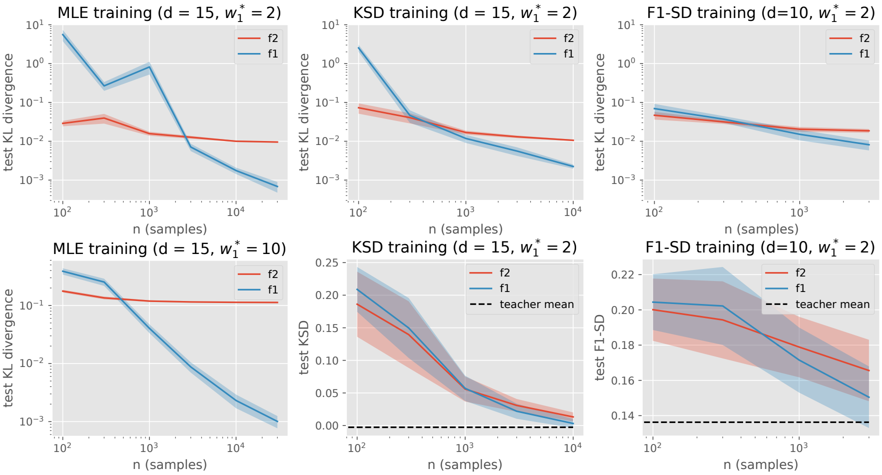


Figure: Test metrics obtained for MLE, KSD and $\mathcal{F}_1$-SD training on a one-neuron teacher with positive output weight.

Experiments in $d = 15$ with two planted neurons

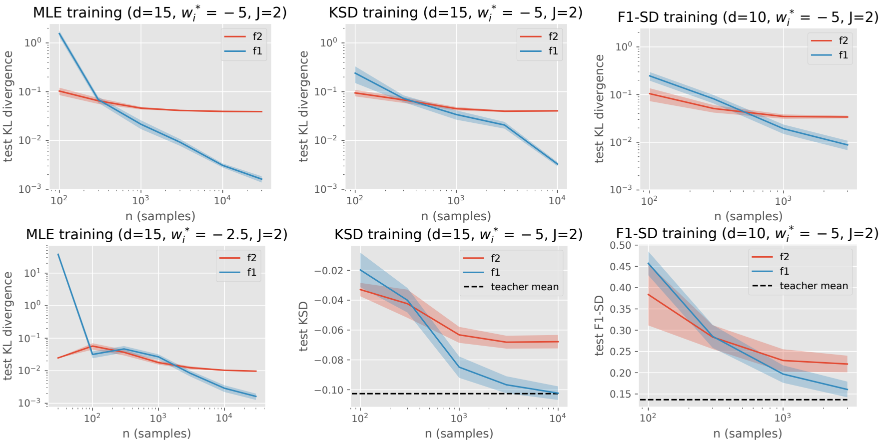


Figure: Test metrics obtained for MLE, KSD and $\mathcal{F}_1$-SD training on a two-neuron teacher with negative output weights.

Experiments in $d = 15$ with four planted neurons

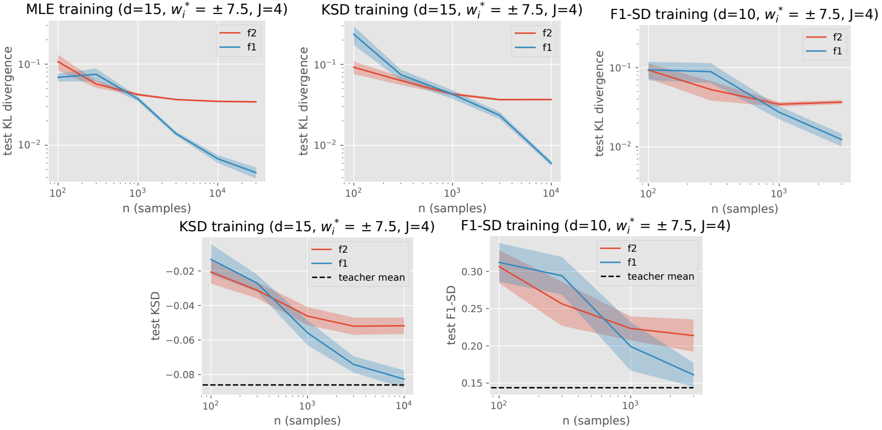


Figure: Test metrics obtained for MLE, KSD and $\mathcal{F}_1$-SD training on a four-neuron teacher with weights $w_1^*, w_2^* = 7.5$ and $w_3^*, w_4^* = -7.5$.

Experiments in $d = 3$ with two planted neurons (1)

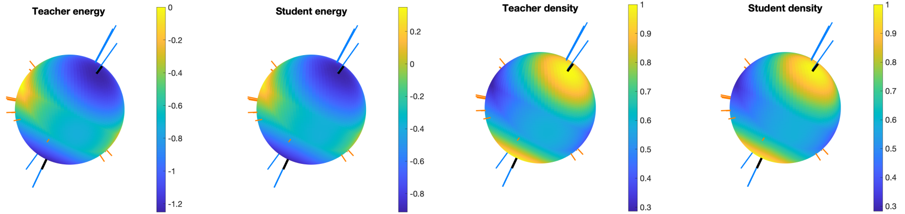


Figure: 3D visualization of the neuron positions, energies and densities, in $d = 3$. The teacher model has two neurons with negative weights $w_1^*, w_2^* = -2.5$, whose positions are represented by black sticks in all the images. The positions of the neurons of the trained model are represented by blue and orange sticks for negative and positive weights, resp.

Experiments in $d = 3$ with two planted neurons (2)

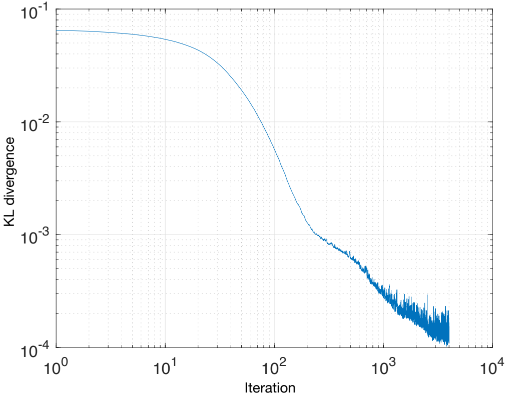


Figure: Log-log plot of the KL divergence between the MLE trained model and the teacher model (same as in 6), versus the iteration number.

Conclusions and discussion

– We provide statistical error bounds for EBMs trained with KL divergence or Stein discrepancies.

– We show adaptivity to low dimensional structures for feature learning overparametrized NN energies.

Conclusions and discussion

– We provide statistical error bounds for EBMs trained with KL divergence or Stein discrepancies.

– We show adaptivity to low dimensional structures for feature learning overparametrized NN energies.

– Possible statistical improvement: show lower bounds for $\mathcal{F}_2$ EBMs to prove theoretical separation.

– Possible computational improvements: computational guarantees for optimization / alternative algorithms (see [Domingo-Enrich et al., 2021]).

References

Bach, F. (2017).
Breaking the curse of dimensionality with convex neural networks.
Journal of Machine Learning Research, 18(19):1–53.

Barron, A. (1993).
Universal approximation bounds for superpositions of a sigmoidal function.
Information Theory, IEEE Transactions on, 39:930 – 945.

Domingo-Enrich, C., Bietti, A., Gabrié, M., Vanden-Eijnden, E., and Bruna, J. (In preparation, 2021).
Dual training of ebms with overparametrized shallow neural networks.

Du, Y. and Mordatch, I. (2019).
Implicit generation and generalization in energy-based models.
In *Advances in Neural Information Processing Systems (NeurIPS)*.