An online passive-aggressive algorithm for difference-of-squares classification

Lawrence Saul

Department of Computer Science and Engineering
University of California, San Diego

NeurIPS Conference
December 7–10, 2021

Motivation

- **Data sets keep growing:**

  Many data sets are too large to fit in memory.
  They are best suited to algorithms for online learning.

- **But which online algorithm?**

  A very large number have been studied, especially for linear classifiers.
  One particularly elegant approach is that of *passive-aggressive learning*:

passive	the model is not updated when it classifies an example correctly with high confidence
aggressive	otherwise the model is changed by the minimum amount required to achieve this goal

- **These updates have some attractive features:**

  They dispense with the need to choose or adapt learning rates.
  There are convergence theorems and regret bounds (for linear classifiers).

Can these updates be extended to nonlinear models of classification?

Background (for linear models)

- **Online setting**

  Algorithm has access to a stream of labeled examples $\{(\mathbf{x}_t, y_t)\}_{t \geq 1}$ with $\mathbf{x}_t \in \mathbb{R}^n$ and $y_t \in \{-1, +1\}$.

- **Linear classification**

  A weight vector $\mathbf{w} \in \mathbb{R}^n$ is initialized to the origin.
  The goal over time is to learn a hyperplane decision boundary:

$$y = \text{sign}(\mathbf{w}^\top \mathbf{x}).$$

- **Passive-aggressive learning**

$$\mathbf{w}_{t+1} = \underset{\mathbf{w} \in \mathbb{R}^n}{\text{argmin}} \, \|\mathbf{w} - \mathbf{w}_t\|^2 \quad \text{such that} \quad y_t(\mathbf{w}^\top \mathbf{x}_t) \geq 1.$$

  [Crammer, Dekel, Keshet, Shalev-Shwartz, & Singer (2006)]

- **Update rule**

  The required optimization is a convex quadratic program (QP).
  It has a simple closed-form solution:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t y_t \mathbf{x}_t \quad \text{with} \quad \alpha_t = \frac{\max(0, 1 - y_t \mathbf{w}^\top \mathbf{x}_t)}{\|\mathbf{x}_t\|^2}.$$

Beyond linear models

Can passive-aggressive (PA) updates be derived for nonlinear models?
One way to do this has been widely studied; another one hasn't.

1. **Updates for kernel-based classifiers**

   The kernel trick converts a linear classifier into a nonlinear one.
   But in the online setting the number of examples is unbounded.
   Most kernelized algorithms budget the number of support vectors.

   [Crammer et al (2004), Weston et al (2005), Cavallanti et al (2007), Dekel et al (2008), Orabona et al (2009), Wang & Vucetic (2010), Wang et al (2012), Lu et all (2016), Wu et all (2020).]

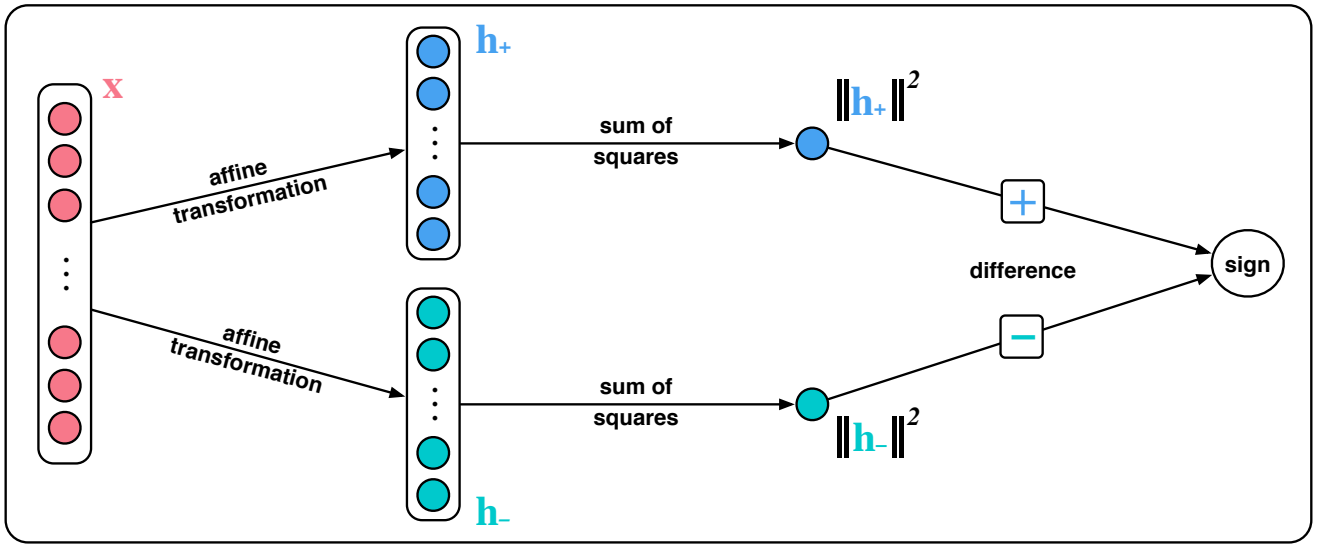
2. **Updates for more richly parameterized models**

   Without the kernel trick we face a different problem.
   PA learning hinges on the ability to derive a minimal update:

$$\boldsymbol{\Theta}_{t+1} \;=\; \operatorname*{argmin}_{\boldsymbol{\Theta}} \|\boldsymbol{\Theta}-\boldsymbol{\Theta}_t\|^2 \quad \text{such that} \quad \textbf{margin}(\mathbf{x}_t, y_t, \boldsymbol{\Theta}) \geq 1.$$

   This problem is a convex QP for hyperplane decision boundaries.
   But is this optimization tractable for any nonlinear models?

Difference-of-squares (DoS) classification



The decision boundary is parameterized by a pair of affine transformations:

$$\begin{aligned} \mathbf{h}_+ &= \mathbf{U}\mathbf{z} \quad \text{where} \quad \mathbf{z} = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} \\ \mathbf{h}_- &= \mathbf{V}\mathbf{z} \\ y &= \operatorname{sign}\left(\|\mathbf{h}_+\|^2 - \|\mathbf{h}_-\|^2 \right) \end{aligned}$$

The maps $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d \times (n+1)}$ define a low-rank model of quadratic classification. The form of the decision boundary is inspired by earlier work:

factorization machines	[Rendle, 2010]	
capsule networks	[Sabour et al, 2017]	(with pose vectors $\mathbf{h}_\pm$)

Passive-aggressive learning

- **Optimization**

$$\min_{\mathbf{U},\mathbf{V}} \left\{ \|\mathbf{U}-\mathbf{U}_t\|_F^2 + \|\mathbf{V}-\mathbf{V}_t\|_F^2 \right\} \quad \text{such that} \quad y_t \left(\|\mathbf{U}\mathbf{z}_t\|^2 - \|\mathbf{V}\mathbf{z}_t\|^2 \right) \geq 1$$

This is a quadratically constrained quadratic program (QCQP).
It is **nonconvex** (due to the constraint) but still **tractable**.

[Moré, 1993; Boyd et al, 1994; Feron, 2000; Ye & Zhang, 2003; Polik & Terlaky, 2007; Feng et al, 2012; Park & Boyd, 2017]

- **Update rule**

Compute a step size α_t by optimizing a convex function over the unit interval:

$$\alpha_t = \operatorname*{arg\,min}_{\alpha \in (0,1)} \left[\left(\frac{1}{1-y_t\alpha} \right) \|\mathbf{U}_t\mathbf{z}_t\|^2 + \left(\frac{1}{1+y_t\alpha} \right) \|\mathbf{V}_t\mathbf{z}_t\|^2 - \alpha \right]$$

Then update $\mathbf{U}$ and $\mathbf{V}$ using this step size:

$$\mathbf{U}_{t+1} = \mathbf{U}_t + \left(\frac{\alpha_t}{y_t - \alpha_t} \right) \frac{(\mathbf{U}_t\mathbf{z}_t)\mathbf{z}_t^\top}{\|\mathbf{z}_t\|^2}$$

$$\mathbf{V}_{t+1} = \mathbf{V}_t - \left(\frac{\alpha_t}{y_t + \alpha_t} \right) \frac{(\mathbf{V}_t\mathbf{z}_t)\mathbf{z}_t^\top}{\|\mathbf{z}_t\|^2}$$

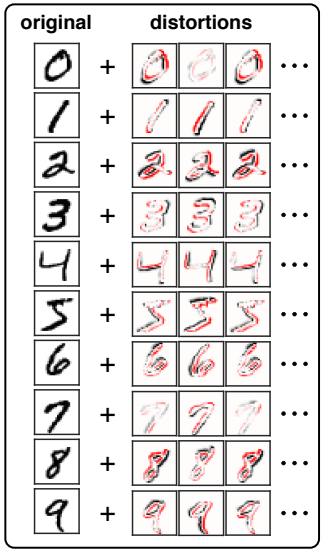
This update is nearly as simple as the one for linear classification.

Experimental results

- **INFIMNIST data set**

  We trained all-vs-all classifiers on 100M training examples of handwritten digits.

  No pairs of digits are linearly separable when there are this many training examples.

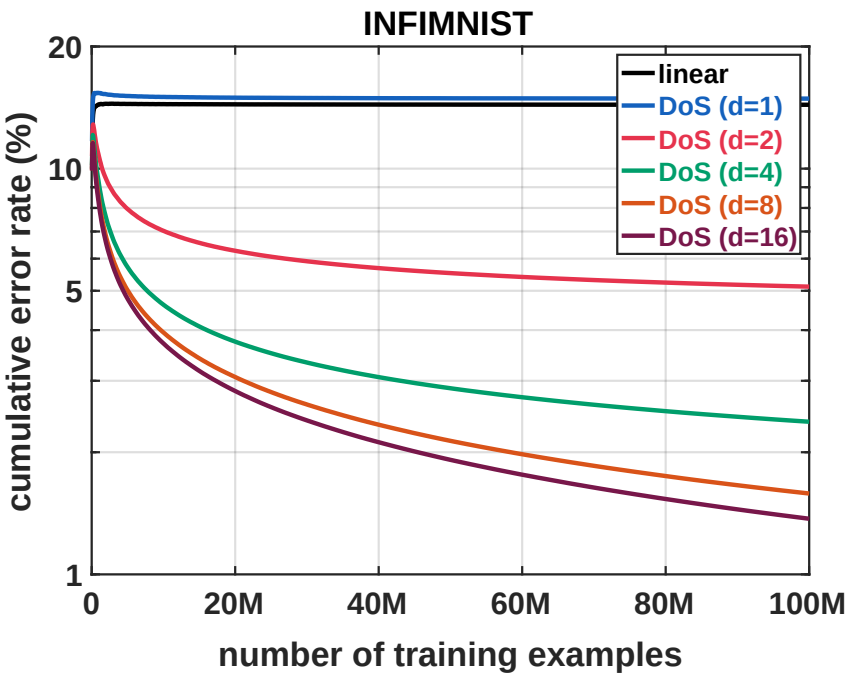


[Loosli et al, 2007]

- **Cumulative error rates**

  The higher the rank d of the DoS classifier, the greater its capacity.

  As expected, classifiers with greater capacity have lower error rates.



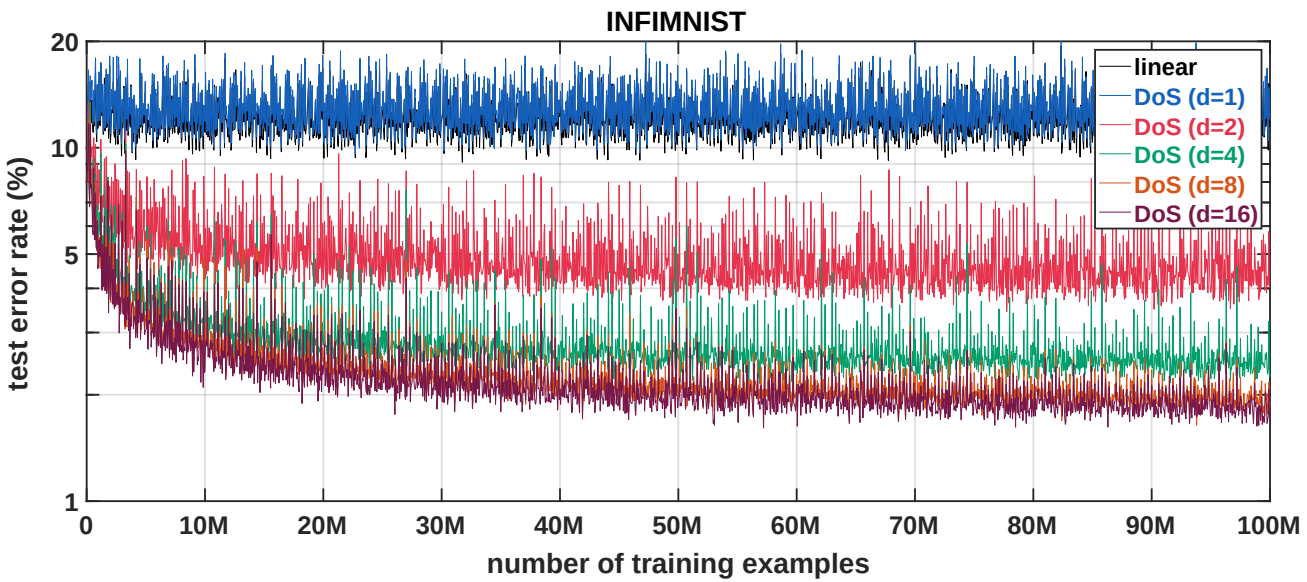
Convergence (or lack thereof)

MODEL	REGULARIZER	CONSTRAINT
linear	$\|\mathbf{w}-\mathbf{w}_t\|^2$	$1 \leq y_t(\mathbf{w}^\top \mathbf{x}_t)$
DoS	$\|\mathbf{U}-\mathbf{U}_t\|_F^2 + \|\mathbf{V}-\mathbf{V}_t\|_F^2$	$1 \leq y_t\left(\|\mathbf{U}\mathbf{z}_t\|^2 - \|\mathbf{V}\mathbf{z}_t\|^2\right)$

- **Model instability**

  PA updates do not converge if the classes are not separable.
  This instability occurs in both linear and nonlinear models.

- **Test error rates fluctuate wildly**



How to fix this?

Model averaging

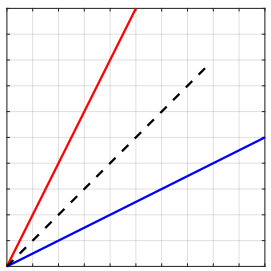
- **Averaging over time:**

  In the linear case, a stable model can be computed by averaging the estimated weight vectors over time: e.g.,

$$\mathbf{w}_{\text{avg}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t.$$

- **Why this makes sense for linear models:**

  If $\mathbf{w}_1$ and $\mathbf{w}_2$ specify hyperplanes, then $\frac{1}{2}(\mathbf{w}_1 + \mathbf{w}_2)$ specifies an intermediate hyperplane sandwiched between $\mathbf{w}_1$ and $\mathbf{w}_2$.



- **Why it doesn't for DoS models:**

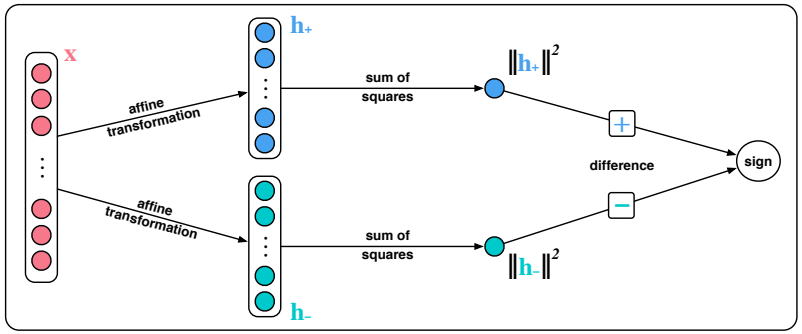
  DoS models are **overparameterized**.
  Simple averages can give nonsensical results.

Overparameterization in DoS models

- **Quadratic decision boundary**

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$$

$$y = \text{sign}\left(\|\mathbf{U}\mathbf{z}\|^2 - \|\mathbf{V}\mathbf{z}\|^2 \right)$$



- **Different parameters, same model**

  Let $\boldsymbol{\Theta}_1 = (\mathbf{U}, \mathbf{V})$ and $\boldsymbol{\Theta}_2 = (-\mathbf{U}, -\mathbf{V})$.
  Then these two models specify the same classifier.

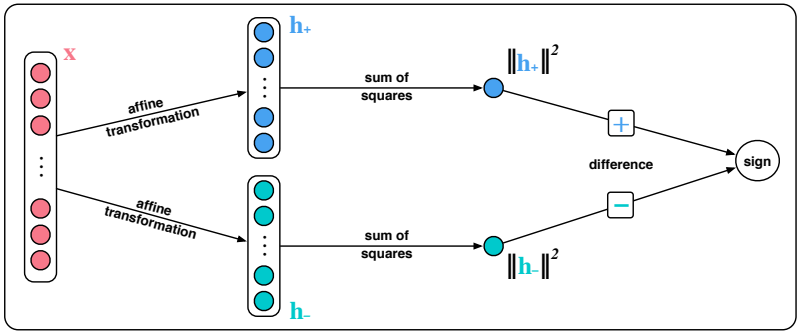
- **Where simple averaging goes wrong**

  The average of two identical classifiers should be the same classifier.
  But note that $\boldsymbol{\Theta}_{\text{avg}} = \frac{1}{2}(\boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2) = (\mathbf{0}, \mathbf{0})$.

Continuous symmetries of the parameter space

The model's decision boundary can be written as:

$$y = \operatorname{sign}\left(\mathbf{z}^\top [\mathbf{U}^\top \mathbf{U} - \mathbf{V}^\top \mathbf{V}]\, \mathbf{z}\right)$$



The model is invariant under the following transformations of its parameters $(\mathbf{U}, \mathbf{V})$:

1. **Orthogonal transformations**

$$\begin{aligned} \mathbf{U} &\mapsto \boldsymbol{\Omega}\,\mathbf{U} \quad \text{where} \quad \boldsymbol{\Omega}^\top \boldsymbol{\Omega} = \mathbf{I} \\ \mathbf{V} &\mapsto \boldsymbol{\Lambda}\,\mathbf{V} \quad \text{where} \quad \boldsymbol{\Lambda}^\top \boldsymbol{\Lambda} = \mathbf{I} \end{aligned}$$

   These transformations preserve the norms $\|\mathbf{U}\|_F^2$ and $\|\mathbf{V}\|_F^2$.

2. **Lorentz transformations**

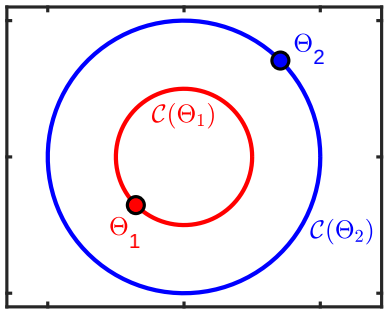
$$\begin{aligned} \mathbf{U} &\mapsto \mathbf{U} \cosh\varphi - \mathbf{V} \sinh\varphi \\ \mathbf{V} &\mapsto \mathbf{V} \cosh\varphi - \mathbf{U} \sinh\varphi \end{aligned}$$

   The norm $\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2$ is minimized by $\varphi = \frac{1}{2} \tanh^{-1}\left(\frac{2\,\langle \mathbf{U}, \mathbf{V} \rangle}{\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2}\right)$.

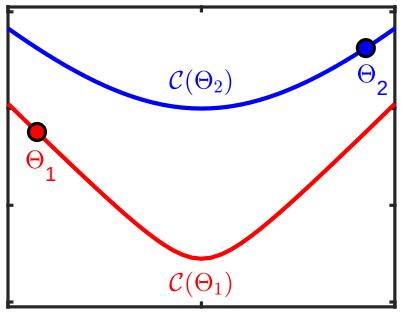
How to average $\mathbf{\Theta}_1 = (\mathbf{U}_1, \mathbf{V}_1)$ and $\mathbf{\Theta}_2 = (\mathbf{U}_2, \mathbf{V}_2)$

- **Manifolds in parameter space**

  Each point $\mathbf{\Theta}_i$ in parameter space belongs to an equivalence class $\mathcal{C}(\mathbf{\Theta}_i)$ of models generated by orthogonal and Lorentz transformations.



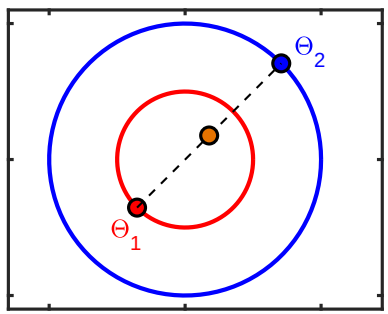
rotational invariance

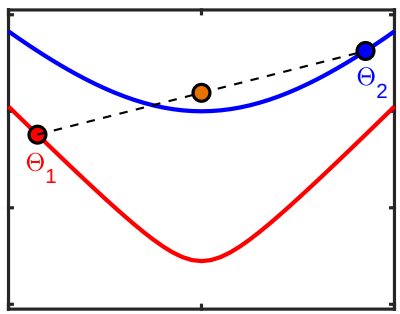


Lorentz invariance

- **Why simple averages fail**

  The midpoint $\frac{1}{2}(\mathbf{\Theta}_1 + \mathbf{\Theta}_2)$ is not parameterization-invariant; it depends on where $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ lie on these manifolds.

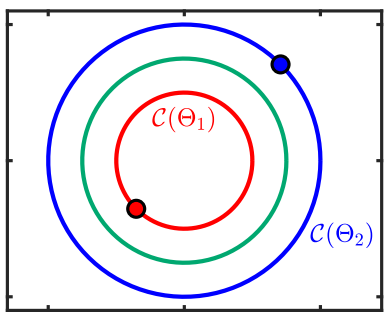


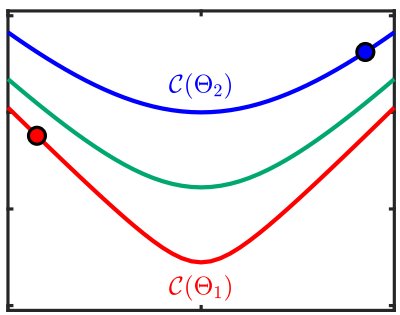


NO!

- **What we want instead**

  An invariant average should yield a model in the **equivalence class** equidistant from $\mathcal{C}(\mathbf{\Theta}_1)$ and $\mathcal{C}(\mathbf{\Theta}_2)$.





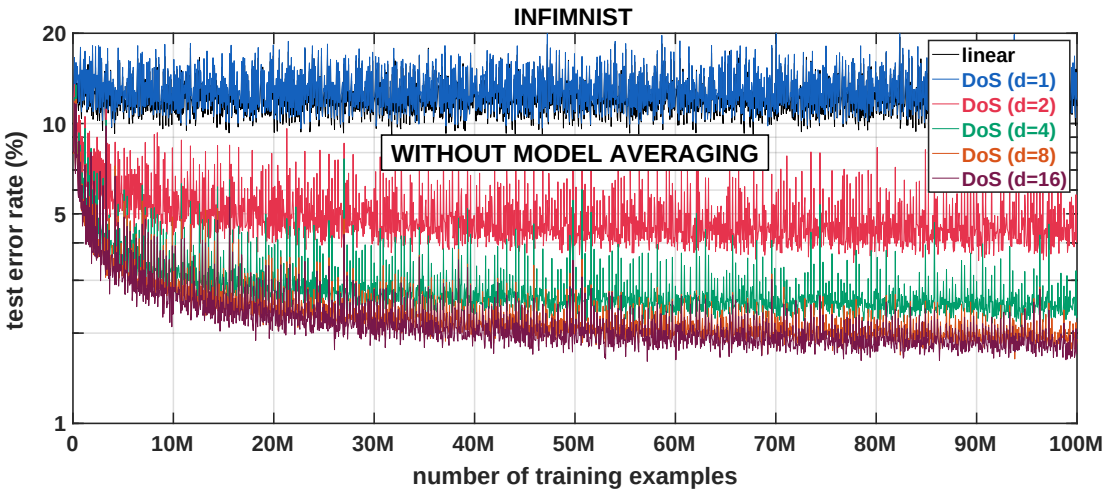
YES!

How to compute parameterization-invariant averages? **See the paper.**

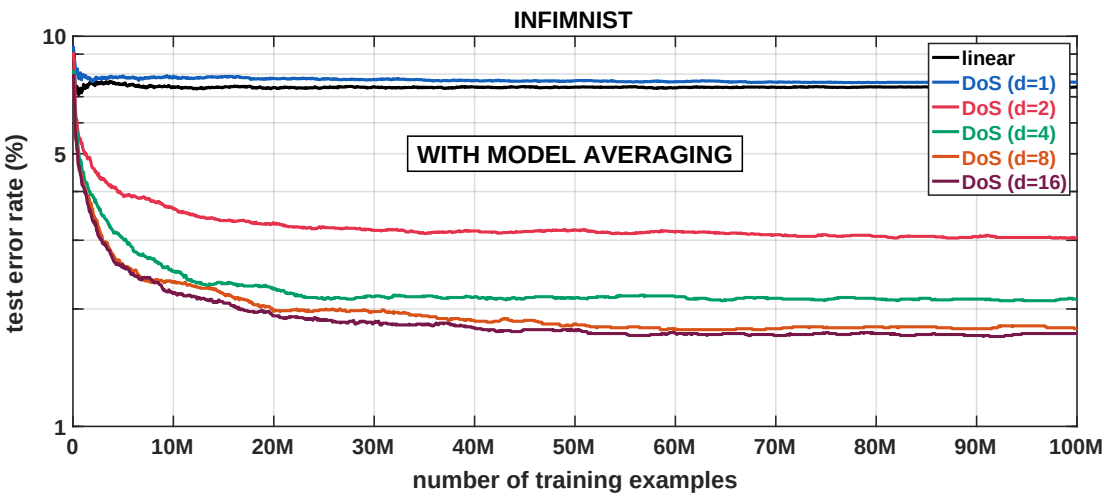
Also: [Bamler & Mandt, 2018]

The effect of model averaging

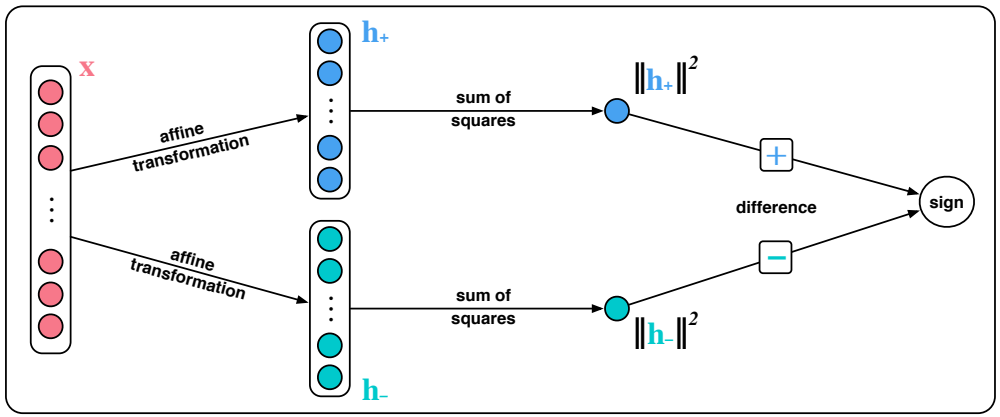
- **Test error rates fluctuate wildly without model averaging**



- **Test error rates stabilize and improve when models are averaged across time**



Conclusion



- **PA learning can be extended to low-rank quadratic classification:**
  Updates are obtained by solving a nonconvex but tractable QCQP.
  The algorithm scales well to very large data sets.

- **The parameter space has continuous symmetries:**
  Different parameters can give rise to the same decision boundary.
  These symmetries suggest a sensible way to average models across time.

- **Are there theoretical guarantees for these PA updates?**
  Convergence theorems and regret bounds exist for linear models.
  It remains an open question to extend these guarantees.