Insensitive Stochastic Gradient Twin Support Vector Machines for Large Scale Problems

Zhen Wang

Inner Mongolia University

imubailan@163.com


February 22, 2018
Outline

1. Introduction
2. Review
3. SGTSVM
4. Experiments
5. Conclusions
1. Introduction

Background

- Standard SVM and TWSVM are powerful classification tools, but they cannot deal with large scale problem because of much more learning time.
- Many approaches were proposed on SVM to improve their learning speed (e.g., SMO, SOR, and DCD for dual problem, Smooth technique and SGD for primal problem).
- These solvers were extended to TWSVM, but they cannot handle large scale problems efficiently because of the inverse of matrix (dual problem) or Hessian matrix (primal problem).
- SGD solver has been applied to SVM (called PEGASOS) with an amazing learning speed, and it has not been applied to TWSVM.
Our work

(1) In this paper, we design an insensitive stochastic gradient descent solver for TWSVM (called SGTSVM);
(2) We prove that the proposed SGTSVM is convergent, instead of almost sure convergence in PEGASOS;
(3) For the uniformly sampling, it is proved that the original objective of the solution to SGTSVM is bounded by the optimum of TWSVM, which indicates the solution to SGTSVM is an approximation of the optimal solution to TWSVM, while PEGASOS only has an opportunity to obtain an approximation of the optimal solution to SVM in theory;
(4) The nonlinear case of SGTSVM is obtained directly based on its original problem;
(5) Each iteration of SGTSVM includes no more than $8n + 4$ multiplications without additional storage, so it is the fastest one among all the existing TWSVM solvers.
(6) Experiments confirm the advantages of our SGTSVM.
A toy example of PEGASOS vs SGTSVM on stochastic sampling

- SVM relies on support vectors (SVs), which may cause problem for SGD.

Figure: PEGASOS on 10 samples from two classes. (i) Training includes all of the 10 samples with 11 iterations, and the circle sample is used twice; (ii) Training includes all of the 10 samples with 28 iterations, and the circle sample is used once; (iii) Training includes 9 samples with 27 iterations, where the circle sample is excluded.
• TWSVM does not rely on any special samples, which indicates it is more suitable for SGD.

![Graphs](image)

Figure: SGTSVM on 10 samples from two classes. (i) Training includes all of the 10 samples with 7 iterations, and the circle sample is used twice; (ii) Training includes all of the 10 samples with 16 iterations, and the circle sample is used once; (iii) Training includes 9 samples with 15 iterations, where the circle sample is excluded.
Notation

In this paper, we consider $m$ data samples $\{x_1, x_2, \ldots, x_m\}$ in the $n$-dimensional real vector space $\mathbb{R}^n$. Assuming these $m$ samples with their corresponding labels $y \in \{+1, -1\}$, and are represented by the matrix $X = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^{n \times m}$. We further organize the samples from $X$ with label $+1$ into the matrix $X_1 \in \mathbb{R}^{n \times m_1}$ and those with the rest labels into the matrix $X_2 \in \mathbb{R}^{n \times m_2}$.
2. Review

SVM SVM searches for a separating hyperplane

\[ w^\top x + b = 0, \] \hspace{1cm} (1)

where \( w \in \mathbb{R}^n \) and \( b \in \mathbb{R} \), by considering following problem

\[
\begin{align*}
\min_{w,b,\xi} & \quad \frac{1}{2} ||w||^2 + \frac{c}{m} e^\top \xi \\
\text{s.t.} & \quad D(X^\top w + b) \geq e - \xi, \quad \xi \geq 0,
\end{align*}
\]

\hspace{1cm} (2)

where \( || \cdot || \) denotes the \( L_2 \) norm, \( c > 0 \) is a parameter with some quantitative meanings, \( e \) is a vector of ones with an appropriate dimension, \( \xi \in \mathbb{R}^m \) is the slack vector, and \( D = \text{diag}(y_1, \ldots, y_m) \).

A new sample \( x \) can be predicted by

\[ y = \text{sign}(w^\top x + b). \] \hspace{1cm} (3)
\[
\min_w \frac{1}{2} \|w\|^2 + \frac{c}{m} e^\top (e - DX^\top w)_+, \tag{4}
\]

where \((\cdot)_+\) replaces negative components of a vector by zeros.

In the \(t\)th iteration \((t \geq 1)\), PEGASOS constructs a temporary function, which is defined by a random sample \(x_t \in X\) as

\[
g_t(w) = \frac{1}{2} \|w\|^2 + c(1 - y_t w^\top x_t)_+. \tag{5}
\]

Then, starting with an initial \(w_1\), PEGASOS iteratively updates \(w_{t+1} = w_t - \eta_t \nabla_{w_t} g_t(w)\) for \(t \geq 1\), where \(\eta_t = 1/t\) is the step size and \(\nabla_{w_t} g_t(w)\) is the sub-gradient of \(g_t(w)\) at \(w_t\),

\[
\nabla_{w_t} g_t(w) = w_t - cy_t x_t \text{sign}(1 - y_t w_t^\top x_t)_+. \tag{6}
\]

When some terminate conditions are satisfied, the last \(w_t\) is outputted as \(w\). And a new sample \(x\) can be predicted by

\[
y = \text{sign}(w^\top x). \tag{7}
\]
It has been proved that the average solution $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$ is bounded by the optimal solution $w^*$ to (4) with $o(1)$, and thus PEGASOS has with a probability of at least $1/2$ to find a good approximation of $w^*$. In order to extend the generalization ability of PEGASOS, the bias term $b$ in SVM can be appended to PEGASOS by replacing $g(w_t)$ of (5) with

$$g(w_t, b) = \frac{1}{2} ||w_t||^2 + C(1 - y_t(w_t^\top x_t + b))_+.$$  

(8)

However, this modification would lead to the function not to be strongly convex and thus yield a slow convergence rate [1].

TWSVM seeks a pair of nonparallel hyperplanes in $R^n$ which can be expressed as

$$w_1^\top x + b_1 = 0 \text{ and } w_2^\top x + b_2 = 0,$$

by considering following two problems

$$\begin{align*}
\min_{w_1, b_1, \xi_1} & \quad \frac{1}{2} (\|w_1\|^2 + b_1^2) + \frac{c_1}{2m_1} \|X_1^\top w_1 + b_1\|^2 + \frac{c_2}{m_2} e^\top \xi_1 \\
\text{s.t.} & \quad X_2^\top w_1 + b_1 - \xi_1 \leq -e, \quad \xi_1 \geq 0,
\end{align*}$$

and

$$\begin{align*}
\min_{w_2, b_2, \xi_2} & \quad \frac{1}{2} (\|w_2\|^2 + b_2^2) + \frac{c_3}{2m_2} \|X_2^\top w_2 + b_2\|^2 + \frac{c_4}{m_1} e^\top \xi_2 \\
\text{s.t.} & \quad X_1^\top w_2 + b_2 + \xi_2 \geq e, \quad \xi_2 \geq 0,
\end{align*}$$

where $c_1, c_2, c_3, \text{ and } c_4$ are positive parameters, $\xi_1 \in R^{m_2}$ and $\xi_2 \in R^{m_1}$ are slack vectors.

A new sample $x \in R^n$ is assigned to which class depends on the distance to the two hyperplanes, i.e.,

$$y = \arg \min_i \frac{|w_i^\top x + b_i|}{\|w_i\|},$$

where $|\cdot|$ is the absolute value.
To find the pair of nonparallel hyperplanes (9), we consider the equivalent problem of TWSVM (10) and (11) as follow

\[
\min_{w_1, b_1} \frac{1}{2} (\|w_1\|^2 + b_1^2) + \frac{c_1}{2m_1} \|X_1^\top w_1 + b_1\|^2 + \frac{c_2}{m_2} e^\top (e + X_2^\top w_1 + b_1)_+,
\]

and

\[
\min_{w_2, b_2} \frac{1}{2} (\|w_2\|^2 + b_2^2) + \frac{c_3}{2m_2} \|X_2^\top w_2 + b_2\|^2 + \frac{c_4}{m_1} e^\top (e - X_1^\top w_2 - b_2)_+.
\]
In order to solve the above two problems, we construct a series of strictly convex functions $f_{1,t}(w_1, b_1)$ and $f_{2,t}(w_2, b_2)$ with $t \geq 1$ as

$$f_{1,t} = \frac{1}{2}(\|w_1\|^2 + b_1^2) + \frac{c_1}{2}\|w_1^\top x_t + b_1\|^2 + c_2(1 + w_1^\top \hat{x}_t + b_1)_+, \quad (15)$$

and

$$f_{2,t} = \frac{1}{2}(\|w_2\|^2 + b_2^2) + \frac{c_3}{2}\|w_2^\top \hat{x}_t + b_2\|^2 + c_4(1 - w_2^\top x_t - b_2)_+, \quad (16)$$

where $x_t$ and $\hat{x}_t$ are selected randomly from $X_1$ and $X_2$, respectively.
Then, the gradients of the above functions are

\[
\nabla_{w_1,t} f_{1,t} = w_1,t + c_1(w_{1,t}^T x_t + b_{1,t})x_t + c_2 \hat{x}_t \text{sign}(1 + w_{1,t}^T \hat{x}_t + b_{1,t})_+,
\]

\[
\nabla_{b_{1,t}} f_{1,t} = b_{1,t} + c_1(w_{1,t}^T x_t + b_{1,t}) + c_2 \text{sign}(1 + w_{1,t}^T \hat{x}_t + b_{1,t})_+,
\]

and

\[
\nabla_{w_2,t} f_{2,t} = w_2,t + c_3(w_{2,t}^T \hat{x}_t + b_{2,t})\hat{x}_t - c_4 x_t \text{sign}(1 - w_{2,t}^T x_t - b_{2,t})_+,
\]

\[
\nabla_{b_{2,t}} f_{2,t} = b_{2,t} + c_3(w_{2,t}^T \hat{x}_t + b_{2,t}) - c_4 \text{sign}(1 - w_{2,t}^T x_t - b_{1,t})_+,
\]

(17)
Our SGTSVM starts from the initial \((w_{1,1}, b_{1,1})\) and \((w_{2,1}, b_{2,1})\). Then, for \(t \geq 1\), the updates are given by

\[
\begin{align*}
    w_{1,t+1} &= w_{1,t} - \eta_t \nabla_{w_{1,t}} f_{1,t}, \\
    b_{1,t+1} &= b_{1,t} - \eta_t \nabla_{b_{1,t}} f_{1,t}, \\
    w_{2,t+1} &= w_{2,t} - \eta_t \nabla_{w_{2,t}} f_{2,t}, \\
    b_{2,t+1} &= b_{2,t} - \eta_t \nabla_{b_{2,t}} f_{2,t},
\end{align*}
\]

(19)

where \(\eta_t\) is the step size and typically is set to \(1/t\). If the terminated condition is satisfied, \((w_{1,t}, b_{1,t})\) is assigned to \((w_1, b_1)\), and \((w_{2,t}, b_{2,t})\) is assigned to \((w_2, b_2)\).
**Algorithm 1: SGTSVM**

**Input:** Training dataset $X_1 \in \mathbb{R}^{n \times m_1}$ as positive class, $X_2 \in \mathbb{R}^{n \times m_2}$ as negative class, positive parameters $c_1, c_2, c_3, c_4$, and a small tolerance $tol$, typically $tol = 1e^{-3}$.

**Output:** $w_1, b_1, w_2, b_2$.

1. Set $w_{1,1}, b_{1,1}, w_{2,1},$ and $b_{2,1}$ be zeros;
2. for $t = 1, \ldots$, 
   (a) Choose a pair of samples $x_t$ and $\hat{x}_t$ from $X_1$ and $X_2$ at random, respectively; 
   (b) Compute the $t$th gradients (17) to update $(w_{1,t+1}, b_{1,t+1})$, and/or (18) to update $(w_{2,t+1}, b_{2,t+1})$, by (19); 
   (c) If $\|w_{1,t+1} - w_{1,t}\| + |b_{1,t+1} - b_{1,t}| < tol$, stop updating $w_{1,t+1}$ and $b_{1,t+1}$; 
   (d) If $\|w_{2,t+1} - w_{2,t}\| + |b_{2,t+1} - b_{2,t}| < tol$, stop updating $w_{2,t+1}$ and $b_{2,t+1}$; 
   (e) If all the $w_{1,t+1}, b_{1,t+1}, w_{2,t+1},$ and $b_{2,t+1}$ are stopped updating, end this loop and goto step 3; 
3. Set $w_1 = w_{1,t+1}, b_1 = b_{1,t+1}, w_2 = w_{2,t+1},$ and $b_2 = b_{2,t+1}.$
Nonlinear SGTSVM

By the kernel trick, we consider

\[
\min_{w_1, b_1} \frac{1}{2} (\|w_1\|^2 + b_1^2) + \frac{c_1}{2m_1} \|K(X_1, X)^\top w_1 + b_1\|^2 \\
+ \frac{c_2}{m_2} e^\top (e + K(X_2, X)^\top w_1 + b_1)_+,
\]

and

\[
\min_{w_2, b_2} \frac{1}{2} (\|w_2\|^2 + b_2^2) + \frac{c_3}{2m_2} \|K(X_2, X)^\top w_2 + b_2\|^2 \\
+ \frac{c_4}{m_1} e^\top (e - K(X_1, X)^\top w_2 - b_2)_+,
\]

where \(K(\cdot, \cdot)\) is a predefined kernel function.

Let \(K_t = K(x_t, X)\) and \(\hat{K}_t = K(\hat{x}_t, X)\). Similar to linear case, we construct a series of functions with \(t \geq 1\) as

\[
h_{1,t} = \frac{1}{2} (\|w_1\|^2 + b_1^2) + \frac{c_1}{2} \|K_t^\top w_1 + b_1\|^2 + c_2 (1 + \hat{K}_t^\top w_1 + b_1)_+,
\]

and

\[
h_{2,t} = \frac{1}{2} (\|w_2\|^2 + b_2^2) + \frac{c_3}{2} \|\hat{K}_t^\top w_2 + b_2\|^2 + c_4 (1 - K_t^\top w_2 - b_2)_+.
\]
Similar to (17), (18), and (19), the sub-gradients and updates are as follow

\[ \nabla_{w_1,t} h_{1,t} = w_{1,t} + c_1(K_t^T w_{1,t} + b_{1,t})K_t + c_2 \hat{K}_t \text{sign}(1 + \hat{K}_t^T w_{1,t} + b_{1,t})_+, \]

\[ \nabla_{b_1,t} h_{1,t} = b_{1,t} + c_1(K_t^T w_{1,t} + b_{1,t}) + c_2 \text{sign}(1 + \hat{K}_t^T w_{1,t} + b_{1,t})_+, \]

(24)

\[ \nabla_{w_2,t} h_{2,t} = w_{2,t} + c_3(\hat{K}_t^T w_{2,t} + b_{2,t})\hat{K}_t - c_4 K_t \text{sign}(1 - K_t^T w_{2,t} - b_{2,t})_+, \]

\[ \nabla_{b_2,t} h_{2,t} = b_{2,t} + c_3(\hat{K}_t^T w_{2,t} + b_{2,t}) - c_4 \text{sign}(1 - K_t^T w_{2,t} - b_{1,t})_+, \]

(25)
and

\[
\begin{align*}
    w_{1,t+1} &= w_{1,t} - \nabla_{w_{1,t}} h_{1,t}/t, \\
    b_{1,t+1} &= b_{1,t} - \nabla_{b_{1,t}} h_{1,t}/t, \\
    w_{2,t+1} &= w_{2,t} - \nabla_{w_{2,t}} h_{2,t}/t, \\
    b_{2,t+1} &= b_{2,t} - \nabla_{b_{2,t}} h_{2,t}/t. \\
\end{align*}
\]

(26)

A new sample \( x \in \mathbb{R}^n \) can be predicted by

\[
y = \arg\min_i \frac{|K(x,X)^T w_i + b_i|}{\|w_i\|}.
\]

(27)
Algorithm 2: Nonlinear SGTSVM

**Input:** Training dataset $X_1 \in \mathbb{R}^{n \times m_1}$ as positive class, $X_2 \in \mathbb{R}^{n \times m_2}$ as negative class, $X$ as the whole training dataset, positive parameters $c_1$, $c_2$, $c_3$, $c_4$, kernel function $K(\cdot, \cdot)$, and $tol = 10^{-3}$.

**Output:** $w_1, b_1, w_2, b_2$.

1. Set $w_{1,1}, b_{1,1}, w_{2,1},$ and $b_{2,1}$ be zeros;
2. for $t = 1, \ldots,$
   (a) Choose a pair of samples $x_t$ and $\hat{x}_t$ respectively from $X_1$ and $X_2$ at random, and compute $K_t = K(x_t, X)$ and $\hat{K}_t = K(\hat{x}_t, X)$;
   (b) Compute the $t$th gradients (24) to update $(w_{1,t+1}, b_{1,t+1})$, and/or (25) to update $(w_{2,t+1}, b_{2,t+1})$, by (26);
   (c) If $\|w_{1,t+1} - w_{1,t}\| + |b_{1,t+1} - b_{1,t}| < tol$, stop updating $w_{1,t+1}$ and $b_{1,t+1}$;
   (d) If $\|w_{2,t+1} - w_{2,t}\| + |b_{2,t+1} - b_{2,t}| < tol$, stop updating $w_{2,t+1}$ and $b_{2,t+1}$;
   (e) If all the $w_{1,t+1}, b_{1,t+1}, w_{2,t+1},$ and $b_{2,t+1}$ are stopped updating, end this loop and goto step 3;
3. Set $w_1 = w_{1,t+1}, b_1 = b_{1,t+1}, w_2 = w_{2,t+1},$ and $b_2 = b_{2,t+1}$. 
For SGTSVM, we have the convergence theorem as follow

**Theorem**

*Our SGTSVM, including linear and nonlinear cases, is convergent.*
The average instantaneous objective of SGTSVM is bounded by the optimal solution to TWSVM, i.e.,

**Theorem**

Suppose $f_t$ ($t = 1, \ldots, T$) is the iterative objective of SGTSVM, $u_t = (w_t, b_t)$, and $u^*$ is the optimal solution corresponding to TWSVM. Then,

(i) there are two constants $G_1$ and $G_2$ (actually, they are the upper bounds of $\|u_t\|$ and $\|\nabla_t\|$, respectively) such that

$$\frac{1}{T} \sum_{t=1}^{T} f_t(u_t) \leq \frac{1}{T} \sum_{t=1}^{T} f_t(u^*) + G_2(G_1 + \|u^*\|) + \frac{1}{2T} G_2^2 (1 + \ln T);$$

(ii) given any $\varepsilon > 0$, then for sufficiently large $T$,

$$\frac{1}{T} \sum_{t=1}^{T} f_t(u_t) \leq \frac{1}{T} \sum_{t=1}^{T} f_t(u^*) + \varepsilon.$$

Based on the above theorem, we immediately obtain following two important corollaries, which indicates that our SGTSVM is a good approximation of TWSVM (i.e., for sufficiently large $T$, $f(u^*) \leq f(u_T) \leq f(u^*) + G_2^2$, where $f(\cdot)$ is the objective of TWSVM).
Corollary

Assume the conditions stated in Theorem 3.1, and \( m_1 = m_2 \), where \( m_1 \) and \( m_2 \) are the sample number of \( X_1 \) and \( X_2 \), respectively. Suppose \( T = km_1 \), where \( k > 0 \) is an integer, and each sample is selected \( k \) times at random. Then

(i) \( f(u_T) \leq f(u^*) + G_2(G_1 + \|u^*\| + G_2) + \frac{1}{2T} G_1^2 (1 + \ln T) \);

(ii) given any \( \varepsilon > 0 \), then for sufficiently large \( T \), \( f(u_T) \leq f(u^*) + G_2^2 + \varepsilon \).

Corollary

Assume the conditions stated in Corollary 3.1, but \( m_1 \neq m_2 \). Suppose \( T = kd(m_1, m_2) \), where \( k > 0 \) is an integer and \( d \) is the least common multiple of \( m_1 \) and \( m_2 \). The sample in \( X_1 \) is selected \( kd/m_1 \) times at random, and the one in \( X_2 \) is \( kd/m_2 \) times at random. Then

(i) \( f(u_T) \leq f(u^*) + G_2(G_1 + \|u^*\| + G_2) + \frac{1}{2T} G_1^2 (1 + \ln T) \);

(ii) given any \( \varepsilon > 0 \), then for sufficiently large \( T \), \( f(u_T) \leq f(u^*) + G_2^2 + \varepsilon \).

The proofs please see section 3.3 in the paper.
Experimental results

Convergence test: benchmark datasets.

Figure: Results of linear classifiers on the four datasets, where the vertical axis denotes the objectives.
Convergence test: benchmark datasets.

Figure: Results of nonlinear classifiers on the four datasets, where the vertical axis denotes the objectives.
**Performance test**: benchmark datasets.

**Table**: Mean accuracy (%) with standard deviation of TWSVM and SGTSVM by 10-fold cross validation.

<table>
<thead>
<tr>
<th>Data</th>
<th>TWSVM(^\dagger)</th>
<th>SGTSVM(^\dagger)</th>
<th>TWSVM(^#)</th>
<th>SGTSVM(^#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross Planes</td>
<td>96.05±0.70</td>
<td>97.71±0.41</td>
<td>99.01±2.24</td>
<td>98.51±2.15</td>
</tr>
<tr>
<td>Australia</td>
<td>86.87±0.38</td>
<td>87.34±0.13</td>
<td>87.10±0.43</td>
<td>85.21±0.16</td>
</tr>
<tr>
<td>Creadit</td>
<td>85.78±0.32</td>
<td>85.72±0.23</td>
<td>86.71±0.33</td>
<td>85.21±0.45</td>
</tr>
<tr>
<td>Hypothyroid</td>
<td>98.21±0.09</td>
<td>97.28±0.01</td>
<td>98.08±0.09</td>
<td>98.07±0.03</td>
</tr>
</tbody>
</table>

\(^\dagger\) linear case; \(^\#\) nonlinear case.
Sampling test: artificial datasets.

Figure: Results of PEGASOS and SGTSVM on 100 artificial datasets, where the 100 upright black solid lines are final classifiers.
Sampling test: artificial datasets.

![Graph](image)

(a) PEGASOS  (b) SGTSVM

**Figure:** Results of PEGASOS and SGTSVM on 100 artificial datasets, where the 100 upright black solid lines are final classifiers, and the samples on the dash line is invisible for sampling.
Performance test: artificial dataset.

**Figure**: Accuracies of PEGASOS and SGTSVM on a normal distribution dataset, where each method is implemented 100 times.
Stability test

Figure: Results of PEGASOS and SGTSVM on a normal distribution dataset, where each method is implemented 10 times. The horizontal axis is the iteration and the vertical one is the classification location.
Convergence vs terminate condition

Figure: Iteration and time of PEGASOS and SGTSVM on a normal distribution dataset, where each method is implemented 100 times.
Performance test: large scale datasets.

Table: The results on the large scale datasets.

<table>
<thead>
<tr>
<th>Data</th>
<th>SVM</th>
<th>PEGASOS</th>
<th>SGTSVM†</th>
<th>SGTSVM♯</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skin validation (%)</td>
<td>78.87</td>
<td>82.46</td>
<td>85.23</td>
<td>84.70</td>
</tr>
<tr>
<td>245,057×3 testing (%)</td>
<td>84.28</td>
<td>85.39</td>
<td>87.70</td>
<td>85.34</td>
</tr>
<tr>
<td>Gashome validation (%)</td>
<td>49.11</td>
<td>70.09</td>
<td>67.50</td>
<td>74.49</td>
</tr>
<tr>
<td>919,438×10 testing (%)</td>
<td>82.57</td>
<td>72.85</td>
<td>76.09</td>
<td>89.13</td>
</tr>
<tr>
<td>Susy validation (%)</td>
<td>78.41</td>
<td>54.11</td>
<td>76.14</td>
<td>69.90</td>
</tr>
<tr>
<td>5,000,000×18 testing (%)</td>
<td>78.52</td>
<td>56.44</td>
<td>75.09</td>
<td>68.61</td>
</tr>
<tr>
<td>Kddcup validation (%)</td>
<td></td>
<td>96.39</td>
<td>95.24</td>
<td>93.19</td>
</tr>
<tr>
<td>4,898,432×41 testing (%)</td>
<td></td>
<td>96.42</td>
<td>97.45</td>
<td>99.20</td>
</tr>
<tr>
<td>Gas validation (%)</td>
<td></td>
<td>69.77</td>
<td>89.73</td>
<td>92.60</td>
</tr>
<tr>
<td>8,386,764×16 testing (%)</td>
<td></td>
<td>50.54</td>
<td>92.45</td>
<td>92.86</td>
</tr>
<tr>
<td>Hepmass validation (%)</td>
<td></td>
<td>80.63</td>
<td>80.80</td>
<td>82.18</td>
</tr>
<tr>
<td>10,500,000×28 testing (%)</td>
<td></td>
<td>80.84</td>
<td>81.10</td>
<td>79.59</td>
</tr>
</tbody>
</table>

† linear case; ♯ nonlinear case; * out of memory.
Learning time test: large scale datasets.

Figure: Learning time of SGTSVM, PEGASOS, and Liblinear on the large scale datasets with the optimal parameters.
Conclusions

- The proposed SGTSVM is more stable on stochastic sampling than PEGASOS.
- SGTSVM is convergent and is an approximation of TWSVM with uniform sampling in theory.
- Experiments show SGTSVM performs better than Liblinear and PEGASOS on the large scale datasets.
- For practical convenience, this slide and the corresponding SGTSVM codes (including Matlab and C language) are uploaded on http://www.optimal-group.org/Resources/Code/SGTSVM.html.

Zhen Wang (IMU)
The end
Thanks