MLRG – Why deep learning works?

The case of separable binary classification with linear models

Fred

The Implicit Bias of Gradient Descent on Separable Data
Soudry, Hoffer, Shpigel Nacson, Gunasekar, Srebro

Previously on MLRG

	Overparametrization and expressivity
Aaron	Bounds for perceptrons
Jason	Neural networks can fitting random noise
	Stochasticity and geometry of minima
Amir	SGD finds shallow minima
Adam	Sharp or flat is not the main story
	Norms, Geometry and Capacity
Will	Exploring generalization with capacity measures
Cathy	Geometry of optimization and regularization: path norm
	Implicit regularization
Betty	Gradient flow for matrix factorization
Fred	Gradient descent for logistic regression

Implicit regularization

Under-parametrized $\longrightarrow$ find **the** global minimum

Over-parametrized $\longrightarrow$ find **a** global minimum

Implicit regularization

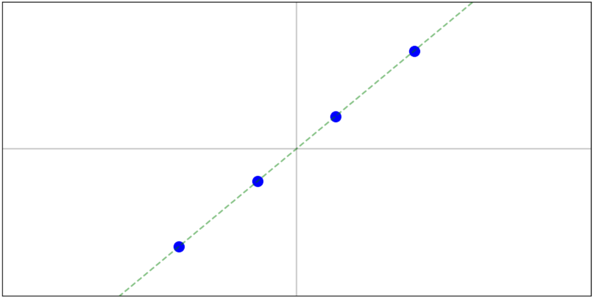
Under-parametrized $\longrightarrow$ find **the** global minimum

optimization: speed

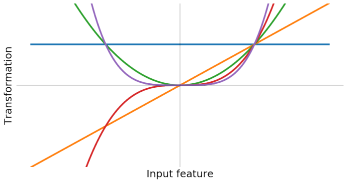
Over-parametrized $\longrightarrow$ find **a** global minimum

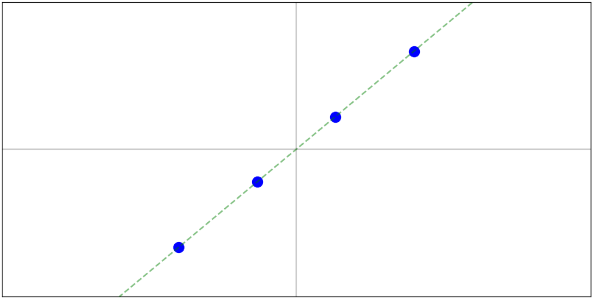
optimization:
speed
generalization
dependent on initialization

Implicit regularization

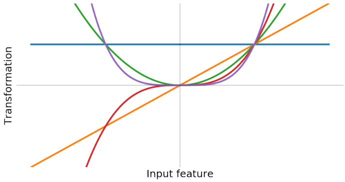


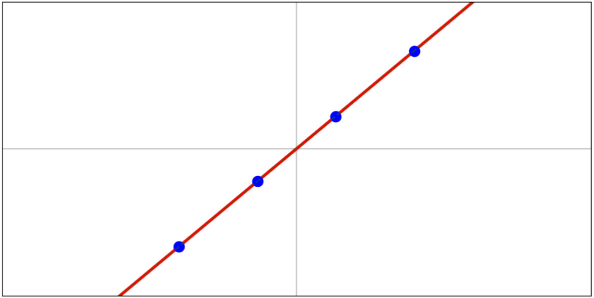
Implicit regularization



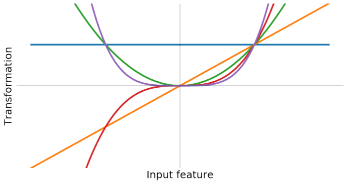


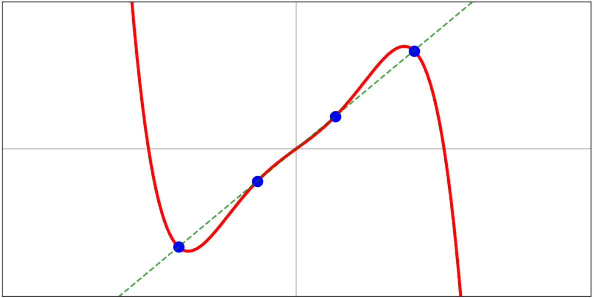
Implicit regularization



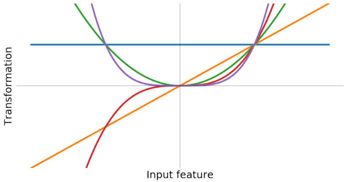


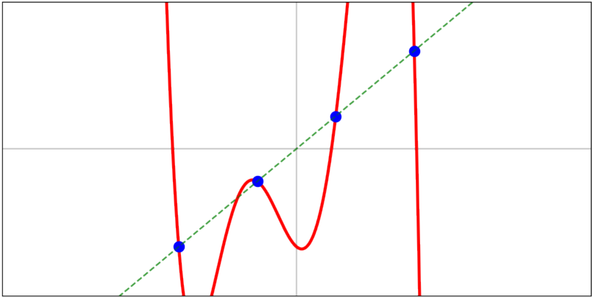
Implicit regularization



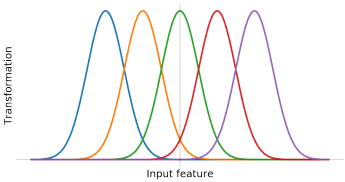


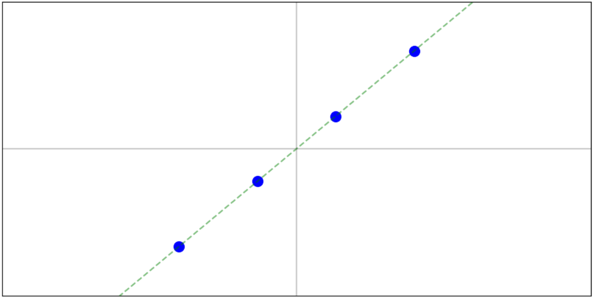
Implicit regularization



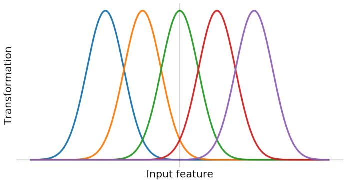


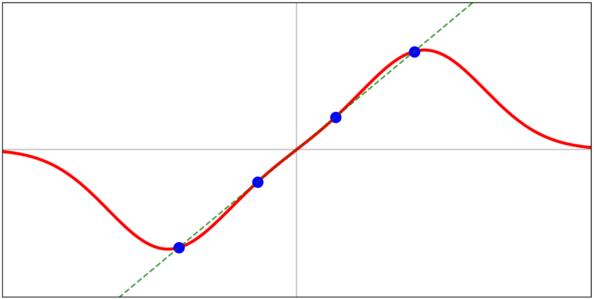
Implicit regularization



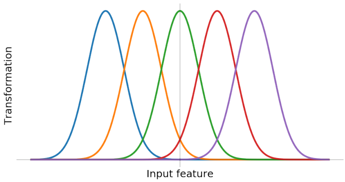


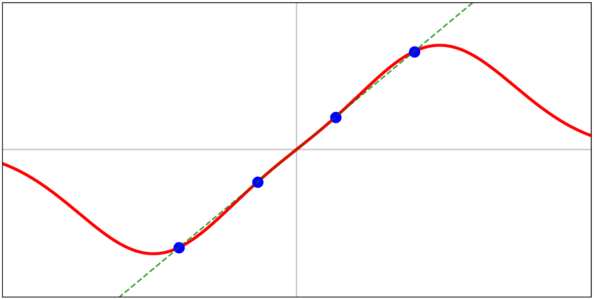
Implicit regularization



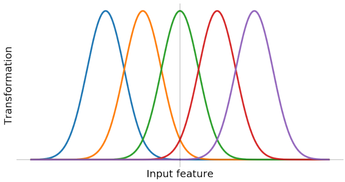


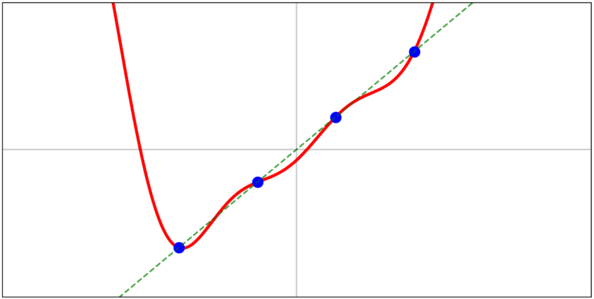
Implicit regularization



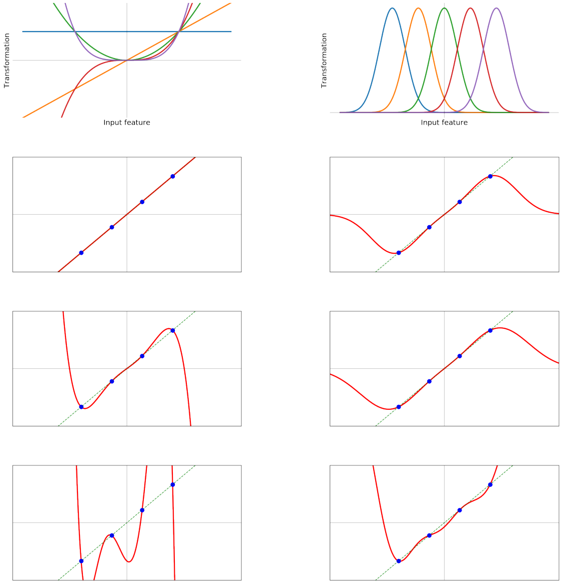


Implicit regularization

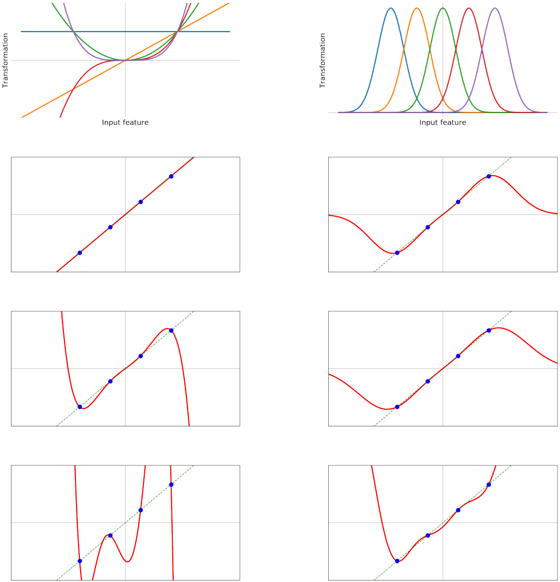




Implicit regularization



Implicit regularization



CD $(w_0 = 0)$

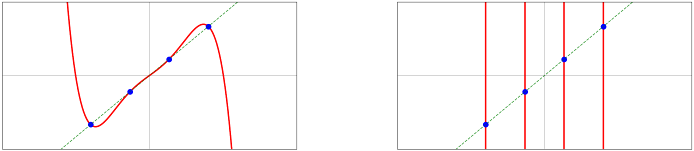
GD $(w_0 = 0)$

CD $(w_0 = 5)$

Implicit regularization

solution space is infinite

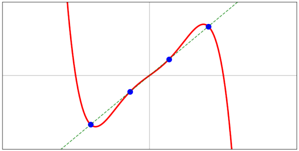
can have 0 train error and ∞ test error

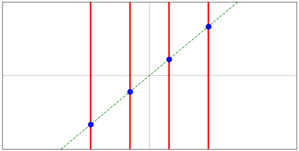


Implicit regularization

solution space is infinite

can have 0 train error and ∞ test error



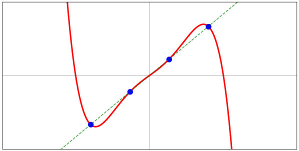


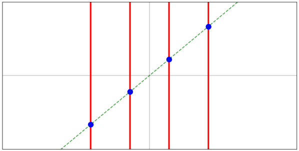
given model
choice of optimizer $\Leftrightarrow$ choice of solution

Implicit regularization

solution space is infinite

can have 0 train error and ∞ test error





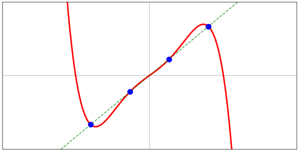
given model
choice of optimizer $\Leftrightarrow$ choice of solution

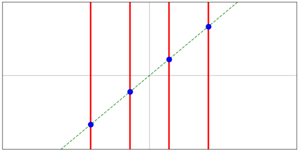
Why ~~deep learning~~ boosting work? ($\approx$ 2000s)

Implicit regularization

solution space is infinite

can have 0 train error and ∞ test error





given model

choice of optimizer $\Leftrightarrow$ choice of solution

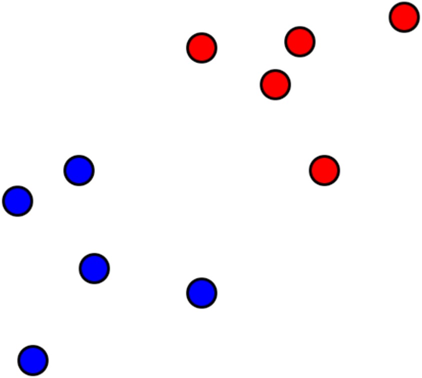
Why ~~deep learning~~ boosting work? ($\approx$ 2000s)

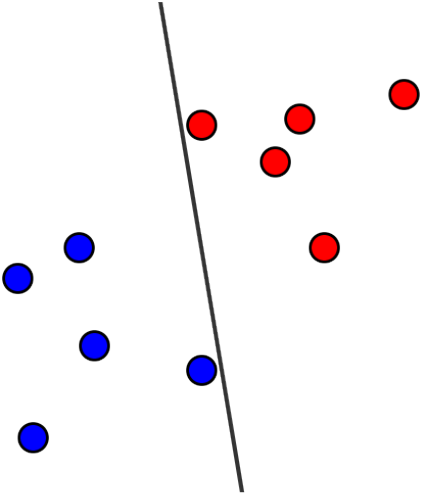
given logistic regression

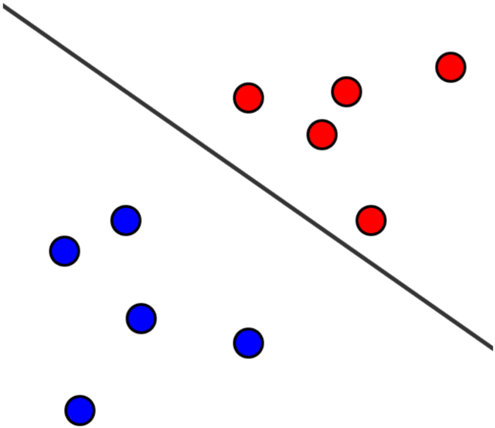
gradient descent $\Leftrightarrow$?

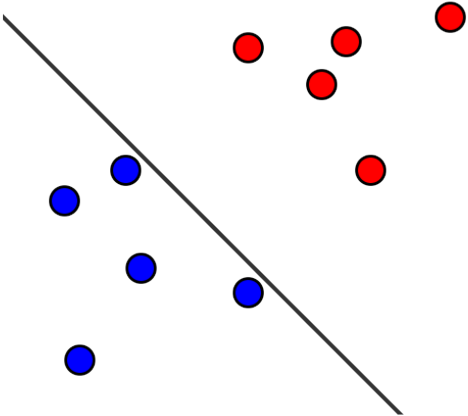
Logistic regression

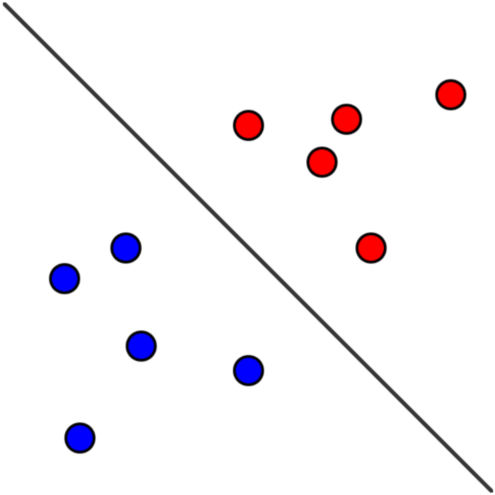
Separating planes Max Margin Confidence Divergence of w

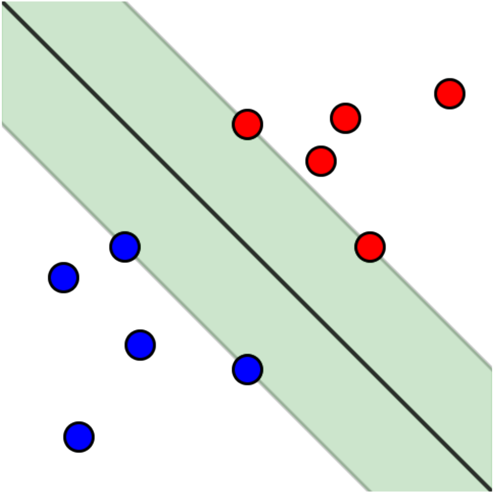


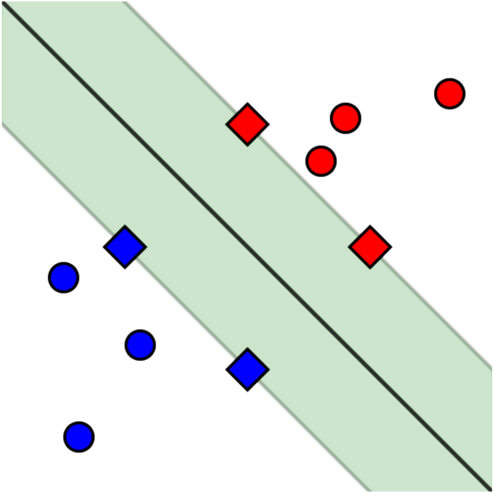


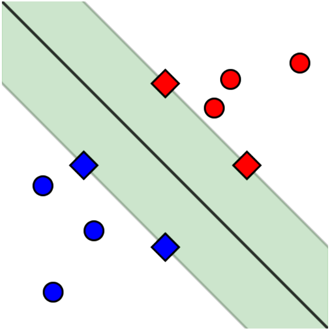


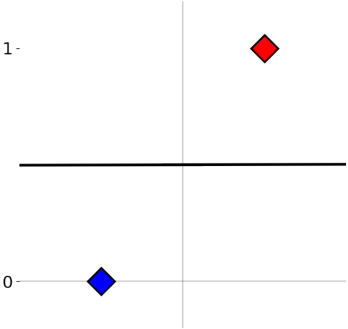




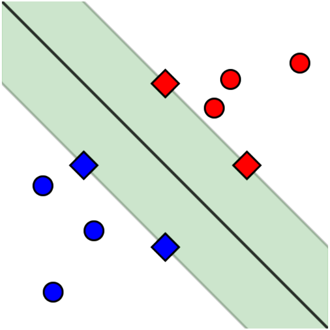


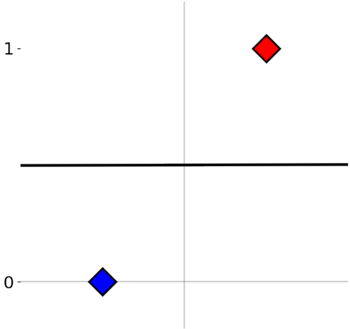




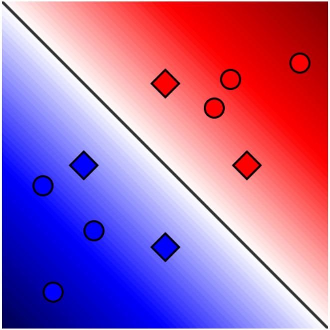

1
0

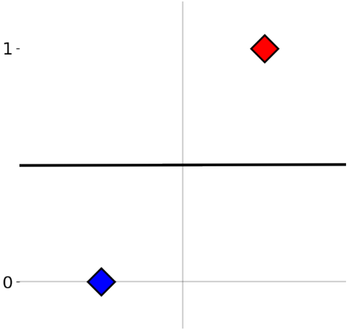
$$p(\bullet|x) = \sigma(w^\top x) = \frac{1}{1 + \exp(-w^\top x)}$$



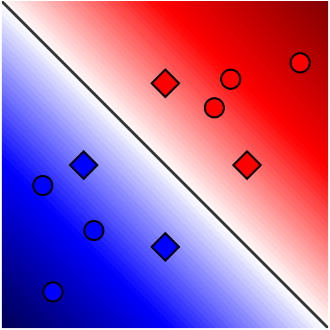


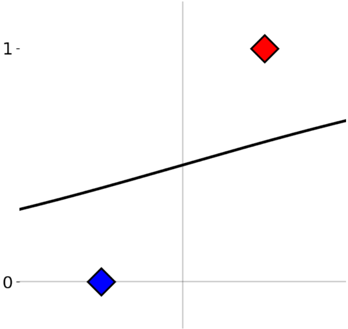
$$p(\bullet|x) = \sigma(w^\top x) = \frac{1}{1+\exp(-w^\top x)}$$



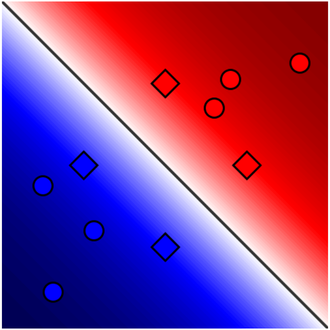


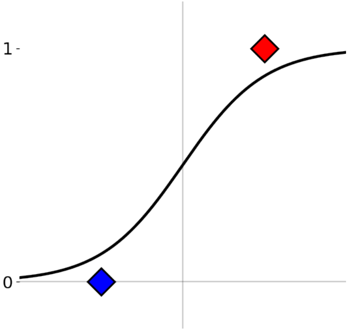
$$p(\bullet|x) = \sigma(w^\top x) = \frac{1}{1 + \exp(-w^\top x)}$$





$$p(\bullet|x) = \sigma(w^\top x) = \frac{1}{1 + \exp(-w^\top x)}$$

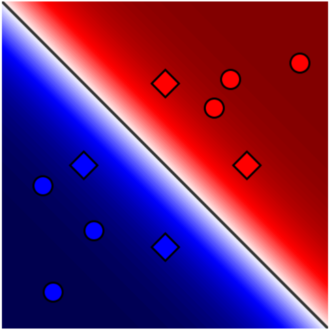


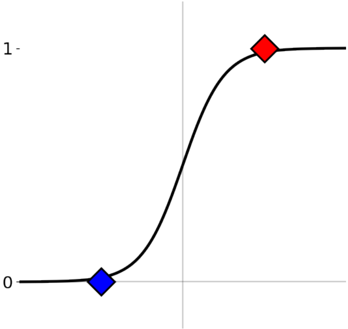


Logistic regression

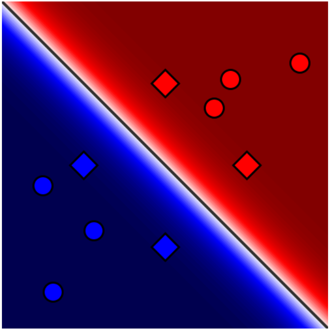
Separating planes Max Margin **Confidence** Divergence of w

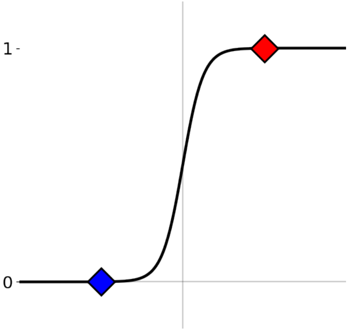
$$p(\bullet|x) = \sigma(w^\top x) = \frac{1}{1+\exp(-w^\top x)}$$





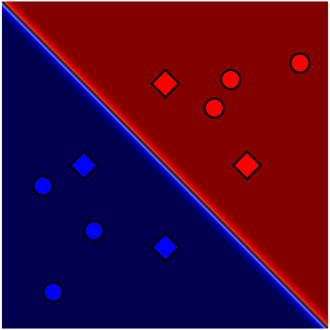
$$p(\bullet|x) = \sigma(w^\top x) = \frac{1}{1+\exp(-w^\top x)}$$

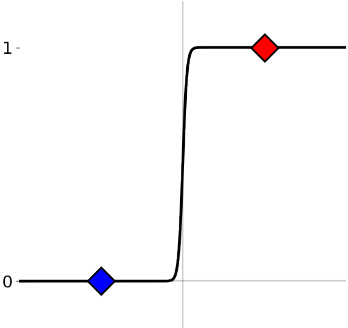




$\frac{w}{\|w\|}$: separating plane

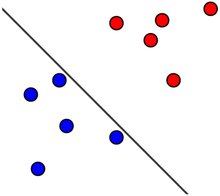
$\|w\|$: confidence





Main results

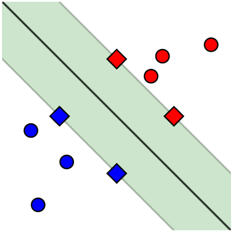
Gradient descent on separable logistic regression



weights diverge $\|w\|$

normalized weights converge $\dfrac{w}{\|w\|}$

Gradient descent on separable logistic regression

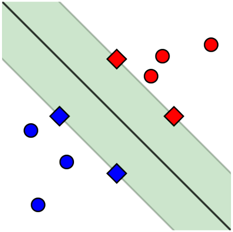


weights diverge $\|w\|$

normalized weights converge $\dfrac{w}{\|w\|}$

converges **very slowly** to the max margin
regardless of the starting point

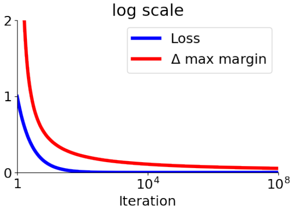
Gradient descent on separable logistic regression

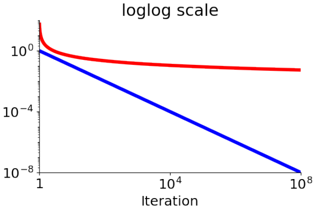


weights diverge $\|w\|$

normalized weights converge $\dfrac{w}{\|w\|}$

converges **very slowly** to the max margin **regardless of the starting point**





Main results

$\hat{w}$	maximum margin/min $\|\cdot\|_2$ solution
w_t	gradient descent iterates

$$\boxed{w_t = \hat{w} \log(t) + \rho(t) \qquad \text{and} \qquad \rho(t) \leq C}$$

Main results

$\hat{w}$ maximum margin/min $\|\cdot\|_2$ solution

w_t gradient descent iterates

$$\boxed{w_t = \hat{w}\log(t) + \rho(t) \qquad \text{and} \qquad \rho(t) \leq C}$$

converges to the max margin: $$\lim_{t\to\infty} \frac{w_t}{\|w_t\|} = \frac{\hat{w}}{\|\hat{w}\|}$$

Main results

$\hat{w}$ maximum margin/min $\|\cdot\|_2$ solution

w_t gradient descent iterates

$$\boxed{w_t = \hat{w}\log(t) + \rho(t) \qquad \text{and} \qquad \rho(t) \leq C}$$

converges to the max margin: $$\lim_{t\to\infty} \frac{w_t}{\|w_t\|} = \frac{\hat{w}}{\|\hat{w}\|}$$

converges slowly: $$\left\| \frac{w_t}{\|w_t\|} - \frac{\hat{w}}{\|\hat{w}\|} \right\| = \tilde{O}\left(\frac{1}{\log t}\right)$$

Main results

$\hat{w}$ maximum margin/min $\|\cdot\|_2$ solution

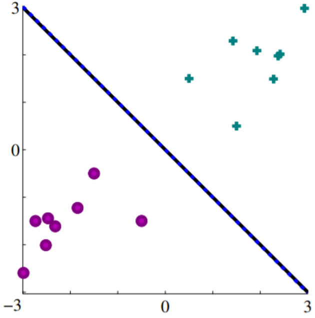
w_t gradient descent iterates

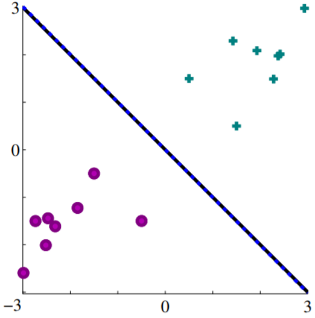
$$\boxed{w_t = \hat{w}\log(t) + \rho(t) \qquad \text{and} \qquad \rho(t) \leq C}$$

converges to the max margin: $$\lim_{t\to\infty} \frac{w_t}{\|w_t\|} = \frac{\hat{w}}{\|\hat{w}\|}$$

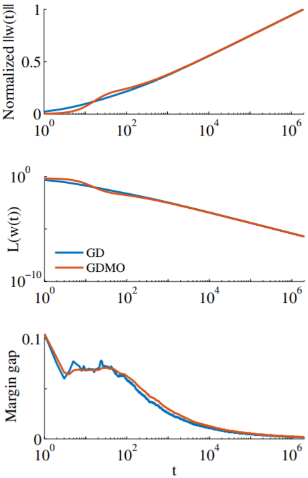
converges slowly: $$\left\| \frac{w_t}{\|w_t\|} - \frac{\hat{w}}{\|\hat{w}\|} \right\| = \tilde{O}\left(\frac{1}{\log t}\right)$$

simple case deep learning? connection to SVM

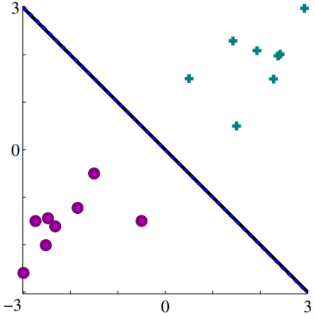

3
0
−3
0
3

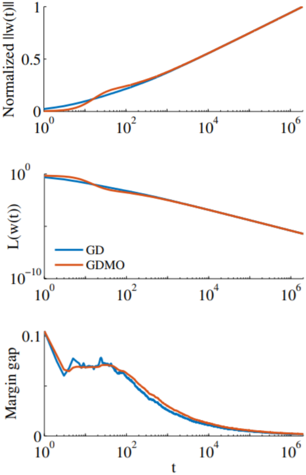


3
0
−3
0
3



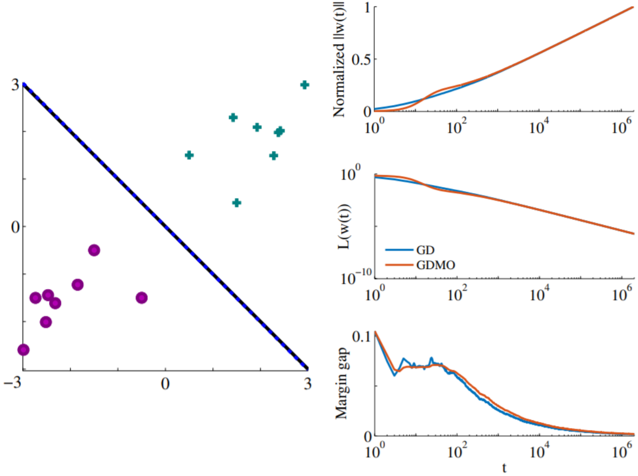
Normalized ||w(t)||
1
0.5
0
10^0
10^2
10^4
10^6
L(w(t))
10^0
10^{-10}
GD
GDMO
Margin gap
0.1
0
t




Normalized ||w(t)||
1
0.5
0
10^0
10^2
10^4
10^6
L(w(t))
10^0
10^{-10}
GD
GDMO
Margin gap
0.1
0
t

$\|w_t\|$

$O(\log t)$



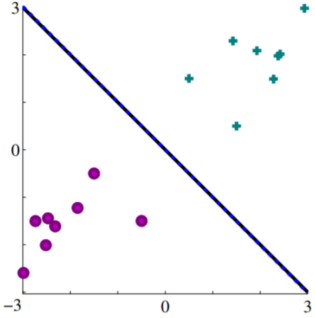
$\|w_t\|$

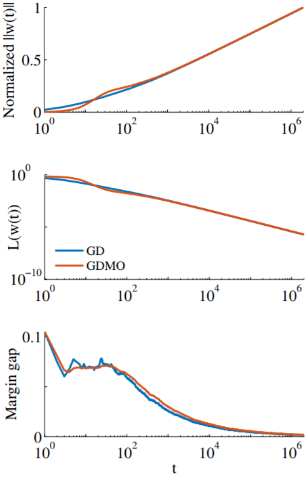
$O(\log t)$

$L(w_t)$

$O(1/t)$

Main results toy data





$\|w_t\|$

$O(\log t)$

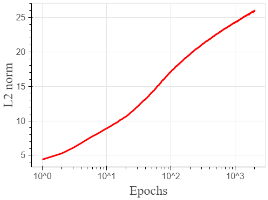
$L(w_t)$

$O(1/t)$

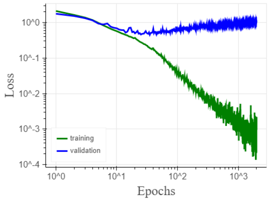
$\left\| \frac{w_t}{\|w_t\|} - \frac{\hat{w}}{\|\hat{w}\|} \right\|$

$\tilde{O}(1/\log t)$

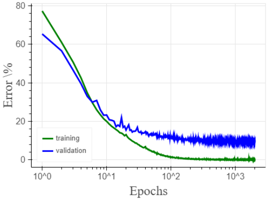
L2 norm, last layer



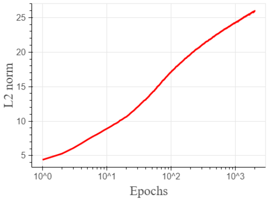
loss



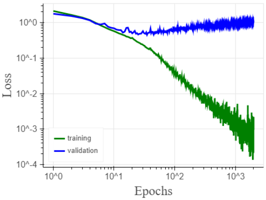
classification error



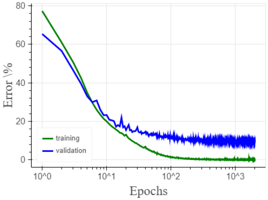
L2 norm, last layer



loss



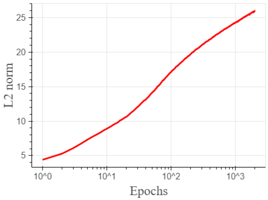
classification error

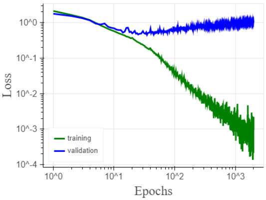


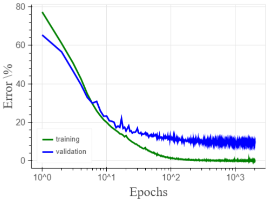
L2 norm, last layer

loss

classification error



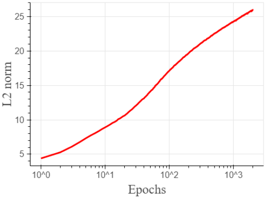


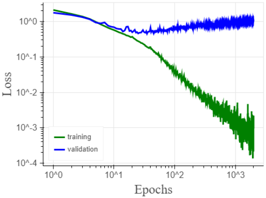


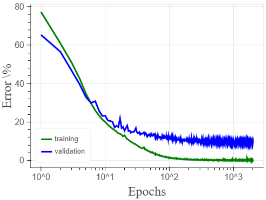
loss $\nearrow$ but error $\searrow$

bad probabilities, good separation

L2 norm, last layer | loss | classification error





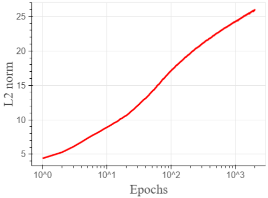


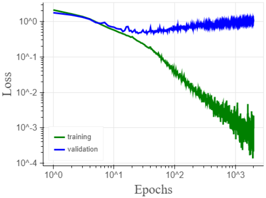
loss $\nearrow$ but error $\searrow$

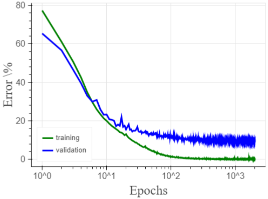
bad probabilities, good separation

Train longer, generalize better: closing the generalization gap in large batch training
Hoffer, Hubara, Soudry – NeurIPS 2017

L2 norm, last layer — loss — classification error







loss $\nearrow$ but error $\searrow$

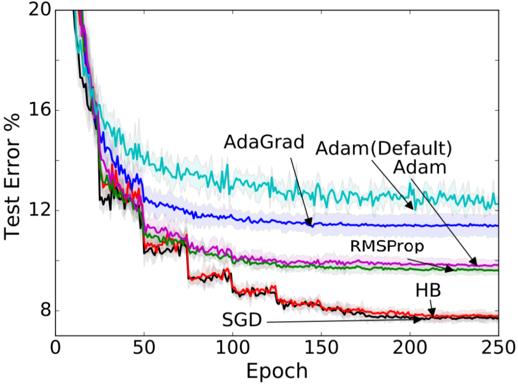
bad probabilities, good separation

Train longer, generalize better: closing the generalization gap in large batch training
Hoffer, Hubara, Soudry – NeurIPS 2017

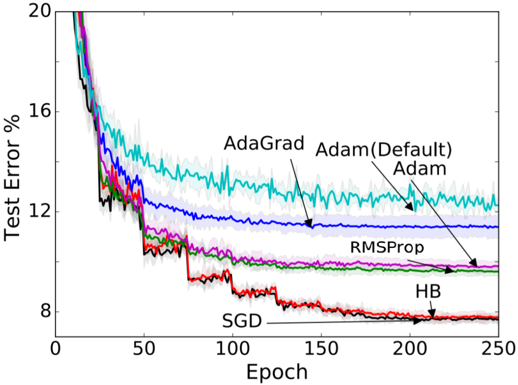
slow convergence to max margin

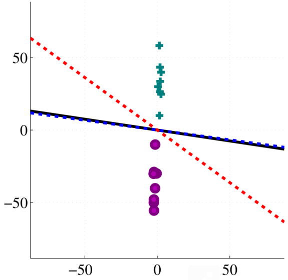
keep improving by training longer

The Marginal Value of Adaptive Gradient Methods in Machine Learning
Wilson, Roelofs, Stern, Srebro and Recht – NeurIPS 2017



The Marginal Value of Adaptive Gradient Methods in Machine Learning
Wilson, Roelofs, Stern, Srebro and Recht – NeurIPS 2017





Adam GD

Proof idea

very high level

- simplify the problem
- connection to support vectors
- converging sequence

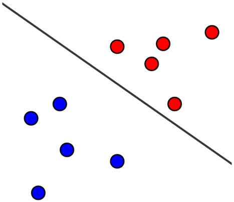
Logistic loss $\quad \log p(\,|x, w) \quad \log(1 + \exp(-w^\top x))$

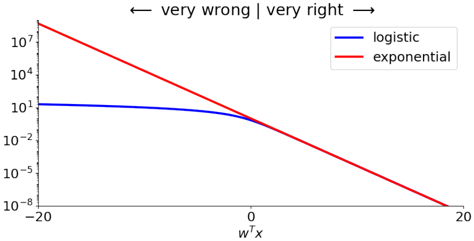
Proof idea | simplification

Logistic loss	$\log p(\bullet \mid x, w)$	$\log(1 + \exp(-w^\top x))$
Exponential loss		$\exp(-w^\top x)$

Proof idea — simplification

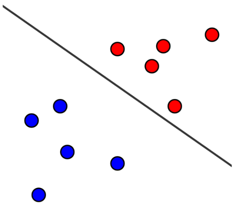
Logistic loss	$\log p(\bullet \mid x, w)$	$\log(1 + \exp(-w^\top x))$
Exponential loss		$\exp(-w^\top x)$

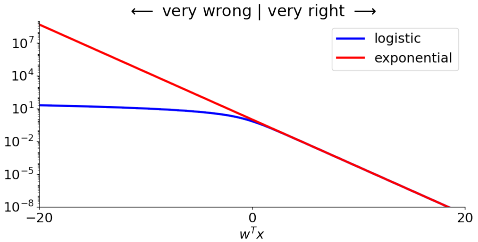




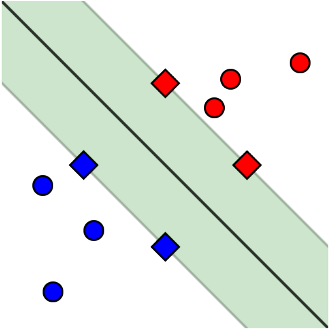
Proof idea — simplification

Logistic loss	$\log p(\bullet \mid x, w)$	$\log(1 + \exp(-w^\top x))$
Exponential loss		$\exp(-w^\top x)$



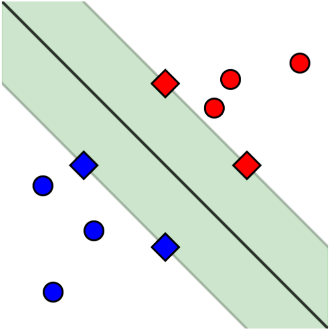


$$L(w) = \sum_n \exp(-w^\top x_n) \qquad \nabla L(w) = \sum_n \exp(-w^\top x_n) x_n$$



exponential loss

$$\nabla L(w) = \sum_n \exp(-w^\top x_n) x_n$$

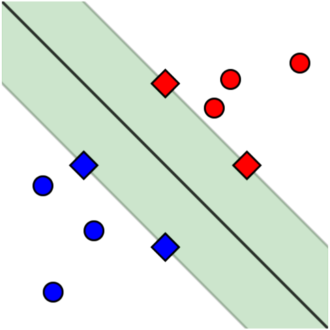


exponential loss

$$\nabla L(w) = \sum_n \exp(-w^\top x_n) x_n$$

max margin = sum of support vectors

$$\hat{w} = \sum_i \alpha_i x_i$$



exponential loss

$$\nabla L(w) = \sum_n \exp(-w^\top x_n) x_n$$

max margin = sum of support vectors

$$\hat{w} = \sum_i \alpha_i x_i$$

$$\underbrace{\text{smallest } |\hat{w}^\top x_n|}_{\text{support vector}} \qquad \underbrace{\text{other } |\hat{w}^\top x_n|}_{\text{not a support}}$$

$$\hat{w} = \sum_i \alpha_i x_i$$ combination of support vectors

$$\nabla L(w_t) = \sum_n \exp(-w_t^\top x_n) x_n$$

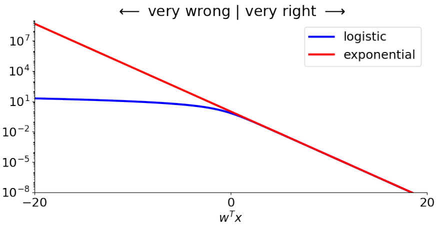
$$\hat{w} = \sum_i \alpha_i x_i$$ combination of support vectors

$$\begin{aligned}\nabla L(w_t) &= \sum_n \exp(-w_t^\top x_n) x_n \\ &= \underbrace{\sum_i \exp(-w_t^\top x_i) x_i}_{\text{data close to the boundary}} + \underbrace{\sum_j \exp(-w_t^\top x_j) x_j}_{\text{the rest}}\end{aligned}$$

$$\hat{w} = \sum_i \alpha_i x_i$$

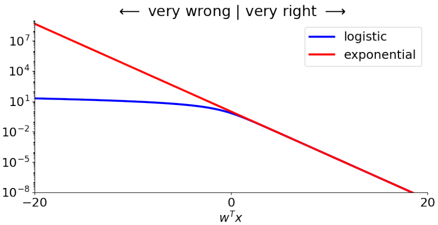
combination of support vectors

$$\begin{aligned}\nabla L(w_t) &= \sum_n \exp(-w_t^\top x_n) x_n \\ &= \underbrace{\sum_i \exp(-w_t^\top x_i) x_i}_{\text{data close to the boundary}} + \underbrace{\sum_j \exp(-w_t^\top x_j) x_j}_{\text{the rest}}\end{aligned}$$



$$\hat{w} = \sum_i \alpha_i x_i$$ combination of support vectors

$$\nabla L(w_t) = \sum_n \exp(-w_t^\top x_n) x_n$$
$$= \underbrace{\sum_i \exp(-w_t^\top x_i) x_i}_{\text{data close to the boundary}} + \underbrace{\sum_j \exp(-w_t^\top x_j) x_j}_{\text{the rest}}$$



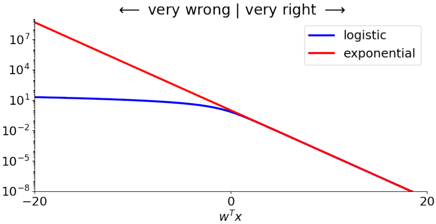
Proof idea

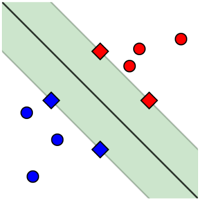
main insight

$$\hat{w} = \sum_i \alpha_i x_i$$

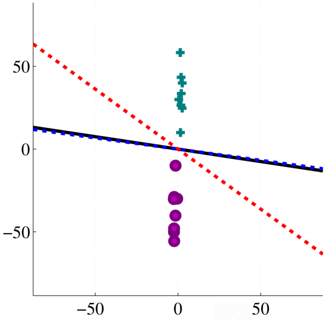
combination of support vectors

$$\begin{aligned}\nabla L(w_t) &= \sum_n \exp(-w_t^\top x_n) x_n \\ &= \underbrace{\sum_i \exp(-w_t^\top x_i) x_i}_{\text{data close to the boundary}} + \underbrace{\sum_j \exp(-w_t^\top x_j) x_j}_{\text{the rest}}\end{aligned}$$





Summary



gradient descent $\longrightarrow$ max margin

very slowly $\Rightarrow$ train longer

choice of optimizer $\Leftrightarrow$ solution

connection to SVM

Next week:
meaning of minimum norm solution and kernels methods with Joey

References, further reading

boosting	A decision-theoretic generalization of on-line learning and an application to boosting 1997 – Freund and Schapire
boosting as CD	Boosting the margin: a new explanation for the effectiveness of voting methods 1998 – Schapire, Freund, Bartlett and Lee
adaptive methods	The Marginal Value of Adaptive Gradient Methods in Machine Learning 2017 – Wilson, Roelofs, Stern, Srebro and Recht
slow conv.	Train longer, generalize better: closing the generalization gap in large batch training 2017 – Hoffer, Hubara, Soudry
	The implicit bias of gradient descent on separable data 2018 – Soudry, Hoffer, Shpigel Nacson, Gunasekar and Srebro
follow-up paper	Characterizing implicit bias in terms of optimization geometry 2018 – Gunasekar, Lee, Soudry and Srebro