–

Supplementary Materials for "A Study on Conjugate Gradient Stopping Criteria in Truncated Newton Methods for Linear Classification"

For experiments we use two-class classification shown in Table I. Except for `yahookr` and `yahoojp`, others are available from LIBSVM Data Sets (2007). Table I shows the statistics of each data set used in our experiments.

Data sets	#instances	#features	density	$\log_2(C_{\text{best}})$	
				LR	L2
`HIGGS`	11,000,000	28	92.1057%	-6	-12
`w8a`	49,749	300	3.8834%	8	2
`rcv1`	20,242	47,236	0.1568%	6	0
`real-sim`	72,309	20,958	0.2448%	3	-1
`news20`	19,996	1,355,191	0.0336%	9	3
`url`	2,396,130	3,231,962	0.0036%	-7	-10
`yahoojp`	176,203	832,026	0.0160%	3	-1
`yahookr`	460,554	3,052,939	0.0111%	6	1
`webspam`	350,000	16,609,143	0.0220%	2	-3
`kdda`	8,407,752	20,216,831	0.0001%	-3	-5
`kddb`	19,264,097	29,890,095	0.0001%	-1	-4
`criteo`	45,840,617	1,000,000	0.0039%	-15	-12
`kdd12`	149,639,105	54,686,452	0.00002%	-4	-11

Table I: Data statistics. The density is the average number of non-zeros per instance. C_{best} is the regularization parameter selected by cross validation.

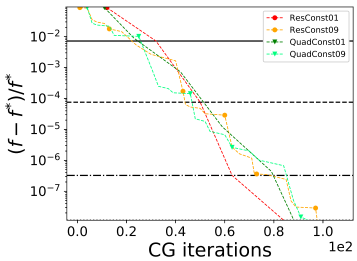
I. Additional Experimental Results

In this section we present complete results for linear classification problems with either

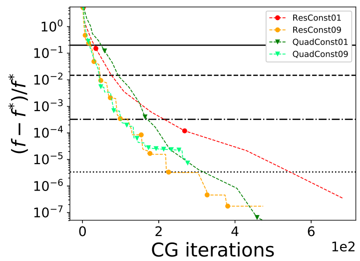
- LR-loss or
- L2-loss.

References

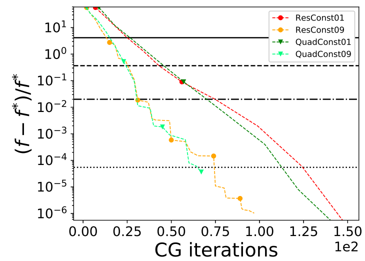
LIBSVM Data Sets, 2007. https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets.



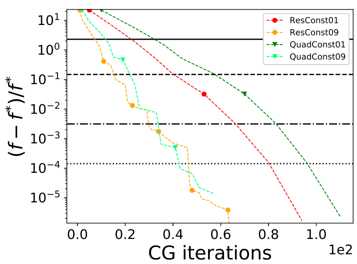
(a) `HIGGS`



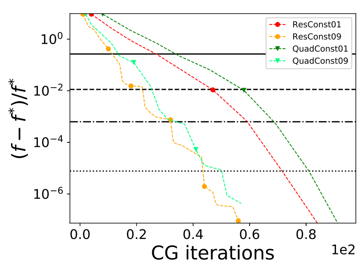
(b) `w8a`



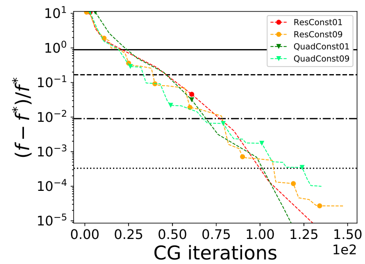
(c) `new20`



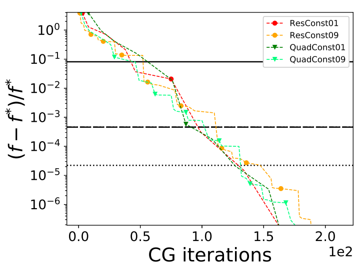
(d) `rcv1`



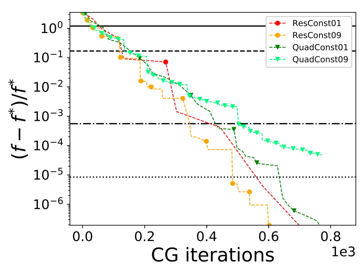
(e) `real-sim`



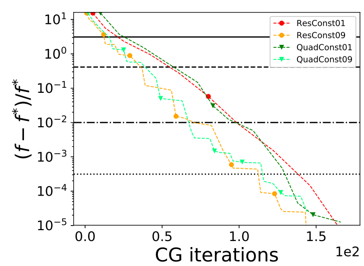
(f) `url`



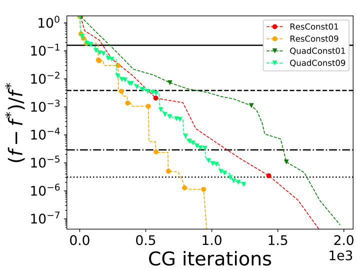
(g) `yahoojp`



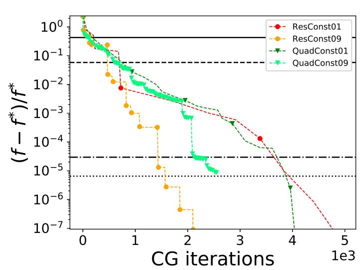
(h) `yahookr`



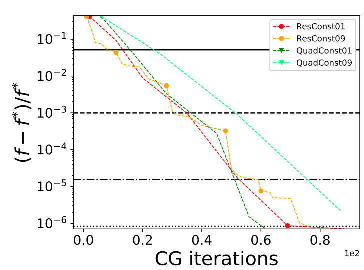
(i) `webspam`



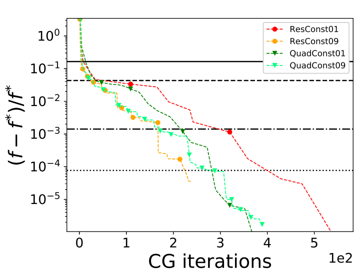
(j) `kdda`



(k) `kddb`

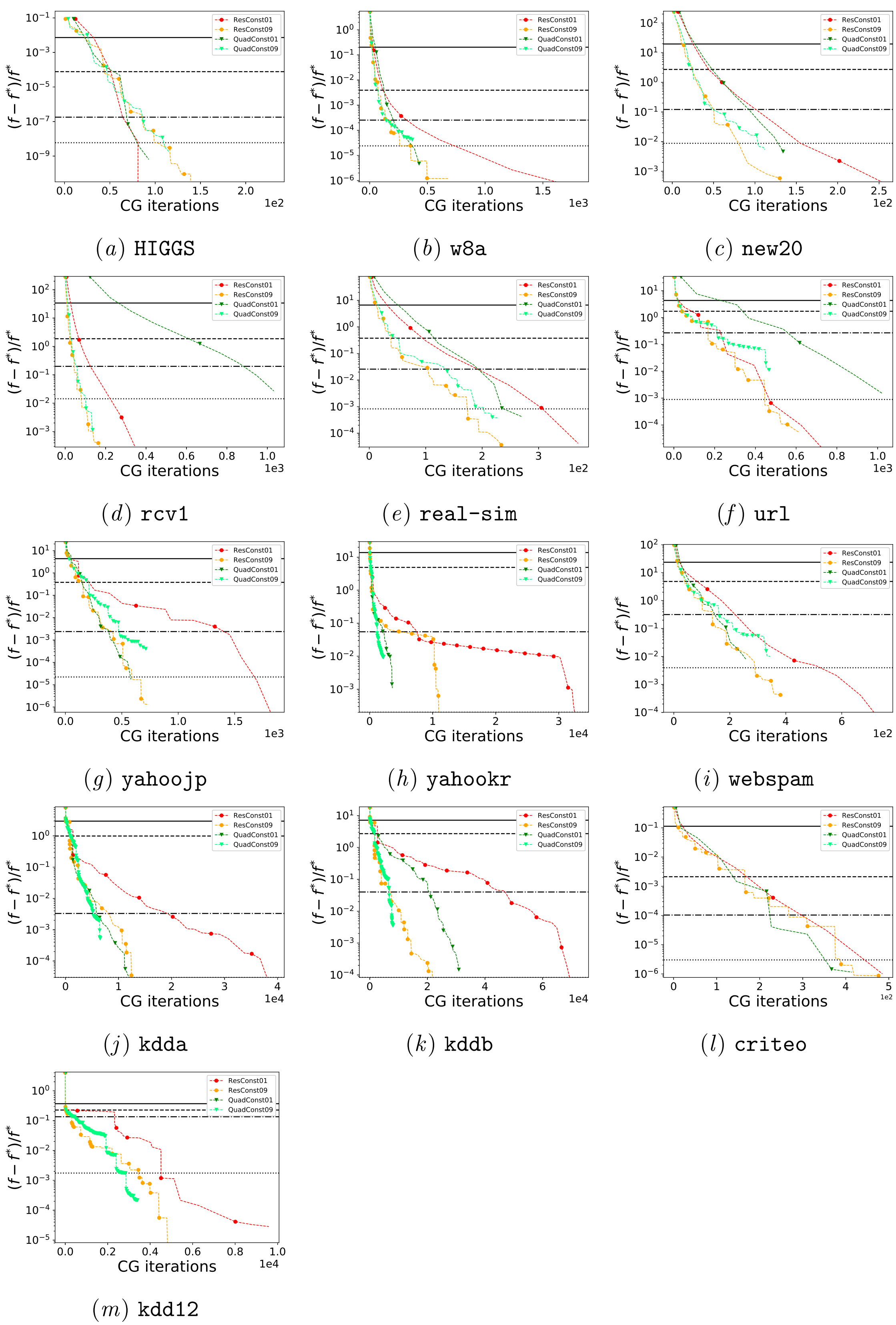


(l) `criteo`



(m) `kdd12`

Figure i: A comparison of different constant thresholds in the inner stopping condition. We show the convergence of a truncated Newton method for logistic regression. The x-axis shows the cumulative number of CG iterations. Loss=LR and $C = C_{\text{best}}$ in each sub-figure.



(a) HIGGS (b) w8a (c) new20

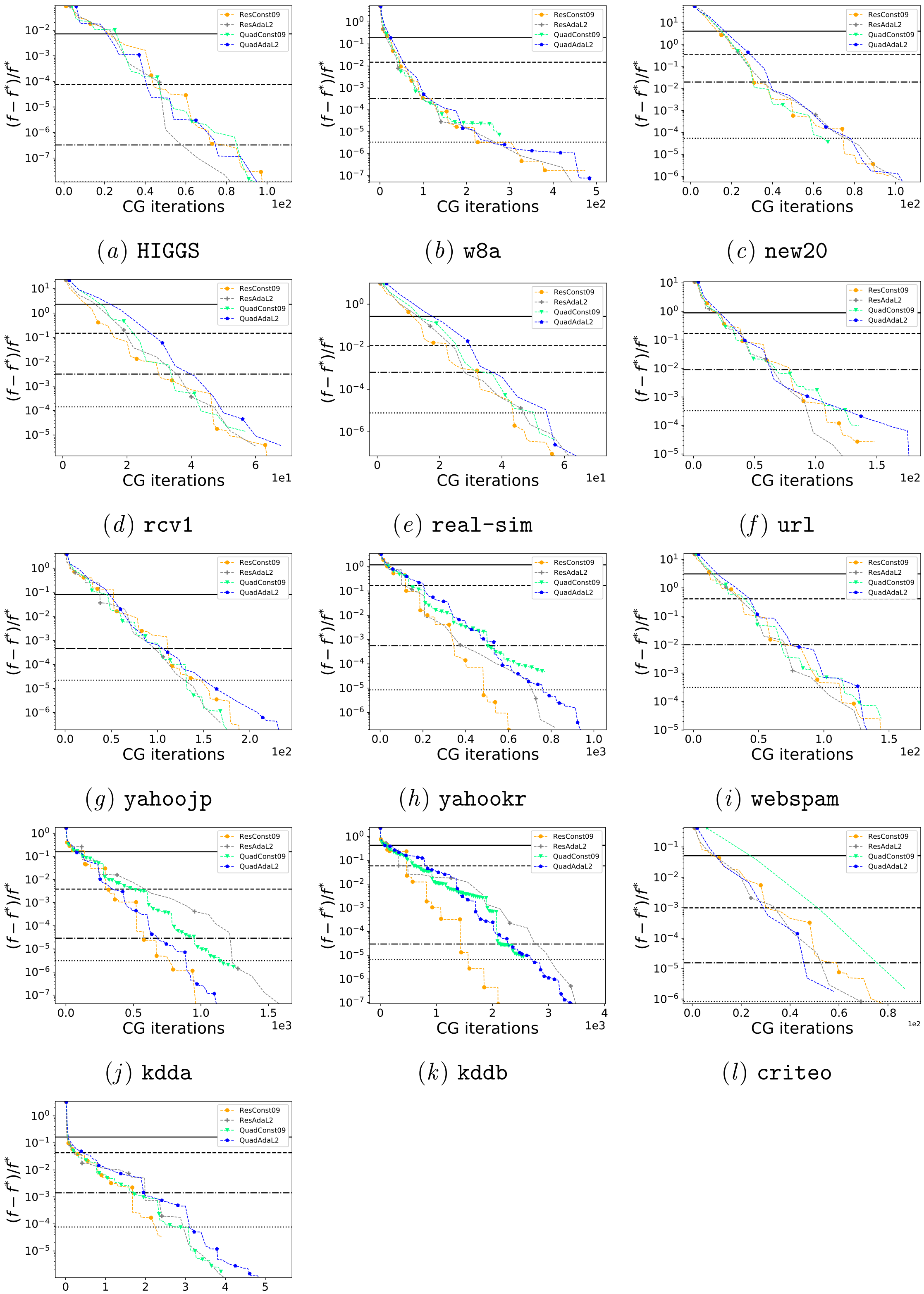
(d) rcv1 (e) real-sim (f) url

(g) yahoojp (h) yahookr (i) webspam

(j) kdda (k) kddb (l) criteo

(m) kdd12

Figure ii: A comparison of different constant thresholds in the inner stopping condition. We show the convergence of a truncated Newton method for logistic regression. The x-axis shows the cumulative number of CG iterations. Loss=LR and $C = 100C_{\text{best}}$ in each sub-figure.



(a) HIGGS (b) w8a (c) new20

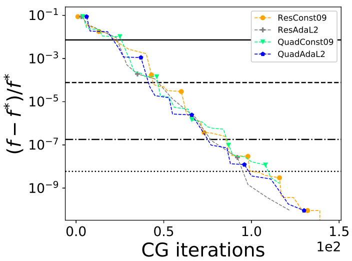
(d) rcv1 (e) real-sim (f) url

(g) yahoojp (h) yahookr (i) webspam

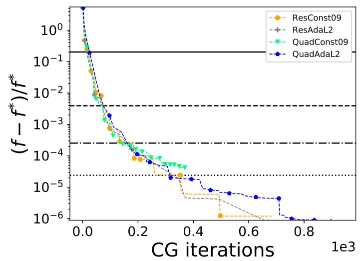
(j) kdda (k) kddb (l) criteo

(m) kdd12

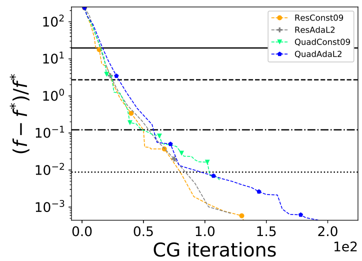
Figure iii: A comparison between adaptive and constant forcing sequences. We show the convergence of a truncated Newton method for logistic regression. The x-axis shows the cumulative number of CG iterations. Loss=LR and $C = C_{\text{best}}$ in each sub-figure.



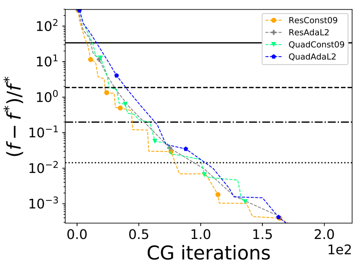
(a) `HIGGS`



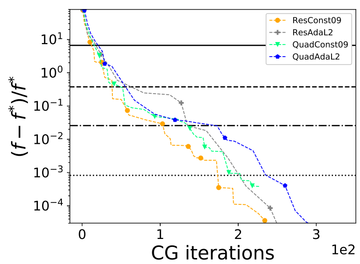
(b) `w8a`



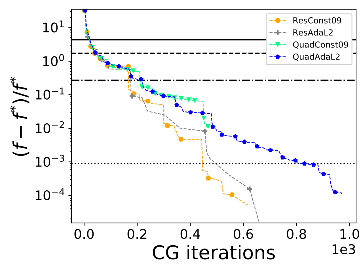
(c) `new20`



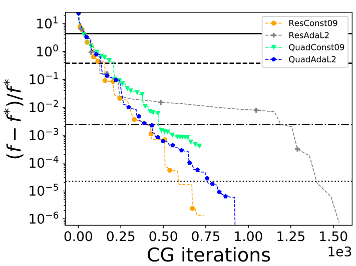
(d) `rcv1`



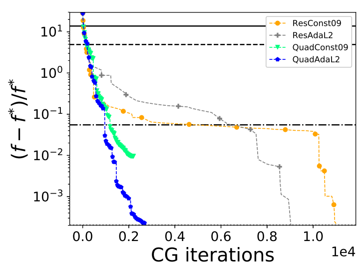
(e) `real-sim`



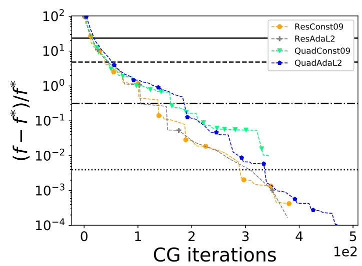
(f) `url`



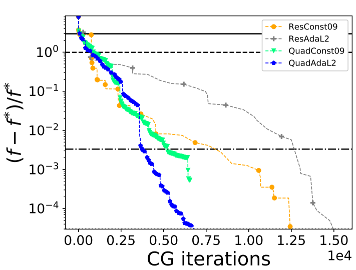
(g) `yahoojp`



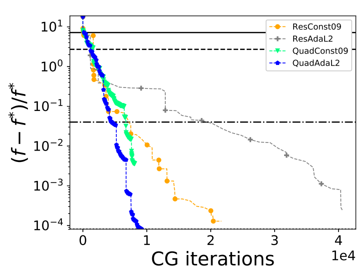
(h) `yahookr`



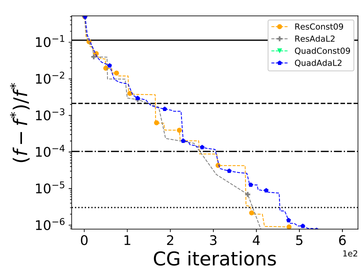
(i) `webspam`



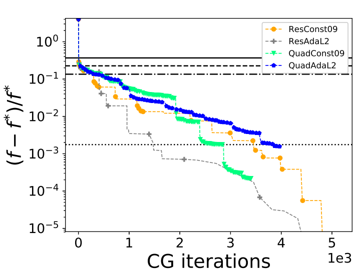
(j) `kdda`



(k) `kddb`

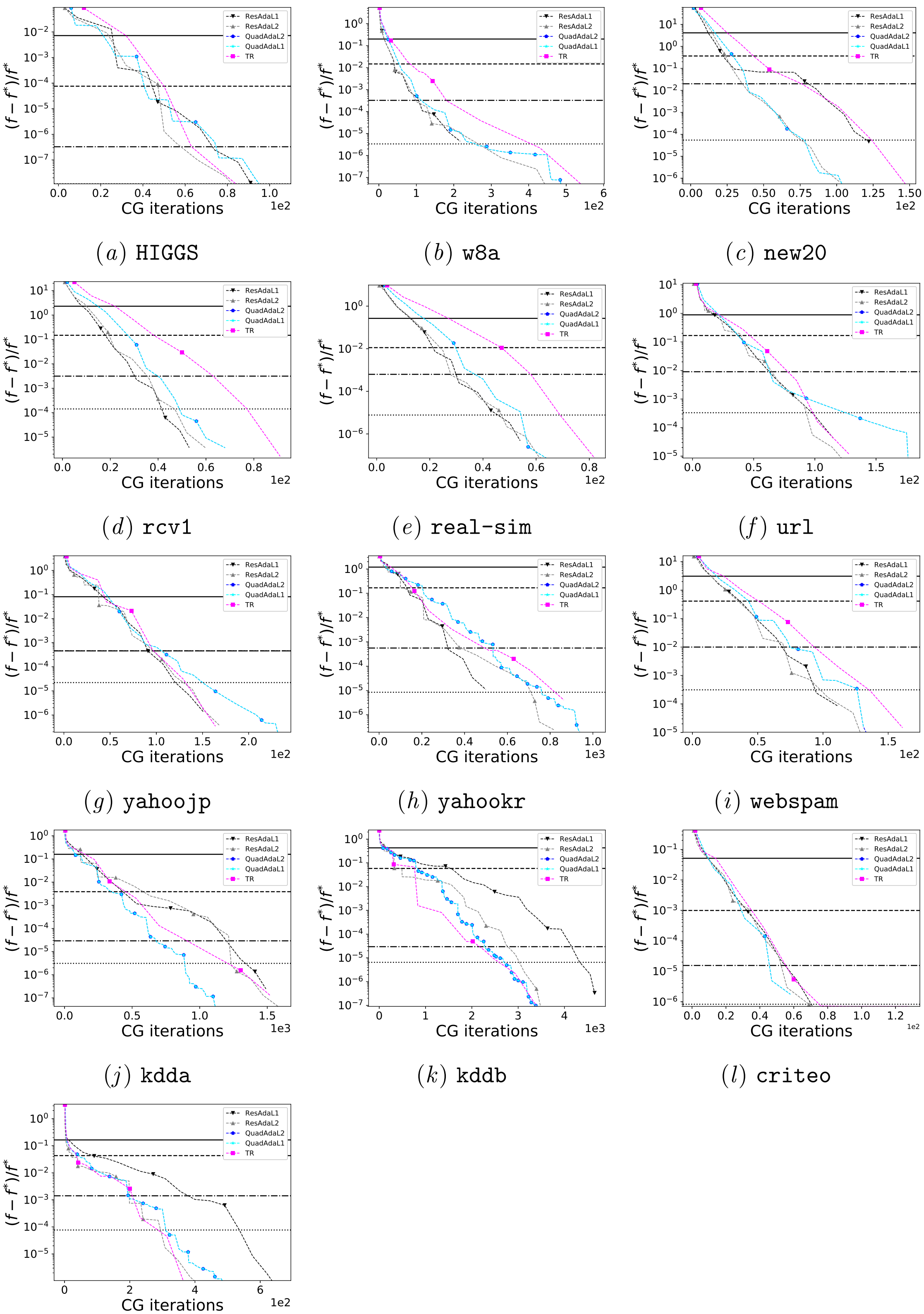


(l) `criteo`



(m) `kdd12`

Figure iv: A comparison between adaptive and constant forcing sequences. We show the convergence of a truncated Newton method for logistic regression. The x-axis shows the cumulative number of CG iterations. Loss=LR and $C = 100C_{\text{best}}$ in each sub-figure.



(a) HIGGS (b) w8a (c) new20

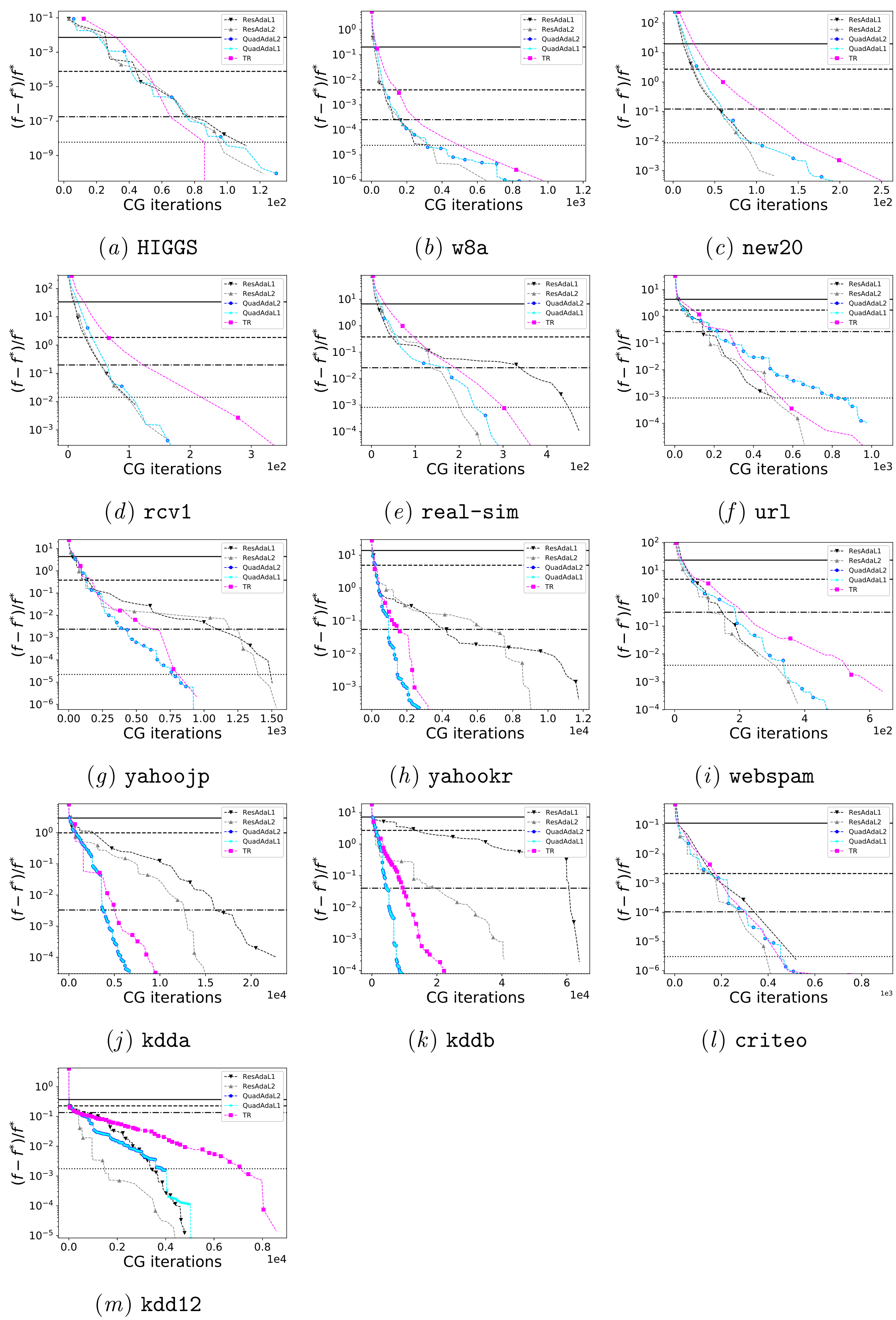
(d) rcv1 (e) real-sim (f) url

(g) yahoojp (h) yahookr (i) webspam

(j) kdda (k) kddb (l) criteo

(m) kdd12

Figure v: An investigation on the robustness of adaptive rules and a comparison with the trust region approach. We show the convergence of a truncated Newton method for logistic regression. The x-axis shows the cumulative number of CG iterations. Loss=LR and $C = C_{\text{best}}$ in each sub-figure.



(a) HIGGS (b) w8a (c) new20

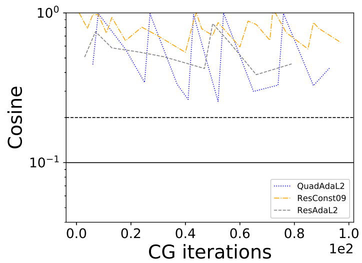
(d) rcv1 (e) real-sim (f) url

(g) yahoojp (h) yahookr (i) webspam

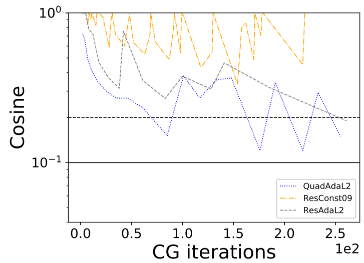
(j) kdda (k) kddb (l) criteo

(m) kdd12

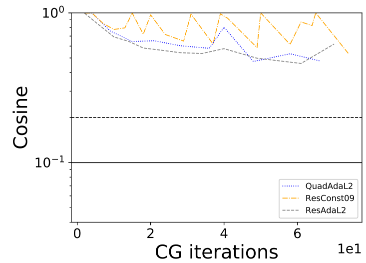
Figure vi: An investigation on the robustness of adaptive rules and a comparison with the trust region approach. We show the convergence of a truncated Newton method for logistic regression. The x-axis shows the cumulative number of CG iterations. Loss=LR and $C = 100C_{\text{best}}$ in each sub-figure.



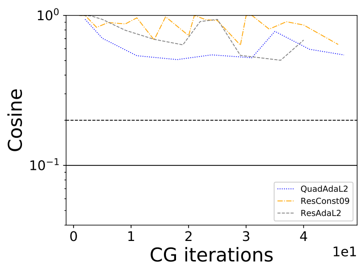
(a) `HIGGS`



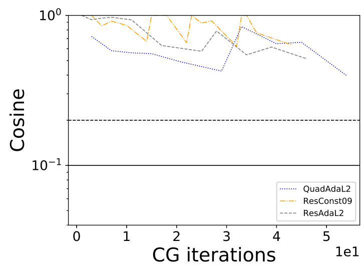
(b) `w8a`



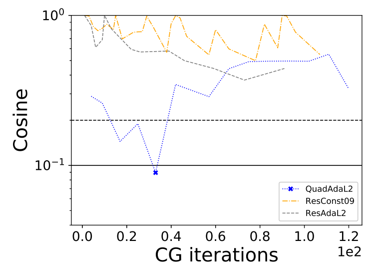
(c) `new20`



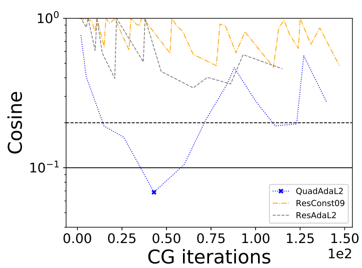
(d) `rcv1`



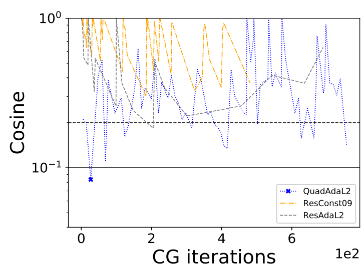
(e) `real-sim`



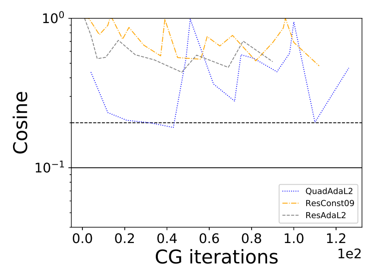
(f) `url`



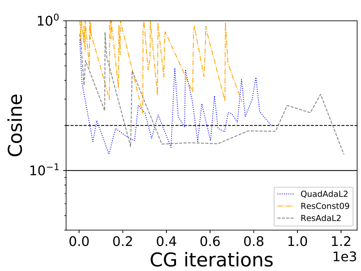
(g) `yahoojp`



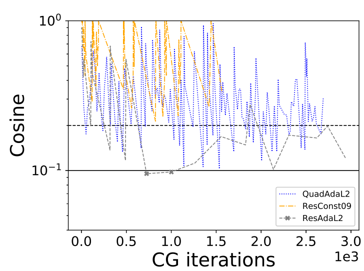
(h) `yahookr`



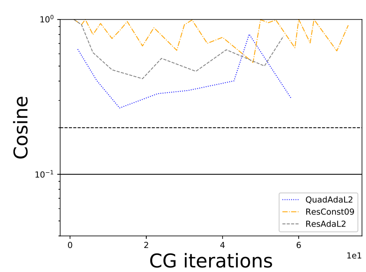
(i) `webspam`



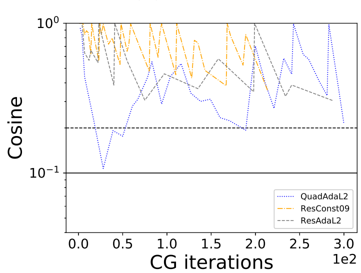
(j) `kdda`



(k) `kddb`

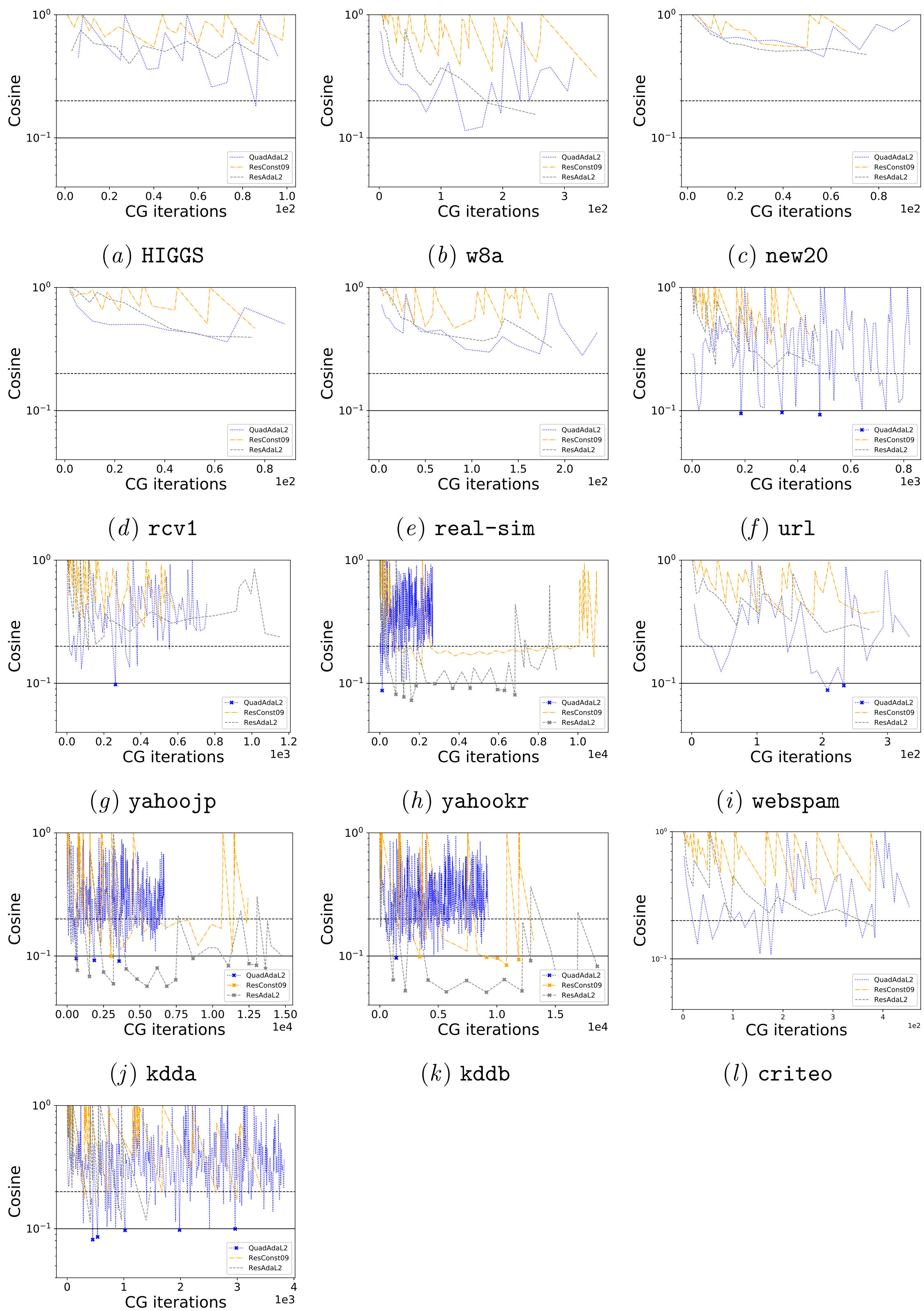


(l) `criteo`



(m) `kdd12`

Figure vii: Cosines between the anti-gradient $-\nabla f(\boldsymbol{w}_k)$ and the resulting direction $\boldsymbol{s}_k$, using different truncation rules. Iterations in which the cosine is below the 0.1 threshold are marked with a ✖. The x-axis shows the cumulative number of CG iterations. Loss=LR and $C = C_{\text{best}}$ in each sub-figure.



(a) HIGGS (b) w8a (c) new20

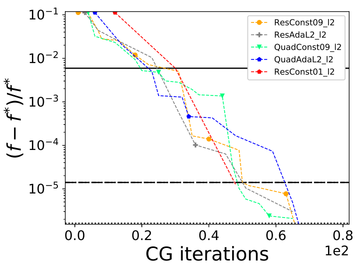
(d) rcv1 (e) real-sim (f) url

(g) yahoojp (h) yahookr (i) webspam

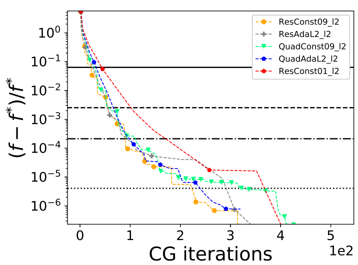
(j) kdda (k) kddb (l) criteo

(m) kdd12

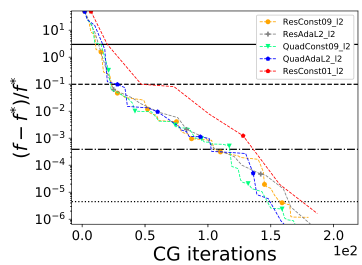
Figure viii: Cosines between the anti-gradient $-\nabla f(\boldsymbol{w}_k)$ and the resulting direction $\boldsymbol{s}_k$, using different truncation rules. Iterations in which the cosine is below the 0.1 threshold are marked with a ✖. The x-axis shows the cumulative number of CG iterations. Loss=LR and $C = 100C_{\text{best}}$ in each sub-figure.



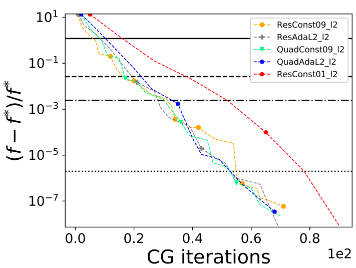
(a) HIGGS



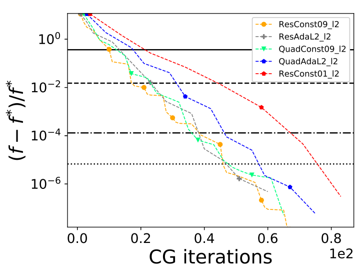
(b) w8a



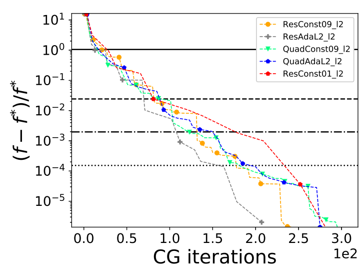
(c) new20



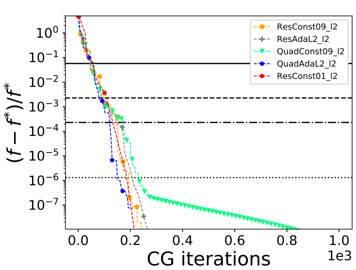
(d) rcv1



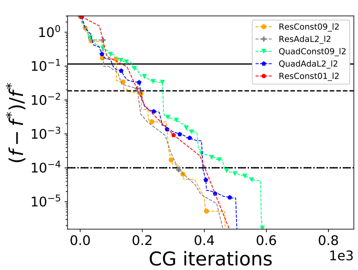
(e) real-sim



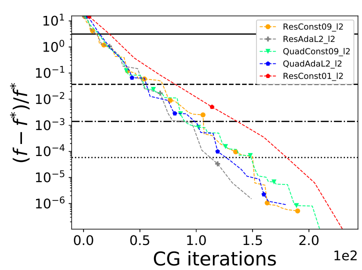
(f) url



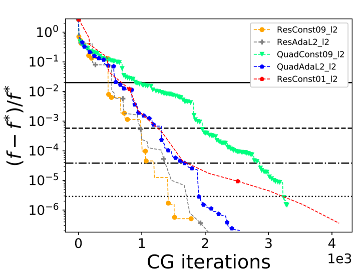
(g) yahoojp



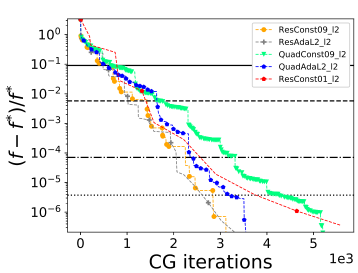
(h) yahookr



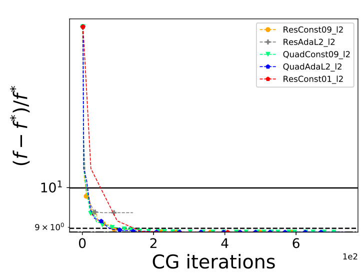
(i) webspam



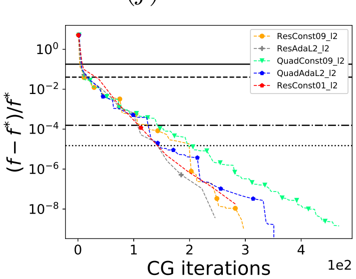
(j) kdda



(k) kddb

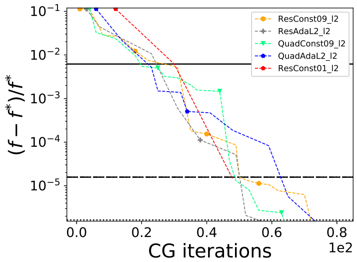


(l) criteo

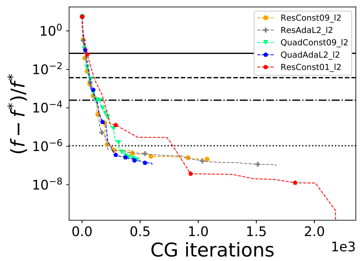


(m) kdd12

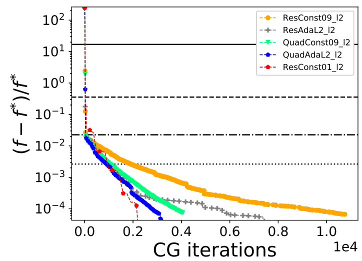
Figure ix: A comparison between adaptive and constant forcing sequences. We show the convergence of a truncated Newton method for linear SVM. The x-axis shows the cumulative number of CG iterations. Loss=L2 and $C = C_{\text{best}}$ in each sub-figure.



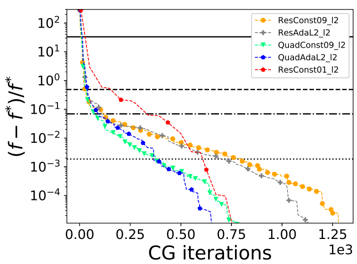
(a) HIGGS



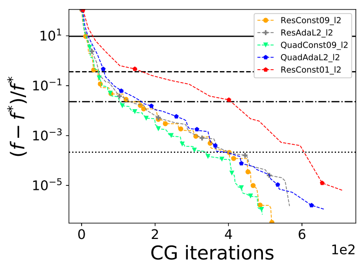
(b) w8a



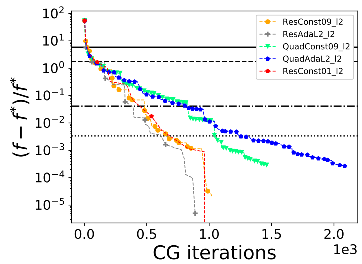
(c) new20



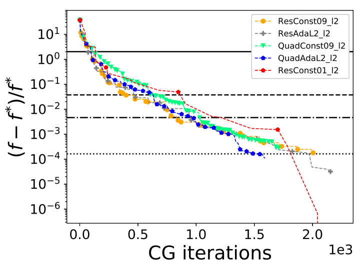
(d) rcv1



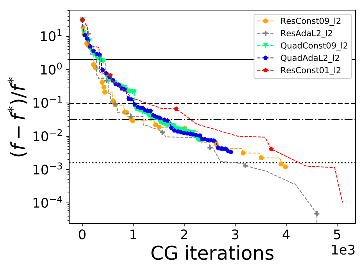
(e) real-sim



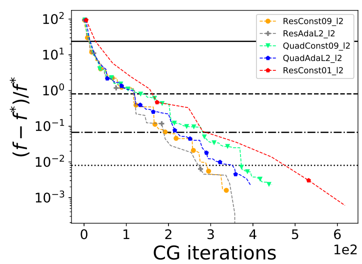
(f) url



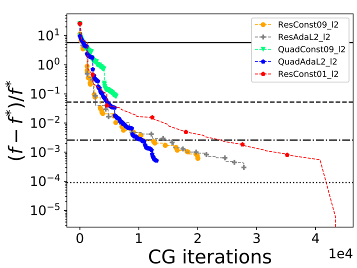
(g) yahoojp



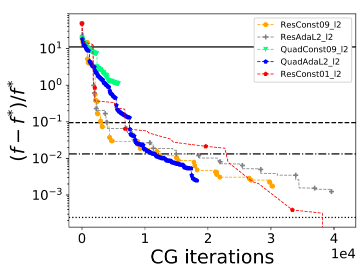
(h) yahookr



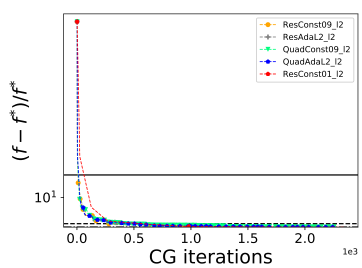
(i) webspam



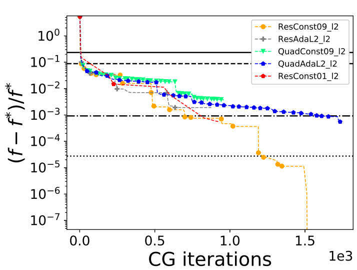
(j) kdda



(k) kddb



(l) criteo



(m) kdd12

Figure x: A comparison between adaptive and constant forcing sequences. We show the convergence of a truncated Newton method for linear SVM. The x-axis shows the cumulative number of CG iterations. Loss=L2 and $C = 100C_{\text{best}}$ in each sub-figure.