Implicit Regularization and Benign Overfitting for Neural Networks in High Dimensions

Youth in High Dimensions, Trieste, June 2023

Spencer Frei (UC Berkeley, Simons Institute $\rightarrow$ UC Davis, Department of Statistics)

Implicit regularization

- A.K.A. "Implicit bias", "algorithmic regularization", "inductive bias", ...
- Optimization algorithms can minimize 'complexity', with no *explicit* regularization.
  - Gradient flow in least squares $\leftrightarrow$ min ℓ^2:

$$\frac{\mathrm{d}}{\mathrm{d}t} w(t) = -\nabla \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w(t), x_i \rangle)^2 \right) \quad \longleftrightarrow \quad w(t) \to \min_{w} \|w\|_2^2 : \langle w, x_i \rangle = y_i \, \forall i.$$

  - Gradient flow/descent on exponential loss $\leftrightarrow$ maximum margin:

$$\frac{\mathrm{d}}{\mathrm{d}t} w(t) = -\nabla \left(\frac{1}{n} \sum_{i=1}^{n} \exp\left(- y_i \langle w(t), x_i \rangle \right) \right) \quad \longleftrightarrow \quad w(t) \to \min_{w} \|w\|_2^2 : y_i \langle w, x_i \rangle \geq 1 \, \forall i.$$

See: [Telgarsky'13; Soudry+'18; Ji-Telgarsky'20; Lyu-Li'20; ...]

Implicit regularization

- A.K.A. “Implicit bias”, “algorithmic regularization”, “inductive bias”, ...
- Optimization algorithms can minimize ‘complexity’, with no *explicit* regularization.
  - Gradient flow in least squares $\leftrightarrow$ min ℓ^2:

$$\frac{\mathrm{d}}{\mathrm{d}t}w(t) = -\nabla\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle w(t), x_i\rangle)^2\right) \quad\longleftrightarrow\quad w(t) \to \min_w \|w\|_2^2 : \langle w, x_i\rangle = y_i\,\forall i.$$

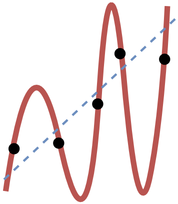
  - Gradient flow/descent on exponential loss $\leftrightarrow$ maximum margin:

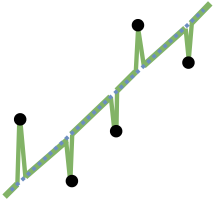
$$\frac{\mathrm{d}}{\mathrm{d}t}w(t) = -\nabla\left(\frac{1}{n}\sum_{i=1}^{n}\exp\left(-y_i\langle w(t), x_i\rangle\right)\right) \quad\longleftrightarrow\quad w(t) \to \min_w \|w\|_2^2 : y_i\langle w, x_i\rangle \geq 1\,\forall i.$$

- For gradient flow/descent on neural nets, story is much more complicated, but conjectured to contribute to success of deep learning

See: [Telgarsky’13; Soudry+’18; Ji-Telgarsky’20; Lyu-Li’20; ...]

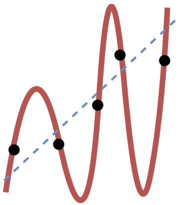
Benign overfitting

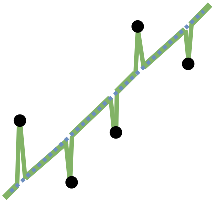




- Benign overfitting refers to settings where there is *noise*, the estimator achieves zero training error (overfits), yet still generalizes well (even optimally).

Benign overfitting





- Benign overfitting refers to settings where there is *noise*, the estimator achieves zero training error (overfits), yet still generalizes well (even optimally).
- Hard to reconcile uniform convergence with interpolation of noisy data:

$$c < \sup_{f \in \mathcal{F}} L(f) = \sup_{f \in \mathcal{F}} L(f) - \widehat{L}_n(f) \overset{(?)}{\lesssim} \sqrt{\text{Complexity}(\mathcal{F})/n}.$$

- Good understanding of mechanisms of benign overfitting in linear regression [Bartlett+'20; Hastie+'22; ...], but little in neural networks

Our contributions

We examine behavior of neural nets when trained on "high-dimensional data" ($d \gg n$, to be made precise shortly).

- **Implicit regularization**: Gradient flow-trained two-layer networks have low rank and a very simple/tractable structure.
- **Benign overfitting from implicit regularization:** in particular distributional settings, this simple structure implies benign overfitting.

Implicit bias in homogeneous neural networks

$$\widehat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i N(x_i;\theta) \big), \qquad \ell(q) = \log(1+\exp(-q)), \qquad \frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = -\nabla \widehat{L}(\theta(t)),$$

$$\min_{\theta} \|\theta\|^2 \quad \text{s.t.} \quad y_i N(x_i;\theta) \geq 1, \text{ for all } i \in [n]. \tag{1}$$

Theorem [Lyu-Li'19; Ji-Telgarsky'20]

Consider gradient flow-trained net. If $N(x;\theta)$ is L-homogeneous $(N(x;\alpha\theta) = \alpha^L N(x;\theta))$ and there exists time t_0 s.t. $\widehat{L}(\theta(t_0)) < 1/n$. Then gradient flow converges in direction to a first-order stationary point (KKT point) of max-margin problem (1), and $\widehat{L}(\theta(t)) \to 0$.

Implicit bias in homogeneous neural networks

$$\widehat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(\boxed{y_i N(x_i;\theta)}\big), \qquad \ell(q) = \log(1+\exp(-q)), \qquad \boxed{\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = -\nabla \widehat{L}(\theta(t))},$$

$$\boxed{\min_{\theta} \|\theta\|^2 \quad \text{s.t.} \quad y_i N(x_i;\theta) \geq 1, \text{ for all } i \in [n].} \tag{1}$$

Theorem [Lyu-Li'19; Ji-Telgarsky'20]

Consider gradient flow -trained net. If $N(x;\theta)$ is L-homogeneous $(N(x;\alpha\theta) = \alpha^L N(x;\theta))$ and there exists time t_0 s.t. $\widehat{L}(\theta(t_0)) < 1/n$. Then gradient flow converges in direction to a first-order stationary point (KKT point) of max-margin problem (1), and $\widehat{L}(\theta(t)) \to 0$.

- There exists θ^* satisfying **K**arush–**K**uhn–**T**ucker conditions of (1) s.t. $\frac{\theta(t)}{\|\theta(t)\|} \to \frac{\theta^*}{\|\theta^*\|}$.
- Satisfying KKT conditions does *not* imply global optimality in general [Vardi-Shamir-Srebro'22].
- Theorem does *not* depend on initialization $\theta(0)$.

Implicit bias in homogeneous neural networks

- By [Lyu-Li'19; Ji-Telgarsky'20], KKT conditions of max-margin problem (1) capture limiting behavior of (homogeneous) neural network training.

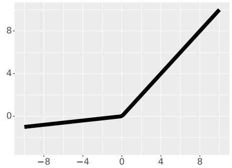
$$\min_{\theta} \|\theta\|^2 \quad \text{s.t.} \quad y_i N(x_i; \theta) \geq 1, \text{ for all } i \in [n]. \tag{1}$$

- We'll show that in some settings, satisfaction of KKT conditions for Problem (1) implies good generalization (and benign overfitting).
  - Any algorithm that produces max-margin neural nets would have same behavior.

The setting: two-layer leaky ReLU networks

- Two-layer nets with leaky ReLU activations ($\phi(z) = \max(z, \gamma z)$ for all z) trained by GD on the logistic loss:

The setting: two-layer leaky ReLU networks



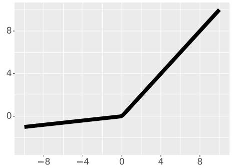
- Two-layer nets with leaky ReLU activations ($\phi(z) = \max(z, \gamma z)$ for all z) trained by GD on the logistic loss:

$$f(x; W) = \sum_{j=1}^{m} a_j \phi(\langle w_j, x \rangle), \quad a_j \in \{\pm 1/\sqrt{m}\},$$

$$\frac{\mathrm{d}}{\mathrm{d}t} W(t) = -\nabla \widehat{L}(W(t)), \quad W(0): \text{ arbitrary},$$

$$\widehat{L}(W) = {}^{1}\!/_{n} \textstyle\sum_{i=1}^{n} \log\left(1 + \exp(-y_i f(x_i; W))\right).$$

The setting: two-layer leaky ReLU networks



- Two-layer nets with leaky ReLU activations ($\phi(z) = \max(z, \gamma z)$ for all z) trained by GD on the logistic loss:

$$f(x; W) = \sum_{j=1}^{m} a_j \phi(\langle w_j, x \rangle), \quad a_j \in \{\pm 1/\sqrt{m}\},$$

$$\frac{\mathrm{d}}{\mathrm{d}t} W(t) = -\nabla \widehat{L}(W(t)), \quad W(0) : \text{ arbitrary},$$

$$\widehat{L}(W) = {}^1\!/_n \textstyle\sum_{i=1}^{n} \log\left(1 + \exp(-y_i f(x_i; W))\right).$$

- Since ϕ is 1-homogeneous, so is $f(x; \cdot)$.
- $\implies$ KKT conditions for margin maximization characterize limiting behavior of trained neural nets.

$$\min \|W\|_F^2 \quad \text{s.t.} \quad y_i f(x_i; W) \geq 1 \text{ for all } i \in [n].$$

The setting: "High-dimensional data"

- We assume data $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$ satisfy,

$$\boxed{\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|}, \quad \max_{i,k} \frac{\|x_i\|}{\|x_k\|} = O(1).$$

- Purely geometric condition on features: no assumptions on labels $\{y_i\}_{i=1}^n$, no probabilistic/distributional assumptions made.

The setting: "High-dimensional data"

- We assume data $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$ satisfy,

$$\boxed{\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|}, \quad \max_{i,k} \frac{\|x_i\|}{\|x_k\|} = O(1).$$

- Purely geometric condition on features: no assumptions on labels $\{y_i\}_{i=1}^n$, no probabilistic/distributional assumptions made.
- Satisfied in many settings w.h.p. when $(x_i, y_i) \overset{\text{i.i.d.}}{\sim}$ P and d is large relative to n:
    - Isotropic Gaussians $x_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d)$ when $d = \tilde{\Omega}(n^2)$.
    - x has independent sub-Gaussian components with $\mathbb{E}[x] = 0$ and $\mathbb{E}[xx^\top] = \Sigma$ where $\frac{\mathrm{trace}(\Sigma)}{\sqrt{\mathrm{trace}(\Sigma^2)}} = \tilde{\Omega}(n)$.

The setting: "High-dimensional data"

- We assume data $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$ satisfy,

$$\boxed{\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|}, \qquad \max_{i,k} \frac{\|x_i\|}{\|x_k\|} = O(1).$$

- Purely geometric condition on features: no assumptions on labels $\{y_i\}_{i=1}^n$, no probabilistic/distributional assumptions made.
- Satisfied in many settings w.h.p. when $(x_i, y_i) \overset{\text{i.i.d.}}{\sim}$ P and d is large relative to n:
  - Isotropic Gaussians $x_i \overset{\text{i.i.d.}}{\sim}$ N$(0, I_d)$ when $d = \tilde{\Omega}(n^2)$.
  - x has independent sub-Gaussian components with $\mathbb{E}[x] = 0$ and $\mathbb{E}[xx^\top] = \Sigma$ where $\frac{\mathrm{trace}(\Sigma)}{\sqrt{\mathrm{trace}(\Sigma^2)}} = \tilde{\Omega}(n)$.
- *Not* satisfied in some high-dimensional settings.
  - $x_i \overset{\text{i.i.d.}}{\sim}$ N$(0, \mathrm{diag}(\lambda))$ where $\lambda = \mathrm{diag}(\mu, 1, \ldots, 1)$ and $\mu \to \infty$.

Implicit bias in leaky ReLU nets for high-dimensional data

Let $f(x;W)$ be two-layer leaky ReLU network, and consider max-margin problem,

$$\min_W \|W\|_F^2 \quad \text{s.t.} \quad y_i f(x_i;W) \geq 1, \text{ for all } i \in [n]. \tag{1}$$

Theorem [F.*-Vardi*-Bartlett-Srebro-Hu ICLR'23]

Suppose $\|x_i\|^2 \gg n \max_{k\neq j} |\langle x_j, x_k\rangle|$ and $\max_{i,k} \frac{\|x_i\|}{\|x_k\|} = O(1)$. Let V be a KKT point of Problem (1). Then the following holds:

Implicit bias in leaky ReLU nets for high-dimensional data

Let $f(x; W)$ be two-layer leaky ReLU network, and consider max-margin problem,

$$\boxed{\min_W \|W\|_F^2 \quad \text{s.t.} \quad y_i f(x_i; W) \geq 1, \text{ for all } i \in [n].} \tag{1}$$

Theorem [F.*-Vardi*-Bartlett-Srebro-Hu ICLR'23]

Suppose $\boxed{\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|}$ and $\max_{i,k} \frac{\|x_i\|}{\|x_k\|} = O(1)$. Let V be a KKT point of Problem (1). Then the following holds:

1. $\boxed{\mathrm{rank}(V) \leq 2}$.

Implicit bias in leaky ReLU nets for high-dimensional data

Let $f(x; W)$ be two-layer leaky ReLU network, and consider max-margin problem,

$$\boxed{\min_W \|W\|_F^2 \quad \text{s.t.} \quad y_i f(x_i; W) \geq 1, \text{ for all } i \in [n].} \tag{1}$$

Theorem [F.*-Vardi*-Bartlett-Srebro-Hu ICLR'23]

Suppose $\boxed{\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|}$ and $\max_{i,k} \frac{\|x_i\|}{\|x_k\|} = O(1)$. Let V be a KKT point of Problem (1). Then the following holds:

1. $\boxed{\mathrm{rank}(V) \leq 2}$.
2. There exists $z \in \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$, $\boxed{\mathrm{sgn}(\langle z, x \rangle) = \mathrm{sgn}(f(x; V))}$.

Implicit bias in leaky ReLU nets for high-dimensional data

Let $f(x; W)$ be two-layer leaky ReLU network, and consider max-margin problem,

$$\boxed{\min_W \|W\|_F^2 \quad \text{s.t.} \quad y_i f(x_i; W) \geq 1, \text{ for all } i \in [n].} \tag{1}$$

Theorem [F.*-Vardi*-Bartlett-Srebro-Hu ICLR'23]

Suppose $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$ and $\max_{i,k} \frac{\|x_i\|}{\|x_k\|} = O(1)$. Let V be a KKT point of Problem (1). Then the following holds:

1. $\text{rank}(V) \leq 2$.
2. There exists $z \in \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$, $\text{sgn}(\langle z, x \rangle) = \text{sgn}(f(x; V))$.
3. This z satisfies $z \propto \sum_{i=1}^n s_i y_i x_i$ for some $s_i > 0$ where $\max_{i,j} s_i / s_j = O(1)$.

Implicit bias in leaky ReLU nets for high-dimensional data

Let $f(x; W)$ be two-layer leaky ReLU network, and consider max-margin problem,

$$\min_W \|W\|_F^2 \quad \text{s.t.} \quad y_i f(x_i; W) \geq 1, \text{ for all } i \in [n]. \tag{1}$$

Theorem [F.*-Vardi*-Bartlett-Srebro-Hu ICLR'23]

Suppose $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$ and $\max_{i,k} \frac{\|x_i\|}{\|x_k\|} = O(1)$. Let V be a KKT point of Problem (1). Then the following holds:

1. $\text{rank}(V) \leq 2$.
2. There exists $z \in \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$, $\text{sgn}(\langle z, x \rangle) = \text{sgn}(f(x; V))$.
3. This z satisfies $z \propto \sum_{i=1}^n s_i y_i x_i$ for some $s_i > 0$ where $\max_{i,j} s_i / s_j = O(1)$.

And for any initialization $W(0)$, gradient flow converges in direction to a net satisfying above.

Implicit bias in leaky ReLU nets for high-dimensional data

Theorem [F.*-Vardi*-Bartlett-Srebro-Hu ICLR'23]

Suppose $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$ and $\max_{i,k} \frac{\|x_i\|}{\|x_k\|} = O(1)$. Let V be a KKT point of Problem (1). Then the following holds:

1. $\mathrm{rank}(V) \leq 2$.
2. There exists $z \in \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$, $\mathrm{sgn}(\langle z, x \rangle) = \mathrm{sgn}(f(x; V))$.
3. This z satisfies $z \propto \sum_{i=1}^n s_i y_i x_i$ for some $s_i > 0$ where $\max_{i,j} s_i / s_j = O(1)$.

And for any initialization $W(0)$, gradient flow converges in direction to a net satisfying above.

Implicit bias in leaky ReLU nets for high-dimensional data

Theorem [F.*-Vardi*-Bartlett-Srebro-Hu ICLR'23]

Suppose $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$ and $\max_{i,k} \frac{\|x_i\|}{\|x_k\|} = O(1)$. Let V be a KKT point of Problem (1). Then the following holds:

1. $\text{rank}(V) \leq 2$.
2. There exists $z \in \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$, $\text{sgn}(\langle z, x \rangle) = \text{sgn}(f(x; V))$.
3. This z satisfies $z \propto \sum_{i=1}^n s_i y_i x_i$ for some $s_i > 0$ where $\max_{i,j} s_i / s_j = O(1)$.

And for any initialization $W(0)$, gradient flow converges in direction to a net satisfying above.

- For $w_j \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2 I)$, $\text{rank}(W(0)) \geq m \wedge d \implies$ *rank reducing* implicit regularization.

Implicit bias in leaky ReLU nets for high-dimensional data

Theorem [F.*-Vardi*-Bartlett-Srebro-Hu ICLR'23]

Suppose $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$ and $\max_{i,k} \frac{\|x_i\|}{\|x_k\|} = O(1)$. Let V be a KKT point of Problem (1). Then the following holds:

1. $\mathrm{rank}(V) \leq 2$.
2. There exists $z \in \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$, $\mathrm{sgn}(\langle z, x \rangle) = \mathrm{sgn}(f(x; V))$.
3. This z satisfies $z \propto \sum_{i=1}^n s_i y_i x_i$ for some $s_i > 0$ where $\max_{i,j} s_i/s_j = O(1)$.

And for any initialization $W(0)$, gradient flow converges in direction to a net satisfying above.

- For $w_j \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2 I)$, $\mathrm{rank}(W(0)) \geq m \wedge d \implies$ *rank reducing* implicit regularization.
- Decision boundary is *linear*, despite nonlinear hypothesis class, and takes simple form.

Proof idea

- Proof is based on analysis of KKT conditions for margin-maximization,

$$f(x;W) = \sum_{j=1}^{m} a_j \phi(\langle w_j, x \rangle), \quad \boxed{\min_{\theta} \|W\|_F^2 \quad \text{s.t.} \quad y_i f(x_i;W) \geq 1, \text{ for all } i \in [n],} \quad ($$

- First step: there exist Lagrange multipliers $\lambda_1, \ldots, \lambda_n \geq 0$ s.t. for every $j \in [m]$,

$$w_j = \sum_{i=1}^{n} \lambda_i \nabla_{w_j} (y_i f(x_i;W)) = \sum_{i=1}^{n} \lambda_i y_i a_j \phi'_{i,j} x_i. \tag{1}$$

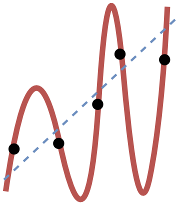
- $\phi'_{i,j} \geq \gamma > 0$ is sub-gradient of ϕ at $\langle w_j, x_i \rangle$, and for all $i \in [n]$, $\lambda_i = 0$ if $y_i f(x_i;W) > 1$.
- Analysis proceeds by showing all λ_i are strictly positive, "not too small" and "not too large", then using (1) to analyze $f(x;W)$.

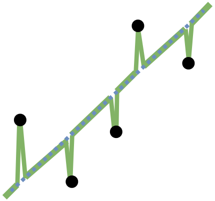
Summary: Implicit regularization of gradient flow

- Provided data is sufficiently high-dimensional ($\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$), gradient flow is biased towards low-rank networks.
- Moreover, decision boundary is linear, and satisfies

$$\mathrm{sgn}(\langle z, x \rangle) = \mathrm{sgn}(f(x; W)), \quad z \propto \textstyle\sum_{i=1}^{n} s_i y_i x_i, \quad \max_{i,j} {s_i}/{s_j} = O(1).$$

- Next: use these results to say something about generalization and *benign overfitting*.





- Benign overfitting refers to settings where there is *noise*, the estimator achieves zero training error (overfits), yet still generalizes well (even optimally).

Benign overfitting

- Minimum-norm least squares interpolation is most well-understood predictor in benign overfitting, aided by the explicit formula for the predictor:

$$\operatorname{argmin}\{\|w\|^2 : y_i = \langle w, x_i \rangle\} = X^\top (XX^\top)^+ y.$$

- Even in linear classification, no explicit formula for max-margin predictor (in general). Ditto NNs trained by GD/GF.

Benign overfitting

- Minimum-norm least squares interpolation is most well-understood predictor in benign overfitting, aided by the explicit formula for the predictor:

$$\text{argmin}\{\|w\|^2 : y_i = \langle w, x_i \rangle\} = X^\top (XX^\top)^+ y.$$

- Even in linear classification, no explicit formula for max-margin predictor (in general). Ditto NNs trained by GD/GF.
- If data is "high-dimensional", then our implicit bias results show that NN trained by GF converges to network satisfying:

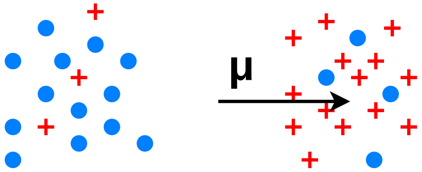
$$\boxed{\text{sgn}(\langle z, x \rangle) = \text{sgn}(f(x; W)), \quad z \propto \textstyle\sum_{i=1}^n s_i y_i x_i}, \ \max_{i,j} s_i/s_j = O(1).$$

- $\implies$ NN trained by gradient flow exhibits benign overfitting if $x \mapsto \text{sgn}(\langle z, x \rangle)$ does.

Distributional setting

- Opposing clusters: for $\mu \in \mathbb{R}^d$ and $z \sim$ P isotropic, independent sub-Gaussian components,

$$\tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x|\tilde{y} = \tilde{y}\mu + z, \quad y = \begin{cases} \tilde{y}, & \text{w.p. } 1-p, \\ -\tilde{y}, & \text{w.p. } p. \end{cases}$$



- Analysis can be extended to additional settings, but key ideas can be seen with opposing clusters, so will focus here

Benign overfitting for τ-uniform classifiers

Say $u \in \mathbb{R}^d$ is τ-uniform w.r.t. $S = \{(x_i, y_i)\}_{i=1}^n$ if $u = \sum_{i=1}^n s_i y_i x_i$ with $\max_{i,j} {}^{s_i}/{}_{s_j} \leq \tau$.

Benign overfitting for τ-uniform classifiers

Say $u \in \mathbb{R}^d$ is τ-uniform w.r.t. $S = \{(x_i, y_i)\}_{i=1}^n$ if $u = \sum_{i=1}^n s_i y_i x_i$ with $\max_{i,j} s_i/s_j \le \tau$.

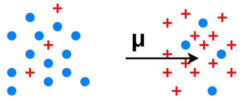
Assumptions: for large $C > 1$,

(A1) Number of samples $n \ge C$.

(A2) Mean separation $\|\mu\| = \Theta(d^\beta)$, $\beta \in (0, 1/2)$.

(A3) Dimension $d \ge C n^{2 \vee \frac{1}{1-2\beta}} \log(n)$.

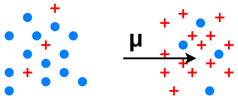
(A2) + (A3) imply $\|x_k\|^2 \gg n \max_{i \ne j} |\langle x_i, x_j \rangle|$.



- $x|\tilde{y} = \tilde{y}\mu + z$, labels flipped w.p. p.

Benign overfitting for τ-uniform classifiers

Say $u \in \mathbb{R}^d$ is τ-uniform w.r.t. $S = \{(x_i, y_i)\}_{i=1}^n$ if $u = \sum_{i=1}^n s_i y_i x_i$ with $\max_{i,j} {}^{s_i}/_{s_j} \le \tau$.
Assumptions: for large $C > 1$,



(A1) Number of samples $n \ge C$.

(A2) Mean separation $\|\mu\| = \Theta(d^\beta)$, $\beta \in (0, {}^1/_2)$.

(A3) Dimension $d \ge C n^{2 \vee \frac{1}{1-2\beta}} \log(n)$.

(A2) + (A3) imply $\|x_k\|^2 \gg n \max_{i \ne j} |\langle x_i, x_j \rangle|$.

- $x|\tilde{y} = \tilde{y}\mu + z$, labels flipped w.p. p.

Theorem [F.*-Vardi-Bartlett-Srebro, COLT'23]

For $\tau \ge 1$, assume noise rate $p < \frac{1}{1+\tau}$. For some absolute $C, C' > 1$, under (A1)-(A3) w.p. at least 99% over P^n, if u is τ-uniform w.r.t. S then: for all $k \in [n]$, $y_k = \operatorname{sgn}(\langle u, x_k \rangle)$, and

$$p \le \mathbb{P}_{(x,y)\sim \mathsf{P}}\left(y \ne \operatorname{sgn}(\langle u, x \rangle)\right) \le p + \exp\left(-{}^{n\|\mu\|^4}/_{C'd}\right).$$

Benign overfitting if $\|\mu\| = \Theta(d^\beta)$ for $\beta \in ({}^1/_4, {}^1/_2)$.

Benign overfitting for τ-uniform classifiers

Say $u \in \mathbb{R}^d$ is τ-uniform w.r.t. $S = \{(x_i, y_i)\}_{i=1}^n$ if $u = \sum_{i=1}^n s_i y_i x_i$ with $\max_{i,j} s_i/s_j \leq \tau$.

Theorem [F.*-Vardi*-Bartlett-Srebro, COLT'23]

For $\tau \geq 1$, assume noise rate $p < \frac{1}{1+\tau}$. For some absolute $C, C' > 1$, under (A1)-(A3) w.p. at least 99% over P^n, if u is τ-uniform w.r.t. S then: for all k, $y_k = \operatorname{sgn}\big(\langle u, x_k\rangle\big)$, and

$$p \leq \mathbb{P}_{(x,y)\sim \mathsf{P}}\big(y \neq \operatorname{sgn}(\langle u, x\rangle)\big) \leq p + \exp\left(-n\|\mu\|^4/C'd\right).$$

- Benign overfitting if $\|\mu\| = \Theta(d^\beta)$ for $\beta \in (1/4, 1/2)$.

Benign overfitting for τ-uniform classifiers

Say $u \in \mathbb{R}^d$ is τ-uniform w.r.t. $S = \{(x_i, y_i)\}_{i=1}^n$ if $u = \sum_{i=1}^n s_i y_i x_i$ with $\max_{i,j} s_i/s_j \leq \tau$.

Theorem [F.*-Vardi*-Bartlett-Srebro, COLT'23]

For $\tau \geq 1$, assume noise rate $p < \frac{1}{1+\tau}$. For some absolute $C, C' > 1$, under (A1)-(A3) w.p. at least 99% over P^n, if u is τ-uniform w.r.t. S then: for all k, $y_k = \operatorname{sgn}(\langle u, x_k \rangle)$, and

$$p \leq \mathbb{P}_{(x,y)\sim \mathsf{P}}\big(y \neq \operatorname{sgn}(\langle u, x\rangle)\big) \leq p + \exp\left(-n\|\mu\|^4/C'd\right).$$

- Benign overfitting if $\|\mu\| = \Theta(d^\beta)$ for $\beta \in (1/4, 1/2)$.
- Sample average $\sum_{i=1}^n y_i x_i$ is 1-uniform and can tolerate noise rates close to $p = 1/2$.

Benign overfitting for τ-uniform classifiers

Say $u \in \mathbb{R}^d$ is τ-uniform w.r.t. $S = \{(x_i, y_i)\}_{i=1}^n$ if $u = \sum_{i=1}^n s_i y_i x_i$ with $\max_{i,j} s_i/s_j \le \tau$.

Theorem [F.*-Vardi*-Bartlett-Srebro, COLT'23]

For $\tau \ge 1$, assume noise rate $p < \frac{1}{1+\tau}$. For some absolute $C, C' > 1$, under (A1)-(A3) w.p. at least 99% over P^n, if u is τ-uniform w.r.t. S then: for all k, $y_k = \operatorname{sgn}\big(\langle u, x_k\rangle\big)$, and

$$p \le \mathbb{P}_{(x,y)\sim\mathsf{P}}\big(y \ne \operatorname{sgn}(\langle u, x\rangle)\big) \le p + \exp\left(-n\|\mu\|^4/C'd\right).$$

- Benign overfitting if $\|\mu\| = \Theta(d^\beta)$ for $\beta \in (1/4, 1/2)$.
- Sample average $\sum_{i=1}^n y_i x_i$ is 1-uniform and can tolerate noise rates close to $p = 1/2$.
- Previous implicit bias results $\implies$ max-margin linear classifiers *and* max-margin two-layer leaky ReLU networks are τ-uniform for $\tau = O(1)$.
- No dependence on number of parameters of network!

Proof idea: signal + overfitting component decomposition

- Will focus on the 1-uniform classifier $u := \sum_{i=1}^n y_i x_i$ and Gaussian data $z \sim \mathsf{N}(0, I_d)$.
- Recall data generated: $\tilde{y} \sim \mathsf{Unif}(\{\pm 1\})$, $x|\tilde{y} = \tilde{y}\mu + z$, then $y = -\tilde{y}$ w.p. p.
- Say $i \in \mathcal{C}$ ('clean') if $y_i = \tilde{y}_i$, $i \in \mathcal{N}$ ('noisy') if $y_i = -\tilde{y}_i$.

Proof idea: signal + overfitting component decomposition

- Will focus on the 1-uniform classifier $u := \sum_{i=1}^n y_i x_i$ and Gaussian data $z \sim \mathsf{N}(0, I_d)$.
- Recall data generated: $\tilde{y} \sim \mathsf{Unif}(\{\pm 1\})$, $x|\tilde{y} = \tilde{y}\mu + z$, then $y = -\tilde{y}$ w.p. p.
- Say $i \in \mathcal{C}$ ('clean') if $y_i = \tilde{y}_i$, $i \in \mathcal{N}$ ('noisy') if $y_i = -\tilde{y}_i$.

$$u = \sum_{i=1}^n y_i x_i = \sum_{i \in \mathcal{C}} (\mu + y_i z_i) + \sum_{i \in \mathcal{N}} (-\mu + y_i z_i) = (|\mathcal{C}| - |\mathcal{N}|)\,\mu + \sum_{i=1}^n y_i z_i$$

$$\propto \boxed{\mu} + \boxed{\frac{1}{|\mathcal{C}| - |\mathcal{N}|} \sum_{i=1}^n y_i z_i} =: \boxed{\mu} + \boxed{\Delta_n}$$

Proof idea: signal + overfitting component decomposition

- Will focus on the 1-uniform classifier $u := \sum_{i=1}^n y_i x_i$ and Gaussian data $z \sim \mathsf{N}(0, I_d)$.
- Recall data generated: $\tilde{y} \sim \mathsf{Unif}(\{\pm 1\})$, $x|\tilde{y} = \tilde{y}\mu + z$, then $y = -\tilde{y}$ w.p. p.
- Say $i \in \mathcal{C}$ ('clean') if $y_i = \tilde{y}_i$, $i \in \mathcal{N}$ ('noisy') if $y_i = -\tilde{y}_i$.

$$u = \sum_{i=1}^n y_i x_i = \sum_{i\in\mathcal{C}} (\mu + y_i z_i) + \sum_{i\in\mathcal{N}} (-\mu + y_i z_i) = (|\mathcal{C}| - |\mathcal{N}|)\,\mu + \sum_{i=1}^n y_i z_i$$

$$\propto \boxed{\mu} + \boxed{\frac{1}{|\mathcal{C}| - |\mathcal{N}|} \sum_{i=1}^n y_i z_i} =: \boxed{\mu} + \boxed{\Delta_n}$$

- Signal component will help with generalization, but hurts overfitting: $\langle \mu, yx \rangle \gg 0$ for clean test examples, $\langle \mu, y_i x_i \rangle \ll 0$ for $i \in \mathcal{N}$

Proof idea: signal + overfitting component decomposition

- Will focus on the 1-uniform classifier $u := \sum_{i=1}^n y_i x_i$ and Gaussian data $z \sim \mathsf{N}(0, I_d)$.
- Recall data generated: $\tilde{y} \sim \mathsf{Unif}(\{\pm 1\})$, $x|\tilde{y} = \tilde{y}\mu + z$, then $y = -\tilde{y}$ w.p. p.
- Say $i \in \mathcal{C}$ ('clean') if $y_i = \tilde{y}_i$, $i \in \mathcal{N}$ ('noisy') if $y_i = -\tilde{y}_i$.

$$u = \sum_{i=1}^n y_i x_i = \sum_{i \in \mathcal{C}} (\mu + y_i z_i) + \sum_{i \in \mathcal{N}} (-\mu + y_i z_i) = (|\mathcal{C}| - |\mathcal{N}|)\,\mu + \sum_{i=1}^n y_i z_i$$

$$\propto \boxed{\mu} + \boxed{\frac{1}{|\mathcal{C}| - |\mathcal{N}|} \sum_{i=1}^n y_i z_i} =: \boxed{\mu} + \boxed{\Delta_n}$$

- Signal component will help with generalization, but hurts overfitting: $\langle \mu, yx \rangle \gg 0$ for clean test examples, $\langle \mu, y_i x_i \rangle \ll 0$ for $i \in \mathcal{N}$
- Overfitting component will help with overfitting, but hurts generalization: $\langle \Delta_n, y_k x_k \rangle \gg 0$ for training, but Δ_n useless for test

Proof idea: signal + overfitting component decomposition

- Will focus on the 1-uniform classifier $u := \sum_{i=1}^n y_i x_i$ and Gaussian data $z \sim \mathsf{N}(0, I_d)$.
- Recall data generated: $\tilde{y} \sim \mathsf{Unif}(\{\pm 1\})$, $x|\tilde{y} = \tilde{y}\mu + z$, then $y = -\tilde{y}$ w.p. p.
- Say $i \in \mathcal{C}$ ('clean') if $y_i = \tilde{y}_i$, $i \in \mathcal{N}$ ('noisy') if $y_i = -\tilde{y}_i$.

$$u = \sum_{i=1}^n y_i x_i = \sum_{i \in \mathcal{C}} (\mu + y_i z_i) + \sum_{i \in \mathcal{N}} (-\mu + y_i z_i) = (|\mathcal{C}| - |\mathcal{N}|)\,\mu + \sum_{i=1}^n y_i z_i$$

$$\propto \boxed{\mu} + \boxed{\frac{1}{|\mathcal{C}| - |\mathcal{N}|} \sum_{i=1}^n y_i z_i} =: \boxed{\mu} + \boxed{\Delta_n}$$

- Signal component will help with generalization, but hurts overfitting: $\langle \mu, yx \rangle \gg 0$ for clean test examples, $\langle \mu, y_i x_i \rangle \ll 0$ for $i \in \mathcal{N}$
- Overfitting component will help with overfitting, but hurts generalization: $\langle \Delta_n, y_k x_k \rangle \gg 0$ for training, but Δ_n useless for test
- Appropriately balanced, they together allow for benign overfitting

Proof idea: signal + overfitting component decomposition

$$u = \sum_{i=1}^n y_i x_i \propto \boxed{\mu} + \boxed{\frac{1}{|\mathcal{C}|-|\mathcal{N}|} \sum_{i=1}^n y_i z_i} \approx \boxed{\mu} + \boxed{\frac{1}{n(1-2p)} \sum_{i=1}^n y_i z_i} =: \mu + \Delta_n.$$

Since $z_i \sim \mathsf{N}(0, I_d)$, for $p = 1/4$ have $\Delta_n \sim \mathsf{N}(0, \frac{1}{4n} I_d)$.

Proof idea: signal + overfitting component decomposition

$$u = \sum_{i=1}^n y_i x_i \propto \boxed{\mu} + \boxed{\frac{1}{|\mathcal{C}|-|\mathcal{N}|} \sum_{i=1}^n y_i z_i} \approx \boxed{\mu} + \boxed{\frac{1}{n(1-2p)} \sum_{i=1}^n y_i z_i} =: \mu + \Delta_n.$$

Since $z_i \sim \mathsf{N}(0, I_d)$, for $p = 1/4$ have $\Delta_n \sim \mathsf{N}(0, \frac{1}{4n} I_d)$.

Signal effect on clean test data:

- $\langle \tilde{y}\mu, \tilde{y}\mu + z \rangle = \|\mu\|^2 + \mathsf{N}(0, \|\mu\|^2)$.

Signal effect on training data:

- $\langle y_k\mu, \tilde{y}_k\mu + z_k \rangle =$
  $y_k \tilde{y}_k \|\mu\|^2 + \mathsf{N}(0, \|\mu\|^2)$.
  $y_k \tilde{y}_k$ is negative for $k \in \mathcal{N}$.

Proof idea: signal + overfitting component decomposition

$$u = \sum_{i=1}^n y_i x_i \propto \boxed{\mu} + \boxed{\frac{1}{|\mathcal{C}|-|\mathcal{N}|} \sum_{i=1}^n y_i z_i} \approx \boxed{\mu} + \boxed{\frac{1}{n(1-2p)} \sum_{i=1}^n y_i z_i} =: \mu + \Delta_n.$$

Since $z_i \sim \mathsf{N}(0, I_d)$, for $p = 1/4$ have $\Delta_n \sim \mathsf{N}(0, \frac{1}{4n} I_d)$.

Signal effect on clean test data:

- $\langle \tilde{y}\mu, \tilde{y}\mu + z \rangle = \|\mu\|^2 + \mathsf{N}(0, \|\mu\|^2)$.

Signal effect on training data:

- $\langle y_k \mu, \tilde{y}_k \mu + z_k \rangle =$
  $y_k \tilde{y}_k \|\mu\|^2 + \mathsf{N}(0, \|\mu\|^2)$.
  $y_k \tilde{y}_k$ is negative for $k \in \mathcal{N}$.

Overfitting component effect on clean test:

- $|\langle \tilde{y}\Delta_n, \tilde{y}\mu + z \rangle| = O\left(\|\mu\|^2/\sqrt{n}\right) + O(\sqrt{d/n})$.

Overfitting component effect on training:

- $2\sqrt{n}\langle \Delta_n, y_k z_k \rangle =$
  $\|z_k\|^2 + \sum_{i \neq k} \langle y_i z_i, y_k z_k \rangle \gtrsim d$ if $d \gg n^2$.
- $\langle y_k \Delta_n, \tilde{y}_k \mu + z_k \rangle \geq \Omega(d/\sqrt{n}) - O(\|\mu\|/\sqrt{n})$.

Proof idea: signal + overfitting component decomposition

$$u = \sum_{i=1}^n y_i x_i \propto \boxed{\mu} + \boxed{\frac{1}{|\mathcal{C}|-|\mathcal{N}|} \sum_{i=1}^n y_i z_i} \approx \boxed{\mu} + \boxed{\frac{1}{n(1-2p)} \sum_{i=1}^n y_i z_i} =: \mu + \Delta_n.$$

Since $z_i \sim \mathsf{N}(0, I_d)$, for $p = 1/4$ have $\Delta_n \sim \mathsf{N}(0, \frac{1}{4n} I_d)$.

Signal effect on clean test data:

- $\langle \tilde{y}\mu, \tilde{y}\mu + z\rangle = \|\mu\|^2 + \mathsf{N}(0, \|\mu\|^2)$.

Signal effect on training data:

- $\langle y_k\mu, \tilde{y}_k\mu + z_k\rangle =$ $y_k\tilde{y}_k\|\mu\|^2 + \mathsf{N}(0, \|\mu\|^2)$. $y_k\tilde{y}_k$ is negative for $k \in \mathcal{N}$.

Overfitting component effect on clean test:

- $|\langle \tilde{y}\Delta_n, \tilde{y}\mu + z\rangle| = O\left(\|\mu\|^2/\sqrt{n}\right) + O(\sqrt{d/n})$.

Overfitting component effect on training:

- $2\sqrt{n}\langle \Delta_n, y_k z_k\rangle =$ $\|z_k\|^2 + \sum_{i\neq k}\langle y_i z_i, y_k z_k\rangle \gtrsim d$ if $d \gg n^2$.
- $\langle y_k\Delta_n, \tilde{y}_k\mu + z_k\rangle \geq \Omega(d/\sqrt{n}) - O(\|\mu\|/\sqrt{n})$.

- If $\|\mu\|^2 \gg \sqrt{d/n}$, Signal dominates effect on clean test data.
- If $d \gg \|\mu\|$, $d/\sqrt{n} \gg \|\mu\|^2$, Overfitting dominates effect on training.
- Simultaneously satisfied if e.g. $\|\mu\| = \Theta(d^\beta)$, $\beta \in (1/4, 1/2)$, and $d \gg n^{\frac{1}{1-2\beta}}$.

Conclusion

- Implicit bias of gradient flow in two-layer leaky ReLU nets when data is 'nearly-orthogonal':

$$\|x_k\|^2 \gg n \max_{i \neq j} |\langle x_i, x_j \rangle|.$$

- KKT points of max-margin problem for two-layer leaky ReLU nets have linear decision boundaries given by τ-uniform classifiers:

$$\operatorname{sgn}\big(f(x;V)\big) = \operatorname{sgn}\big(\langle z, x\rangle\big), \quad z = \textstyle\sum_{i=1}^n s_i y_i x_i, \quad \max_{i,j} {}^{s_i}/_{s_j} = O(1).$$

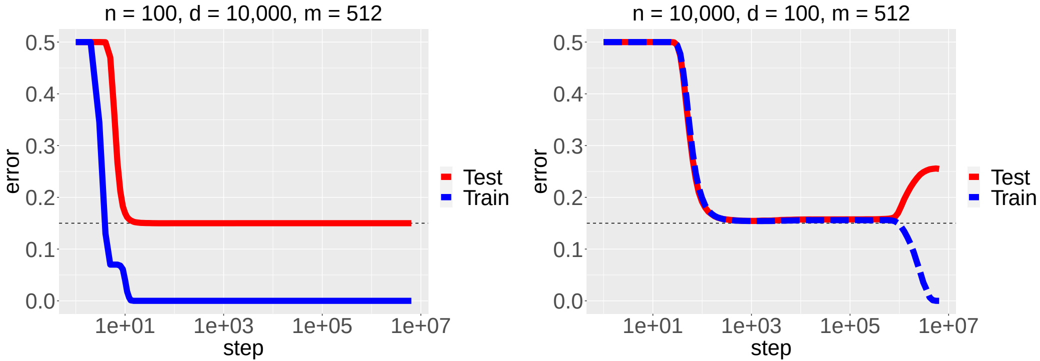
- Under certain distributional assumptions and if $d \gg n$, τ-uniform classifiers exhibit benign overfitting.
- In opposing cluster setting, such classifiers decomposed into 'signal' and 'overfitting' components which are in tension but can be balanced.

Surprises in neural networks trained by gradient descent

- $d \gg n$ necessary for benign overfitting in *linear* models, but unknown if necessary for neural networks
- What happens in two-layer leaky nets on opposing cluster data when $n \gg d$?

Surprises in neural networks trained by gradient descent

- $d \gg n$ necessary for benign overfitting in *linear* models, but unknown if necessary for neural networks
- What happens in two-layer leaky nets on opposing cluster data when $n \gg d$?



- Learning dynamics different in $n > d$ setting; overfitting less 'benign'
  $\longrightarrow$ "Blessing of dimensionality"?

Benign overfitting for leaky ReLU networks

Let $f(x;W) = \sum_{j=1}^{m} a_j \phi(\langle w_j, x\rangle)$, $\phi(q) = \max(q, \gamma q)$, and max-margin problem,

$$\min_{W} \|W\|_F^2 \quad \text{s.t.} \quad y_i f(x_i; W) \geq 1, \text{ for all } i \in [n]. \tag{1}$$

Theorem [F.*-Vardi*-Bartlett-Srebro, COLT'23]

Let V be a KKT point of (1). For opposing cluster data, under (A1)-(A3), w.p. at least 99%:

1. There exists $z \in \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$, $\mathrm{sgn}(\langle z, x\rangle) = \mathrm{sgn}(f(x;V))$.
2. $z \propto \sum_{i=1}^{n} s_i y_i x_i$ where $\max_{i,j} s_i/s_j \leq \frac{51}{49}\gamma^{-2}$, i.e. z is τ-uniform for $\tau \leq \frac{51}{49}\gamma^{-2}$.
3. For noise rate $p \leq 0.49\gamma^2$, $p \leq \mathbb{P}_{(x,y)\sim \mathsf{P}}\big(y \neq \mathrm{sgn}(f(x;V))\big) \leq p + \exp\left(-n\|\mu\|^4/C'd\right)$.

And for any initialization $W(0)$, gradient flow converges in direction to a net satisfying above.

- Test error does not depend on number of neurons.
- $m = 1$, $\gamma \to 1$: leaky ReLU net becomes linear max-margin, tolerates close to $p = 1/2$.