Hard labels sampled from sparse targets misleads rotation invariant algorithms

Bin Yu

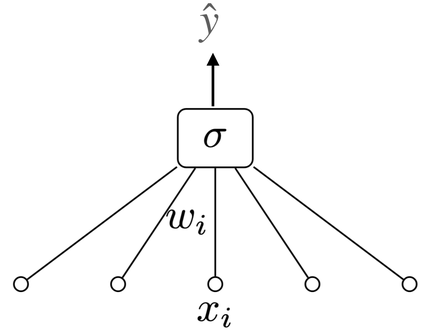
Manfred Warmuth

Peter Bartlett

Avrajit Ghosh

Logistic Regression

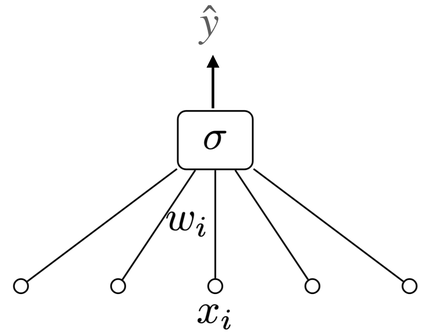
- Fundamental machine learning problem:
- Logistic loss on example $(\mathbf{x}, y)$
- Logistic Loss: $y \log \frac{y}{\hat{y}} + (1-y) \log \frac{1-y}{1-\hat{y}}$ $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x})$



Single neuron predictor

Logistic Regression

- Fundamental machine learning problem:
- Logistic loss on example $(\mathbf{x}, y)$
- Logistic Loss: $y \log \frac{y}{\hat{y}} + (1-y) \log \frac{1-y}{1-\hat{y}}$ $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x})$



Single neuron predictor

Choice-1

Soft Labels: *Distillation*

Match $y \in (0,1)$ with $\sigma(\mathbf{w} \cdot \mathbf{x})$

Choice-2

Hard Labels: *Classification*

Receive $y = 0$ or $y = 1$

Match $y \in \{0,1\}$ with $\sigma(\mathbf{w} \cdot \mathbf{x})$

Well specified logistic model

- Logistic Loss: $y \log \frac{y}{\hat{y}} + (1-y) \log \frac{1-y}{1-\hat{y}}$ $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x})$

Well specified logistic model

- Logistic Loss: $y \log \frac{y}{\hat{y}} + (1-y) \log \frac{1-y}{1-\hat{y}}$ $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x})$

Choice-1

Soft Labels: *Distillation*

Target: $y = \sigma\left(\mathbf{x}^\top \mathbf{w}^\star\right)$

Well specified logistic model

- Logistic Loss: $y \log \frac{y}{\hat{y}} + (1-y) \log \frac{1-y}{1-\hat{y}}$ $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x})$

Choice-1

Soft Labels: *Distillation*

Target: $y = \sigma\left(\mathbf{x}^\top \mathbf{w}^\star\right)$

- When n (#samples) > d (#dimension)
- There is a unique solution.
- GD on logistic loss recovers $\mathbf{w}^\star$

Well specified logistic model

- Logistic Loss: $y \log \frac{y}{\hat{y}} + (1-y) \log \frac{1-y}{1-\hat{y}}$ $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x})$

Choice-1

Soft Labels: *Distillation*

Target: $y = \sigma\left(\mathbf{x}^\top \mathbf{w}^\star\right)$

- When n (#samples) > d (#dimension)
- There is a unique solution.
- GD on logistic loss recovers $\mathbf{w}^\star$

Choice-2

Hard Labels: *Classification*

Target: $y \in \{0,1\},\ \mathbb{E}[y] = \sigma\left(\mathbf{x}^\top \mathbf{w}^\star\right)$

Well specified logistic model

- Logistic Loss: $y \log \frac{y}{\hat{y}} + (1-y) \log \frac{1-y}{1-\hat{y}}$ $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x})$

Choice-1

Soft Labels: *Distillation*

Target: $y = \sigma\left(\mathbf{x}^\top \mathbf{w}^\star\right)$

- When n (#samples) > d (#dimension)
- There is a unique solution.
- GD on logistic loss recovers $\mathbf{w}^\star$

Choice-2

Hard Labels: *Classification*

Target: $y \in \{0,1\},\ \mathbb{E}[y] = \sigma\left(\mathbf{x}^\top \mathbf{w}^\star\right)$

- When $n = \Theta(d)$
- $\mathbf{w}^\star$ is s-sparse. $(s \ll \frac{d}{\log d})$
- GD fails to recover $\mathbf{w}^\star$ properly!

Also for Poisson regression (experimental)

$$y \mid \mathbf{x} \sim \text{Poisson}\big(\exp(\mathbf{x}^\top \mathbf{w}^\star)\big), \quad y \in \{0,1,2,\dots\}.$$

Learning with hard labels mislead GD algorithms when:

1) Model relies only on dot product $\langle \mathbf{x}, \mathbf{w} \rangle$

2) Targets are hard labels.

3) $\mathbf{w}^*$ is sparse.

4) Algorithm is GD. $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla L(\mathbf{w}_t)$

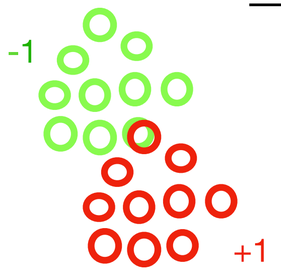
How large is this gap?

- GD even with early stopping on logistic loss.
- Excess risk gap is $\Omega\left(\frac{d-1}{n}\right)$
- Minimax excess risk rate for sparse oracle $\frac{s \log d}{n}$.
- Large gap for $n = \Theta(d), \quad s \ll \frac{d}{\log d}$!

GD on logistic regression is Rotation-Invariant

$$y \mid \mathbf{x} \sim \text{Bernoulli}\left(\sigma(\mathbf{x}^\top \mathbf{w}^\star)\right)$$

$$\text{Logistic-loss}(\langle \mathbf{x}, \hat{\mathbf{w}} \rangle) = \text{Logistic-loss}\left(\langle \boldsymbol{U}\mathbf{x}, \boldsymbol{U}\hat{\mathbf{w}} \rangle\right)$$



$$D_1 = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$

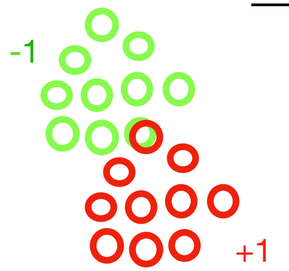
Learning by dot product $\Longrightarrow$ Rotation Invariant

GD on logistic regression is Rotation-Invariant

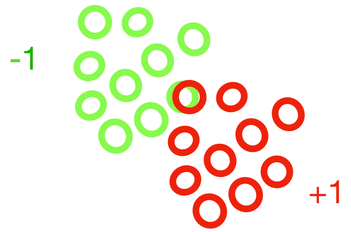
$$y \mid \mathbf{x} \sim \text{Bernoulli}\left(\sigma(\mathbf{x}^\top \mathbf{w}^\star)\right)$$

$$\text{Logistic-loss}(\langle \mathbf{x}, \hat{\mathbf{w}} \rangle) = \text{Logistic-loss}\left(\langle \boldsymbol{U}\mathbf{x}, \boldsymbol{U}\hat{\mathbf{w}} \rangle\right)$$

Learning by dot product $\Longrightarrow$ Rotation Invariant



$$D_1 = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$



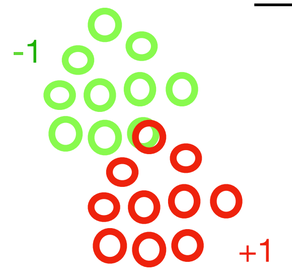
$$D_2 = \{(\boldsymbol{U}_1\mathbf{x}_i, y_i)\}_{i=1}^n$$

GD on logistic regression is Rotation-Invariant

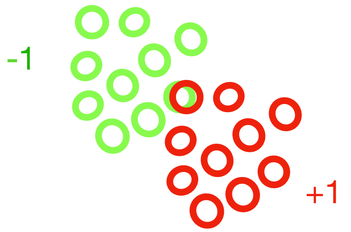
$$y \mid \mathbf{x} \sim \text{Bernoulli}\left(\sigma(\mathbf{x}^\top \mathbf{w}^\star)\right)$$

$$\text{Logistic-loss}(\langle \mathbf{x}, \hat{\mathbf{w}} \rangle) = \text{Logistic-loss}\left(\langle \boldsymbol{U}\mathbf{x}, \boldsymbol{U}\hat{\mathbf{w}} \rangle\right)$$

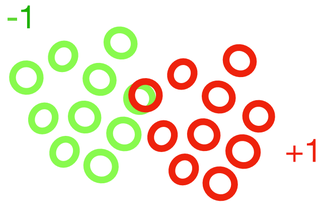
Learning by dot product $\Longrightarrow$ Rotation Invariant



$$D_1 = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$



$$D_2 = \{(\boldsymbol{U}_1\mathbf{x}_i, y_i)\}_{i=1}^n$$



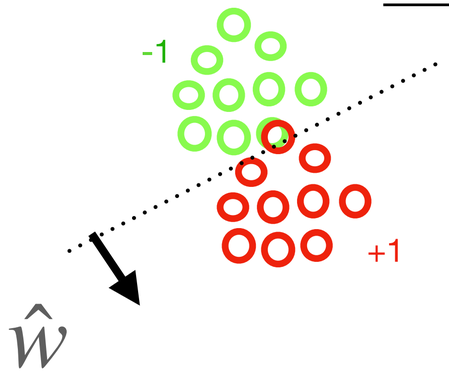
$$D_3 = \{(\boldsymbol{U}_2\mathbf{x}_i, y_i)\}_{i=1}^n$$

GD on logistic regression is Rotation-Invariant

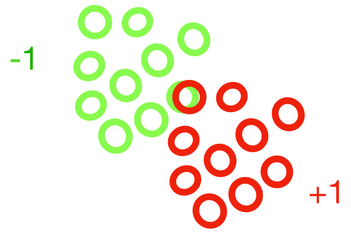
$$y \mid \mathbf{x} \sim \text{Bernoulli}\left(\sigma(\mathbf{x}^\top \mathbf{w}^\star)\right)$$

$$\text{Logistic-loss}(\langle \mathbf{x}, \hat{\mathbf{w}} \rangle) = \text{Logistic-loss}\left(\langle \boldsymbol{U}\mathbf{x}, \boldsymbol{U}\hat{\mathbf{w}} \rangle\right)$$

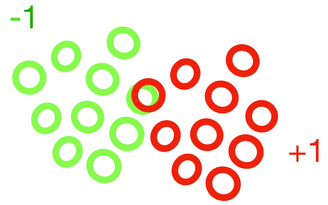
Learning by dot product $\Longrightarrow$ Rotation Invariant



$$D_1 = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$



$$D_2 = \{(\boldsymbol{U}_1\mathbf{x}_i, y_i)\}_{i=1}^n$$



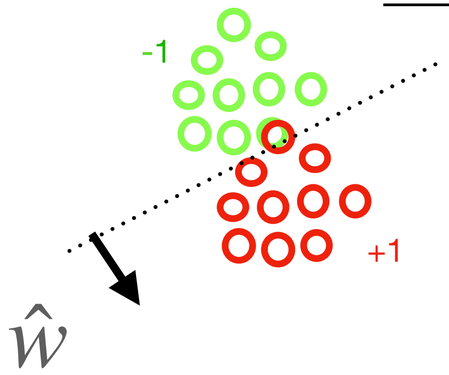
$$D_3 = \{(\boldsymbol{U}_2\mathbf{x}_i, y_i)\}_{i=1}^n$$

GD on logistic regression is Rotation-Invariant

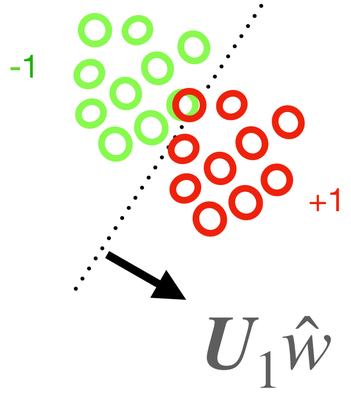
$$y \mid \mathbf{x} \sim \text{Bernoulli}\left(\sigma(\mathbf{x}^\top \mathbf{w}^\star)\right)$$

$$\text{Logistic-loss}(\langle \mathbf{x}, \hat{\mathbf{w}} \rangle) = \text{Logistic-loss}\left(\langle \boldsymbol{U}\mathbf{x}, \boldsymbol{U}\hat{\mathbf{w}} \rangle\right)$$

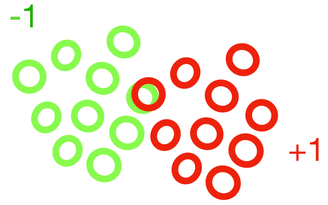
Learning by dot product $\Longrightarrow$ Rotation Invariant



$$D_1 = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$



$$D_2 = \{(\boldsymbol{U}_1\mathbf{x}_i, y_i)\}_{i=1}^n$$



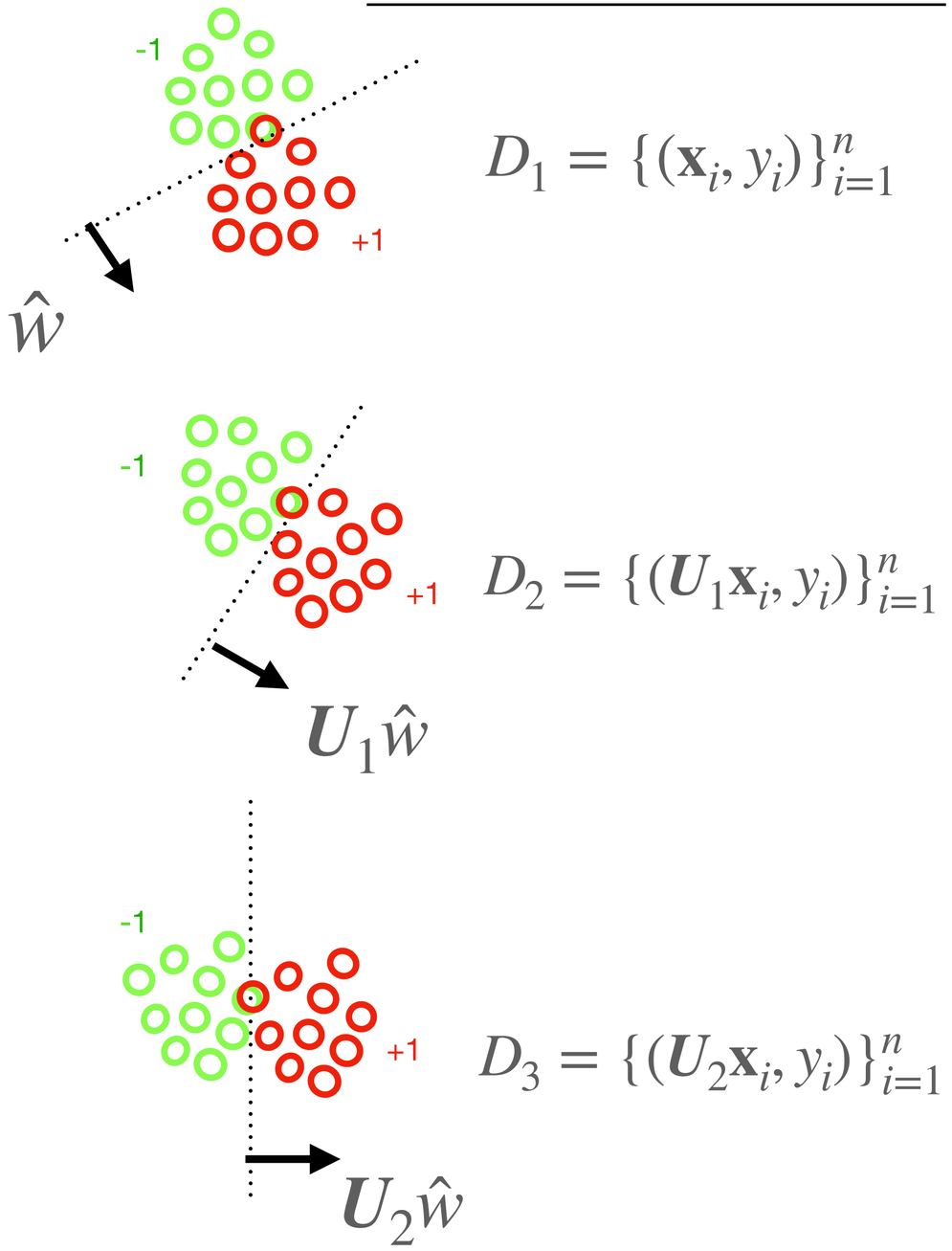
$$D_3 = \{(\boldsymbol{U}_2\mathbf{x}_i, y_i)\}_{i=1}^n$$

GD on logistic regression is Rotation-Invariant

$$y \mid \mathbf{x} \sim \text{Bernoulli}\left(\sigma(\mathbf{x}^\top \mathbf{w}^\star)\right)$$

$$\text{Logistic-loss}(\langle \mathbf{x}, \hat{\mathbf{w}} \rangle) = \text{Logistic-loss}\left(\langle \boldsymbol{U}\mathbf{x}, \boldsymbol{U}\hat{\mathbf{w}} \rangle\right)$$

Learning by dot product $\Longrightarrow$ Rotation Invariant



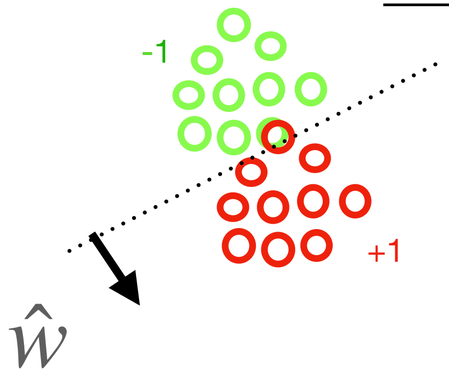
GD on logistic regression is Rotation-Invariant

$$y \mid \mathbf{x} \sim \text{Bernoulli}\left(\sigma(\mathbf{x}^\top \mathbf{w}^\star)\right)$$

$$\text{Logistic-loss}(\langle \mathbf{x}, \hat{\mathbf{w}} \rangle) = \text{Logistic-loss}\left(\langle \boldsymbol{U}\mathbf{x}, \boldsymbol{U}\hat{\mathbf{w}} \rangle\right)$$

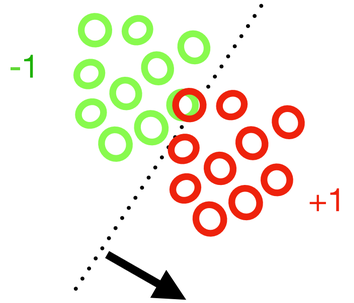
Learning by dot product $\Longrightarrow$ Rotation Invariant

Magnitude information preserved.
Direction information lost.



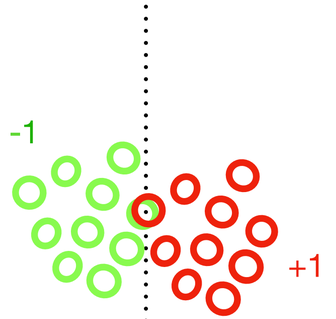
$$D_1 = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$

$\hat{w}$



$$D_2 = \{(\boldsymbol{U}_1\mathbf{x}_i, y_i)\}_{i=1}^n$$

$\boldsymbol{U}_1\hat{w}$



$$D_3 = \{(\boldsymbol{U}_2\mathbf{x}_i, y_i)\}_{i=1}^n$$

$\boldsymbol{U}_2\hat{w}$

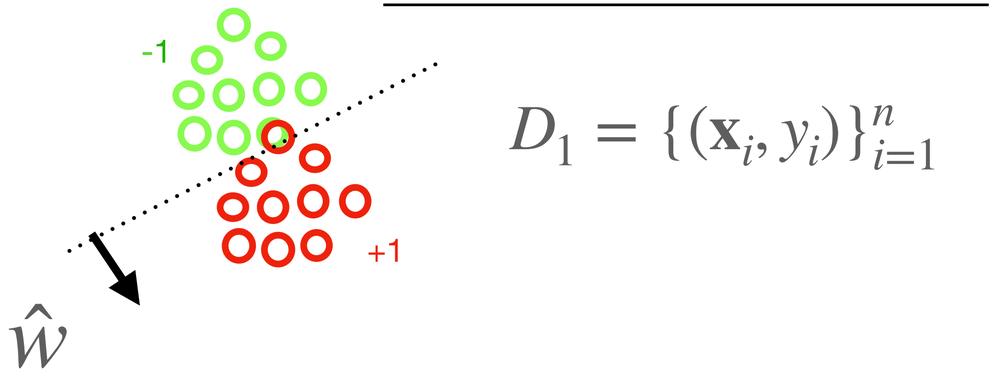
GD on logistic regression is Rotation-Invariant

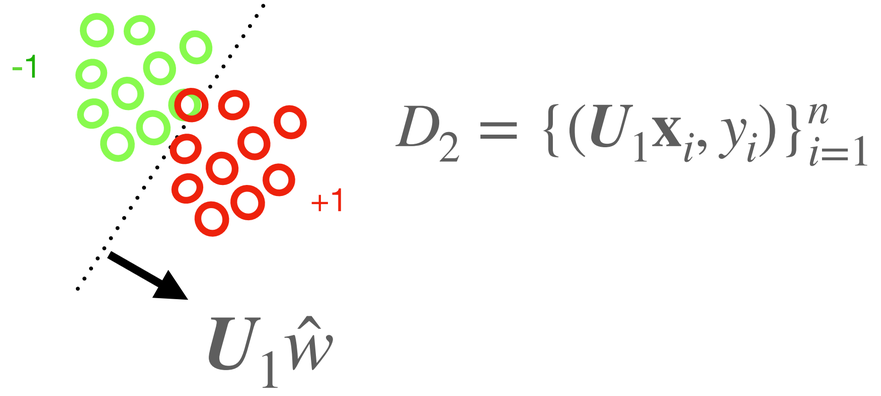
$$y \mid \mathbf{x} \sim \text{Bernoulli}\left(\sigma(\mathbf{x}^\top \mathbf{w}^\star)\right)$$

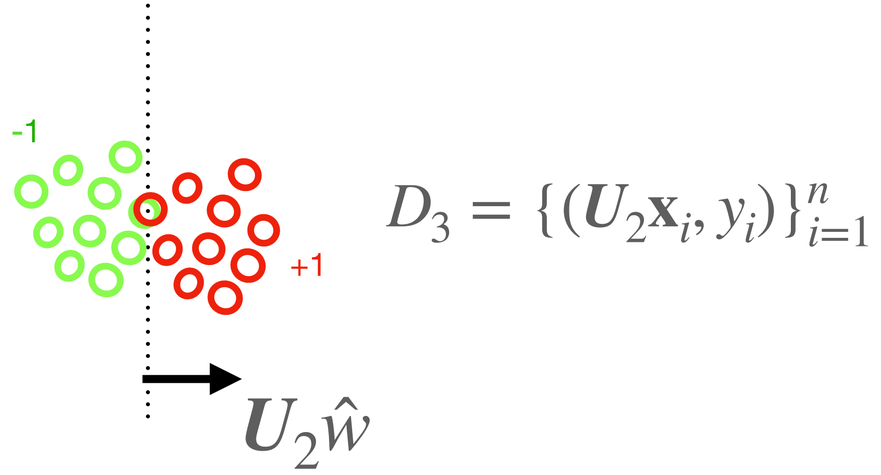
$$\text{Logistic-loss}(\langle \mathbf{x}, \hat{\mathbf{w}} \rangle) = \text{Logistic-loss}\left(\langle \boldsymbol{U}\mathbf{x}, \boldsymbol{U}\hat{\mathbf{w}} \rangle\right)$$

Learning by dot product $\Longrightarrow$ Rotation Invariant

Magnitude information preserved.
Direction information lost.







Proof sketch: Lower- bound by Rotational Symmetrization:

$$\mathbb{E}_D\left[L(\hat{\mathbf{w}}) - L(\mathbf{w}^\star)\right]$$

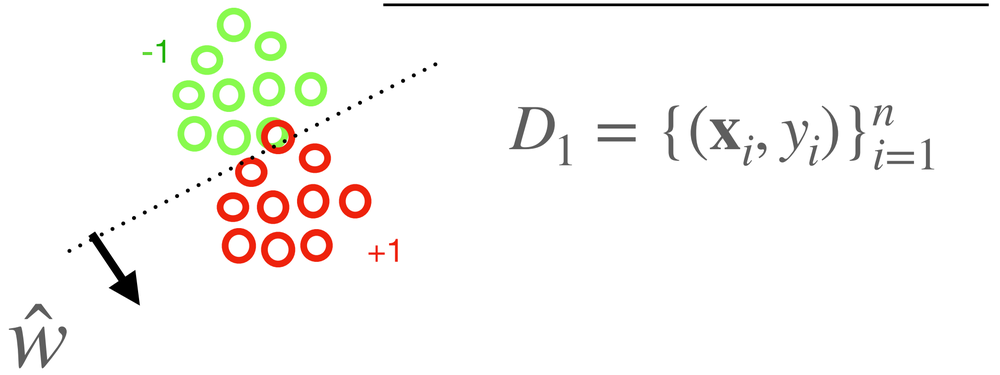
GD on logistic regression is Rotation-Invariant

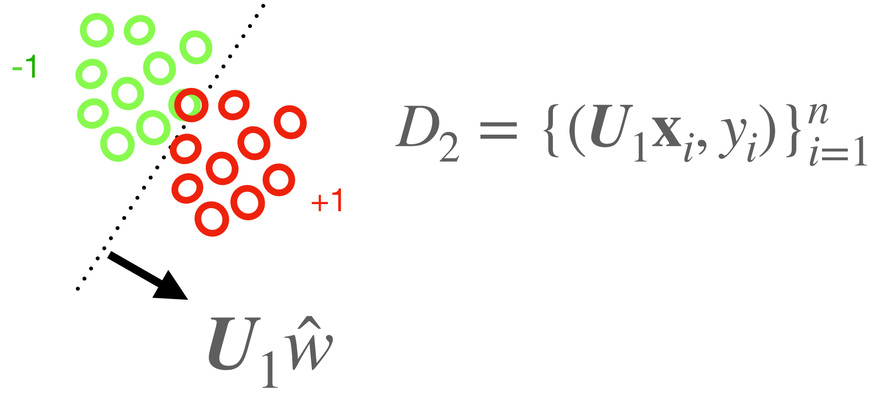
$$y \mid \mathbf{x} \sim \text{Bernoulli}\left(\sigma(\mathbf{x}^\top \mathbf{w}^\star)\right)$$

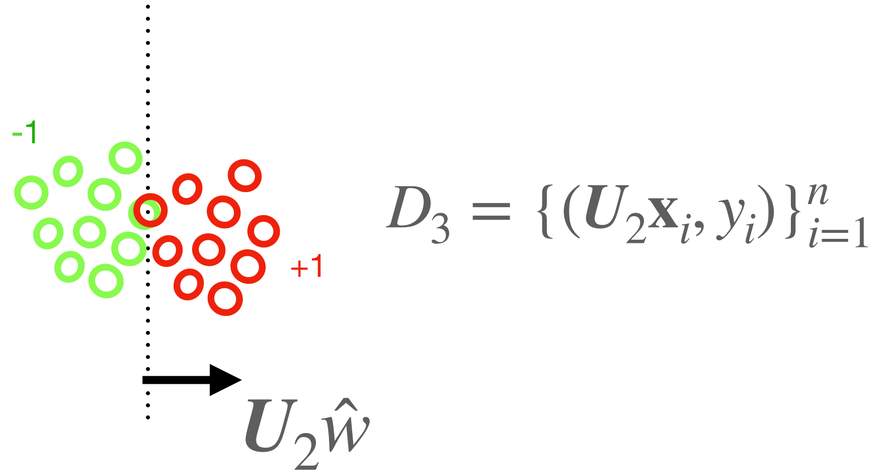
$$\text{Logistic-loss}(\langle \mathbf{x}, \hat{\mathbf{w}} \rangle) = \text{Logistic-loss}\left(\langle \boldsymbol{U}\mathbf{x}, \boldsymbol{U}\hat{\mathbf{w}} \rangle\right)$$

Learning by dot product $\Longrightarrow$ Rotation Invariant

Magnitude information preserved.
Direction information lost.







Proof sketch: Lower- bound by Rotational Symmetrization:

$$\mathbb{E}_D\left[L(\hat{\mathbf{w}}) - L(\mathbf{w}^\star)\right] \geq \mathbb{E}_D\, \mathbb{E}_{\mathbf{w}\sim\exp(-n\hat{L}(\mathbf{w}))\,\mathbf{1}_{\|\mathbf{w}\|_2=\|\mathbf{w}^*\|_2}}\left[L(\mathbf{w}) - L(\mathbf{w}^\star)\right]$$

Direction information averaged out

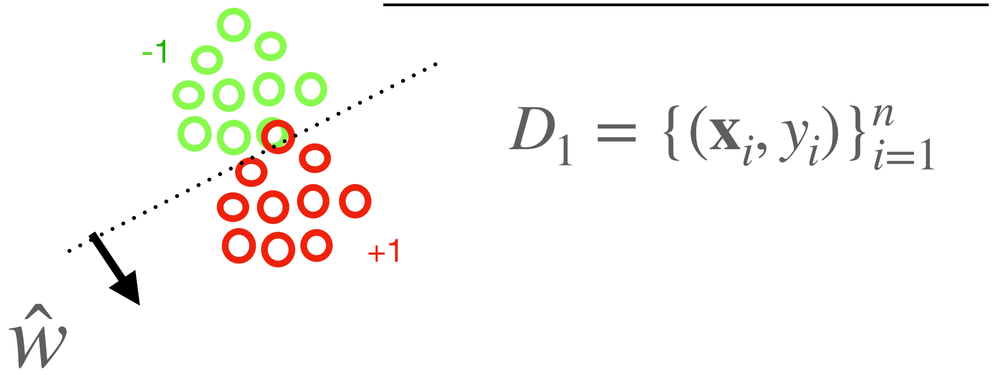
GD on logistic regression is Rotation-Invariant

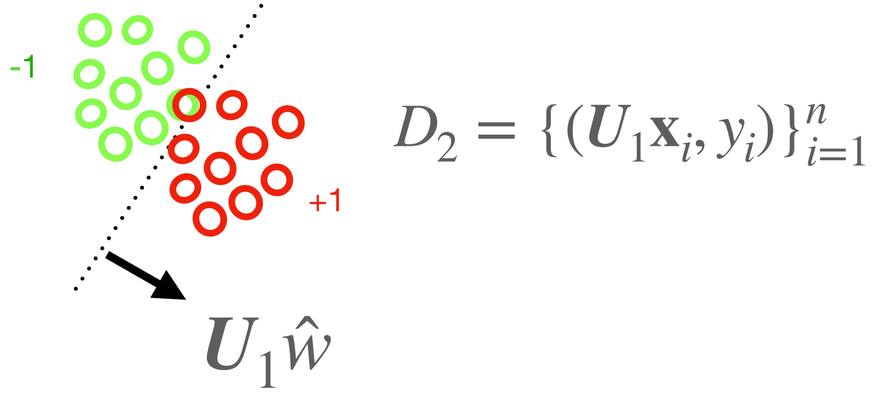
$$y \mid \mathbf{x} \sim \text{Bernoulli}\left(\sigma(\mathbf{x}^\top \mathbf{w}^\star)\right)$$

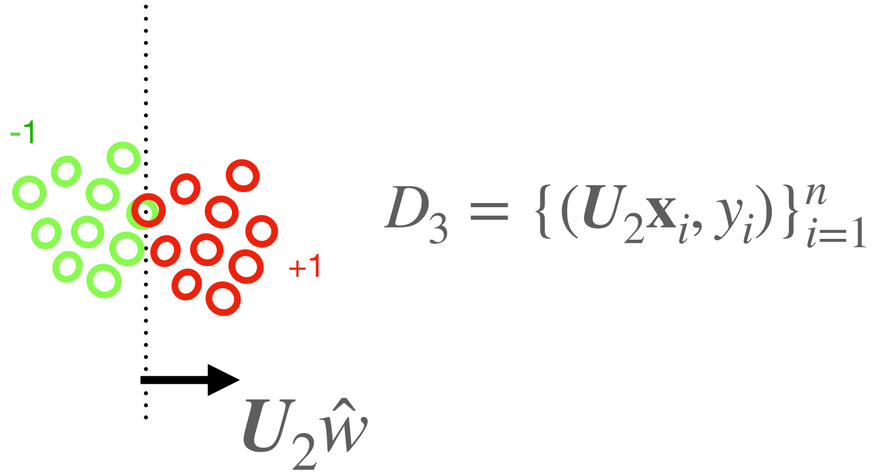
$$\text{Logistic-loss}(\langle \mathbf{x}, \hat{\mathbf{w}} \rangle) = \text{Logistic-loss}\left(\langle \boldsymbol{U}\mathbf{x}, \boldsymbol{U}\hat{\mathbf{w}} \rangle\right)$$

Learning by dot product $\Longrightarrow$ Rotation Invariant

Magnitude information preserved.
Direction information lost.







Proof sketch: Lower- bound by Rotational Symmetrization:

$$\mathbb{E}_D\left[L(\hat{\mathbf{w}}) - L(\mathbf{w}^\star)\right] \geq \mathbb{E}_D\, \mathbb{E}_{\mathbf{w}\sim\exp(-n\hat{L}(\mathbf{w}))\,\mathbf{1}_{\|\mathbf{w}\|_2=\|\mathbf{w}^*\|_2}}\left[L(\mathbf{w}) - L(\mathbf{w}^\star)\right]$$

Direction information averaged out

$$\geq \frac{c(d-1)}{n}$$

(Symmetrization spread error across (d-1) directions)

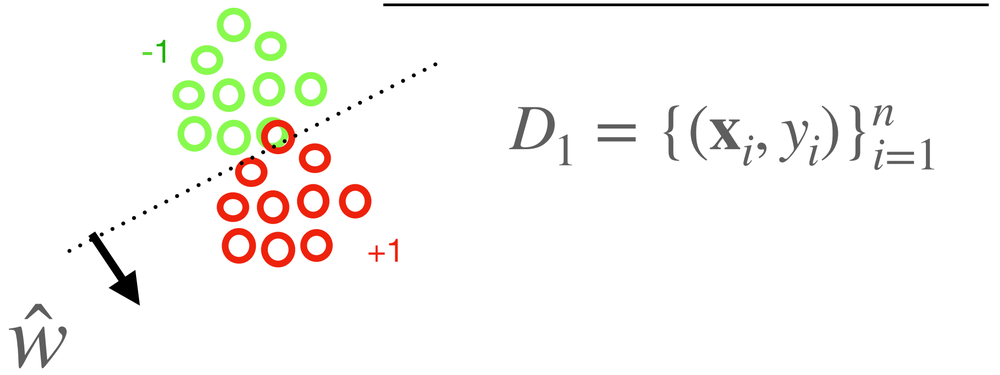
GD on logistic regression is Rotation-Invariant

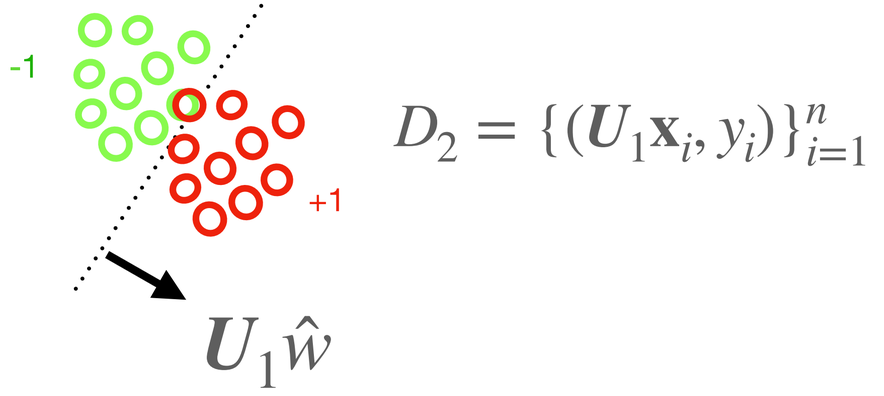
$$y \mid \mathbf{x} \sim \text{Bernoulli}\left(\sigma(\mathbf{x}^\top \mathbf{w}^\star)\right)$$

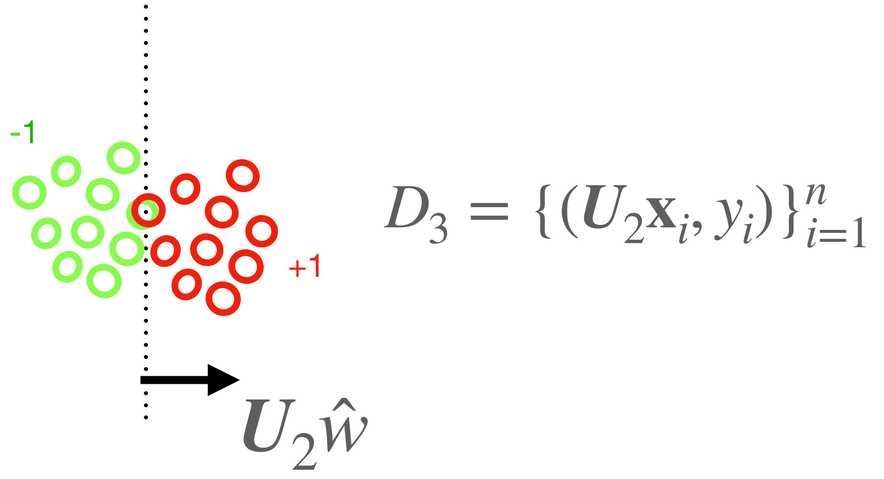
$$\text{Logistic-loss}(\langle \mathbf{x}, \hat{\mathbf{w}} \rangle) = \text{Logistic-loss}\left(\langle \boldsymbol{U}\mathbf{x}, \boldsymbol{U}\hat{\mathbf{w}} \rangle\right)$$

Learning by dot product $\Longrightarrow$ Rotation Invariant

Magnitude information preserved.
Direction information lost.







Proof sketch: Lower- bound by Rotational Symmetrization:

$$\mathbb{E}_D\left[L(\hat{\mathbf{w}}) - L(\mathbf{w}^\star)\right] \geq \mathbb{E}_D\, \mathbb{E}_{\mathbf{w}\sim\exp(-n\hat{L}(\mathbf{w}))\,\mathbf{1}_{\|\mathbf{w}\|_2=\|\mathbf{w}^*\|_2}}\left[L(\mathbf{w}) - L(\mathbf{w}^\star)\right]$$

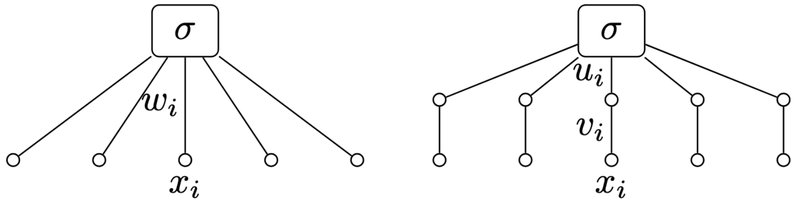
Direction information averaged out

$$\geq \frac{c(d-1)}{n}$$

(Symmetrization spread error across (d-1) directions)

Fix: Break the rotation-invariance. Keep GD.

Non-rotation Invariant: Gradient flow on Spindle



Reparameterize: $w_i = u_i v_i$

Linear Regression

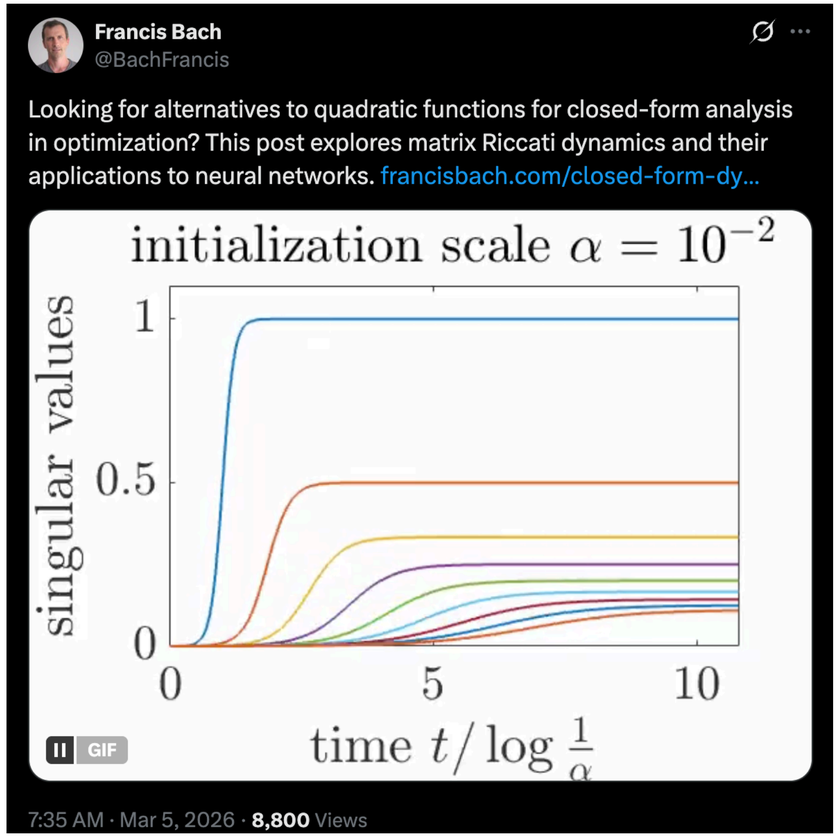
$$\dot{w}_i = bw_i - aw_i^2$$

- Riccati-ODE
- Closed form solution

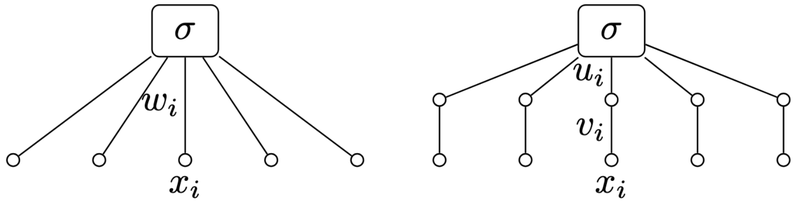
Logistic Regression

$$\dot{w}_i = bw_i - A(\mathbf{w})w_i^2$$

- State-dependent Riccati-ODE
- No exact closed-form
- Enveloping ODE argument



Non-rotation Invariant: Gradient flow on Spindly



Reparameterize: $w_i = u_i v_i$

Linear Regression

$$\dot{w}_i = bw_i - aw_i^2$$

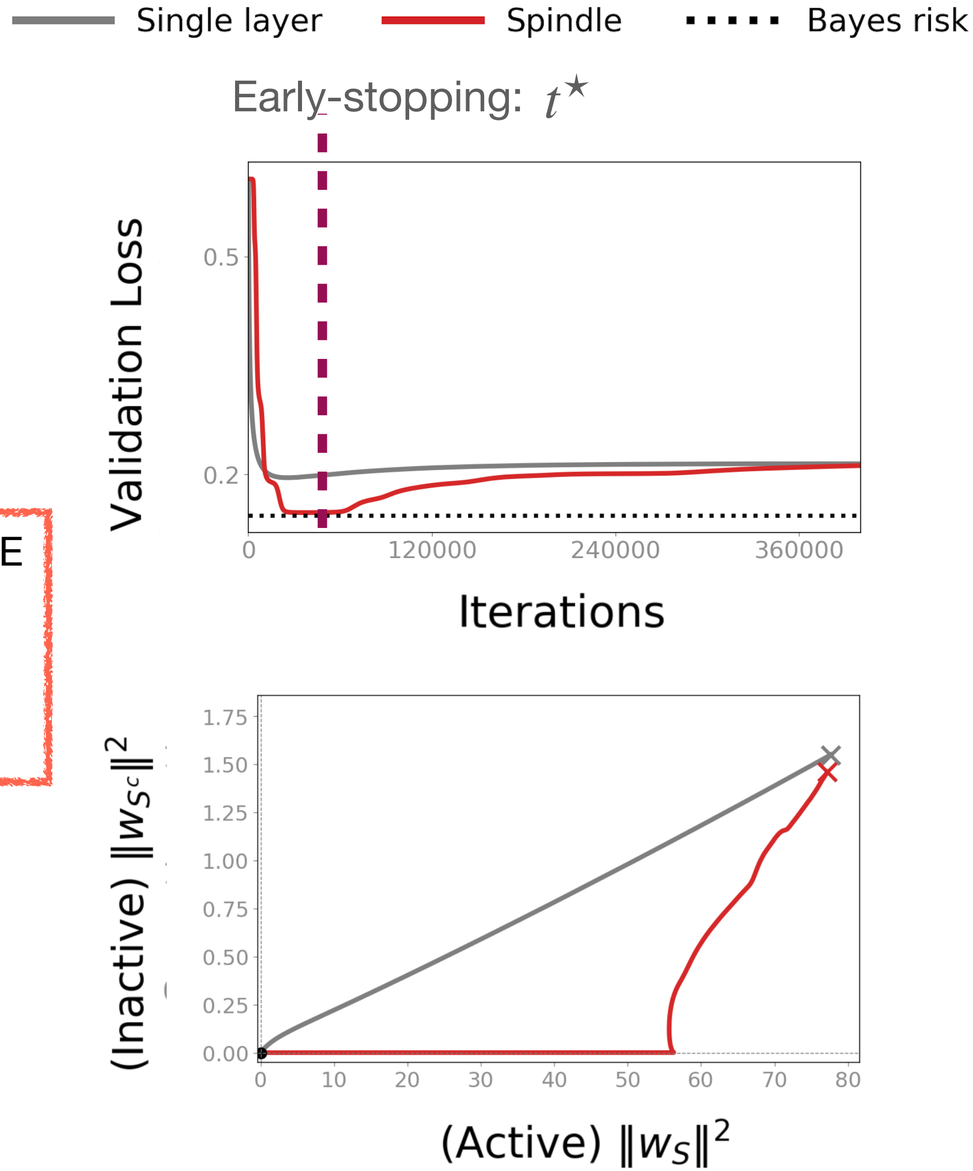
- Riccati-ODE
- Closed form solution

Logistic Regression

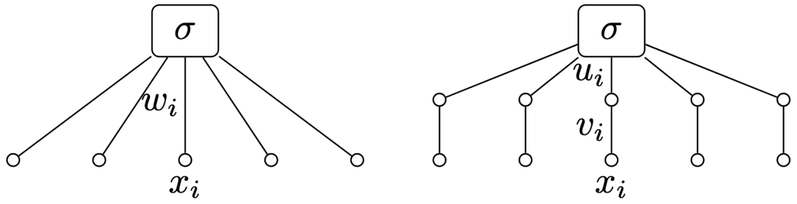
$$\dot{w}_i = bw_i - A(\mathbf{w})w_i^2$$

- State-dependent Riccati-ODE
- No exact closed-form
- Enveloping ODE argument

$d = 50, \quad n = 1000$



Non-rotation Invariant: Gradient flow on Spindly

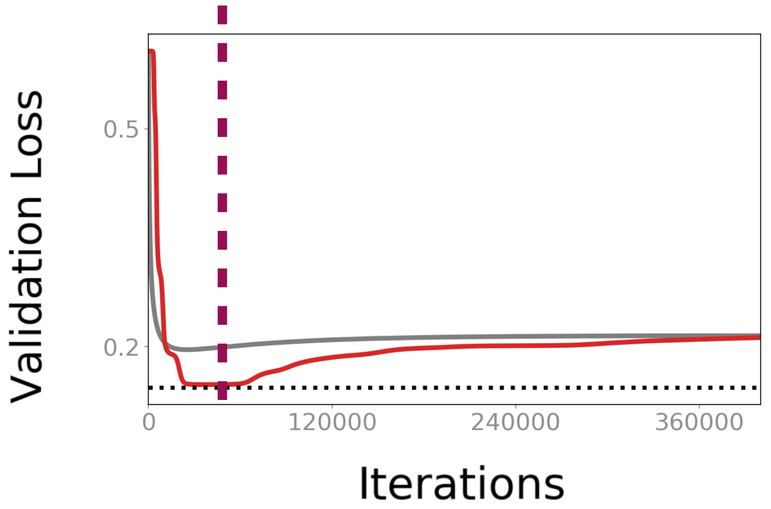


Reparameterize: $w_i = u_i v_i$

$d = 50, \quad n = 1000$

Single layer Spindle Bayes risk

Early-stopping: $t^\star$



Linear Regression

$$\dot{w}_i = bw_i - aw_i^2$$

- Riccati-ODE
- Closed form solution

Logistic Regression

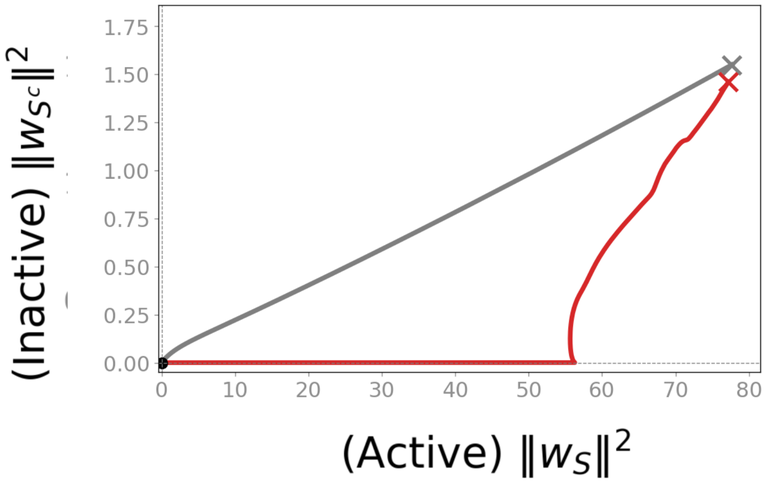
$$\dot{w}_i = bw_i - A(\mathbf{w})w_i^2$$

- State-dependent Riccati-ODE
- No exact closed-form
- Enveloping ODE argument

Theorem: (Early-stopping on spindle dynamics induce signal-noise separation)

Small-init+Early-stopping: Init: $\mathbf{w}(0) = \frac{1}{d}\mathbf{1} \quad \Rightarrow \quad \exists\, t^\star \sim \log d$

Excess risk: $\Rightarrow \quad L(\mathbf{w}(t^\star)) - L(\mathbf{w}^\star) \lesssim \frac{s \log d}{n}$



Hard labels mislead rotation invariant GD

- Learning through $< \mathbf{x}, \mathbf{w} >$ fails to capture direction information.
- Open problems:

★ Extension to general exponential family distribution.

★Limitation of GD for various data distribution models.

Arxiv: https://arxiv.org/abs/2603.20967