Model: $F(w) = h_w$ **Model Class:** $\mathcal{H} = \text{range}(F)$

$f(w, x) = h_w(x) =$ prediction on x with params ("weights") w

Linear models: $f(w, x) = \langle \beta_w, x \rangle$ $F(w) = \beta_w$

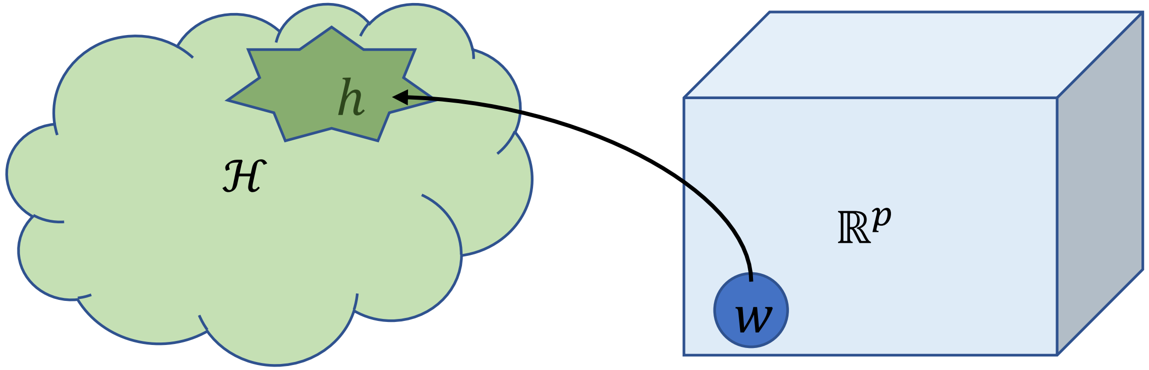
Loss: $L_S(w) = \frac{1}{m} \sum_i \ell(f(w, x_i), y_i)$ e.g. $\ell(\hat{y}, y) = (\hat{y} - y)^2$

GD on $L_s(w)$: $w_{k+1} = w_k - \eta \nabla_w L_S(w)$ $F(w_k) \to ???$

With $\eta \to 0$: $\dot{w}(t) = -\nabla_w L_S(w)$ $F\big(w(t)\big) \to ???$

D-homogenous: $F(cw) = c^D F(w)$, i.e. $f(cw, x) = c^D f(w, x)$

- 1-homogenous: standard linear $F(w) = w$, $f(w, x) = \langle w, x \rangle$
- 2-homogenous:
  - Matrix factorization $F(U, V) = UV$
  - 2-Layer ReLU: $f(W, x) = \sum_j w_{2,j} \left[\langle w_{1,j}, x \rangle\right]_+$
- D-homogenous:
  - D layer linear network
  - D layer linear conv net
  - D layer ReLU net



Model: $\boldsymbol{F(w) = h_w}$ **Model Class:** $\boldsymbol{\mathcal{H} = \mathrm{range}(F)}$

$f(w, x) = h_w(x)$ = prediction on x with params ("weights") w

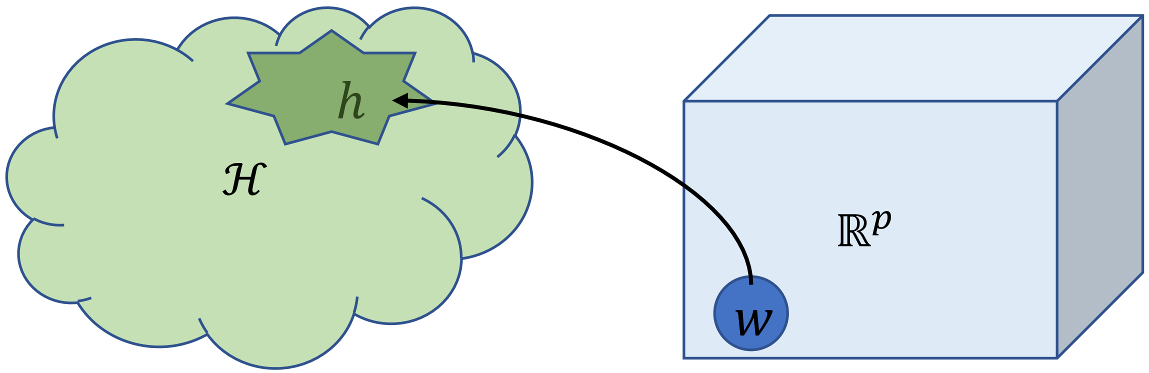
Linear models: $f(w, x) = \langle \beta_w, x \rangle$ $\quad F(w) = \beta_w$

Loss: $L_S(w) = \frac{1}{m} \sum_i \ell(f(w, x_i), y_i)$ $\quad$ e.g. $\ell(\hat{y}, y) = (\hat{y} - y)^2$

GD on $L_s(w)$: $w_{k+1} = w_k - \eta \nabla_w L_S(w)$ $\quad F(w_k) = h_{w_k} \rightarrow ???$

With $\eta \rightarrow 0$: $\dot{w}(t) = -\nabla_w L_S(w)$ $\quad F\big(w(t)\big) = h_{w(t)} \rightarrow ???$

- How is the optimization geometry and dynamics on h (or β), and the implicit bias effected by the parametrization?
- How is it related, or different, from explicitly $\|w\|_2$ regularization?
- How is it effect by optimization choices?
- How it is related to the Kernel regime?



Model: $\boldsymbol{F(w) = h_w}$ **Model Class:** $\boldsymbol{\mathcal{H} = \mathrm{range}(F)}$

$f(w,x) = h_w(x) =$ prediction on x with params ("weights") w

Linear models: $f(w,x) = \langle \beta_w, x \rangle$ $F(w) = \beta_w$

Loss: $L_S(w) = \frac{1}{m}\sum_i \ell(f(w,x_i), y_i)$

<u>**Kernel Regime**</u>: training behaves according to 1st order approximation about $w^{(0)}$,

$$\boldsymbol{f(w,x) \approx f(w_0,x) + \langle w - w_0, \phi_0(x) \rangle}$$

where: $\phi_o(x) = \nabla_w f(w_0, x)$ corresponding to the ***tangent kernel***

$$K_0(x,x') = \langle \nabla_w f(w_0,x), \nabla_w f(w_0,x') \rangle$$

(we will focus on "unbiased initialization": $f(w^{(0)}, x) = 0$, i.e. $h_0 = 0$)

In this regime, $\boldsymbol{w(t) \rightarrow \mathrm{argmin}_{L_S(w)=0} \|w - w_0\|_2}$ **and** $\boldsymbol{h_{w(t)} \rightarrow \min_{h(x_i)=y_i} \|h - h_0\|_{K_0}}$

[Jacot et al 2018]: Width→∞ leads to Kernel Regime

[Chizat Bach 2018]: Scale→∞ leads to Kernel Regime

Kernel Regime and Scale of Init

- For D-homogenous model, $f(cw,x) = c^D f(w,x)$, consider gradient flow with:

  $\dot{w}_\alpha = -\nabla L_S(w)$ and $w_\alpha(0) = \alpha w_0$ with unbiased $f(w_0,x) = 0$

  We are interested in $w_\alpha(\infty) = \lim_{t\to\infty} w_\alpha(t)$

- For squared loss, under some conditions [Chizat and Bach 18]:

$$\lim_{\alpha\to\infty} \sup_t \left\| w_\alpha\left(\frac{1}{\alpha^{D-1}} t\right) - w_K(t) \right\| = 0$$

Gradient flow of linear least squares w.r.t tangent kernel K_0 at initialization
$\dot{w}_K = -\nabla_w \hat{L}\left(x \mapsto \langle w, \phi_{K_0}(x)\rangle\right)$

and so $\boldsymbol{f}(\boldsymbol{w}_{\boldsymbol{\alpha}}(\infty), \boldsymbol{x}) \xrightarrow{\boldsymbol{\alpha}\to\infty} \widehat{\boldsymbol{h}}_{\boldsymbol{K}}(\boldsymbol{x})$ where $\widehat{\boldsymbol{h}}_{\boldsymbol{K}} = \mathbf{arg\,min} \|\boldsymbol{h}\|_{\boldsymbol{K}_0}$ $\mathbf{s.t.}\ \boldsymbol{h}(\boldsymbol{x}_i) = \boldsymbol{y}_i$

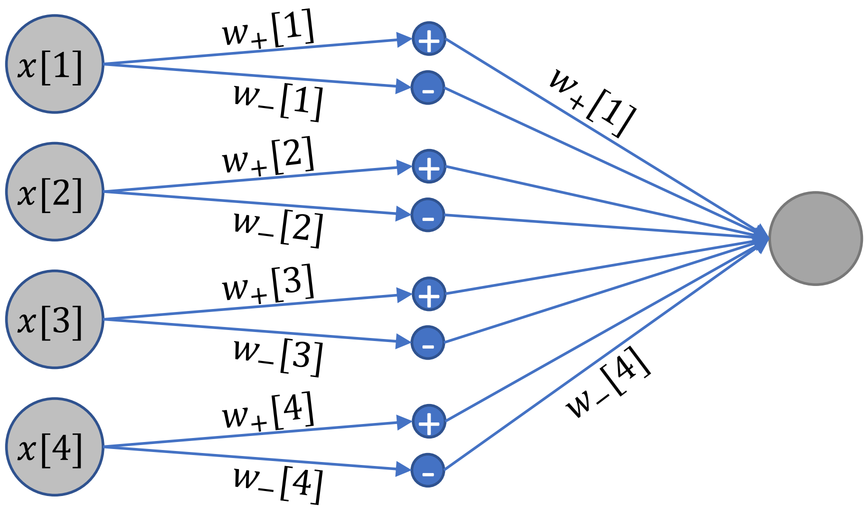
Consider a "linear diagonal net" (ie linear regression with squared parametrization):

$$f(w,x) = \sum_j (w_+[j]^2 - w_-[j]^2)x[j] = \langle \beta(w), x \rangle \qquad \text{with } \beta(w) = w_+^2 - w_-^2$$

And initialization $w_\alpha(0) = \alpha \mathbf{1}$ (so that $\beta(w_\alpha(0)) = 0$).

What's the implicit bias of grad flow w.r.t square loss $L_s(w) = \sum_i (f(w,x_i) - y_i)^2$?

$$\beta_\alpha(\infty) = \lim_{t\to\infty} \beta(w_\alpha(t))$$



$$f(w,x) = w^\top diag(w) \begin{bmatrix} +x \\ -x \end{bmatrix}$$

$$\beta(t) = w_+(t)^2 - w_-(t)^2$$

$$L = \|X\beta - y\|_2^2$$

$$\dot{w}_+(t) = -\nabla_{w_+} L(t) = -2X^\top r(t) \circ \frac{d\beta}{dw_+}$$

$$\beta(t) = w_+(t)^2 - w_-(t)^2 \qquad L = \|X\beta - y\|_2^2$$

$$\dot{w}_+(t) = -\nabla_{w_+} L(t) = -2X^\top r(t) \circ 2w_+(t) \qquad w_+(t) = w_+(0) \circ \exp\left(-2X^\top \int_0^t r(\tau)\, d\tau\right)$$

$$\dot{w}_-(t) = -\nabla_{w_-} L(t) = +2X^\top r(t) \circ 2w_-(t) \qquad w_-(t) = w_-(0) \circ \exp\left(+2X^\top \int_0^t r(\tau)\, d\tau\right)$$

$$\beta(t) = \alpha^2\left(e^{-4X^\top \int_0^t r(\tau)\, d\tau} - e^{4X^\top \int_0^t r(\tau)\, d\tau}\right) \qquad r(t) = X\beta(t) - y$$

$$s = 4\int_0^\infty r(\tau) d\tau \in \mathbb{R}^m$$

$$\beta(\infty) = \alpha^2\left(e^{-X^\top s} - e^{X^\top s}\right) = 2\alpha^2 \sinh X^\top s$$

$$X\beta(\infty) = y$$

$$\min Q(\beta) \quad s.t.\ X\beta = y$$

$$\nabla Q(\beta^*) = X^\top \nu$$

$$X\beta^* = y$$

$$\beta(\infty) = \alpha^2 \left(e^{-X^T s} - e^{X^\top s}\right) = 2\alpha^2 \sinh X^\top s$$

$$X\beta(\infty) = y$$

$$\nabla Q(\beta) = \sinh^{-1}\frac{\beta}{2\alpha^2}$$

$$Q(\beta) = \sum_i \int \sinh^{-1}\frac{\beta[i]}{2\alpha^2} = \alpha^2 \sum_i \left(\frac{\beta[i]}{\alpha^2}\sinh^{-1}\frac{\beta[i]}{2\alpha^2} - \sqrt{4 + \left(\frac{\beta[i]}{\alpha^2}\right)^2} \right)$$

$$\min Q(\beta) \quad s.t. \; X\beta = y$$

$$\nabla Q(\beta^*) = X^\top \nu$$

$$X\beta^* = y$$

$$\sinh^{-1}\frac{\beta(\infty)}{2\alpha^2} = X^\top s$$

$$X\beta(\infty) = y$$

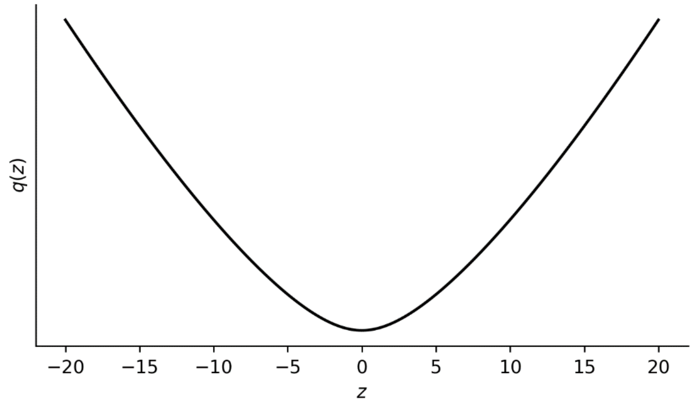
$$f(w,x) = \sum_j (w_+[j]^2 - w_-[j]^2)x[j] = \langle \beta(w), x\rangle \quad \text{with } \beta(w) = w_+^2 - w_-^2$$

$$\beta_\alpha(\infty) = \lim_{t\to\infty} \beta(w_\alpha(t)) \quad \text{where } w_\alpha(0) = \alpha\mathbf{1} \text{ and } \quad \dot{w}_\alpha = -\nabla \sum_i (f(w_\alpha, x_i) - y_i)^2$$

$$\beta_\alpha(\infty) = \arg\min_{X\beta=y} Q_\alpha(\beta)$$

where $Q_\alpha(\beta) = \sum_j q\left(\frac{\beta[j]}{\alpha^2}\right)$ and

$$q(b) = 2 - \sqrt{4+b^2} + b\sinh^{-1}\left(\frac{b}{2}\right)$$



$$\beta_\alpha(\infty) \xrightarrow{\alpha\to\infty} \hat{\beta}_{L2} = \arg\min_{X\beta=y} \|\beta\|_2$$

"Kernel Regime" with NTK $K_0(x,x') = 4\langle x, x'\rangle$

$$\alpha \geq \sqrt{2(1+\epsilon)\left(1+\frac{2}{\epsilon}\right)\|\boldsymbol{\beta}^*_{L2}\|_2} \implies \left\|\hat{\boldsymbol{\beta}}_\alpha\right\|_2^2 \leq (1+\epsilon)\|\boldsymbol{\beta}^*_{L2}\|_2^2$$

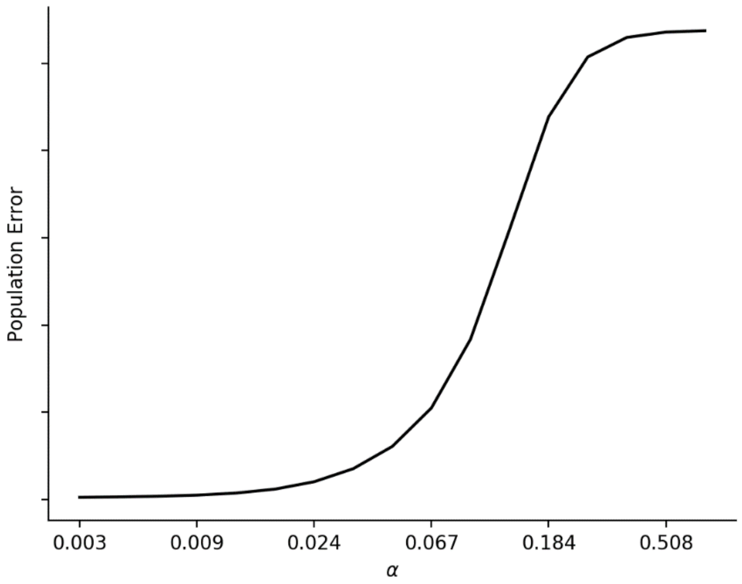
$$\beta_\alpha(\infty) \xrightarrow{\alpha\to 0} \hat{\beta}_{L1} = \arg\min_{X\beta=y} \|\beta\|_1$$

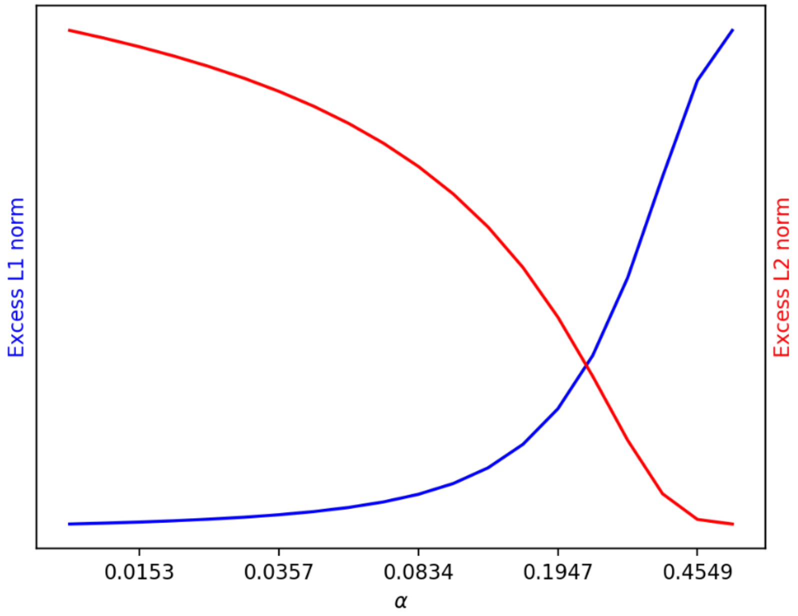
"Rich Regime"

$$\alpha \leq \min\left\{(2(1+\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1)^{-\frac{2+\epsilon}{2\epsilon}}, \exp\left(-\frac{d}{\epsilon\|\boldsymbol{\beta}^*_{L1}\|_1}\right)\right\} \implies \left\|\hat{\boldsymbol{\beta}}_\alpha\right\|_1 \leq (1+\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1$$

Sparse Learning

$$y_i = \langle \beta^*, x_i \rangle + N(0{,}0.01)$$
$$d = 1000, \qquad \|\beta^*\|_0 = 5, \qquad m = 100$$

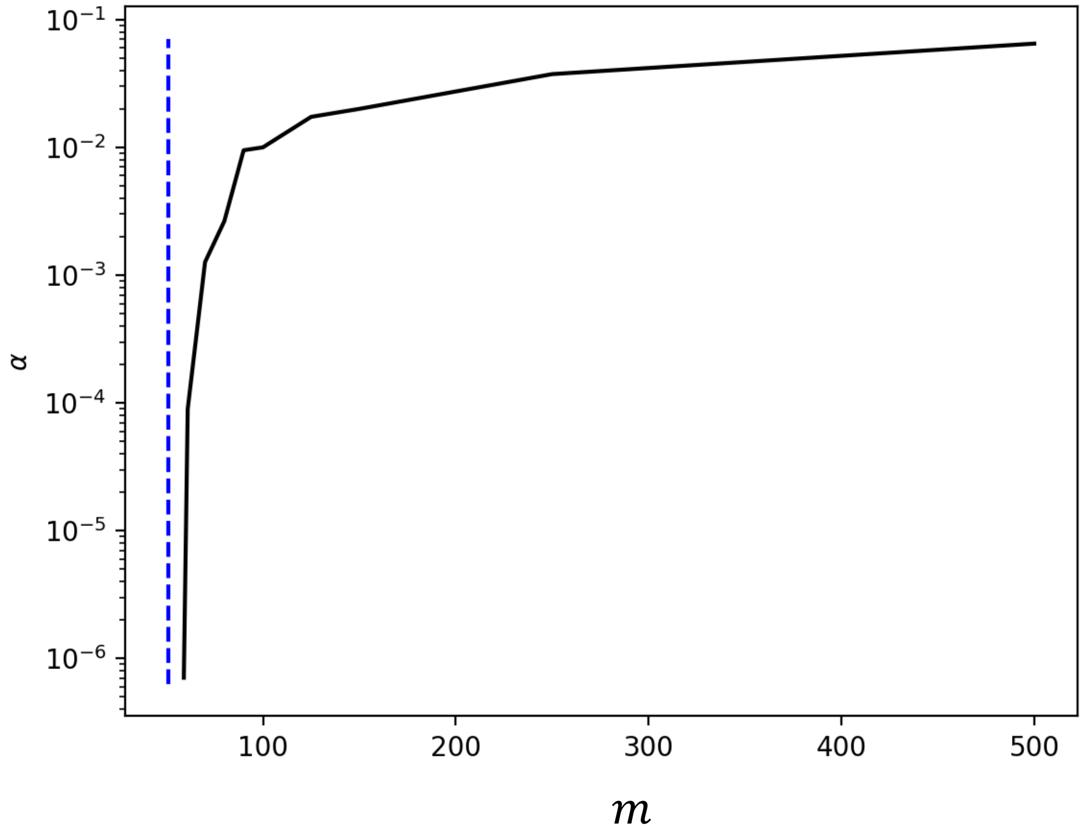




Sparse Learning

$$y_i = \langle \beta^*, x_i \rangle + N(0{,}0.01)$$
$$d = 1000, \qquad \|\beta^*\|_0 = k$$

How small does α need to be to get $L\big(\beta_\alpha(\infty)\big) < 0.025$



$$\beta = F(w) = w_+^2 - w_-^2 \qquad \dot{w} = \nabla_w L(\beta) \qquad w(0) = \alpha\mathbf{1}$$

$$\dot{\beta} = \frac{d\beta}{dw}\dot{w} = -\nabla F\big(w(t)\big)^\top \Big(\nabla F\big(w(t)\big)\nabla L\big(\beta(t)\big)\Big) = -\rho^{-1}\nabla L(\beta)$$

$$\rho = \Big(\nabla F\big(w(t)\big)^\top \nabla F\big(w(t)\big)\Big)^{-1} = diag(w_+^2 + w_-^2)^{-1}$$

$$\nabla F^\top = [diag(w_+)\; diag(w_-)]$$

Problem 1: $w(t)$ as a function of $\beta(t)$

Claim: $\frac{d}{dt}\big(w_+(t)w_-(t)\big) = -2X^\top r(t)\big(w_+(t)\circ w_-(t) - w_-(t)\circ w_+(t)\big) = 0$

➔ $w_+(t)w_-(t) = \alpha^2$, combined with $\beta = w_+^2 - w_-^2$ ➔ $w_\pm^2(t) = \frac{\sqrt{\beta^2+4\alpha^4}\pm\beta}{2}$

➔ $\rho^{-1} = diag(w_+^2 + w_-^2) = diag\left(\sqrt{\beta^2 + 4\alpha^4}\right)$

Induced dynamics: $\dot{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = -\sqrt{\boldsymbol{\beta}_{\boldsymbol{\alpha}}^2 + 4\boldsymbol{\alpha}^4} \odot \boldsymbol{\nabla L_s(\beta_\alpha)}$

Problem 2: Is $\rho = diag(\beta^2 + 4\alpha^4)^{-\frac{1}{2}}$ a Hessian map? Solve $\rho = \nabla^2\Psi$

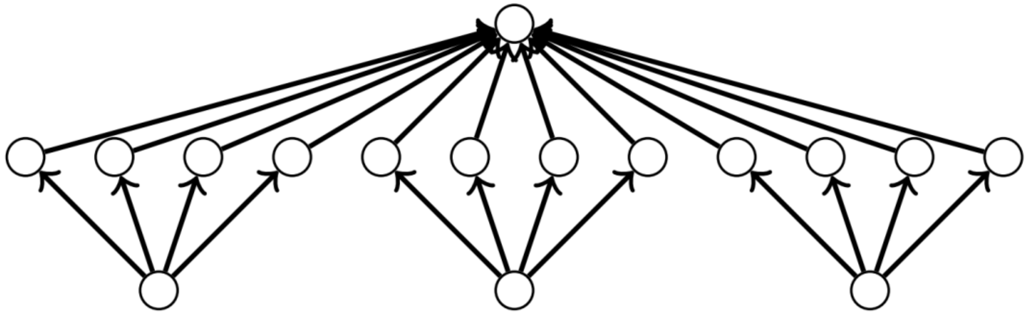
$$\Psi = \sum_i \int\int \left(\beta_i^2 + 4\alpha^4\right)^{-\frac{1}{2}} d\beta_i d\beta_i = \alpha^2 \sum_i \left(const - \sqrt{4 + \left(\frac{\beta}{\alpha^2}\right)^2} + \frac{\beta}{\alpha^2}\sinh^{-1}\left(\frac{\beta}{2\alpha^2}\right)\right)$$

Width and Initialization Scale

$$f\big((U,V),x\big) = \sum_{i=1..d,j=1..\boldsymbol{k}} u_{i,j} v_{i,j} x[i] = \langle UV^\top, diag(x)\rangle$$

$$U,V \in \mathbb{R}^{d\times \boldsymbol{k}} \quad \beta_{U,V} = diag(UV^\top)$$

- Initialization: $u_{i,j}, v_{i,j} \sim iid\ N\left(0, \sigma^2 = \frac{\alpha^2}{\sqrt{\boldsymbol{k}}}\right)$ s.t. $\mathbf{Var}\big[\beta(0)[i]\big] = \alpha^2$



Width and Initialization Scale

$$f\big((U,V),x\big) = \sum_{i=1..d,j=1..\boldsymbol{k}} u_{i,j} v_{i,j} x[i] = \langle UV^\top, diag(x)\rangle$$

$$U,V \in \mathbb{R}^{d\times \boldsymbol{k}} \quad \beta_{U,V} = diag(UV^\top)$$

- Initialization: $u_{i,j}, v_{i,j} \sim iid\ N\left(0, \sigma^2 = \frac{\alpha^2}{\sqrt{\boldsymbol{k}}}\right)$ s.t. $\mathbf{Var}\big[\beta(0)[i]\big] = \alpha^2$
- Symmetrized problem: $\tilde{f}(W,x) = \langle WW^\top, \tilde{X}\rangle$

$W = \begin{bmatrix} U \\ V \end{bmatrix}$ so $WW^\top = \begin{bmatrix} UU^\top & UV^\top \\ VU^\top & VV^\top \end{bmatrix}$

$\tilde{X} := \frac{1}{2}\begin{bmatrix} 0 & diag(x) \\ diag(x) & 0 \end{bmatrix}$

- Relevant scale: $WW^\top \approx \sqrt{\boldsymbol{k}}\alpha^2 I$
  - $\beta(\infty) \xrightarrow{k\to\infty,\alpha\to 0} \arg\min_{X\beta=y} Q_{\boldsymbol{\mu}}(\beta)$

    $\mu = \lim \alpha\sqrt[4]{k} = \lim \sigma\sqrt{k}$
  - $\alpha = o\big(1/\sqrt[4]{k}\big)$, i.e. $\sigma = o\big(1/\sqrt{k}\big)$ ➔ ℓ_1
  - $\alpha = \omega\big(1/\sqrt[4]{k}\big)$, i.e. $\sigma = \omega\big(1/\sqrt{k}\big)$ ➔ ℓ_2
  - $\sqrt{\boldsymbol{k}}\alpha^2 \to 0$ leads to the kernel regime, even if $|\beta(0)| \approx \alpha^2 \to 0$

Width and Initialization Scale

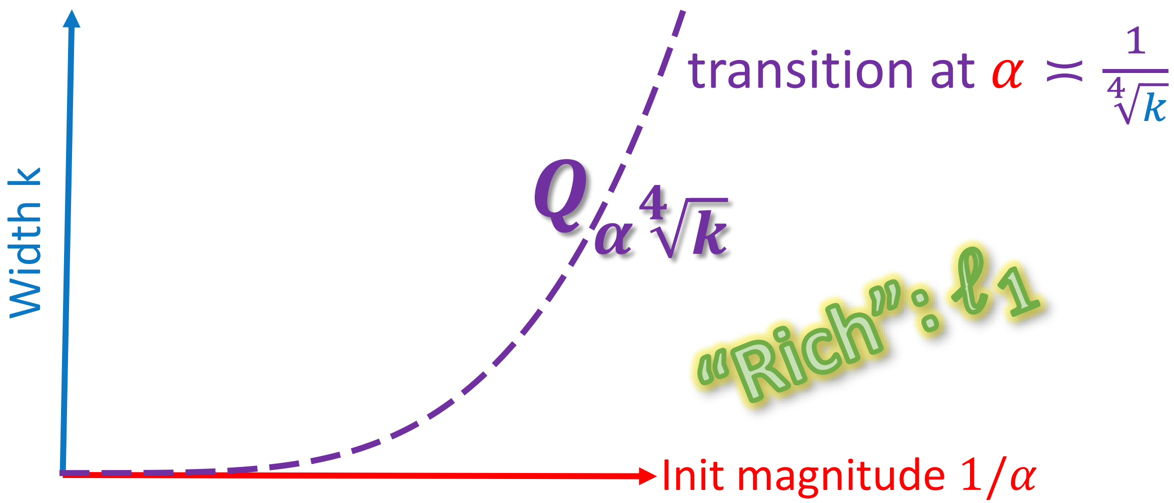
$$f\big((U,V),x\big) = \sum_{i=1..d,\, j=1..k} u_{i,j} v_{i,j} x[i] = \langle UV^\top, diag(x) \rangle$$

$U,V \in \mathbb{R}^{d\times k}$ $\quad \beta_{U,V} = diag(UV^\top)$

- Initialization: $u_{i,j}, v_{i,j} \sim iid\ N\left(0, \sigma^2 = \frac{\alpha^2}{\sqrt{k}}\right)$ s.t. $\mathbf{Var}\big[\beta(0)[i]\big] = \alpha^2$
- Symmetrized problem: $\tilde{f}(W,x) = \langle WW^\top, \tilde{X} \rangle$

$W = \begin{bmatrix} U \\ V \end{bmatrix}$ so $WW^\top = \begin{bmatrix} UU^\top & UV^\top \\ VU^\top & VV^\top \end{bmatrix}$

$\tilde{X} := \frac{1}{2}\begin{bmatrix} 0 & diag(x) \\ diag(x) & 0 \end{bmatrix}$

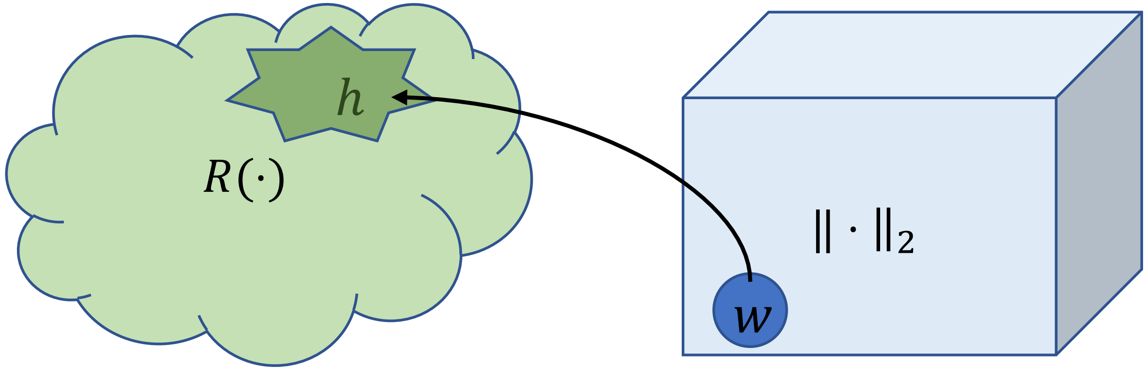


Is implicit bias of GD just ℓ_2 in param space + mapping to func space?

Is initializing to $w(0) = \alpha\mathbf{1}$ the same as regularizing distance to $\alpha\mathbf{1}$?

$$\beta_\alpha^R = F\left(\arg\min_{L_S(w)=0} \|w - \alpha\mathbf{1}\|_2^2\right) = \arg\min_{X\beta=y} R_\alpha(\beta)$$

Where $R_\alpha(\beta) = \min_{F(w)=\beta} \|w - \alpha\mathbf{1}\|_2^2$



Is implicit bias of GD just ℓ_2 in param space + mapping to func space?

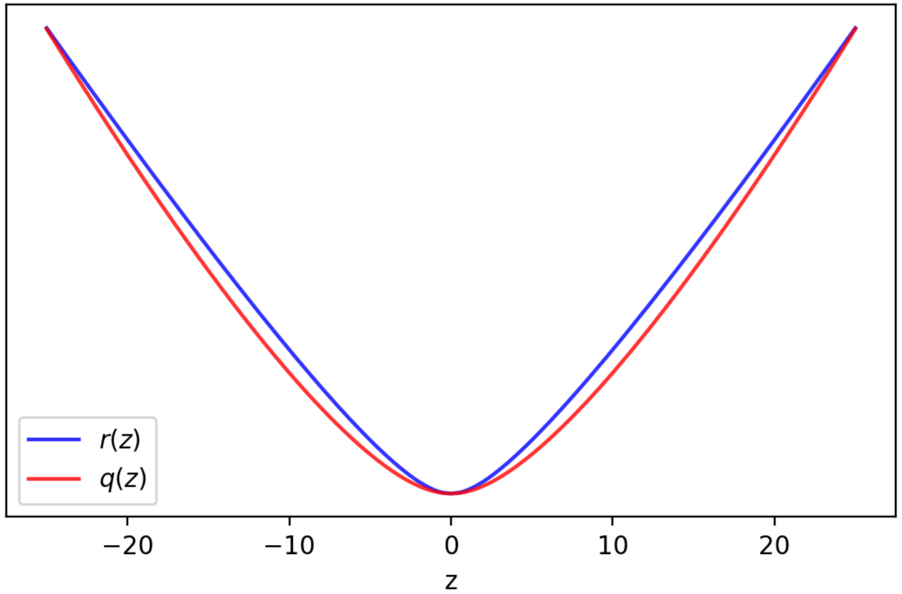
Is initializing to $w(0) = \alpha\mathbf{1}$ the same as regularizing distance to $\alpha\mathbf{1}$?

$$\beta_\alpha^R = F\left(\arg\min_{L_S(w)=0} \|w - \alpha\mathbf{1}\|_2^2\right) = \arg\min_{X\beta=y} R_\alpha(\beta)$$

Where $R_\alpha(\beta) = \min_{F(w)=\beta} \|w - \alpha\mathbf{1}\|_2^2$

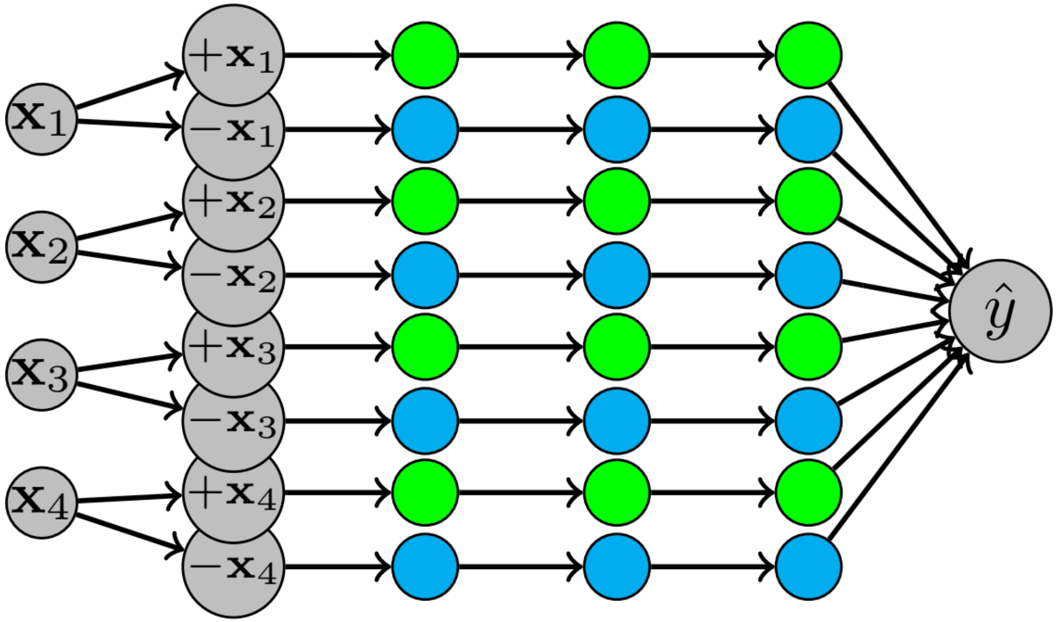
$R_\alpha(\beta) = \sum_j r\left(\frac{\beta[j]}{\alpha^2}\right)$ where $r(b)$ is solution of quartic equation:

$$r^4 - 6r^3 + (12 - 2b^2)r^2 - (8 + 10b^2)r + b^2 + b^4 = 0$$



Deep Diagonal Linear Net

$\beta(t) = w_+(t)^D - w_-(t)^D$



Deep Diagonal Linear Net

$$\beta(t) = w_+(t)^D - w_-(t)^D$$

$$r(t) = X\beta(t) - y$$

$$\beta(t) = \alpha^D \left(\left(1 + \alpha^{D-2} D(D-2) X^\top \int_0^t r(\tau)\, d\tau \right)^{\frac{-1}{D-2}} - \left(1 - \alpha^{D-2} D(D-2) X^\top \int_0^t r(\tau)\, d\tau \right)^{\frac{-1}{D-2}} \right)$$

KKT for $\min Q(\beta)\ \ s.t.\ X\beta = y$:

$$s = \int_0^\infty r(\tau) d\tau \in \mathbb{R}^m$$

$$\nabla Q(\beta^*) = X^\top v$$

$$X\beta^* = y$$

$$\beta(\infty) = \alpha^D h_D(X^\top s)$$

$$X\beta(\infty) = y$$

$$L = \|X\beta - y\|_2^2$$

$$\frac{\mathrm{d}w}{\mathrm{d}t} = -\nabla L$$

$$h_D(z) = \alpha^D \left((1 + \alpha^{D-2} D(D-2) z)^{\frac{-1}{D-2}} - (1 - \alpha^{D-2} D(D-2) z)^{\frac{-1}{D-2}} \right)$$

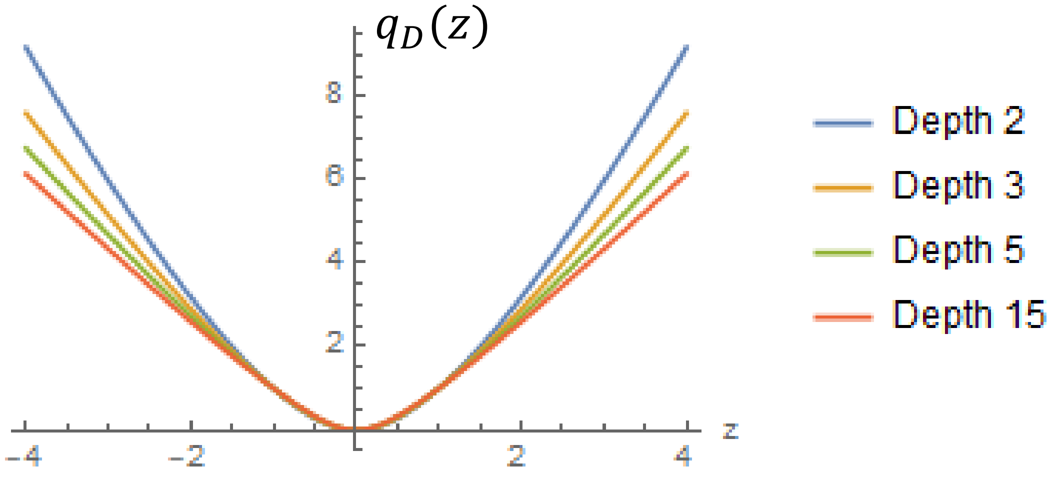
$$\frac{\mathrm{d}\beta}{\mathrm{d}t} = \frac{\mathrm{d}\beta}{\mathrm{d}w} \cdot \frac{\mathrm{d}w}{\mathrm{d}t}$$

$$q_D = \int h_D^{-1}$$

$$Q_D(\beta) = \sum_i q_D\left(\frac{\beta[i]}{\alpha^D}\right)$$

Deep Diagonal Linear Net

$$\beta(t) = w_+(t)^D - w_-(t)^D \qquad \beta(\infty) = \arg\min Q_D\left({}^{\beta}/_{\alpha^D}\right) \;\; s.t.\; X\beta = y$$



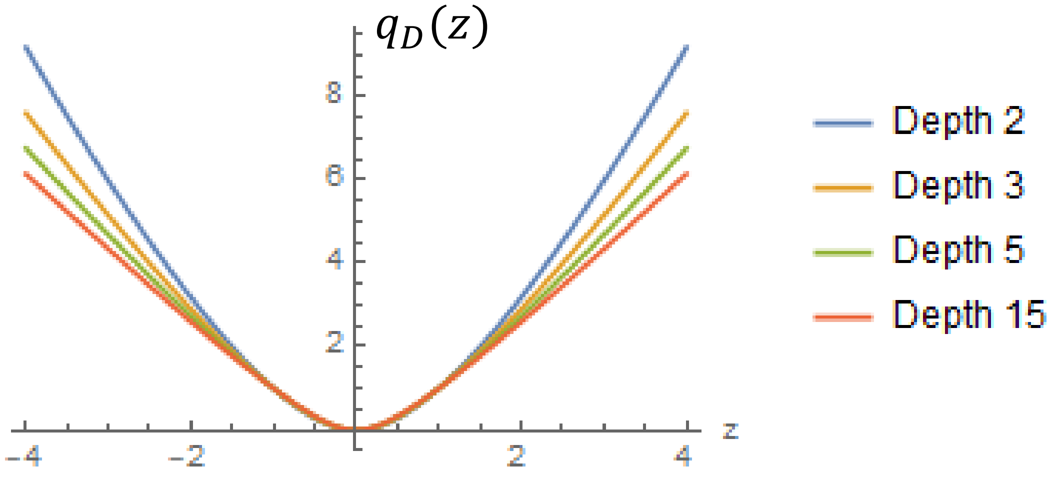
$$h_D(z) = \alpha^D\left((1+\alpha^{D-2}D(D-2)z)^{\frac{-1}{D-2}} - (1-\alpha^{D-2}D(D-2)z)^{\frac{-1}{D-2}}\right)$$

$$q_D = \int h_D^{-1}$$

$$Q_{D,\alpha}(\beta) = \sum_i q_D\left(\frac{\beta[i]}{\alpha^D}\right)$$

Deep Diagonal Linear Net

$$\beta(t) = w_+(t)^D - w_-(t)^D \qquad \beta(\infty) = \arg\min Q_D\left({}^{\beta}/_{\alpha^D}\right) \; s.t.\; X\beta = y$$



For all depth $D \geq 2$, $\beta(\infty) \xrightarrow{\alpha \to 0} \arg\min_{X\beta=y} \|\beta\|_1$

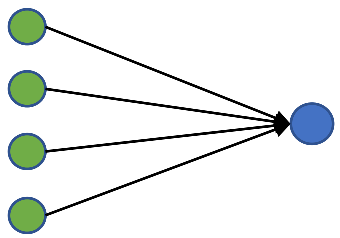
- Contrast with explicit reg: For $R_\alpha(\beta) = \min_{\beta = w_+^D - w_-^D} \|w - \alpha \mathbf{1}\|_2^2$, $R_\alpha(\beta) \xrightarrow{\alpha \to 0} \|\beta\|_{2/D}$

also observed by [Arora Cohen Hu Luo 2019]

- Also with **logistic loss**, $\beta(\infty) \to \propto SOSP\ of\ \|\beta\|_{2/D}$

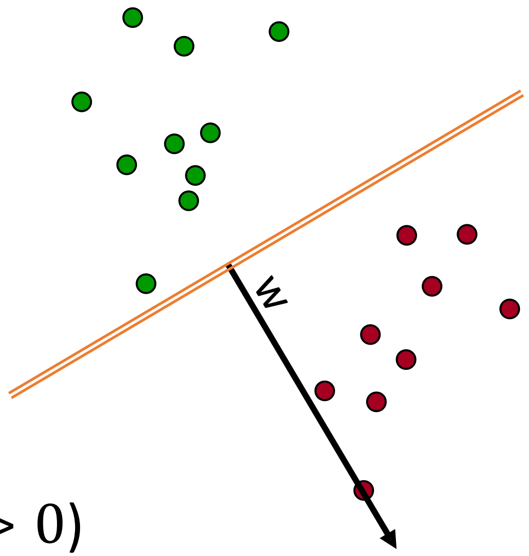
[Gunasekar Lee Soudry Srebro 2018]
[Lyu Li 2019]

Implicit Bias in Logistic Regression



$$\arg\min_{w \in \mathbb{R}^n} \mathcal{L}(w) = \sum_{i=1}^{m} \ell(y_i \langle w, x_i \rangle)$$

$$\ell(z) = \log(1 + e^{-z})$$

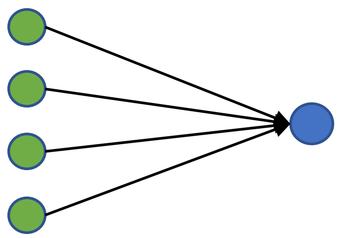


- Data $\{(x_i, y_i)\}_{i=1}^{m}$ linearly separable $(\exists_w \forall_i y_i \langle w, x_i \rangle > 0)$
- Where does gradient descent converge?

$$w(t) = w(t) - \eta \nabla \mathcal{L}(w(t))$$

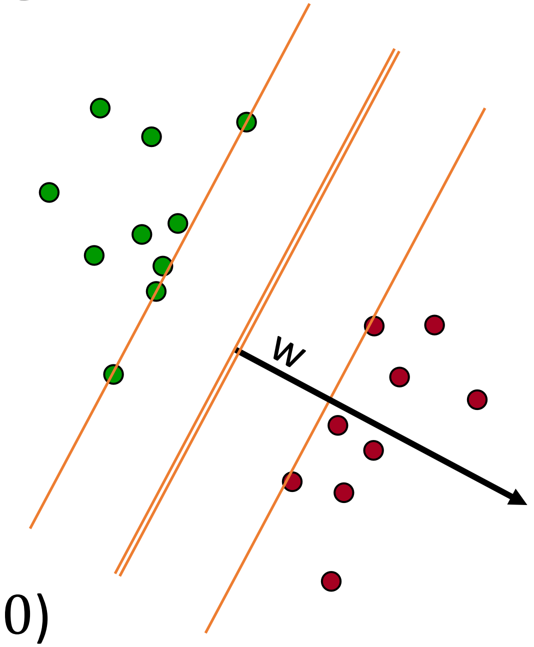
  - $\inf \mathcal{L}(w) = 0$, but minima unattainable
  - GD diverges to infinity: $w(t) \to \infty$, $\mathcal{L}\big(w(t)\big) \to 0$
- **In what direction?** What does $\frac{w(t)}{\|w(t)\|}$ converge to?

Implicit Bias in Logistic Regression



$$\arg\min_{w\in\mathbb{R}^n} \mathcal{L}(w) = \sum_{i=1}^{m} \ell(y_i\langle w, x_i\rangle)$$

$$\ell(z) = \log(1 + e^{-z})$$



- Data $\{(x_i, y_i)\}_{i=1}^m$ linearly separable ($\exists_w \forall_i y_i\langle w, x_i\rangle > 0$)
- Where does gradient descent converge?
$$w(t) = w(t) - \eta\nabla\mathcal{L}(w(t))$$
  - $\inf \mathcal{L}(w) = 0$, but minima unattainable
  - GD diverges to infinity: $w(t) \to \infty$, $\mathcal{L}\big(w(t)\big) \to 0$
- **In what direction?** What does $\frac{w(t)}{\|w(t)\|}$ converge to?
- **<u>Theorem</u>**: $\frac{w(t)}{\|w(t)\|_2} \to \frac{\hat{w}}{\|\hat{w}\|_2}$ $\quad \hat{w} = \arg\min\|w\|_2 \;\; s.t. \forall_i y_i\langle w, x_i\rangle \geq 1$

Implicit Bias in Logistic Regression

$$\arg\min_{w\in\mathbb{R}^n} \mathcal{L}(w) = \sum_{i=1}^{m} \ell(y_i\langle w, x_i\rangle)$$

$$\ell(z) = \log(1 + e^{-z})$$

Theorem: $w(t) = \hat{w}\log t + \rho(t)$, with $\rho(t)$ bounded*

$$\hat{w} = \arg\min\|w\|_2 \;\; s.t.\, \forall_i y_i\langle w, x_i\rangle \geq 1$$

- Holds for any initial point $w(0)$ and stepsize $\eta \leq 2$
- Holds for any monotonically decreasing strictly positive smooth loss s.t. $-\ell'(z)$ has a tight exponential tail. Asymptotically, all behave as:

$$\ell(z) = e^{-z}$$

*For data in general position. With degenerate data, $\rho(t) = O(\log\log t)$

[Soudry Hoffer S ICLR 2018]

Proof sketch: ($y_i = 1$ w.l.og.)

Write $w(t) = g(t)w_\infty + \rho(t)$ with $g(t) \to \infty$ and $\rho(t) = o\big(g(t)\big)$.

Since we converge to zero error, $\forall_i \langle w_\infty, x_i \rangle > 0$

Since the loss derivative has an exponential tail:

$$-\nabla\mathcal{L}(w) \approx \sum_i e^{-\langle w(t), x_i\rangle} x_i^\top = \sum_i e^{-g(t)\langle w_\infty, x_i\rangle - \langle \rho(t), x_i\rangle} x_i^\top$$

As $g(t) \to \infty$, only points with minimal $\langle w_\infty, x_i \rangle$ (points on the margin, "support vectors") will dominate gradient

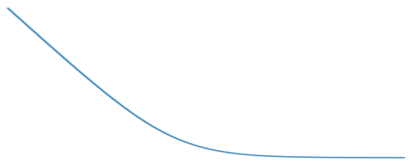
➔ $\nabla\mathcal{L}(w)$ spanned by support vectors

➔ $w(t)$ spanned by support vectors

Define $\widehat{w} = \frac{w_\infty}{\min_i \langle w_\infty, x_i\rangle}$. We have:

$\widehat{w} = \sum \alpha_i w_i \qquad \forall_i (\alpha_i \geq 0 \text{ and } \langle \widehat{w}, x_i \rangle = 1) \text{ OR } (\alpha_i = 0 \text{ and } \langle \widehat{w}, x_i \rangle > 1)$

$$\ell_{\text{logistic}}(h(w),y) = \log\left(1+e^{-yh(w)}\right) \approx e^{-yh(w)} = \ell_{\exp}(h(w),y)$$



Consider gradient descent w.r.t. logistic loss $L_s(w) = \sum_i \ell(f(w,x_i);y_i)$ (or other exp-tail loss) on a D-homogenous model $f(w,x)$:

Theorem [Nacson Gunasekar Lee S Soudry 2019][Lyu Li 2019]:
If $L_s(w) \to 0$, and small enough stepsize (ensuring convergence in direction):

$\boldsymbol{w}_\infty \propto$ **first order stationary point of**

$$\arg\min \|\boldsymbol{w}\|_2 \;\; s.t. \forall_i y_i f(\boldsymbol{w},x_i) \geq 1 \qquad (*)$$

Suggests implicit bias defined by $\boldsymbol{R_F(h)} = \arg\min_{F(w)=h} \|\boldsymbol{w}\|_2$ and

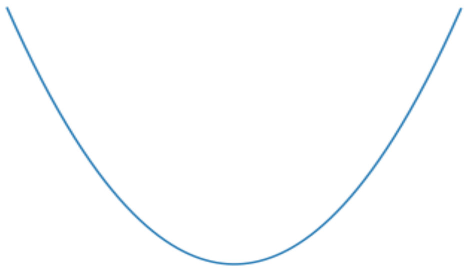
$\boldsymbol{h}_\infty = \boldsymbol{F}(\boldsymbol{w}_\infty) \propto$ **first order stationary point of**

$$\arg\min \boldsymbol{R_F(h)} \; s.t. \, y_i f(x_i) \geq 1 \qquad (**)$$

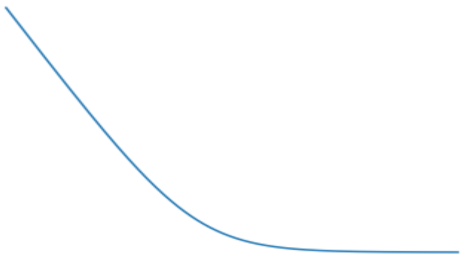
But need to be careful: f.o.s.p of **(*)** does ***not*** imply f.o.s.p of **(**)**

Different Asymptotics

- For least squares (or any other loss with attainable minimum):
  - w_∞ depends on initial point w_0 and stepsize η
  - To get clean characterization, need to take $\eta \to 0$
  - If 0 is a saddle point, need to take $w_0 \to 0$



- For monotone decreasing loss (eg logistic)
  - w_∞ does NOT depend on initial w_0 and stepsize η
  - Don't need $\eta \to 0$ and $w_0 \to 0$
  - What happens at the beginning doesn't effect w_∞



Squared Loss vs Logistic/Exp Loss

$$\ell_{\text{logistic}}(h(w), y) = \log\left(1 + e^{-yh(w)}\right) \approx e^{-yh(w)} = \ell_{\exp}(h(w), y)$$

When $\ell \to 0$, ie $yh(w) \to \infty$

- For squared loss, under some conditions [Chizat and Bach 18]:

$$\lim_{\alpha\to\infty} \sup_t \left\| w_\alpha\left(\frac{1}{\alpha^{D-1}} t\right) - w_K(t) \right\| = 0$$

$$\Rightarrow h_\alpha(\infty) \to \hat{h}_K = \arg\min \|h\|_K \ \text{s.t.}\ h(x_i) = y_i$$

- For logistic:

$$\forall_t \lim_{\alpha\to\infty} \sup_{\boldsymbol{\tau<t}} \left\| w_\alpha\left(\frac{1}{\alpha^{D-1}} \tau\right) - w_K(\tau) \right\| = 0$$

Contrast with [Nacson Gunasekar Lee S Soudry 2019][Lyu Li 2019]:

$$\forall_\alpha \lim_{t\to\infty} \frac{w_\alpha(t)}{\|w_\alpha(t)\|} \propto \text{f.o.s.p of } \arg\min \|w\|_2 \ s.t.\, y_i h(x_i) \geq 1$$

For our model $\beta = w_+^2 - w_-^2$ with logistic loss:

$\beta(\epsilon) = \beta(t)\ \ s.t.\ L_S(\beta) = \epsilon$

Uniquely defined since $L_S\big(beta(t)\big)$ monotonically decreases from 1 to 0.

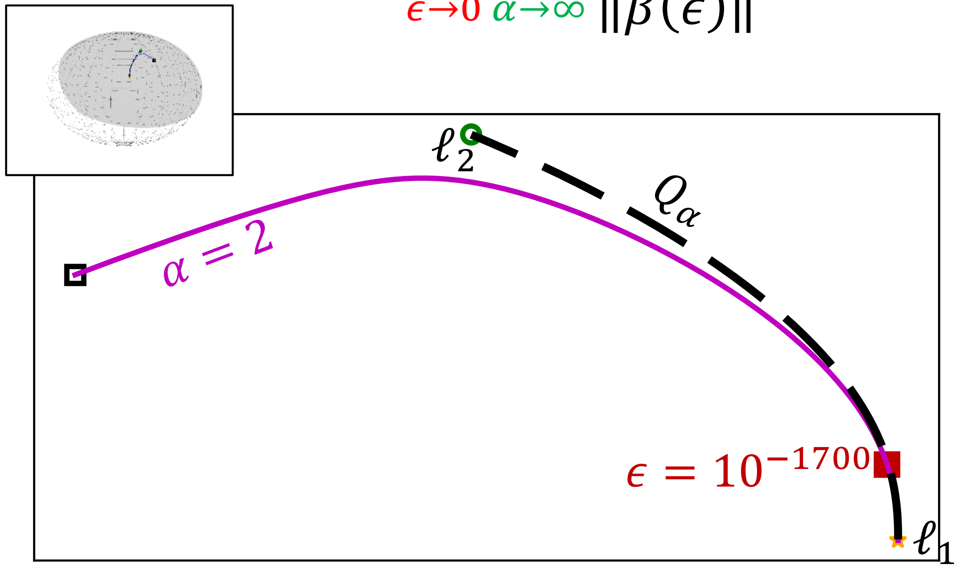
$$\forall_\alpha \lim_{\epsilon \to 0} \frac{\beta(\epsilon)}{\|\beta(\epsilon)\|} \propto \arg\min \|\beta\|_1\ \ s.t.\ y_i x_i^\top \beta \geq 1$$ **Rich**

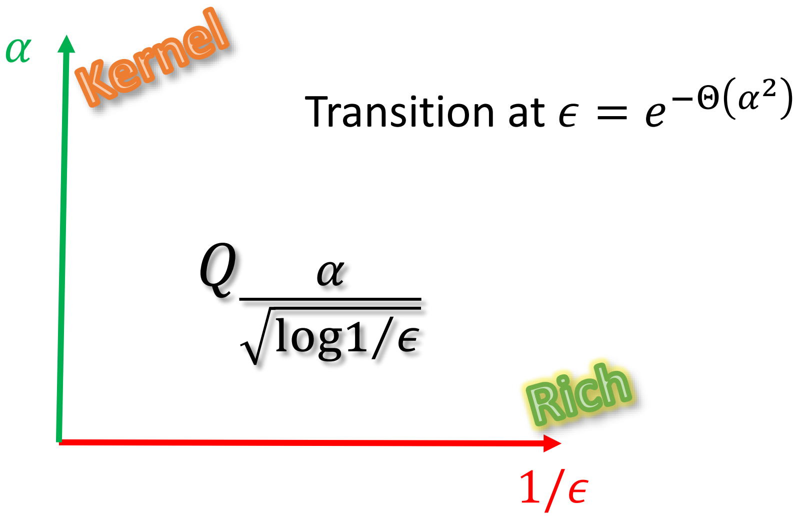
➔

$$\lim_{\alpha \to \infty} \lim_{\epsilon \to 0} \frac{\beta(\epsilon)}{\|\beta(\epsilon)\|} \propto \arg\min \|\beta\|_1\ \ s.t.\ y_i x_i^\top \beta \geq 1$$ **Rich**

Contrast with:

$$\lim_{\epsilon \to 0} \lim_{\alpha \to \infty} \frac{\beta(\epsilon)}{\|\beta(\epsilon)\|} \propto \arg\min \|\beta\|_2\ \ s.t.\ y_i x_i^\top \beta \geq 1$$ **Kernel**

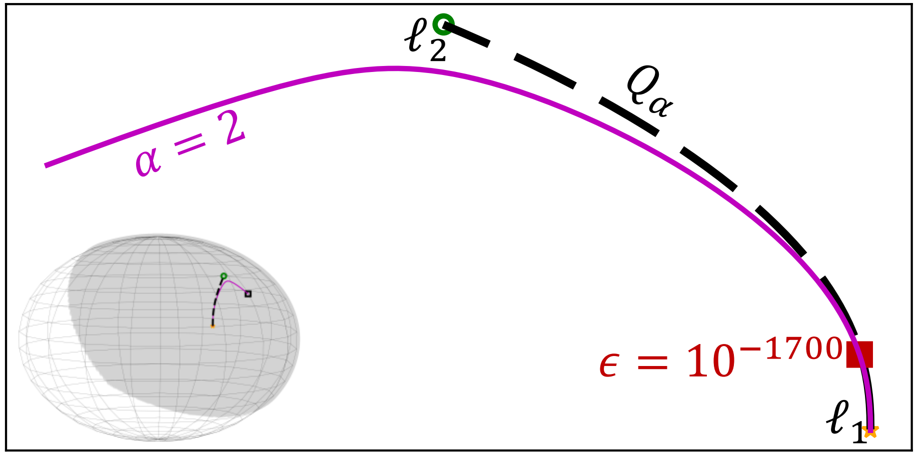




Logistic Loss vs Squared Loss

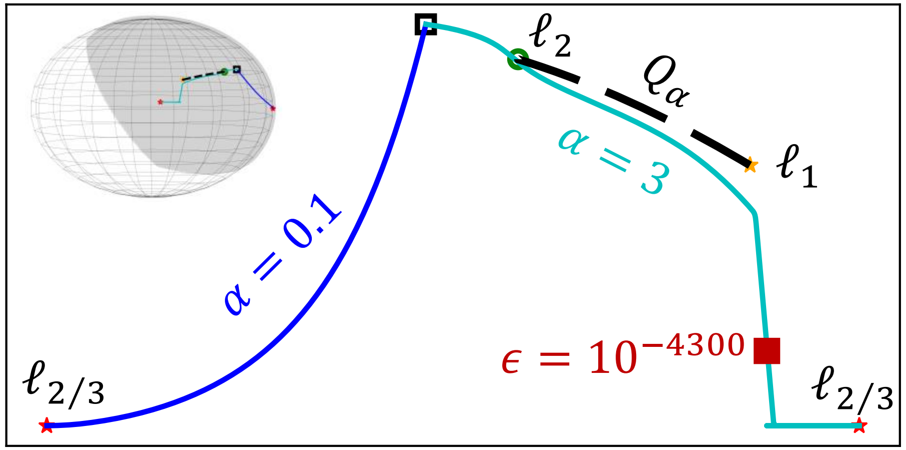
Depth two:

- Square loss: $\beta(\infty) \propto \arg\min_{X\beta=y} Q_\alpha(\beta)$
- Logistic loss: $\forall_\alpha \beta(\infty) \propto \arg\min_{X\beta=y} \|\beta\|_1$



Deeper Diagonal Nets:

- Squared loss, $\beta(\infty) \xrightarrow{\alpha\to 0} \propto \arg\min_{X\beta=y} \|\beta\|_1$
- Logistic loss, $\beta(\infty) \propto SOSP\ of\ \|\beta\|_{2/D}$



Depth=3

[Moroshko Gunasekar Woodworth Lee S Soudry 2020 "Implicit Bias in Deep Linear Classification: Initialization Scale vs Training Accuracy"]

Other Control Choices

- Early Stopping (and not so early stopping)
- Shape/relative scale [Azulay Moroshko Nacson Woodworth S Globerson Soudry 2021]
- Stepsize [Nacson Ravichandran S Soudry 2022]
- **Stochasticity**
  - Batchsize [Pesme Pillaud-Vivien Flammarion 2021]
  - Label noise [HaoChen, Wei, Lee, Ma 2020][Blanc, Gupta, Valiant, Valiant 2020]

…

$$f(w,x) = \sum_j (w_+[j]^2 - w_-[j]^2)x[j] = \langle \beta(w), x \rangle$$

with $\beta(w) = w_+^2 - w_-^2$

Simplest architecture displaying comlex implicit bias phenomena

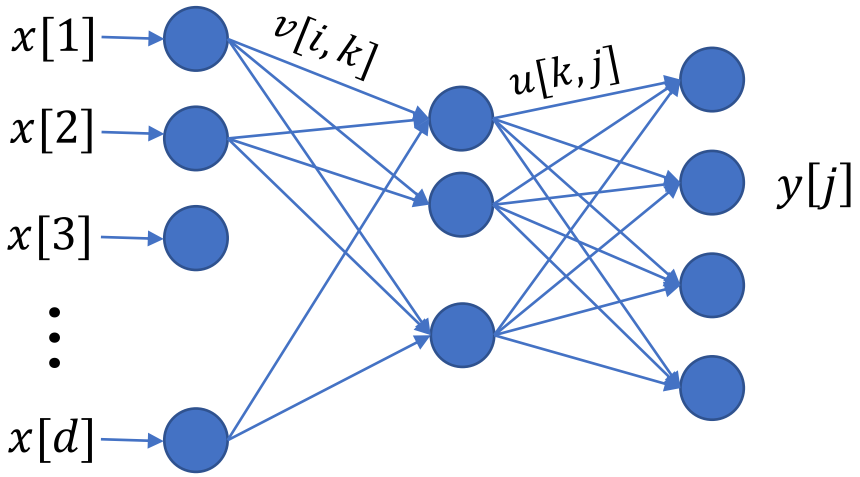
$$\min_{\beta\in\mathbb{R}^{d_1\times d_2}} \hat{L}(\beta) = \|\mathcal{X}(\beta) - y\|_2^2 \quad \mathcal{X}(\beta)_i = \langle X_i, \beta\rangle \qquad X_1,\dots,X_m \in \mathbb{R}^{d_1\times d_2}, y \in \mathbb{R}^m$$

$$\beta = F(U,V) = UV^T, \quad U,V \in \mathbb{R}^{n\times n}$$

GD on U,V: $\dot{U}(t) = -\nabla_U \hat{L}(UV^\top), \dot{V}(t) = -\nabla_V \hat{L}(UV^\top)$

➔ $\dot{\beta} = -(\nabla F^\top \nabla F)\nabla\hat{L}(\beta) = -\left(UU^\top\nabla\hat{L}(\beta) + \nabla\hat{L}(\beta)VV^\top\right)$

- Matrix completion (X_i is indicator matrix)
- Matrix reconstruction from linear measurements
- Multi-task learning ($X_i = e_{task\ of\ example\ i} \cdot \phi(example\ i)^\top$)



$$\min_{\beta\in\mathbb{R}^{d_1\times d_2}} \hat{L}(\beta) = \|\mathcal{X}(\beta) - y\|_2^2 \quad \mathcal{X}(\beta)_i = \langle X_i, \beta\rangle \qquad X_1, \dots, X_m \in \mathbb{R}^{d_1\times d_2}, y \in \mathbb{R}^m$$

$$\beta = F(U,V) = UV^T, \quad U,V \in \mathbb{R}^{n\times n}$$

GD on U,V: $\dot{U}(t) = -\nabla_U \hat{L}(UV^\top), \dot{V}(t) = -\nabla_V \hat{L}(UV^\top)$

$$\Rightarrow \qquad \dot{\beta} = -(\nabla F^\top \nabla F)\nabla\hat{L}(\beta) = -\left(UU^\top \nabla\hat{L}(\beta) + \nabla\hat{L}(\beta)VV^\top\right)$$

$$W = \begin{bmatrix} U \\ V \end{bmatrix}, \ \tilde{\beta} = WW^\top = \begin{bmatrix} UU^\top & UV^\top \\ VU^\top & VV^\top \end{bmatrix} = \begin{bmatrix} UU^\top & \beta \\ \beta^\top & VV^\top \end{bmatrix}$$

$$\min_{\tilde{\beta}\succcurlyeq 0} \hat{L}(\tilde{\beta}) = \left\|\tilde{\mathcal{X}}(\tilde{\beta}) - y\right\|_2^2 \qquad \tilde{\mathcal{X}}(\tilde{\beta})_i = \langle \tilde{X}_i, \beta\rangle \qquad \tilde{X}_i = \begin{bmatrix} 0 & X_i \\ X_i^\top & 0 \end{bmatrix} \in \mathbb{S}_d, y \in \mathbb{R}^m$$

$$\tilde{\beta} = \tilde{F}(W) = WW^T, W \in \mathbb{R}^{d\times d}$$

$$\dot{\tilde{\beta}} = -(\nabla\tilde{F}^\top \nabla\tilde{F})\nabla\hat{L}(\tilde{\beta}) = -\left(WW^\top \nabla\hat{L}(\tilde{\beta}) + \nabla\hat{L}(\tilde{\beta})WW^\top\right) = -\left(\tilde{\beta}\nabla\hat{L}(\tilde{\beta}) + \nabla\hat{L}(\tilde{\beta})\tilde{\beta}\right)$$

$$\min_{\beta \succcurlyeq 0} \hat{L}(\beta) = \|\mathcal{X}(\beta) - y\|_2^2 \qquad \mathcal{X}(\beta)_i = \langle X_i, \beta \rangle \qquad X_i \in \mathbb{S}_d, y \in \mathbb{R}^m$$

$$\dot{\beta} = -\left(\beta \nabla \hat{L}(\beta) + \nabla \hat{L}(\beta)\beta\right) = \left(-\beta \mathcal{X}^*(r(t)) - \mathcal{X}^*\left(r(t)\right)\beta\right)$$

$r(t) = \mathcal{X}(\beta) - y$

If X_i, β_0 commute: $$\beta(t) = e^{\mathcal{X}^*(s_t)}\beta_0 e^{\mathcal{X}^*(s_t)} \qquad s_t = -\int r_t dt \in \mathbb{R}^m$$
$$\in \left\{ e^{\mathcal{A}^*(s)}\beta_0 e^{\mathcal{A}^*(s)} \middle| s \in \mathbb{R}^m \right\}$$

- Independent of “steering” r_t
- Can use other loss, weights, or sample X_i; But finite steps, as well as (infinitesimal) momentum, will fall off $\mathcal{M}$!
- Restricting to joint diagonalization $\beta = U\tilde{\beta}U^\top$, $\rho(\beta) = (A \mapsto \beta A + A\beta)^{-1} = \frac{1}{2}\beta^{-1}$ is a Hessian map:

$$\Psi(\beta) = \sum_i \tilde{\beta}[i] \log \frac{\tilde{\beta}[i]}{e} \;\Rightarrow\; D_\Psi(\beta \| \alpha I) = \sum_i \tilde{\beta}[i] \left(\log \frac{1}{e\alpha} + \tilde{\beta}[i] \right)$$

If X_i don’t commute, solution given by “time ordered exponential”:

$$\beta(t) = \left(\lim_{\epsilon \to 0} \prod_{\tau = t/\epsilon}^{0} e^{-\epsilon \mathcal{X}^*(r_\tau)} \right) \beta_0 \left(\lim_{\epsilon \to 0} \prod_{\tau = 0}^{t/\epsilon} e^{-\epsilon \mathcal{X}^*(r_\tau)} \right)$$

- With arbitrary (crazy) steering, can move in any direction and get to any psd matrix (even with $m = 2$ random measurement matrices)
- $\rho(\beta) = (A \mapsto \beta A + A\beta)^{-1}$ is not a Hessian map

The "complexity measure" approach

Identify $c(h)$ s.t.

- Optimization algorithm biases towards low $c(h)$
- $\mathcal{H}_{c(reality)} = \{h | c(h) \leq c(reality)\}$ has low capacity
- Reality is well explained by low $c(h)$

Can optimization bias can be described as $\mathbf{\arg\min c(h)\ \ s.t. L_S(h) = 0}$??

- Not always [Dauber Feder Koren Livni 2020]
- Approximately? Enough to explain generalization??

Ultimate Question: What is the true Inductive Bias? What makes reality *efficiently* learnable by fitting a (huge) neural net with a specific algorithm?