# Supplement Materials for "Incremental and Decremental Training for Linear Classification"

Cheng-Hao Tsai
Dept. of Computer Science
National Taiwan Univ., Taiwan
r01922025@csie.ntu.edu.tw

Chieh-Yen Lin
Dept. of Computer Science
National Taiwan Univ., Taiwan
r01944006@csie.ntu.edu.tw

Chih-Jen Lin
Dept. of Computer Science
National Taiwan Univ., Taiwan
cjlin@csie.ntu.edu.tw

## Details of Deriving (25)

From Section 3.2, $\Delta_D\bar{\boldsymbol{\alpha}}$ is the dual initial solution. Then

$$
\begin{aligned}
&\text{dual initial value}\\
=&\sum_{i=k+1}^{l} h(\Delta_D\alpha_i^*, \Delta_D C)-\\
&\frac{1}{2}\sum_{i=k+1}^{l}\sum_{j=k+1}^{l}\Delta_D^2\alpha_i^*\alpha_j^*K(i,j)-\sum_{i=k+1}^{l}(\Delta_D\alpha_i^*)^2\frac{d}{2\Delta_D}\\
=&\Delta_D\Bigg(\sum_{i=k+1}^{l}h(\alpha_i^*, C)\\
&-\frac{\Delta_D}{2}\Big(\sum_{i=k+1}^{l}\sum_{j=k+1}^{l}\alpha_i^*\alpha_j^*K(i,j)+\sum_{i=k+1}^{l}(\alpha_i^*)^2\frac{d}{\Delta_D}\Big)\Bigg)\\
=&\Delta_D\Bigg(\frac{1}{2}\bar{\boldsymbol{w}}^T\bar{\boldsymbol{w}}+C\sum_{i=k+1}^{l}\xi(\bar{\boldsymbol{w}};\boldsymbol{x}_i,y_i)-\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\alpha_i^*\alpha_j^*K(i,j)\\
&+\frac{1}{2}\Big(\sum_{i=k+1}^{l}\sum_{j=k+1}^{l}\alpha_i^*\alpha_j^*K(i,j)+\sum_{i=k+1}^{l}(\alpha_i^*)^2 d\Big)\\
&-\frac{\Delta_D}{2}\Big(\sum_{i=k+1}^{l}\sum_{j=k+1}^{l}\alpha_i^*\alpha_j^*K(i,j)+\sum_{i=k+1}^{l}(\alpha_i^*)^2\frac{d}{\Delta_D}\Big)\Bigg)\\
=&\frac{1}{2}\bar{\boldsymbol{w}}^T\bar{\boldsymbol{w}}+\Delta_D C\sum_{i=k+1}^{l}\xi(\bar{\boldsymbol{w}};\boldsymbol{x}_i,y_i)-\frac{\Delta_D}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\alpha_i^*\alpha_j^*K(i,j)\\
&-\frac{\Delta_D^2-\Delta_D}{2}\sum_{i=k+1}^{l}\sum_{j=k+1}^{l}\alpha_i^*\alpha_j^*K(i,j)+\frac{\Delta_D-1}{2}\bar{\boldsymbol{w}}^T\bar{\boldsymbol{w}}.
\end{aligned}
$$

The second and third equalities respectively use (31) and (23).

## Experiments for Logistic regression with Larger $C$

In the paper we present figures of comparing optimization methods using logistic regression with $C = 1$. Figures I and II give result of using $C = 128$.

## Experiments for L2-loss Support Vector Machine

We have conducted experiments on L2-loss SVM by following the same settings for logistic regression. Table I shows the relative difference between initial and optimal function values for incremental and decremental learning, respectively. The comparison of the three optimization methods are shown in Figures III and IV for $C = 1$ and Figures V and VI for $C = 128$.

**Table I: Relative objective value difference between the initial point and the optimal solution. L2-SVM is considered. wo-ws: without warm start. The better value between primal and dual is boldfaced.**

| Data set | Formulation | $C = 1$ | | | | $C = 128$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | wo-ws | $r = 5$ | $r = 50$ | $r = 500$ | wo-ws | $r = 5$ | $r = 50$ | $r = 500$ |
| ijcnn | Primal | 3.2e+00 | **3.3e−04** | **2.4e−05** | **1.7e−06** | 3.2e+00 | **3.4e−04** | **2.5e−05** | **1.7e−06** |
| | Dual | **1.0e+00** | 1.9e−01 | 1.9e−02 | 1.3e−03 | **1.0e+00** | 1.9e−01 | 1.9e−02 | 1.3e−03 |
| webspam | Primal | 2.9e+00 | **2.4e−04** | **2.6e−05** | **3.8e−06** | 3.0e+00 | **3.2e−04** | **3.8e−05** | **4.3e−06** |
| | Dual | **1.0e+00** | 2.0e−01 | 2.0e−02 | 2.4e−03 | **1.0e+00** | 2.0e−01 | 2.0e−02 | 2.4e−03 |
| news20 | Primal | 9.3e+00 | 1.8e−01 | 1.5e−02 | 1.3e−03 | 2.0e+02 | 5.1e+00 | 4.4e−01 | 3.3e−02 |
| | Dual | **1.0e+00** | **1.6e−01** | **1.2e−02** | **1.0e−03** | **1.0e+00** | **2.3e−01** | **2.3e−02** | **2.1e−04** |
| rcv1 | Primal | 1.3e+01 | **1.8e−01** | **1.4e−02** | 2.6e−03 | 3.1e+02 | 9.0e+00 | 7.5e−01 | 1.7e−01 |
| | Dual | **1.0e+00** | 1.8e−01 | 1.4e−02 | **2.0e−03** | **1.0e+00** | **2.3e−01** | **2.6e−02** | **9.9e−04** |
| real-sim | Primal | 1.5e+01 | **1.1e−01** | **1.1e−02** | **1.4e−03** | 1.5e+02 | 3.9e+00 | 4.5e−01 | 3.7e−02 |
| | Dual | **1.0e+00** | 1.8e−01 | 1.7e−02 | 2.0e−03 | **1.0e+00** | **3.0e−01** | **3.3e−02** | **8.0e−04** |
| yahoo-japan | Primal | 5.8e+00 | **1.3e−01** | **1.1e−02** | **1.1e−03** | 5.5e+01 | 4.6e+00 | 4.4e−01 | 4.9e−02 |
| | Dual | **1.0e+00** | 2.2e−01 | 2.1e−02 | 2.3e−03 | **1.0e+00** | **2.7e−01** | **2.7e−02** | **2.9e−03** |

(a) Incremental learning

| Data set | Formulation | $C = 1$ | | | | $C = 128$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | wo-ws | $r = 5$ | $r = 50$ | $r = 500$ | wo-ws | $r = 5$ | $r = 50$ | $r = 500$ |
| ijcnn | Primal | 3.2e+00 | **3.3e−04** | **3.7e−05** | **2.5e−06** | 3.2e+00 | **3.4e−04** | **3.7e−05** | **2.6e−06** |
| | Dual | **1.0e+00** | 6.8e−01 | 8.0e−02 | 3.5e−03 | **1.0e+00** | 8.5e+01 | 8.3e+00 | 4.4e−01 |
| webspam | Primal | 2.9e+00 | **2.3e−04** | **3.4e−05** | **5.8e−06** | 3.0e+00 | **3.2e−04** | **5.1e−05** | **6.8e−06** |
| | Dual | **1.0e+00** | 5.1e−02 | 9.6e−03 | 1.4e−03 | **1.0e+00** | 3.1e+00 | 7.6e−01 | 1.9e−01 |
| news20 | Primal | 8.5e+00 | **7.7e−02** | 8.6e−03 | 7.4e−04 | 2.1e+02 | **1.0e−01** | **1.7e−02** | 3.2e−04 |
| | Dual | **1.0e+00** | 9.1e−02 | **6.1e−03** | **3.3e−04** | **1.0e+00** | 1.6e+01 | 1.4e+00 | **1.6e−04** |
| rcv1 | Primal | 1.3e+01 | **8.6e−02** | 8.1e−03 | 1.4e−03 | 3.2e+02 | **1.9e−01** | **2.4e−02** | 1.3e−03 |
| | Dual | **1.0e+00** | 1.0e−01 | **7.5e−03** | **7.7e−04** | **1.0e+00** | 2.1e+01 | 2.9e−01 | **7.3e−04** |
| real-sim | Primal | 1.5e+01 | **6.3e−02** | **7.8e−03** | **1.2e−03** | 1.7e+02 | **1.9e−01** | **1.8e−02** | **1.4e−03** |
| | Dual | **1.0e+00** | 1.5e−01 | 1.6e−02 | 2.4e−03 | **1.0e+00** | 2.4e+01 | 1.3e−01 | 4.2e−02 |
| yahoo-japan | Primal | 5.9e+00 | **7.5e−02** | **8.5e−03** | **8.7e−04** | 5.9e+01 | **2.0e−01** | **2.0e−02** | **2.8e−03** |
| | Dual | **1.0e+00** | 2.0e−01 | 2.4e−02 | 2.8e−03 | **1.0e+00** | 3.2e+01 | 5.3e+00 | 2.4e−01 |

(b) Decremental learning

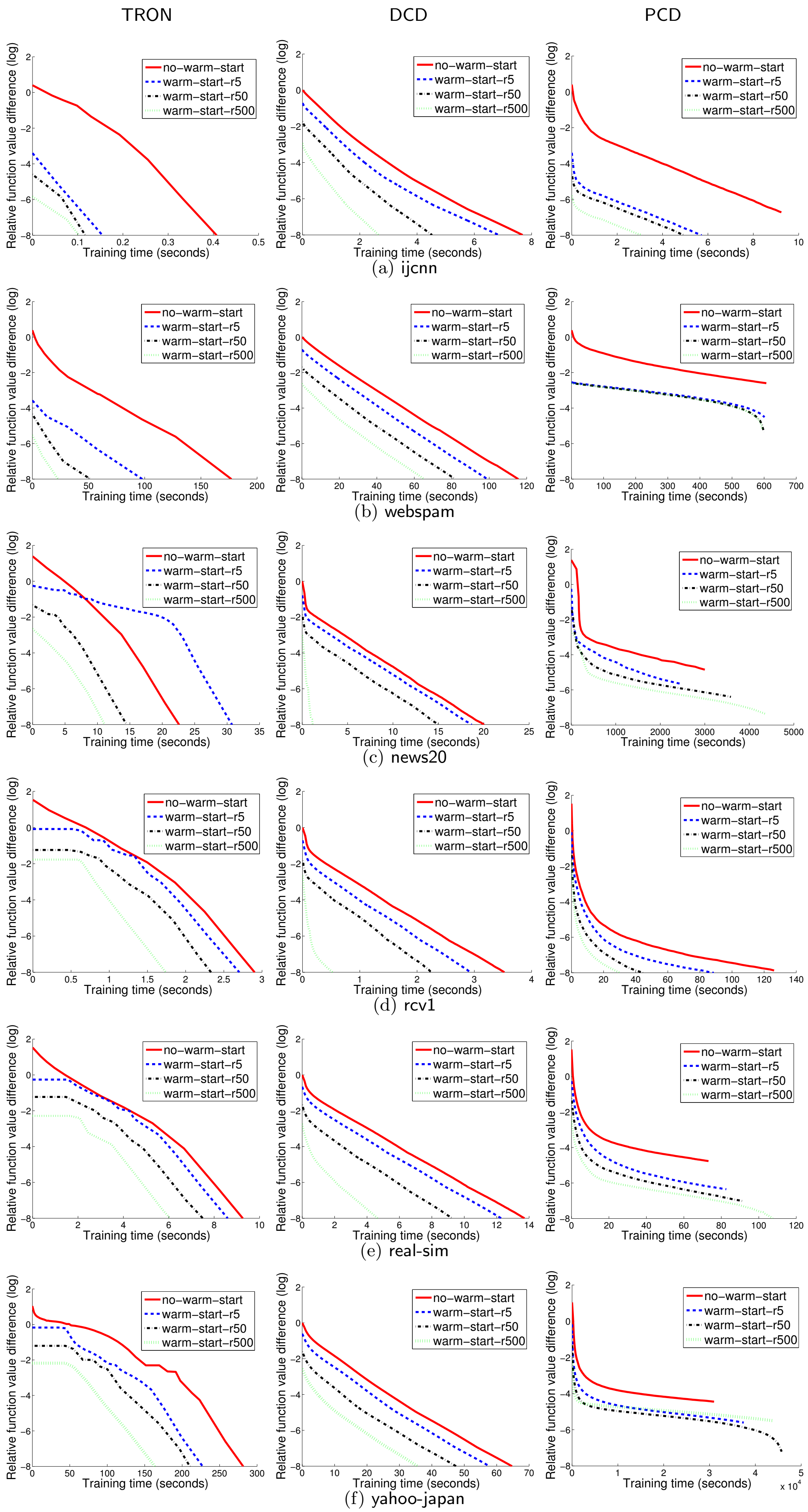
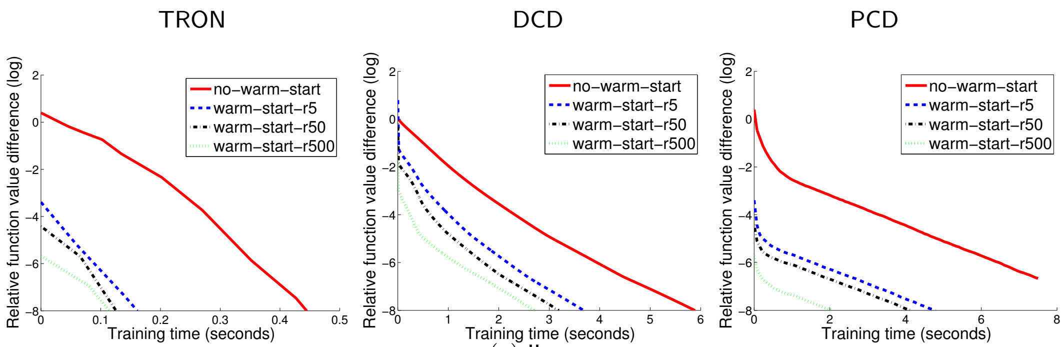
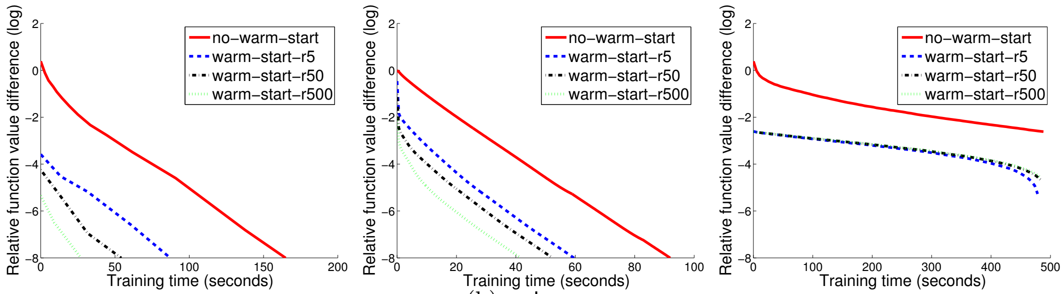


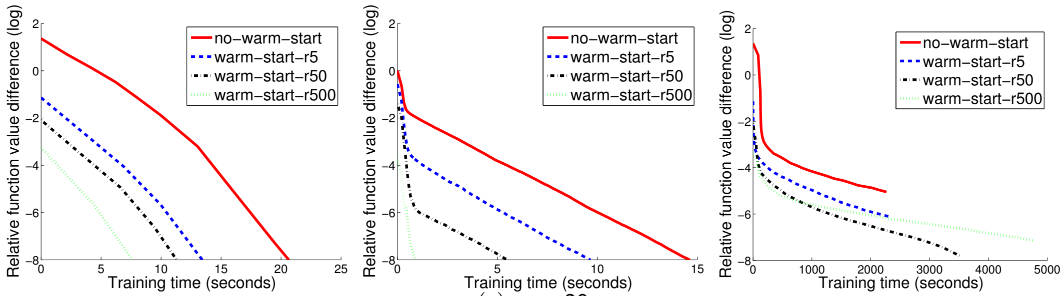

**Figure I: Incremental learning: running time (in seconds) versus the relative objective value difference. Logistic regression with $C = 128$ is used.**



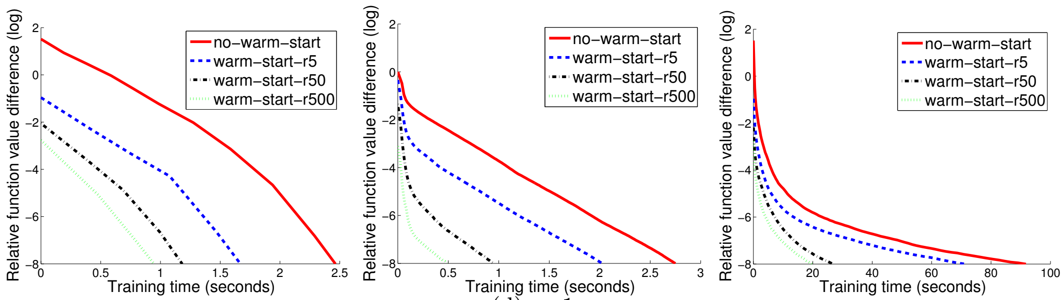

(a) ijcnn



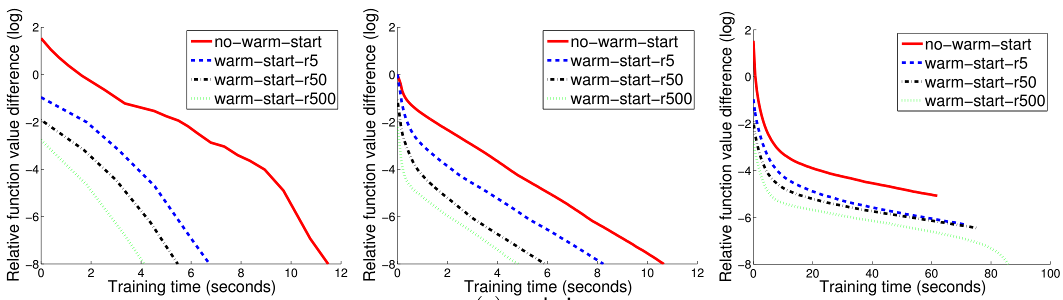

(b) webspam



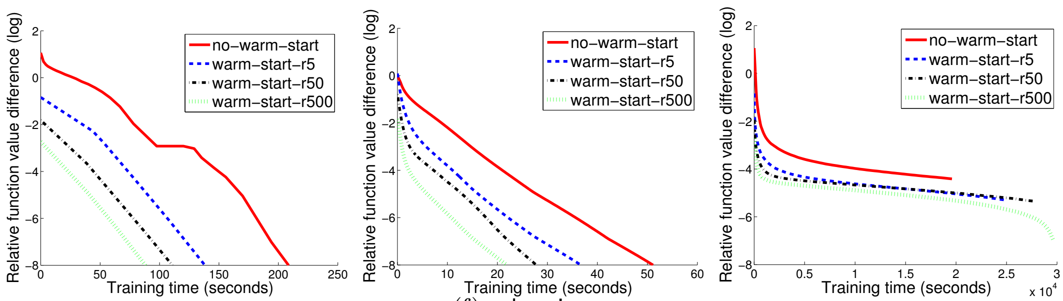

(c) news20




(d) rcv1




(e) real-sim




(f) yahoo-japan

**Figure II: Decremental learning: running time (in seconds) versus the relative objective value difference. Logistic regression with $C = 128$ is used.**

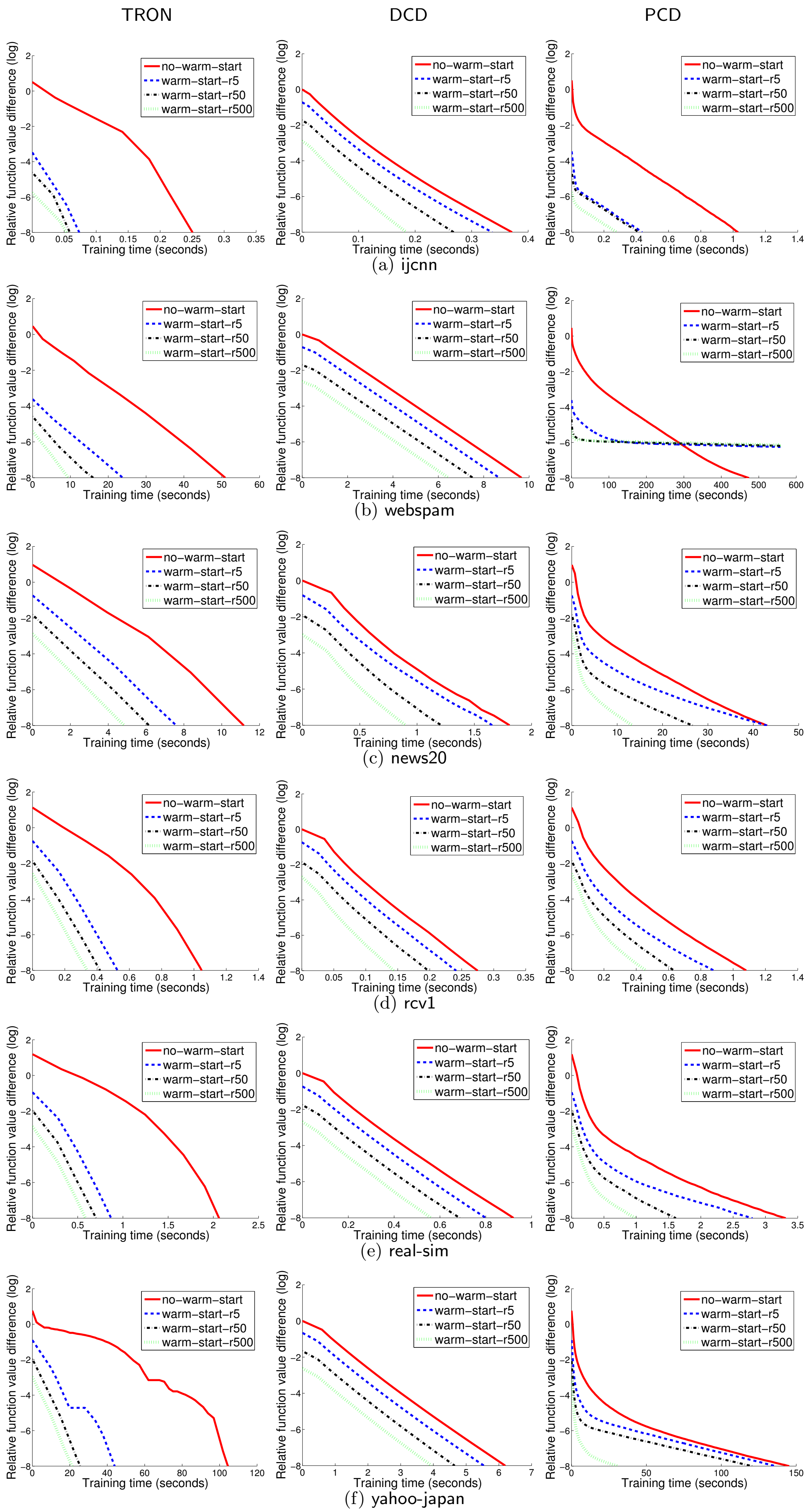
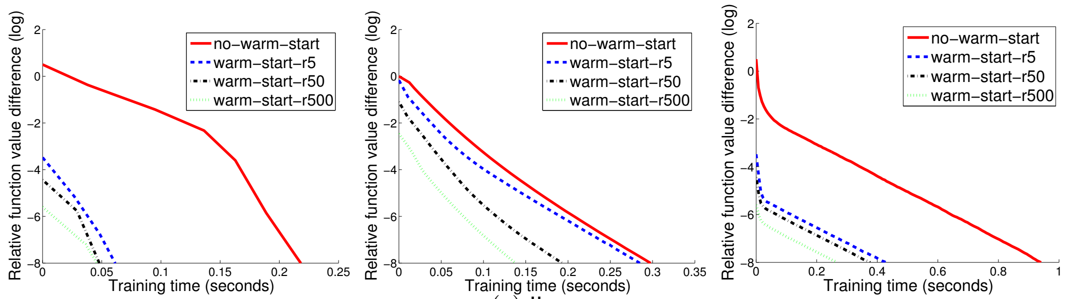
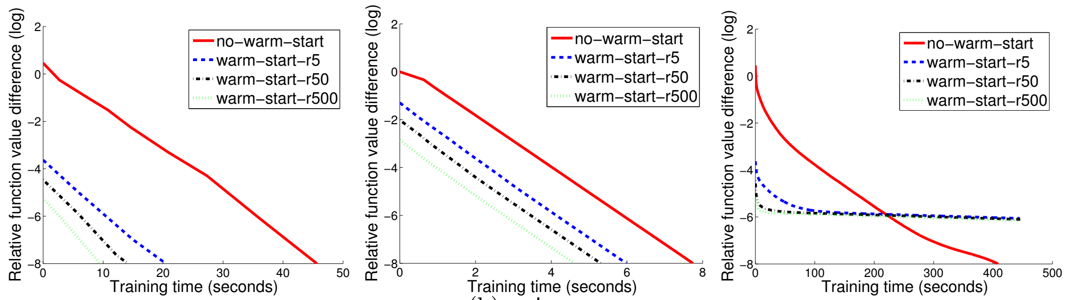



**Figure III: Incremental learning: running time (in seconds) versus the relative objective value difference. L2-SVM with $C = 1$ is used.**

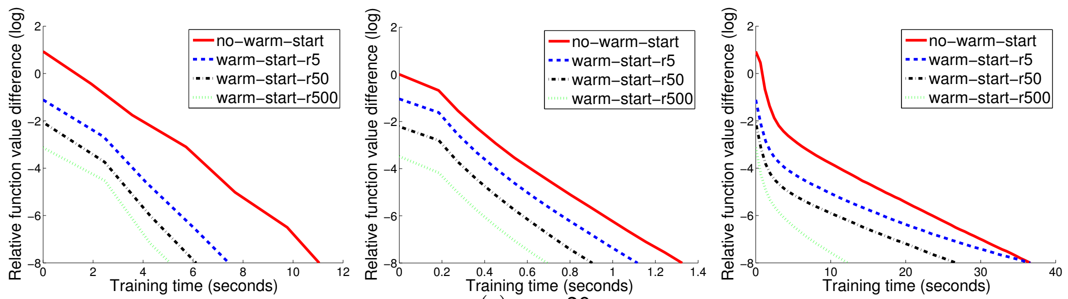
TRON DCD PCD



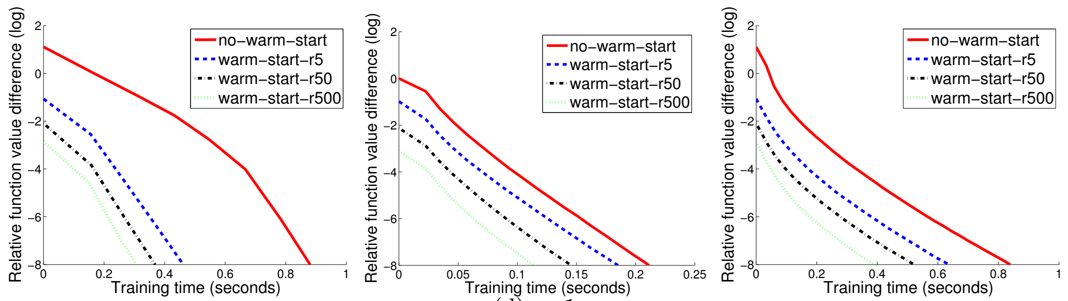

(a) ijcnn



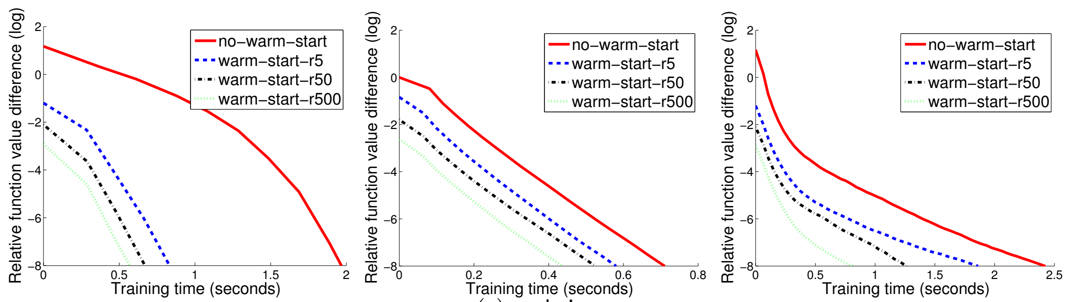

(b) webspam



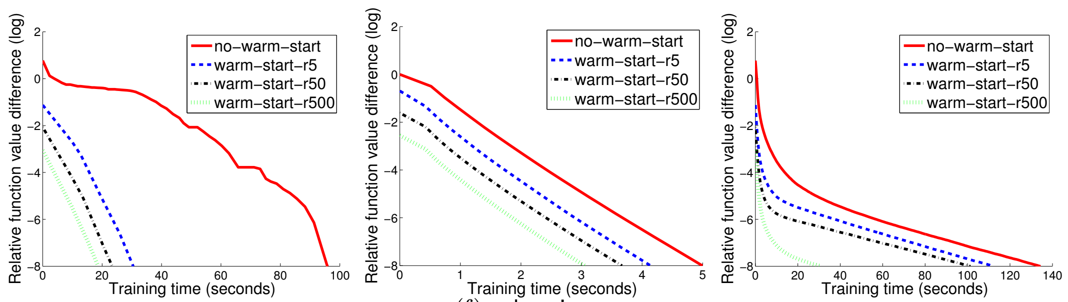

(c) news20




(d) rcv1

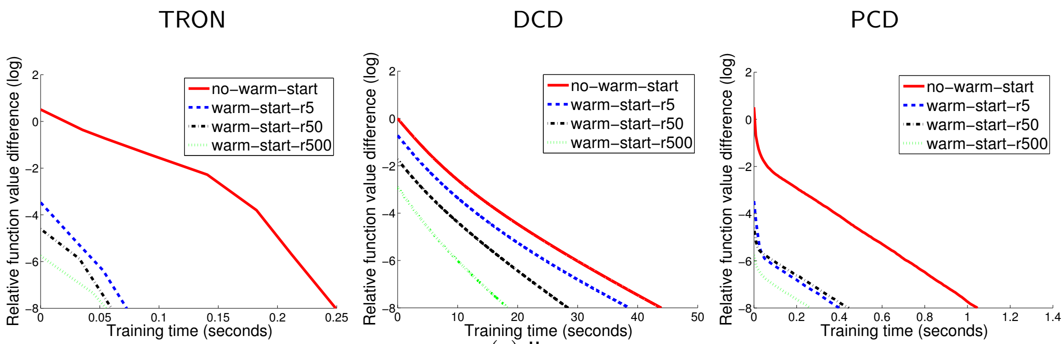



(e) real-sim

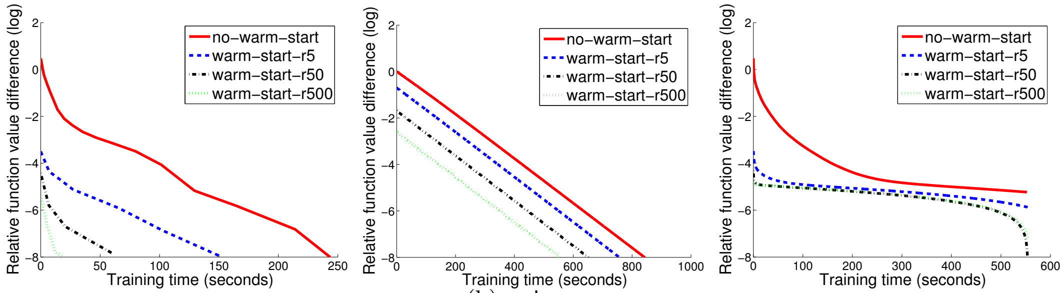



(f) yahoo-japan

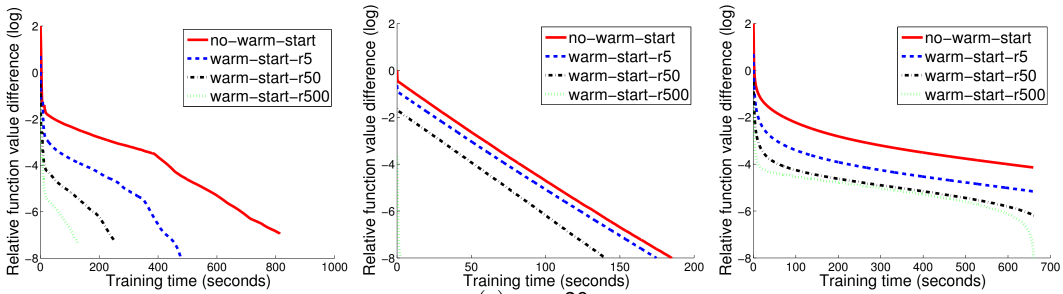
**Figure IV: Decremental learning: running time (in seconds) versus the relative objective value difference. L2-SVM with $C = 1$ is used.**



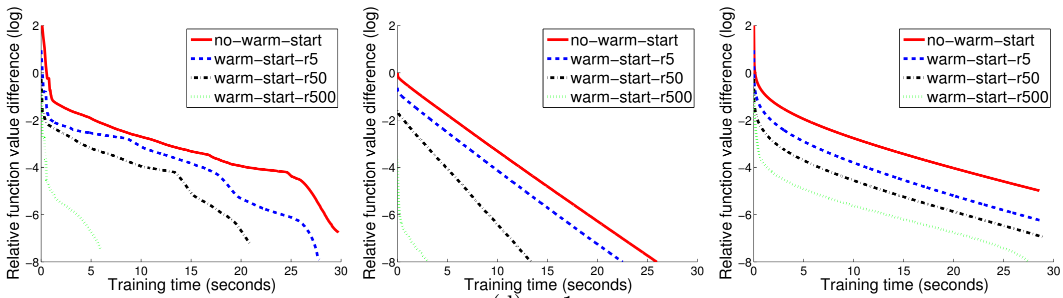

(a) ijcnn



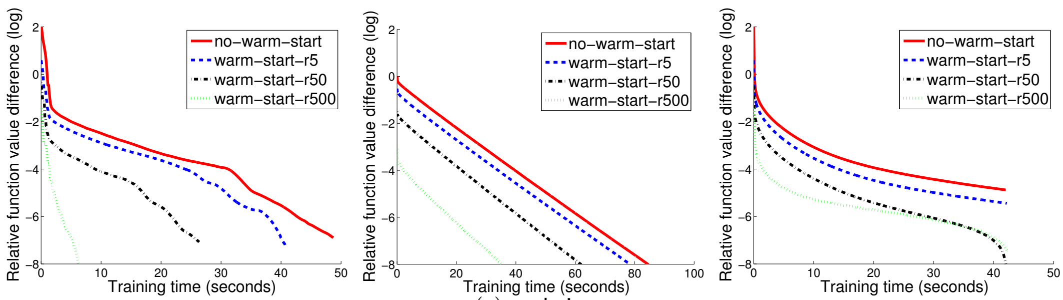

(b) webspam



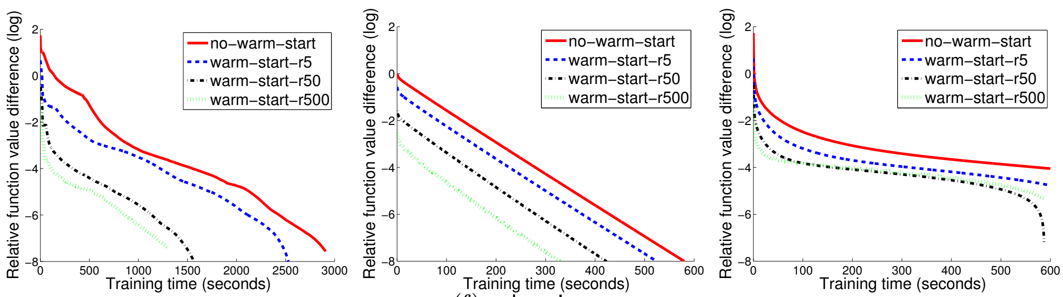

(c) news20




(d) rcv1




(e) real-sim




(f) yahoo-japan

**Figure V: Incremental learning: running time (in seconds) versus the relative objective value difference. L2-SVM with $C = 128$ is used.**

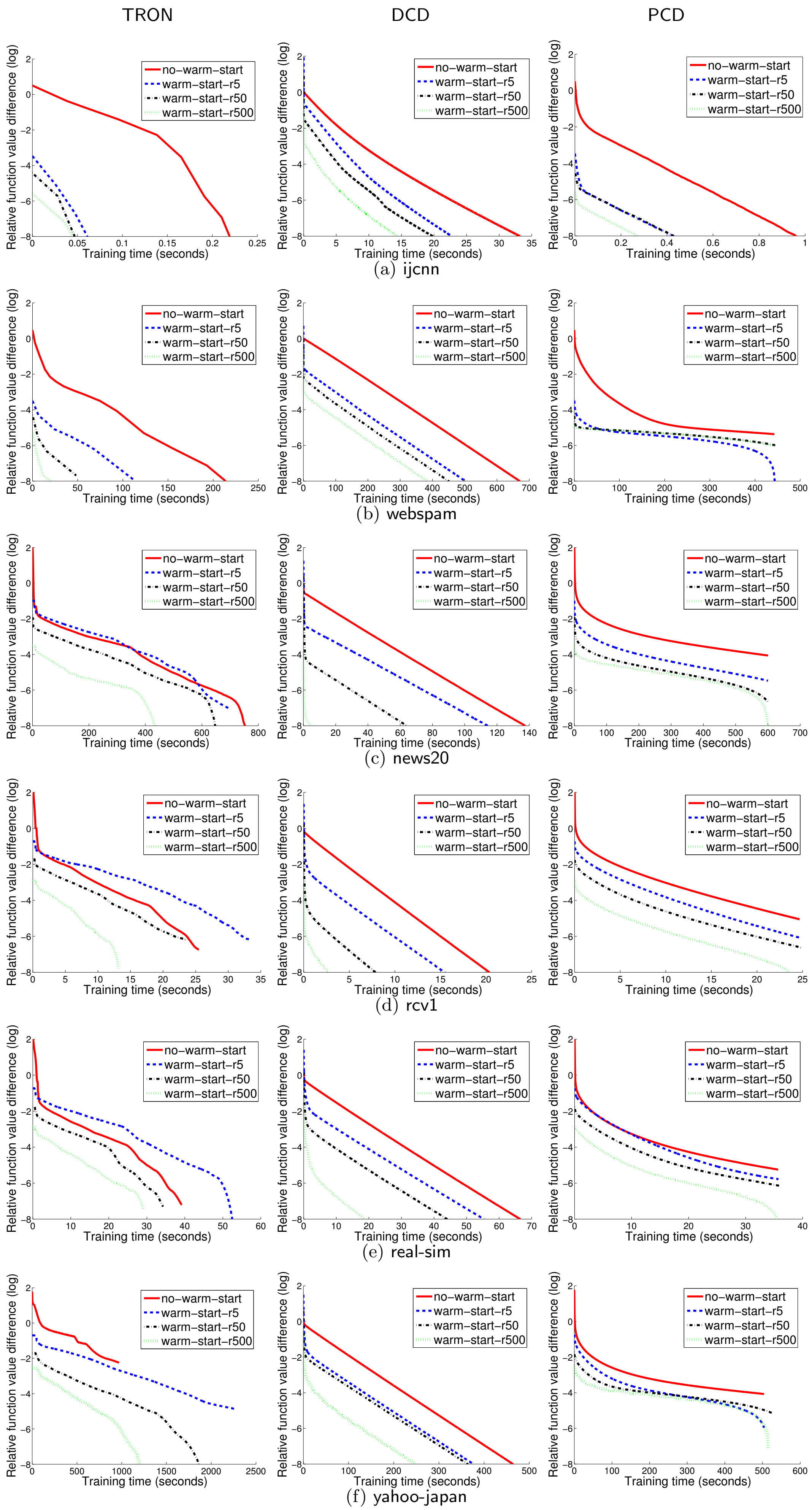



**Figure VI: Decremental learning: running time (in seconds) versus the relative objective value difference. L2-SVM with $C = 128$ is used.**