Pegasos: Primal Estimated sub-Gradient Solver for SVM

Zhimo Shen

University of California

shenzhimo@hotmail.com

February 25, 2014
Definition (SVM)

Given a training set $B = \{(x_i, y_i)\}_{i=1}^{m}$, where $x_i \in \mathbb{R}$ and $y_i \in \{+1, -1\}$, we would like to find the minimizer of the problem

$$f(w) = \min_w \frac{\sigma}{2} \|w\|^2 + \frac{1}{m} \sum_{(x, y) \in B} l(w; (x, y))$$

where

$$l(w; (x, y)) = \max\{0, 1 - y \langle w, x \rangle\}$$
On iteration $t$ of the algorithm, we first choose a set $A_t \subseteq B$ of size $k$. Then, we replace the objective with an approximate objective function $f(w; A_t) = \min_w \frac{\sigma}{2} \|w\|^2 + \frac{1}{k} \sum_{(x,y) \in A_t} l(w; (x, y))$.

Definition (Gradient)

$$\nabla_t = \sigma w_t - \frac{1}{|A_t|} \sum_{(x,y) \in A_t^+} y x$$
Pseudo code

INPUT: \( B, \sigma, T, k \)

INITIALIZE: Choose \( w_1 \) s.t. \( \|w_1\| \leq \frac{1}{\sqrt{\sigma}} \)

FOR \( t=1,2,...T \)
- Choose \( A_t \subseteq B \), where \( |A_t| = k \)
- Set \( A_t^+ = \{ (x, y) \in A_t : y \langle w_t, x \rangle < 1 \} \)
- Set \( \eta_t = \frac{1}{\sigma t} \)
- Set \( w_{t+\frac{1}{2}} = w_t - \eta_t \nabla t \)
- Set \( w_{t+1} = \min\{1, \frac{1/\sqrt{\sigma}}{\|w_{t+\frac{1}{2}}\|} \} w_{t+\frac{1}{2}} \)

OUTPUT: \( w_{T+1} \)
Basic definition

**Definition (sub-gradient)**

A vector $\lambda$ is a sub-gradient of a function $f$ at $v$ if

$$\forall u \in S, f(u) - f(v) \geq \langle u - v, \lambda \rangle$$

The differential set of $f$ at $v$, denoted $\partial f(v)$, is the set of all sub-gradients of $f$ at $v$.

**Definition (convex)**

A function $f$ is convex iff $\partial f(v)$ is non-empty for all $v \in S$. If $f$ is convex and differentiable at $v$ then $\partial f(v)$ consists of a single vector which amounts to $\nabla f(v)$

As a consequence we obtain that a differential function $f$ is convex iff for all $v, u \in S$ we have that

$$\forall u \in S, f(u) - f(v) - \langle u - v, \nabla f(v) \rangle \geq 0$$
Basic Definition

Definition (Bregman divergence)

\[ B_f(u||v) = f(u) - f(v) - \langle u - v, \nabla f(v) \rangle \]

if \( f(v) = \frac{1}{2} \| v \|^2 \), then \( B_f(u||v) = \frac{1}{2} \| u - v \|^2 \)

Definition (Fenchel conjugate)

\[ f^*(\theta) = \sup_{w \in S} (\langle w, \theta \rangle - f(w)) \]

if \( f(w) = \frac{1}{2} \| w \|^2 \), then

\[ f^*(\theta) = \max_{w \in S} \langle w, \theta \rangle - f(w) = \frac{1}{2} \| \theta \|^2 - \min_{w \in S} \frac{1}{2} \| w - \theta \|^2 \]

\[ \nabla f^*(\theta) = \arg\max_{w \in S} \langle w, \theta \rangle - f(w) = \arg\min_{w \in S} \| w - \theta \|^2 \]

Definition (strong convex)

A closed and convex function \( f \) is \( \sigma \)-strongly convex over \( S \) with respect to a convex and differentiable function \( f \) if

\[ \forall u, v \in S, \forall \lambda \in \partial g(v), g(u) - g(v) - \langle u - v, \lambda \rangle \geq \sigma B_f(u||v) \]
Lemma (1)

Assume that $f$ is a differentiable and convex function and let $g = \sigma f + h$ where $h$ is also a convex function. Then $g$ is $\sigma$-strongly convex w.r.t $f$.

proof: Let $v, u \in S$ and choose a vector $\lambda \in \partial g(v)$. Since $\partial g(v) = \partial h(v) + \sigma \partial f(v)$, we have that there exists $\lambda_1 \in \partial h(v)$ s.t. $\lambda = \lambda_1 + \sigma \nabla f(v)$. Thus

$$g(u) - g(v) - \langle u - v, \lambda \rangle = \sigma B_f(u\|v) + h(u) - h(v) - \langle u - v, \lambda_1 \rangle \geq \sigma B_f(u\|v)$$
Lemma (2)

Let $f(w) = \frac{1}{2} \|w\|^2$ over $S$, we can get that
\[ \forall \theta \in \mathbb{R}^n, \forall u \in S, \langle u - v, \theta - \nabla f(v) \rangle \leq 0 \]
where $v = \nabla f^*(\theta)$

Proof: Let $P(w) = \langle w, \theta \rangle - f(w)$
By the definition of $v$, we can easily get that
\[ \forall u, P(u) - P(v) \leq 0 \]
and $P(u) - P(v) \geq \langle u - v, \nabla P(v) \rangle$
so $\langle u - v, \nabla P(v) \rangle \leq 0$
which concludes our proof since $\nabla P(v) = \theta - \nabla f(v)$
Lemma (3)

Let $f(w) = \frac{1}{2} \|w\|^2$. $\sigma > 0$ is a scalar. $g_1, g_2, \ldots, g_T$ to be a sequence of $\sigma$-strongly convex functions w.r.t over $S$. $w_1, w_2, \ldots, w_T$ to be a sequence of vector that $w_1 \in S$ and $w_{t+1} = \nabla f^*(w_t - \eta_t \lambda_t)$ where $\eta_t = 1/(\sigma t)$ and $\lambda_t \in \partial g_t(w_t)$. we can get

$$\forall u \in S, \langle w_t - u, \lambda_t \rangle \leq \frac{B_f(u\|w_t) - B_f(u\|w_{t+1})}{\eta_t} + \eta_t \frac{\|\lambda_t\|^2}{2}$$

proof: donate $\Delta_t = B_f(u\|w_t) - B_f(u\|w_{t+1})$.

$$\Delta_t = \langle u - w_{t+1}, \nabla f(w_{t+1}) - \nabla f(w_t) \rangle + B_f(w_{t+1}\|w_t)$$

$$= \langle u - w_{t+1}, w_{t+1} - w_t \rangle + \frac{1}{2} \|w_{t+1} - w_t\|^2$$

we denote by $\theta_t$ the term $w_t - \eta_t \lambda_t$, so $w_{t+1} = \nabla f^*(\theta_t)$
by lemma 2 we can get:

\[ 0 \geq \langle u - w_{t+1}, \theta_t - \nabla f(w_{t+1}) \rangle \]
\[ = \langle u - w_{t+1}, w_t - \eta_t \lambda_t - w_{t+1} \rangle \]

so

\[ \langle u - w_{t+1}, w_t - w_{t+1} \rangle \geq \eta_t \langle w_{t+1} - u, \lambda_t \rangle \]

by combining them we can get

\[ \Delta_t \geq \eta_t \langle w_{t+1} - u, \lambda_t \rangle + \frac{1}{2} \| w_{t+1} - w_t \|^2 \]
\[ = \eta_t \langle w_t - u, \lambda_t \rangle - \langle w_{t+1} - w_t, \eta_t \lambda_t \rangle + \frac{1}{2} \| w_{t+1} - w_t \|^2 \]
\[ = \eta_t \langle w_t - u, \lambda_t \rangle - \frac{1}{2} \| w_{t+1} - w_t \|^2 - \frac{1}{2} \| \eta_t \lambda_t \|^2 + \frac{1}{2} \| w_{t+1} - w_t \|^2 \]
\[ = \eta_t \langle w_t - u, \lambda_t \rangle - \frac{\eta_t^2}{2} \| \lambda_t \|^2 \]
Lemma (4)

Let $G$ be a scalar such that $\|\lambda_t\| \leq G$ for all $t$. Then the following bound holds for all $T \geq 1$

$$\sum_{t=1}^{T} g_t(w_t) - \sum_{t=1}^{T} g_t(u) \leq \frac{G^2}{2\sigma} (1 + \log(T))$$

Proof:

$$\langle w_t - u, \lambda_t \rangle \geq g_t(w_t) - g_t(u) + \sigma B_f(u\|w_t)$$

Combining with lemma 3 and using $\|\lambda_t\| \leq G$ we get that

$$g_t(w_t) - g_t(u) \leq \left( \frac{1}{\eta_t} - \sigma \right) B_f(u\|w_t) - \frac{1}{\eta_t} B_f(u\|w_{t+1}) + \frac{\eta_t G^2}{2}$$

Summing over $t$ we obtain

$$\sum_{t=1}^{T} (g_t(w_t) - g_t(u)) \leq \left( \frac{1}{\eta_1} - \sigma \right) B_f(u\|w_1) - \frac{1}{\eta_T} B_f(u\|w_{T+1}) +$$

$$\sum_{t=2}^{T} B_f(u\|w_t)\left( \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \sigma \right) + \frac{G^2}{2} \sum_{t=1}^{T} \eta_t$$

Plugging the value of $\eta_t$ we obtain first and third summands of right-hand side vanish and second summand is negative. We therefore get

$$\sum_{t=1}^{T} (g_t(w_t) - g_t(u)) \leq \frac{G^2}{2} \sum_{t=1}^{T} \eta_t \leq \frac{G^2}{2\sigma} (1 + \log(T))$$
Lemma (5)

The norm of optimal solution of optimization problem of SVM is bounded by $1/\sqrt{\sigma}$

proof: Let us denote the optimal solution by $w^*$. The Lagrange dual problem of the optimization problem is

$$\max_{\alpha \in [0, 1/m]^m} \sum_{i=1}^m \alpha_i - \frac{1}{2\sigma} \| \sum_{i=1}^m \alpha_i y_i x_i \|^2$$

denote $\alpha^*$ be an optimal solution of the dual problem. we get

$$\frac{\sigma}{2} \| w^* \|^2 + \frac{1}{m} \sum_{(x,y) \in B} \max\{0, 1 - y \langle w^*, x \rangle\} =$$

$$\sum_{i=1}^m \alpha_i^* - \frac{1}{2\sigma} \| \sum_{i=1}^m \alpha_i^* y_i x_i \|^2$$

In addition, at the optimum we have that $w^* = \frac{1}{\sigma} \sum_{i=1}^m \alpha_i^* y_i x_i$

Plugging this and rearranging terms

$$\sigma \| w^* \|^2 = \sum_{i=1}^m \alpha_i^* - \max\{0, 1 - y \langle w^*, x \rangle\} \leq 1$$
Theorem

**Theorem (1)**

In the pegasos algorithm. Let \( S = \{ w : \| w \| \leq 1/\sqrt{\sigma} \} \). Assume that for all \((x, y) \in S\) the norm of \( x \) is at most \( R \). Denote \( w^* = \arg\min_{w \in S} \) and let \( c = (\sqrt{\sigma} + R)^2 \). Then for \( T \geq 3 \),

\[
\frac{1}{T} \sum_{t=1}^{T} f(w_t; A_t) \leq \frac{1}{T} \sum_{t=1}^{T} f(w^*; A_t) + \frac{c \ln(T)}{\sigma T}
\]

**proof:** We use shorthand \( f_t(w) = f(w; A_t) \).

By lemma 5 we know that the \( \min_w \frac{\sigma}{2} \| w \|^2 + \frac{1}{m} \sum_{(x, y) \in B} l(w; (x, y)) = \min_{w \in S} \frac{\sigma}{2} \| w \|^2 + \frac{1}{m} \sum_{(x, y) \in B} l(w; (x, y)) \)

Because of \( \frac{\sigma}{2} \| w \|^2 \) is a \( \sigma \)-strongly convex function w.r.t to \( \frac{1}{2} \| w \|^2 \) and the average hinge-loss function is convex. So by Lemma 1 we can get to know that \( f_t \) is a \( \sigma \)-strongly convex function w.r.t to \( \frac{1}{2} \| w \|^2 \)

The projection step is to do the \( \nabla f^* \)
By the facts that $\|w_t\| \leq 1/\sqrt{\sigma}$ and that $\|x\| \leq R$ we can get that
$\|\nabla_t\| \leq \sigma \|w_t\| + \|x\| \leq \sqrt{\sigma} + R$
In condition $T \geq 3, \frac{1+\ln(T)}{2} \leq \ln(T)$
Now we can use the Lemma 4 and we can get our conclusion:
$\frac{1}{T} \sum_{t=1}^{T} f(w_t; A_t) \leq \frac{1}{T} \sum_{t=1}^{T} f(w^*; A_t) + \frac{c \ln(T)}{\sigma T}$
Corollary

Assume the conditions stated in Thm. 1 and that $A_t = B$ for all $t$. Let $\overline{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$. Then,

$$f(\overline{w}) \leq f(w^*) + \frac{c \ln(T)}{\sigma T}$$

Note that the convexity of $f$ implies that

$$f(\overline{w}) \leq \frac{1}{T} \sum_{t=1}^{T} f(w_t)$$

Based on the above corollary, the number of iterations required for achieving a solution of accuracy $\epsilon$ is $O(c/(\sigma \epsilon))$ and the complexity of single iteration is $O(md)$.
Theorem (2)

Assume that the conditions stated in Thm. 1 hold for all $t, A_t$ is chosen i.i.d from $B$. Let $r$ be an integer picked uniformly at random from 1 to $T$. Then,

$$\mathbb{E}_{A_1, A_2, ..., A_T} \mathbb{E}_r[f(w_r)] \leq f(w^*) + \frac{c \ln(T)}{\sigma T}$$

**proof:** We denote by $A^i$ the sequence of sets $(A_i, ..., A_j)$. From Thm. 1, we obtain

$$\mathbb{E}_{A^i} \left[ \frac{1}{T} \sum_{t=1}^{T} f(w_t; A_t) \right] \leq \mathbb{E}_{A^i} \left[ \frac{1}{T} \sum_{t=1}^{T} f(w^*; A_t) \right] + \frac{c \ln(T)}{\sigma T}$$

and $w^*$ does not depend on the choice of $A^T_1$, we have,

$$\mathbb{E}_{A^i} \left[ \frac{1}{T} \sum_{t=1}^{T} f(w^*; A_t) \right] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{A_t} [f(w^*; A_t)] = f(w^*)$$

Recall that the $\mathbb{E}[f(X)] = \mathbb{E}_Y \mathbb{E}_X[f(X)|Y]$ and $w_t$ only depends on $A^{t-1}_1$, we get

$$\mathbb{E}_{A^i} \left[ \frac{1}{T} \sum_{t=1}^{T} f(w_t; A_t) \right] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{A^i} [f(w_t; A_t)] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{A^t_{t-1}} [\mathbb{E}_{A^i} [f(w_t; A_t)|A^{t-1}_1]] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{A^t_{t-1}} [f(w_t)] = \mathbb{E}_{A^i} \left[ \frac{1}{T} \sum_{t=1}^{T} f(w_t) \right]$$
Theorem (3)

Assume that the conditions stated in Thm. 2 hold. Let \( \delta \in (0, 1) \), Then, with probability of at least \( 1 - \delta \) we have that

\[
f(w_r) \leq f(w^*) + \frac{c \ln(T)}{\delta \sigma T}
\]

Proof: Let \( Z \) be the random variable \( f(w_r) - f(w^*) \), from the definition we can know \( Z \) is non-negative. Thus, from Markov inequality

\[
P[Z > a] \leq \frac{\mathbb{E}[Z]}{a}.
\]

Setting \( \frac{\mathbb{E}[Z]}{a} = \delta \) and using Thm. 2 we obtain that

\[
a \leq \frac{c \ln(T)}{\delta \sigma T}
\]

From Thm. 3 we obtain that to achieve accuracy \( \epsilon \) with confidence \( 1 - \delta \) we need \( O\left(\frac{1}{\sigma \delta \epsilon}\right) \) iterations.
Figure: fix $T$
Figure: fix $kT$
The End