Stochastic Algorithms for One-Pass Learning

Léon Bottou

Question

SGD often needs to process each example many times.

– Can we reduce the number of epochs to just one?
– How much do we need to increase the cost of each iteration?

- In general, reducing the number of epochs k times is attractive if we increase the cost per iteration less than k times.

- When the data does not fit in memory, the cost per iteration includes loading and preprocessing the data. We can do a lot of computation while the data is loading.
  (e.g. VW, Langford et al.)

Refining the question

Can we reduce the number of epochs to just one?

– What does this mean anyway?
– We can always perform a single epoch and stop.
– But we expect a certain level of accuracy.
– Test set accuracy (generalization.)

Which level of accuracy do we expect?

A – The "optimal" statistical generalization accuracy?
B – The test set accuracy we could reach with batch optimization?
C – The test set accuracy we could reach by performing more epochs?

All three criteria often mean the same thing.
Papers often look very different.

Answers

Optimal estimation using stochastic approximations

A (Sakrison, *Trans. Information Theory*, 12(1). 1966)
A (Anbar, Ph.D. dissertation, Berkeley, 1971)
A (Abdelhamid, Ph.D. dissertation, Michigan State, 1971)
A (Nevel'son and Has'minskij, *Stochastic Approximation...*, Nauka, 1973)
A (Fabian, *Annals of Statistics*, 1(3), 1973)
A (Fabian, *Annals of Statistics*, 1(3), 1978)
A (Amari, *Neural Computation*, 10(2), 1998)
BC (Bottou and LeCun, *Applied Stoch. Models in Bus. and Ind.*, 21(2), 2004)

Averaged Stochastic Gradient Descent

C (Ruppert, *Efficient Estimators from a Slowly Convergent RM Process*, 1988)
AC (Polyak and Juditsky, *Automation and Remote Control*, 30(4), 1992)
BC (Xu, NEC Tech Report (ArXiV.1107.2490), 2010)
C (Bach and Moulines, Nips 2011)

Summary

i. Asymptotically efficient stochastic algorithms (A).

ii. Two diagrams and one experiment (BC).

iii. Averaged Stochastic Gradient Descent.

I. Asymptotically Efficient Stochastic Algorihms

– Estimating a mean.
– Estimating a regression.
– Estimating a density.

Estimating a mean

Setup

– Sample $x_1, x_2, x_3, ...$ from an unknown normal distribution.
– Estimate the mean.

The least-square estimator ...

$$\hat{\mu}(x_1 \dots x_n) \stackrel{\triangle}{=} \arg\min_{\mu} \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

... is also the best unbiased estimator.

$$\hat{\mu}(x_1 \dots x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

– Note: slightly shrinking $\hat{\mu}$ gives a better *biased* estimator (James-Stein).

SGD for estimating the mean

Minimize: $\min_w \frac{1}{n}\sum_{i=1}^{n}(w-x_i)^2$.

Iterate

– Pick a sample x_t

– Update $w_t = w_{t-1} - 2\eta_t(w_{t-1}-x_t)$.

Optimal gain

– Choose η_t such that $2\eta_t = 1/t$. Then

$$t\,w_t = (t-1)\,w_{t-1} + x_t = (t-2)\,w_{t-2} + x_{t-1} + x_t = \ldots$$
$$w_t = \hat{\mu}(x_1,\ldots,x_t)$$

– Optimal unbiased estimator in one pass!

Linear Regression

Setup

– Let $\ell(x, y, w) = \frac{1}{2}(w^\top x - y)^2$.

– Assume (X, Y) follow a multivariate normal distribution.

– Find w^* that minimizes $C(w) \stackrel{\triangle}{=} \mathbb{E}\left[\ell(X, Y, w)\right]$.

Least Square Estimator

$$w_n^* \stackrel{\triangle}{=} \hat{w}(x_1, y_1, \ldots, x_n, y_n) \stackrel{\triangle}{=} \arg\min_w \frac{1}{2n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

– With $H \stackrel{\triangle}{=} \mathbb{E}\left[\frac{\partial^2 \ell}{\partial w^2}(X, Y, w^*)\right]$ and $G \stackrel{\triangle}{=} \mathrm{Cov}\left[\frac{\partial \ell}{\partial w}(X, Y, w^*)\right]$,
$w_n^* - w^*$ is asymptotically normal with variance $\mathrm{tr}(H^{-1} G\, H^{-1})\, n^{-1}$.

Note: slightly biasing w_n^* with a regularization term gives a better estimator...

SGD for the regression

Minimize $\mathbb{E}[\ell(X, Y, w)]$ with SGD.

Iterate

– Pick a sample (x_t, y_y) of (X, Y).

– Update $w_t = w_{t-1} - \eta_t \dfrac{\partial \ell}{\partial w}(x_t, y_t, w_{t-1})$

– Gain η_t can be a scalar or a positive matrix.

Gradient descent with noise

– $\dfrac{\partial \ell}{\partial w}(x_t, y_t, w)$ is a noisy estimate of the gradient $C'(w)$.

– $\dfrac{\partial \ell}{\partial w}(x_t, y_t, w) = C'(w) + \varepsilon_t$ with $\mathbb{E}[\varepsilon_t] = 0$, and $\mathrm{Cov}[\varepsilon_t] \to G$.

– Gradient descent with noise: $w_t = w_{t-1} - \eta_t C'(w_{t-1}) - \eta_t \varepsilon_t$

SGD for the regression (ii)

Rewrite

– Let $u_t \stackrel{\triangle}{=} w_t - w^*$.

$$\begin{aligned} u_t &= u_{t-1} - \eta_t C'(w_{t-1}) - \eta_t \varepsilon_t \\ &= u_{t-1} - \eta_t H u_{t-1} - \eta_t \varepsilon_t \\ &= (I - \eta_t H) u_{t-1} + \eta_t \varepsilon_t . \end{aligned}$$

Recurse

$$u_t = \left(\prod_{i=1}^{t} (I - \eta_i H) \right) u_0 - \sum_{i=1}^{t} \left(\prod_{j=i+1}^{t} (I - \eta_j H) \right) \eta_i \varepsilon_i$$

– The green term gives the mean of $w_t - w^*$.
– The variance of the red term gives the variance of $w_t - w^*$.
– Both can be upper- and lower-bounded with calculus.

SGD for the regression (iii)

After lenghty derivations:

η_t	Green	Var[Red]
B with $\lambda \ll 1$	$\equiv e^{-\lambda t}$	$\not\to 0$
$Bt^{-\alpha}$ with $1/2<\alpha<1$	$\equiv e^{-\frac{\lambda}{1-\alpha} t^{1-\alpha}}$	$\equiv t^{-\alpha}$
Bt^{-1} with $\lambda > 1/2$	$\equiv t^{-\lambda}$	$\sim \mathrm{tr}(BGB)\,(2\lambda-1)^{-1}\,t^{-1}$
$H^{-1}\,t^{-1}$	$\equiv t^{-1}$	$\sim \mathrm{tr}(H^{-1}G\,H^{-1})\,t^{-1}$

λ denotes the min/max eigenvalue of $(BH+HB)/2$ for upper and lower bounds.

Second order SGD

– Choosing η_t such that $\eta_t H = 1/t$ yields the same variance as $\hat{w}$.

– Although $w_t \neq \hat{w}(x_1, y_1 \ldots x_t, y_t)$, w_t is as good an estimator as $\hat{w}$.

(Sakrison, 1966; Anbar 1971, Abdemhamid, 1971, Fabian, 1973; . . .)

Parametric Density Estimation

Setup

– Let $\phi(x;\theta)$ be a parametric family of densities.
– Let's recover θ^* using an i.i.d. sample of $\phi(\cdot;\theta^*)$

Fisher Information

$$I(\theta) \triangleq \mathbb{E}_\theta\left[\left(\frac{\partial \log \phi(X;\theta)}{\partial \theta}\right)\left(\frac{\partial \log \phi(X;\theta)}{\partial \theta}\right)^\top\right] = \mathbb{E}_\theta\left[\frac{\partial^2 \log \phi(X;\theta)}{\partial \theta^2}\right]$$

Cramer-Rao bound

– An estimator $\hat{\theta}(x_1,\ldots,x_n)$ is unbiased if $\mathbb{E}_{\theta^*}[\hat{\theta}(x_1,\ldots,x_n)] = \theta^* \ \ (\forall \theta^*)$.
– Unbiased estimators satisfy $\mathrm{Var}[\hat{\theta}(x_1,\ldots,x_n)] \geq I(\theta^*)^{-1}\, n^{-1}$.
– *Efficient estimators* achieve this bound (best unbiased estimators.)
– MLE estimator is *asymptotically unbiased* and *asymptotically efficient*.

Note: With $I(\theta^*) = G = H$, observe $I(\theta^*)^{-1} = H^{-1} G\, H^{-1}$.

SGD for Density Estimation

Maximize $C(\theta) = \mathbb{E}_{\theta^*}[\log \phi(X;\theta)]$ with SGD.

Iterate

– Pick a sample x_t of X, distributed according to $\phi(\cdot;\theta^*)$.

– Update $\theta_t = \theta_{t-1} + \eta_t \frac{\partial \log \phi}{\partial \theta}(x_t;\theta_{t-1})$.

Optimal gains

– Second order SGD : $\eta_t = H^{-1}t^{-1}$.

– Natural gradient : $\eta_t = I(\theta_{t-1})^{-1}t^{-1}$

– In both cases, we find $\mathbb{E}[\theta_t - \theta^*] \equiv t^{-1}$ and $\mathrm{Var}[\theta_t - \theta^*] \sim I(\theta^*)^{-1}t^{-1}$.

Stochastic algorithm with optimal gains reaches the Cramer-Rao bound.
Stochastic algorithm with optimal gains is *asymptotically efficient*.

(Nevel'son and Has'minskij, 1973; Fabian, 1978; Amari, 1998)

II. Two diagrams and one experiment

The two cultures

"Statistical modeling: the two cultures" (Breiman, 2001)

The *data modeling culture*:

- Assume responsevar $= f($predictorvars, randomnoise, parameters$)$
- Justify the model and noise assumptions by hand waving.
- Estimate parameters and draw conclusions.
- Validate using goodness-of-fit, etc.

The *algorithmic modeling culture*:

- Assume data comes from a complex and unknown black box.
- Choose a family of functions $f($predictordata, parameters$)$
- Pick the function that best approximate the data.
- Validate answer using cross-validation, out-of-sample testing, etc.

This split also exists in machine learning.

Stream of consciousness

My intuition belongs to the algorithmic modeling culture.

"*Statistics from the seventies: data modeling culture.*"

"*I can follow the maths, but I do not get what they mean.*"

"*What's this business about unbiased estimators anyway.*"

"*Don't we always bias our results with priors and regularizers?*"

"*These people are not idiots. They surely mean something.*"

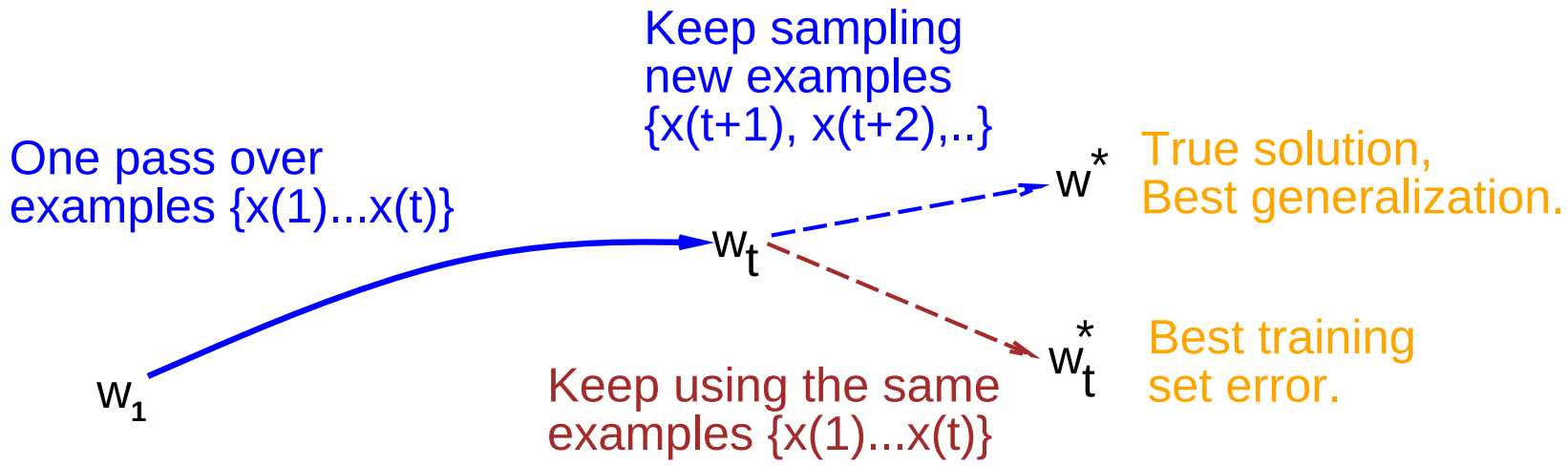
"*How can I understand what they mean, but in my world?*"

Stochastic gradient paths

How do we pick the SGD training examples?

– If we pick the training examples from a distribution p,
then SGD optimizes the expectation $\mathbb{E}_p[\ell(X, w)]$ for this distribution.

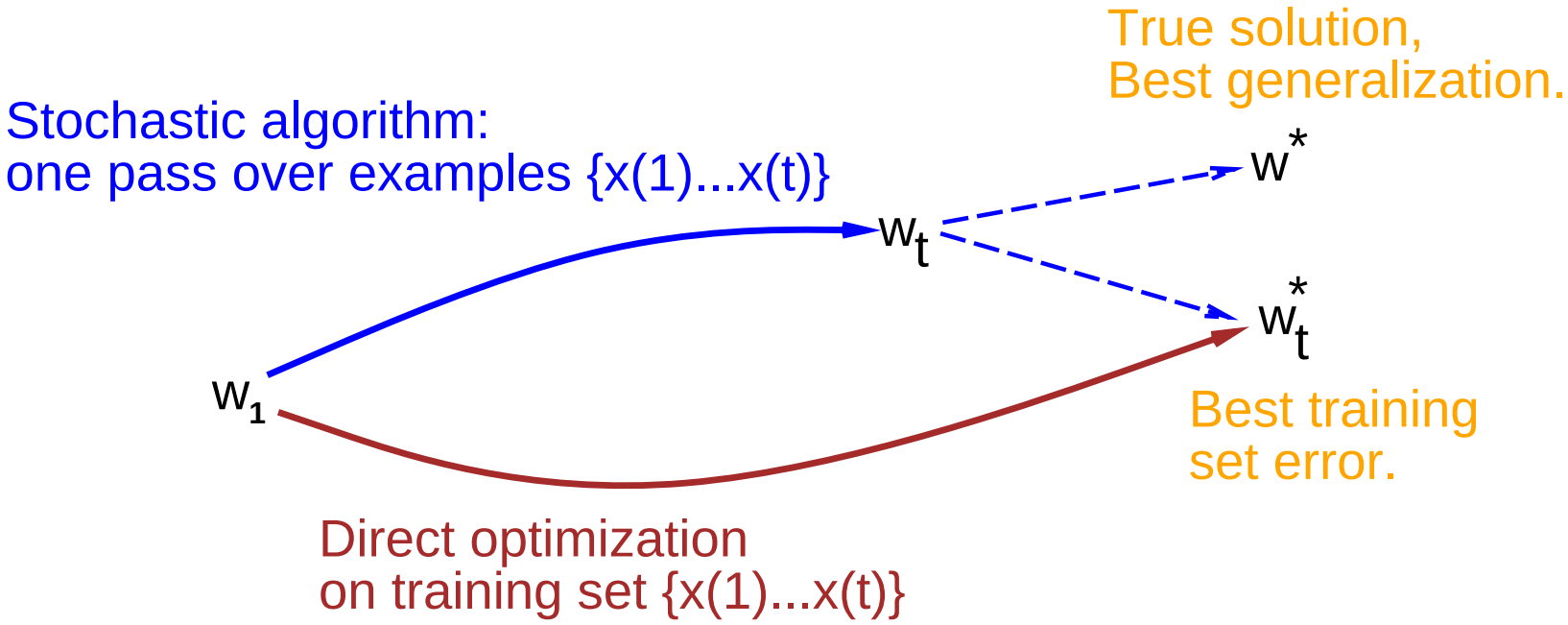
– The true distribution or the discrete training set distribution?



Batch and online paths

Stochastic algorithm $w_t = w_{t-1} - \eta_t \frac{\partial \ell}{\partial w}(x_t, w_{t-1})$

Batch optimization $w_t^* = \arg\min_w \frac{1}{t} \sum_{i=1}^{t} \ell(x_t, w)$



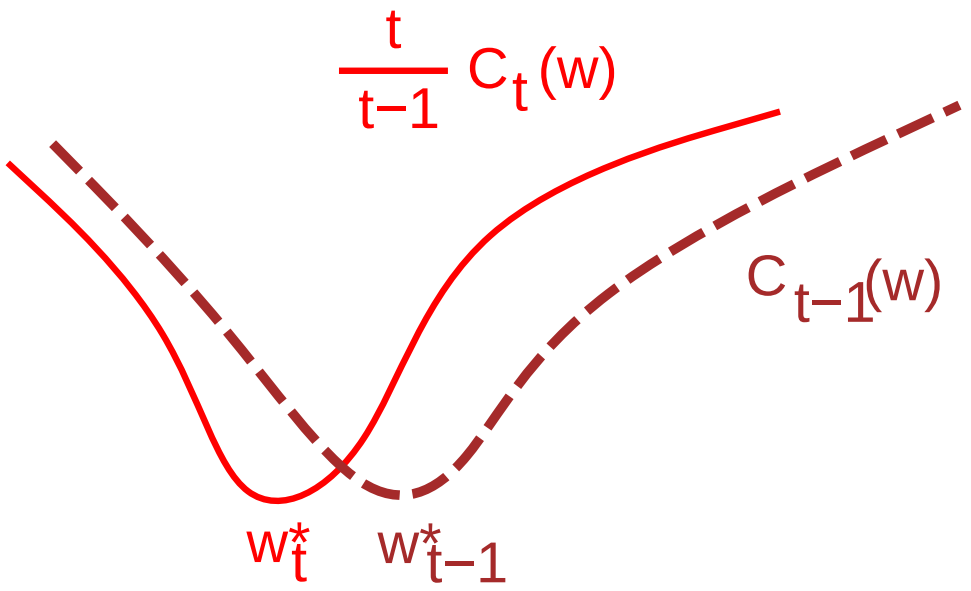
– This diagram represents criteria B and C.

The Direct Optimization Path

Let $C_t(w) \triangleq \frac{1}{t} \sum_{i=1}^{t} \ell(x_i, w)$ and compare

$$w_{t-1}^* = \arg\min_w C_{t-1}(w)$$

$$w_t^* = \arg\min_w C_t(w) = \arg\min_w \left[C_{t-1}(w) + \frac{1}{t-1} \ell(x_t, w) \right]$$



The Direct Optimization Path (ii)

First Order Calculation

$$w_t^* = w_{t-1}^* - \frac{1}{t} H_t^{-1} \frac{\partial \ell}{\partial w}(x_t, w_{t-1}) + \mathcal{O}\left(\frac{1}{t^2}\right)$$

where H_t is the empirical Hessian on t examples.

$$H_t = \frac{1}{t} \sum_{i=1}^{t} \frac{\partial^2 \ell}{\partial w^2}(x_i, w_{t-1}) \quad \stackrel{t\to\infty}{\longrightarrow} H$$

Compare with Second Order Stochastic Gradient Descent

$$w_t = w_{t-1} - \frac{1}{t} H^{-1} \frac{\partial \ell}{\partial w}(x_t, w_{t-1})$$

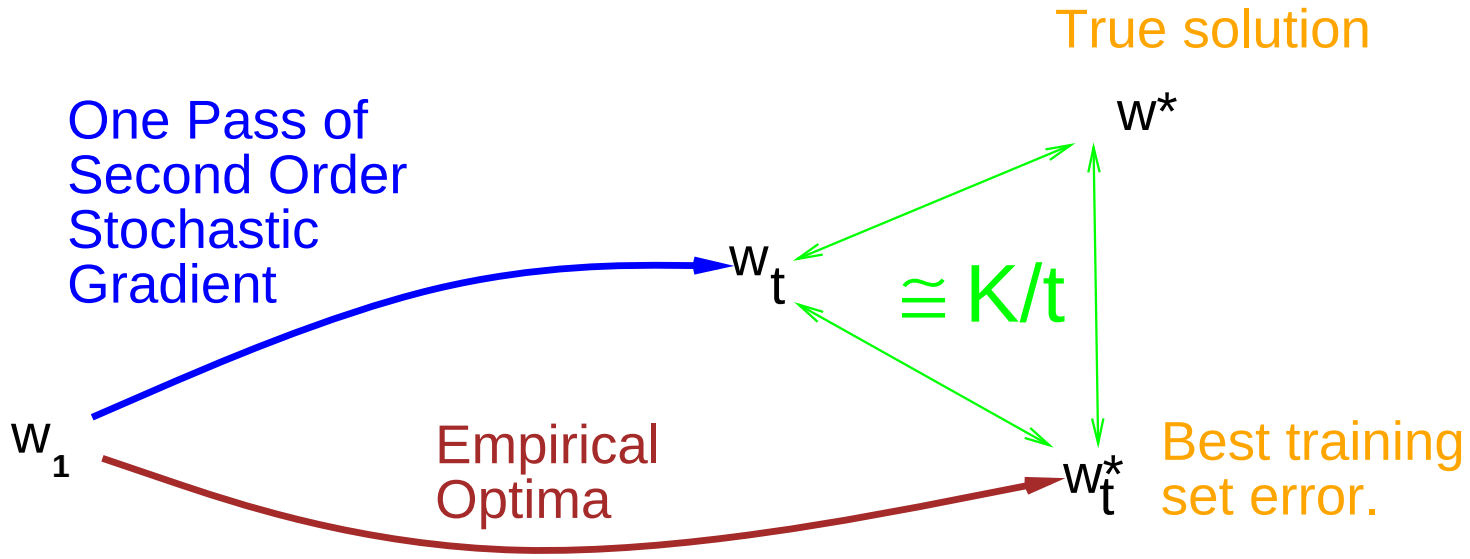
– Could w_t^* and w_t converge with the same speed despite the high order differences?

Result

Under ordinary differentiability assumptions:

$$\mathbb{E}[(w_t - w^*)^2] \sim \mathrm{tr}(H^{-1}G\,H^{-1})\,t^{-1} \qquad \mathbb{E}[C(w_t)] \sim \mathrm{tr}(H^{-1}G)\,t^{-1}$$

$$\mathbb{E}[(w_t^* - w^*)^2] \sim \mathrm{tr}(H^{-1}G\,H^{-1})\,t^{-1} \qquad \mathbb{E}[C(w_t^*)] \sim \mathrm{tr}(H^{-1}G)\,t^{-1}$$

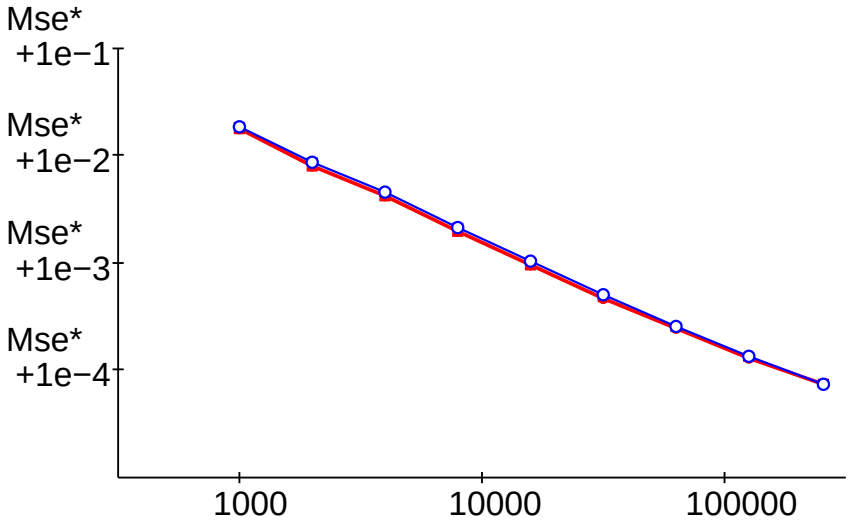


(Murata and Amari, 1998; Bottou and LeCun, 2004)

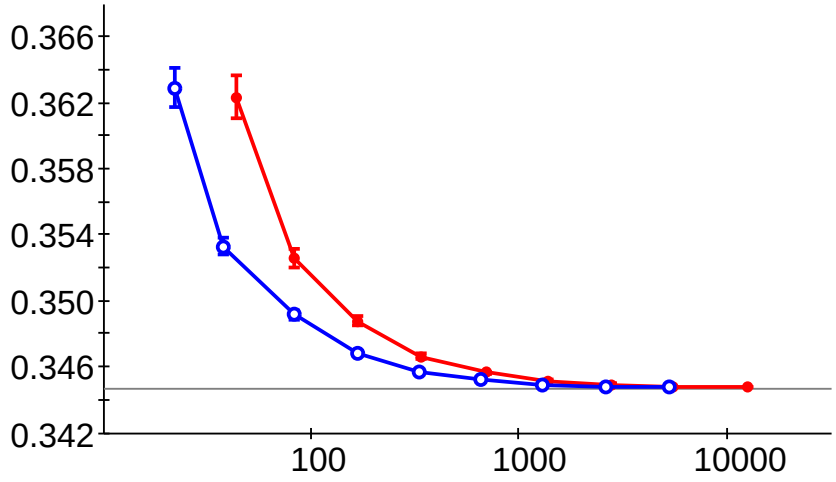
Optimal Learning in One Pass

A Single Pass of Second Order Stochastic Gradient generalizes as well as the Empirical Optimum.

Experiments on synthetic data



Number of examples



Milliseconds

Unfortunate Issue

Second order SGD

Repeat: (a) Pick random example x_t, y_t

(b) $w \leftarrow w - \frac{1}{t} H^{-1} \frac{\partial \ell}{\partial w}(x_t, w)$

Very costly in high dimension

– Estimate and store $d \times d$ matrix H^{-1}.

– Multiply the gradient for each example by this matrix H^{-1}.

Works well for low dimension problems

– But these are easy anyway...

III. Averaged Stochastic Gradient Descent

Intuition

Consider SGD with $\eta_t = Bt^{-\alpha}$.

α	Green	Var[Red]
$\alpha = 0$	Fast ($\equiv e^{-\lambda t}$)	No convergence
$0 < \alpha < 1$	Fast ($\equiv e^{-\kappa t^{1-\alpha}}$)	Slow ($\equiv t^{-\alpha}$)
$\alpha = 1$	Poly ($\equiv t^{-\lambda}$)	Poly ($\equiv t^{-1}$)
$\alpha > 1$	No convergence	Poly ($\equiv t^{-1}$)

– Use the case $\alpha < 1$.

– Further reduce the noise by averaging: $\bar{w}_t = \frac{1}{t}\sum_{i=1}^{t} w_i$.

Averaged Stochastic Gradient

Iterate

– Pick random example x_t

– SGD update $w_t = w_{t-1} - \eta_t \frac{\partial \ell}{\partial w}(x_t, w_{t-1})$

– Averaging $\bar{w}_t = \bar{w}_{t-1} - \frac{1}{t}(w_t - \bar{w}_{t-1})$

Computational cost

– Less than twice the cost of SGD per iteration.

(Ruppert, 1991; Polyak and Juditsky, 1992)

Analysis

– With $u_t = w_t - w^*$ and $\bar{u}_t = \bar{w}_t - w^*$

$$u_t = u_{t-1} - \eta_t H u_{t-1} - \eta_t \varepsilon_t \qquad \bar{u}_t = t^{-1} \sum_{i=1}^{t} u_t$$

– Recurse

$$\bar{u}_t = \left(t^{-1} \sum_{i=1}^{t} \prod_{j=1}^{i} (I - \eta_j H) \right) u_0 - t^{-1} \sum_{i=1}^{t} \left(\sum_{j=1}^{i} \prod_{k=j+1}^{i} (I - \eta_k H) \right) \eta_i \, \varepsilon_i$$

– The green term gives the mean of $\bar{w}_t - w^*$.
– The variance of the red term gives the variance of $\bar{w}_t - w^*$.
– Both can be characterized with calculus.

Result

Result

– When $\eta_t \equiv t^{-\alpha}$ with $\frac{1}{2} < \alpha < 1$,

$$\mathbb{E}[w_t - w^*] \equiv e^{-\kappa t^{1-\alpha}}$$

$$\mathrm{Var}[w_t - w^*] \equiv t^{-\alpha}$$

$$\mathbb{E}[\bar{w}_t - w^*] \equiv t^{-1}$$

$$\mathrm{Var}[\bar{w}_t - w^*] \sim H^{-1} G H^{-1} t^{-1}$$

Optimal convergence rate

– Same asymptotic convergence rate as second order SGD.
– Learning in a single pass?
– With such a low computational cost?

(Polyak and Juditsky, 1992)

One-Pass Learning with ASGD. . . ?

Too little computation?

– A linear system of d equations can be viewed as a least square problem.

– Single pass ASGD learning would mean that we can find an approximate solution in $\mathcal{O}\left(d^2\right)$ operations instead of $\mathcal{O}\left(d^3\right)$ operations.

– Meaning of "approximate".

One-Pass Learning with ASGD. . . ?

Too little computation?

– A linear system of d equations can be viewed as a least square problem.

– Single pass ASGD learning would mean that we can find an approximate solution in $\mathcal{O}\left(d^2\right)$ operations instead of $\mathcal{O}\left(d^3\right)$ operations.

– Meaning of "approximate".

The dangers of asymptotic analysis

– How many iterations before reaching the asymptotic regime?

– When the data set is limited, this can take more than one epoch!

Further analysis can help.

– More precise analyzes give guidance in selecting η_t and accelerating the onset of the asymptotic regime.

  – (Xu, 2010) extends the asymptotic analysis with more high order terms.

  – (Bach and Moulines, 2010) give non-asymptotic bounds.

The initialization problem

More dangers of asymptotic analysis

– The green term decreases in t^{-1}, i.e. $\left(\mathbb{E}[\bar{w}_t - w^*]\right)^2 \equiv t^{-2}$.
– But averaging makes the constant less favorable than SGD.
– It is useful to first optimize with plain SGD.
– Start averaging at the second epoch?

Bag of Ideas

– Why starting with plain SGD?
– Initialize with batch optimization on a training set subsample?
– Smooth transition from batch to stochastic:

$$w_t = \arg\max_w \; C_{\text{minibatch}}(w) + (1/2\eta_t)\,(w - w_{t-1})^2$$

– Same asymptotic guarantees as SGD or (with averaging) ASGD.
– How to choose the mini-batch size?
– What stopping criterion for the mini-batch optimization.

Experiments

Problems

– L2-Regularized LogLoss SVM.
– L2-Regularized CRF.

$$\ell(x, y, w) = \frac{\lambda}{2} w^2 + \log P_w(y \mid x)$$

Algorithms

– SGD: $\eta_t = \eta_0(1 + \eta_0 \lambda t)^{-1}$
– ASGD: $\eta_t = \eta_0(1 + \eta_0 \lambda t)^{-0.75}$ (Xu, 2010)
– Initial gain η_0 determined by automated trial-and-errors.
– Averaging starts after the first epoch!

Datasets

– Alpha, Webspam, RCV1, CONLL2000.

`http://leon.bottou.org/projects/sgd` (sgd-2.0)

ASGD For Sparse Training Data

Only the green term below is sparse

$$w_t = (1 - \eta_t \lambda) w_{t-1} - \eta_t \nabla P_{w_{t-1}}(Y = y_t | X = x_t)$$
$$\bar{w}_t = \bar{w}_{t-1} - \mu_t (w_t - \bar{w}_{t-1})$$

The well known SGD solution (Shalev-Shwartz et al., 2007)

– Define $w_t = W_t a_t^{-1}$ and rewrite the SGD update as

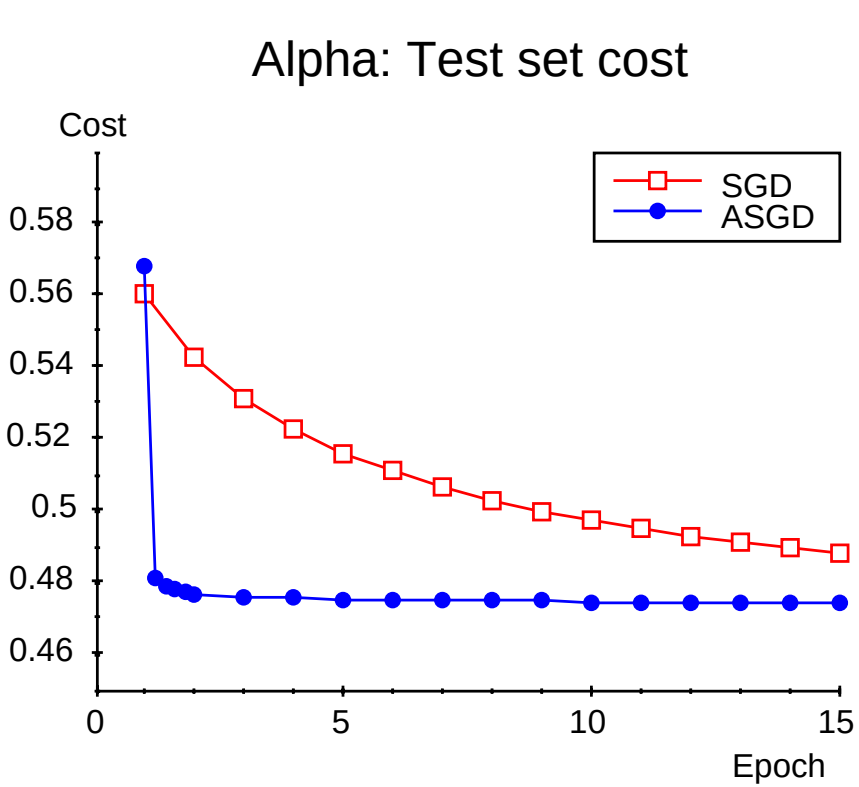
$$a_t = a_{t-1} (1 - \eta_t \lambda)^{-1}$$
$$W_t = W_{t-1} - \eta_t a_t \nabla P_{w_{t-1}}(Y = y_t | X = x_t)$$

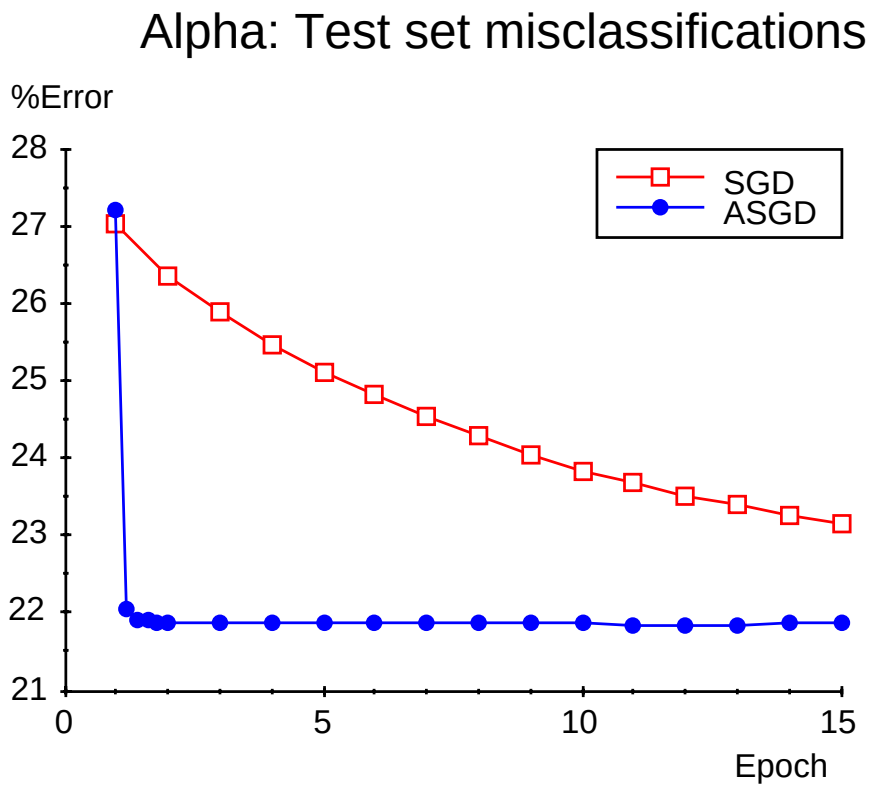
The wonderful ASGD solution (Xu et al., 2010)

– Define $\bar{w}_t = (V_t + b_t W_t) c_t^{-1}$ and continue with

$$V_t = V_{t-1} + \eta_t a_t b_{t-1} \nabla P_{w_{t-1}}(Y = y_t | X = x_t)$$
$$c_t = c_t (1 - \mu_t)^{-1}$$
$$b_t = b_t + \mu_t c_t a_t^{-1}$$

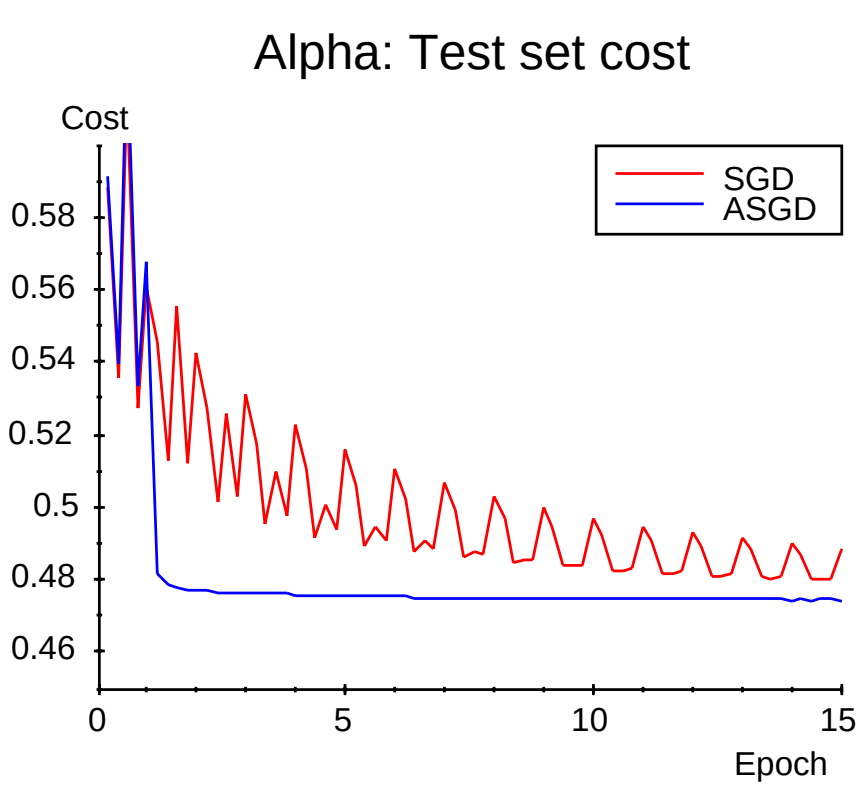
Alpha

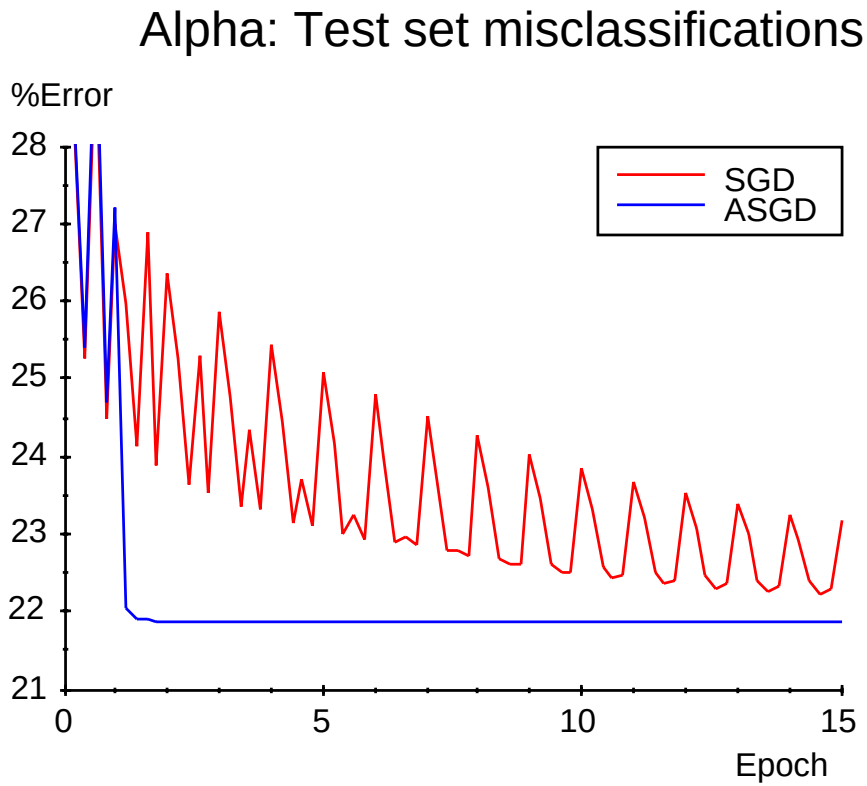




L2-Regularized LogLoss SVM.
ALPHA dataset from the Pascal Large-Scale Challenge.
Train=250K. Test=250K. Dim=500. $\lambda = 10^{-6}$.

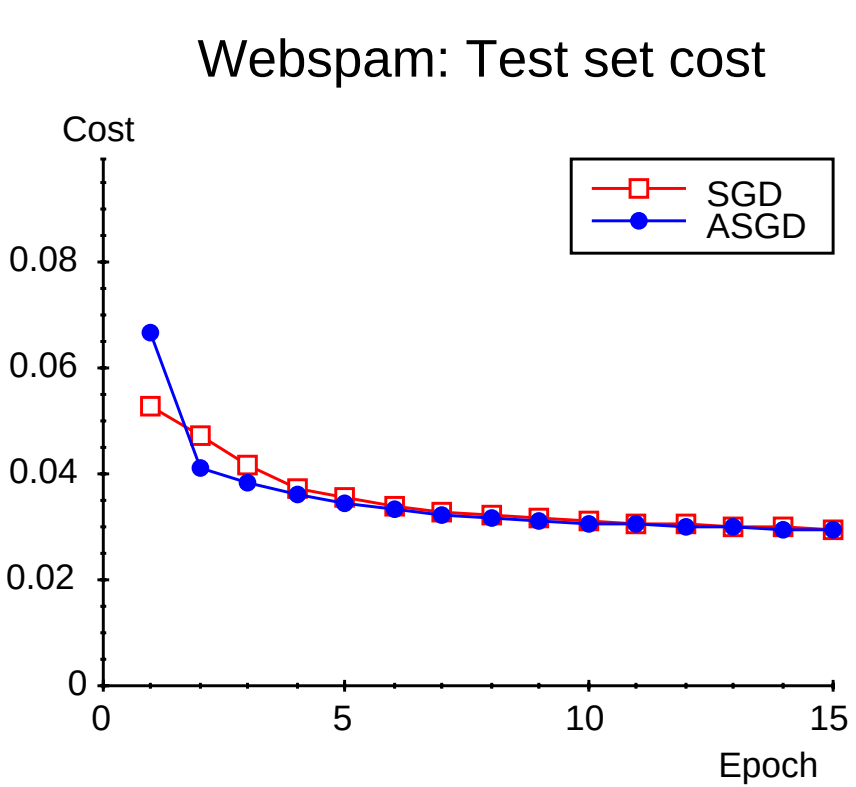
Alpha

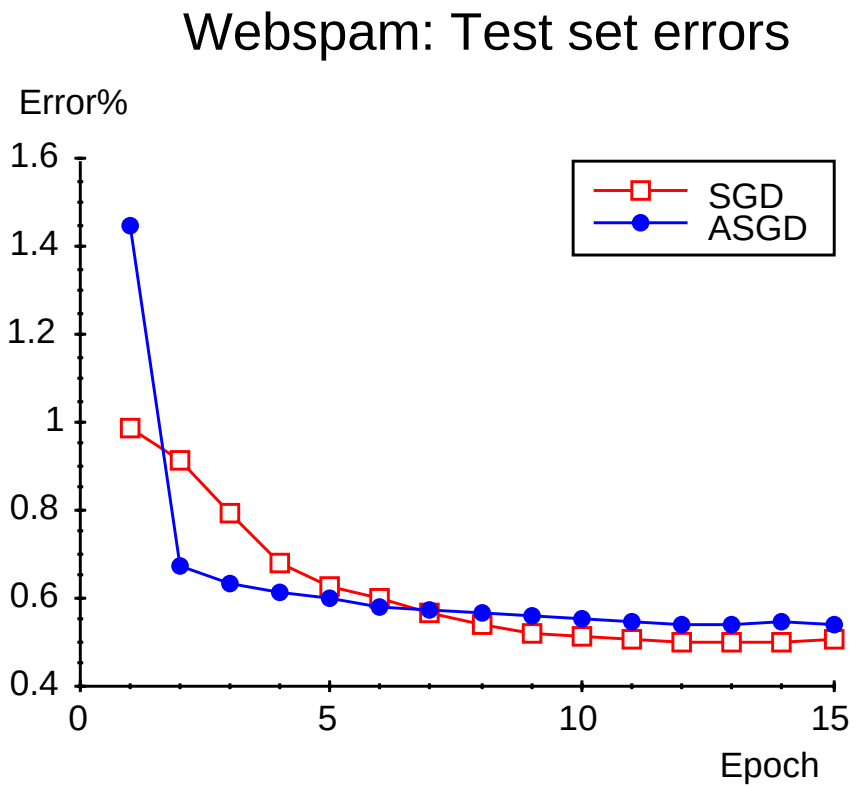




L2-Regularized LogLoss SVM.
ALPHA dataset from the Pascal Large-Scale Challenge.
Train=250K. Test=250K. Dim=500. $\lambda = 10^{-6}$.

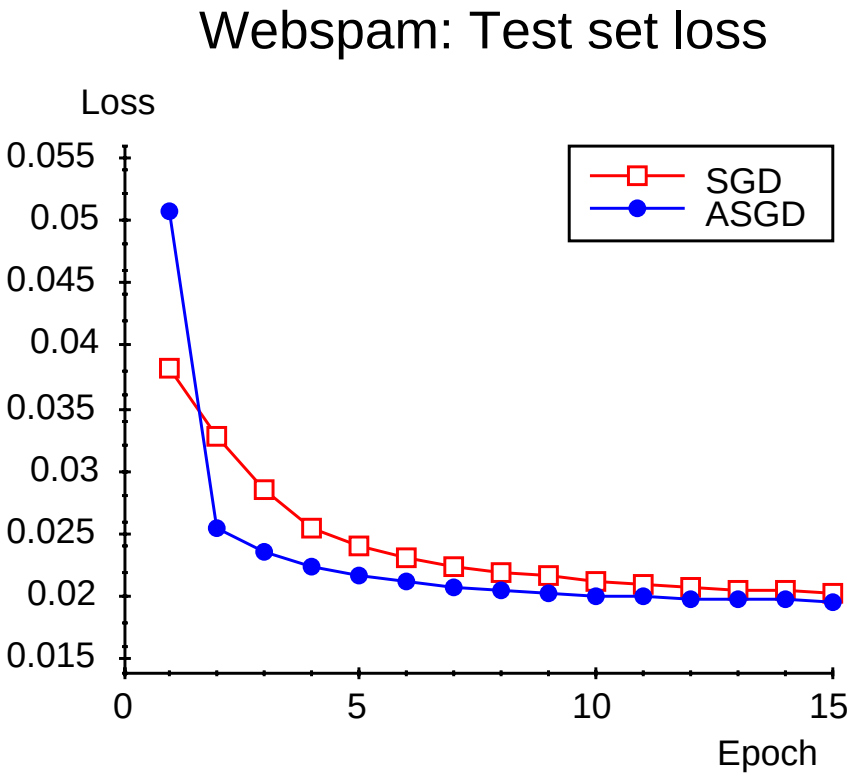
Webspam

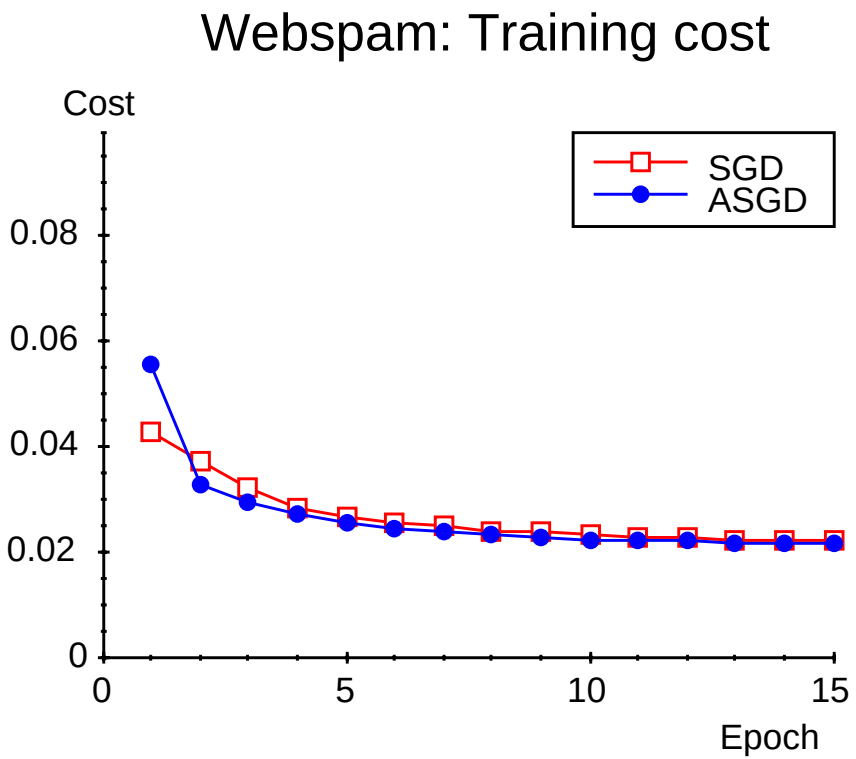




L2-Regularized LogLoss SVM.
WEBSPAM dataset from the Pascal Large-Scale Challenge.
Train=250K. Test=100K. Dim=16M. $\lambda = 10^{-7}$.

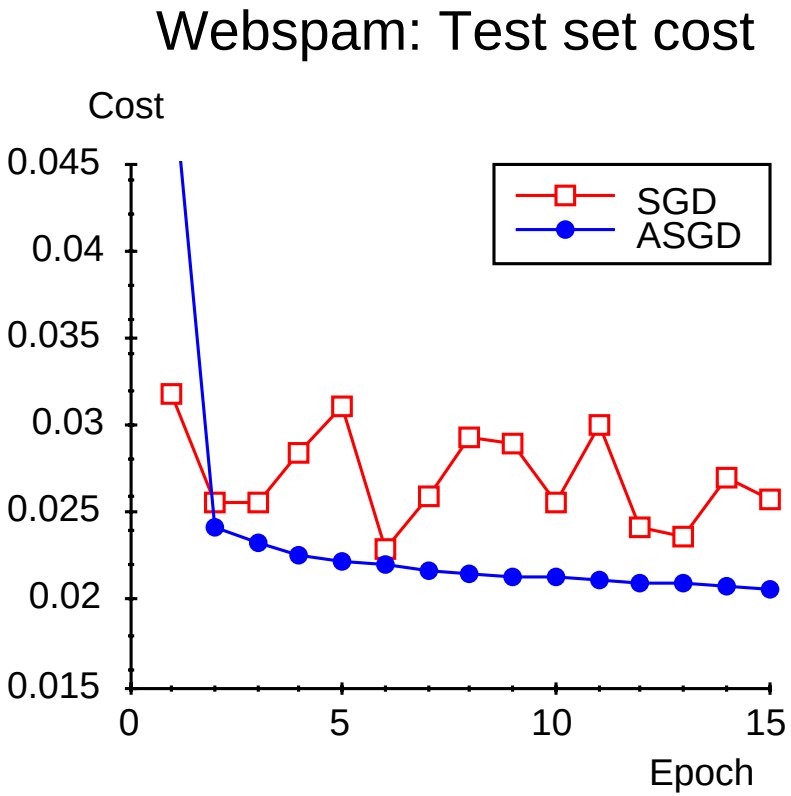
Webspam

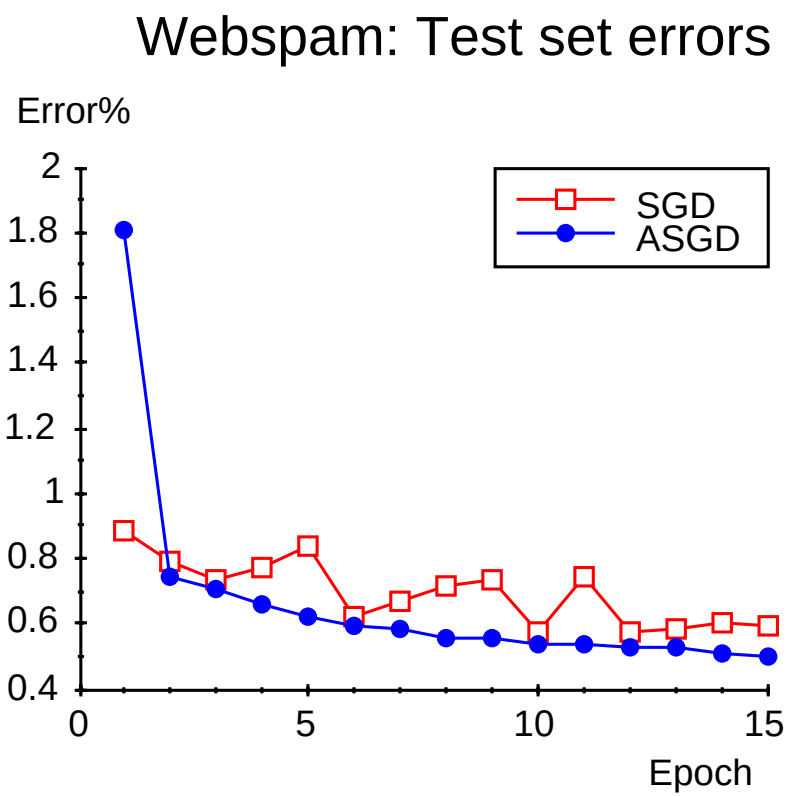




L2-Regularized LogLoss SVM.
WEBSPAM dataset from the Pascal Large-Scale Challenge.
Train=250K. Test=100K. Dim=16M. $\lambda = 10^{-7}$

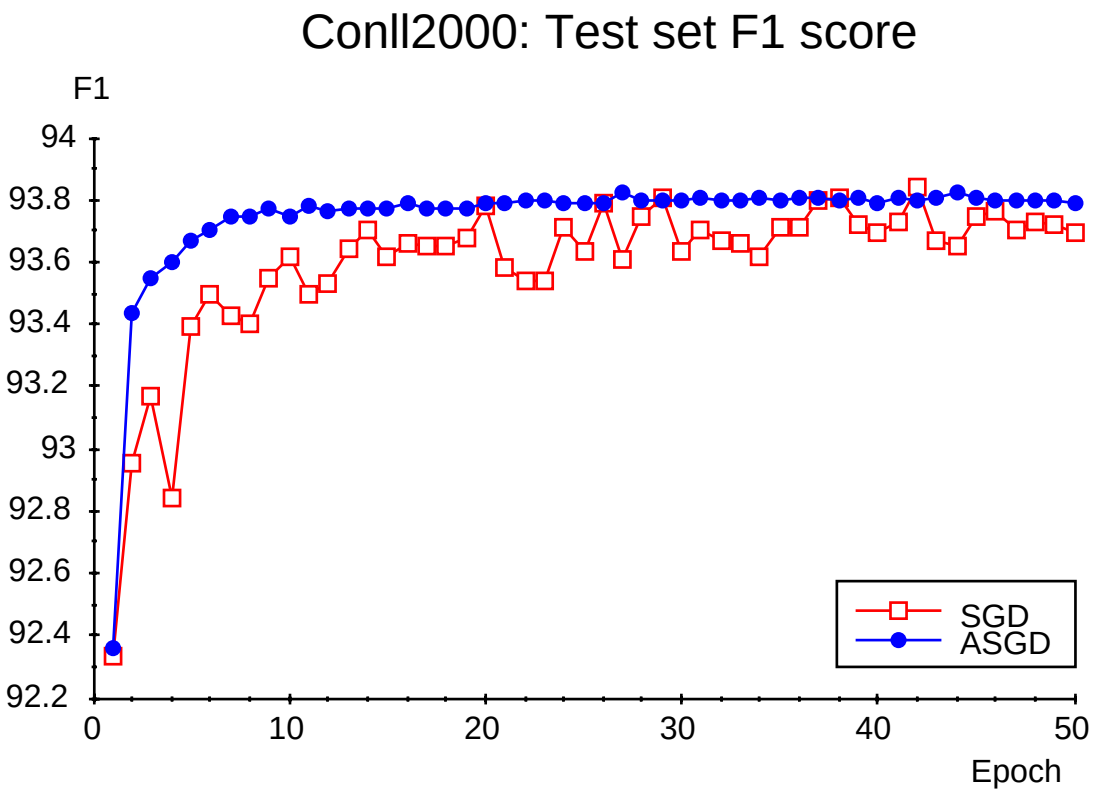
Webspam – Hinge Loss





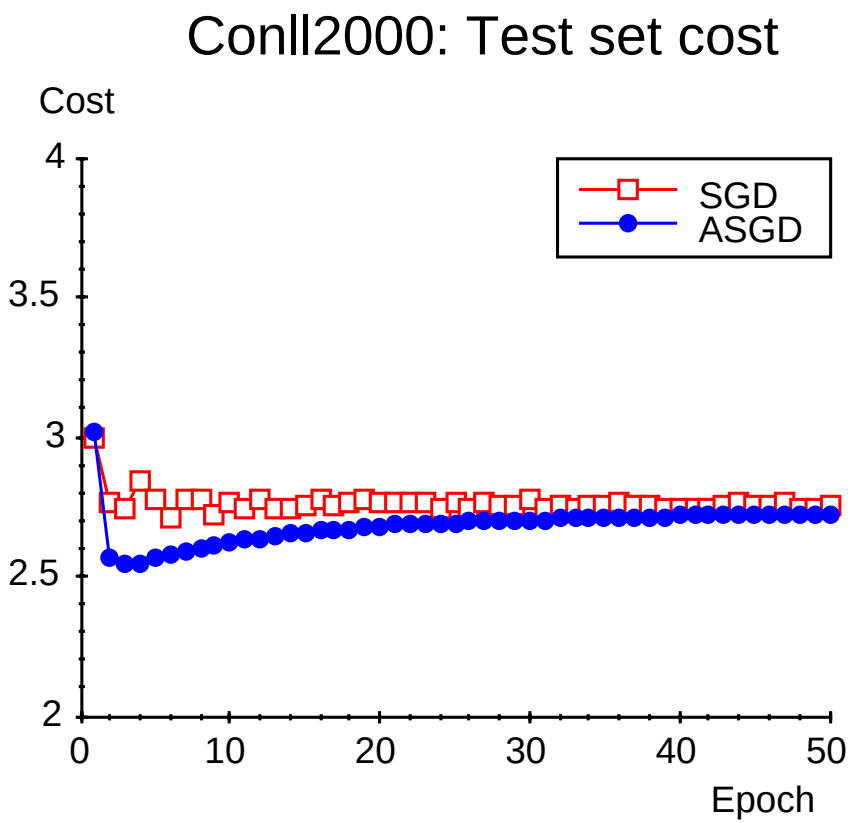
L2-Regularized HingeLoss SVM.
WEBSPAM dataset from the Pascal Large-Scale Challenge.
Train=250K. Test=100K. Dim=16M. $\lambda = 10^{-7}$.

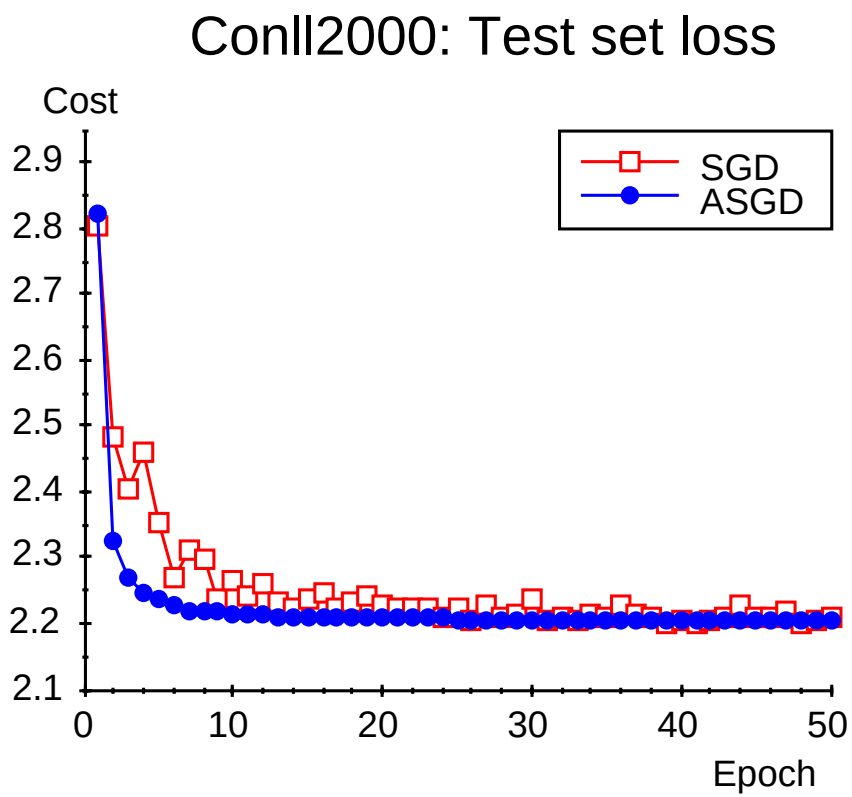
CONLL2000 Chunking



L2-Regularized CRF. CONLL2000 Chunking dataset.
9k+2k sentences. 100k+24k chunks. Dim=16M. C=1

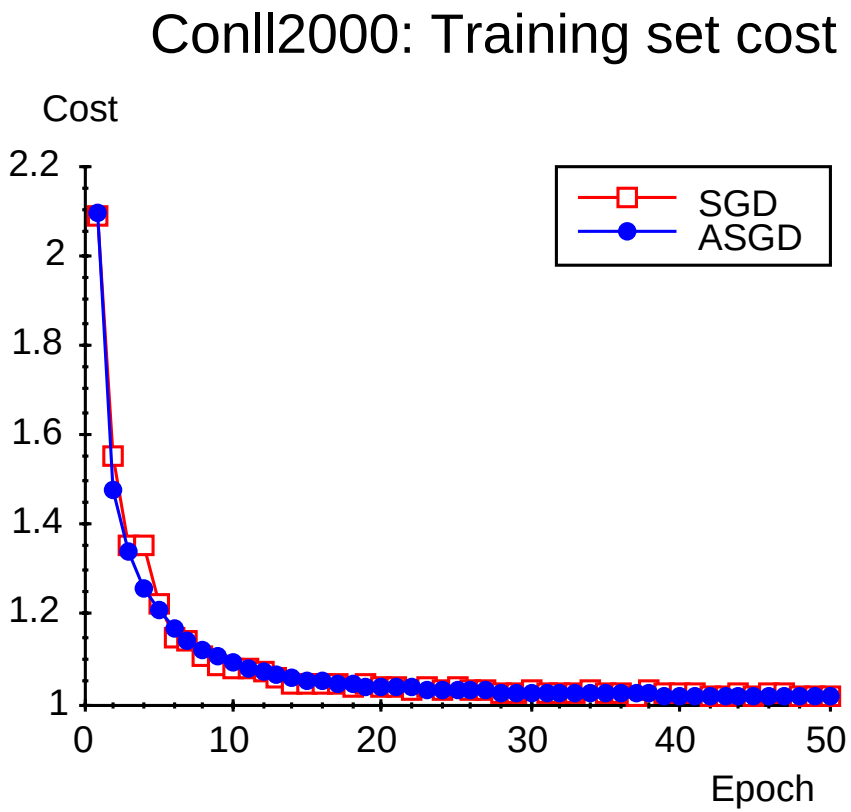
CONLL2000 Chunking

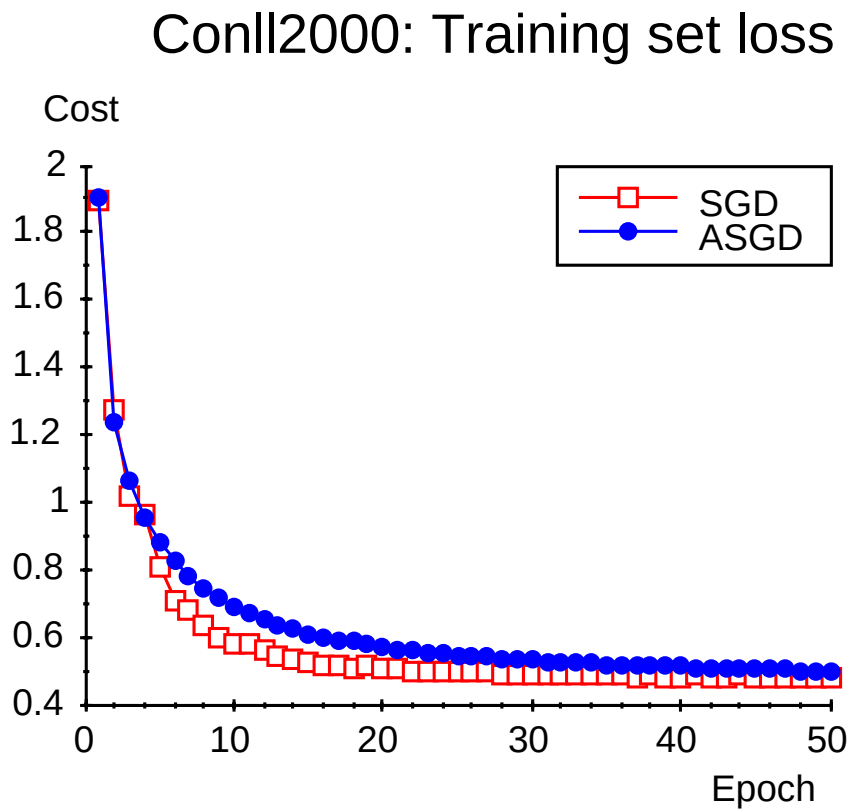




L2-Regularized CRF. CONLL2000 Chunking dataset.
9k+2k sentences. 100k+24k chunks. Dim=16M. C=1

CONLL2000 Chunking





L2-Regularized CRF. CONLL2000 Chunking dataset.
9k+2k sentences. 100k+24k chunks. Dim=16M. C=1

IV. Conclusions

Conclusions

– Many old papers on "asymptotically efficient" stochastic algorithms.
– Direct interpretation without the "unbiased estimator" assumption.
– Second order SGD works but is too costly.
– ASGD sometimes achieves one-pass learning.
– ASGD handles high dimensional problems in fewer epochs.
– Robust one-pass learning seems within reach.