A Proofs and other technical material

The generic statement behind Theorem 1 is as follows; note that it neither makes specific requirements upon the form of the features (i.e., they need not be monomials), nor upon how support set $S_{(t+1)}$ is derived from $S_{(t)}$ (i.e., it only needs to satisfy the containment $S_{(t+1)} \supseteq S_{(t)}$).

Theorem 2. *Let convex function f be given with respective strong convexity and strong smoothness parameters $\lambda > 0$ and $\beta < \infty$. Let $(\boldsymbol{w}_t, \boldsymbol{g}_t)_{t\geq 1}$ be as specified by* apple *with step size $\eta_t := 1/(\lambda(t+1))$, where $\mathbb{E}_t([\boldsymbol{g}_t]_{S_{(t)}}) = [\nabla f(\boldsymbol{w}_t)]_{S_{(t)}}$, and $S_{(t)}$ is the support set corresponding to epoch k_t at time t in* apple*, with $S_{(t)} \subseteq S_{(t+1)}$ and $w_1 \in S_{(0)}$. Then for any comparator sequence $(\boldsymbol{u}_t)_{t=1}^{\infty}$ satisfying $\boldsymbol{u}_t \in S_{(t)}$, for any fixed $T \geq 1$,*

$$f(\boldsymbol{w}_{T+1}) - \frac{\sum_{t=1}^{T}(t+2)f(\boldsymbol{u}_t)}{\sum_{t=1}^{T}(t+2)} \leq \frac{1}{T+1}\left(\frac{\beta G^2}{2\lambda^2} + \frac{1}{\lambda}\sum_{t=1}^{T}\mathrm{dev}_t\right),$$

where $G := \max_{t\leq T}\|\boldsymbol{g}_t\|$, and the random variable

$$\mathrm{dev}_t := \left(\frac{t+2}{T+2}\right)\left\langle [\nabla f(\boldsymbol{w}_t)]_{S_{(t)}} - [\boldsymbol{g}_t]_{S_{(t)}}, \nabla f(\boldsymbol{w}_t)\right\rangle$$

satisfies $\mathbb{E}(\sum_{t=1}^{T}\mathrm{dev}_t) = 0$ and, with probability at least $1-\delta$ over the draw of $\{\boldsymbol{g}_t\}_{t=1}^{T}$, $\sum_{t=1}^{T}\mathrm{dev}_t \leq 4G^2\sqrt{T\ln(1/\delta)}$.

As discussed in the main text, the existence of a single target function f (rather than a new function each round) suggests its use in tracking the progress of the algorithm; indeed, the proof directly decreases $f(\boldsymbol{w}_{T+1}) - \sum_{t=1}^{T}(t+2)f(\boldsymbol{u}_t)/\sum_{t=1}^{T}(t+2)$, rather than passing through a surrogate such as measuring parameter distance $\|\boldsymbol{w}_t - \boldsymbol{u}_t\|$. The invocation of smoothness and strong convexity at the core of the argument (see the display with eq. (4)) is similar to the analogous invocation of smoothness and boundedness of the domain in the convergence guarantee for the Frank-Wolfe method [26, Theorem 3.4]. This bound is on the last iterate, whereas the standard proof scheme for subgradient descent, most naturally stated for averaged iterates [26, Theorem 3.1], requires some work for the last iterate [27, Theorem 1]; on the other hand, the approach here incurs an extra factor β/λ.

Proof of Theorem 2. Let $r_T \in \mathbb{R}$ be a parameter (dependent on T) left temporarily unspecified, and set the quantities

$$\begin{aligned}
\varepsilon_t^{(1)} &:= 2\lambda\left(f(\boldsymbol{u}_t) - r_T\right), && \forall t \geq 1,\\
\varepsilon_t^{(2)} &:= \left\langle [\nabla f(\boldsymbol{w}_t)]_{S_{(t)}} - [\boldsymbol{g}_t]_{S_{(t)}}, \nabla f(\boldsymbol{w}_t)\right\rangle, && \forall t \geq 1,\\
c &:= \frac{\beta G^2}{2\lambda^2}.
\end{aligned}$$

To prove the desired bound, it will first be shown, for any $t \geq 1$, that

$$f(\boldsymbol{w}_{t+1}) - r_T \leq \left(\frac{t-1}{t+1}\right)\left(f(\boldsymbol{w}_t) - r_T\right) + \frac{\eta_t^2\beta G^2}{2} + \eta_t\left(\varepsilon_t^{(1)} + \varepsilon_t^{(2)}\right). \tag{3}$$

Let $t \geq 1$ be arbitrary, and note by strong convexity, for any $\boldsymbol{w}$ with $\boldsymbol{w} \in S_{(t)}$, since $\boldsymbol{w}_t \in S_{(t-1)} \subseteq S_{(t)}$ and thus $\boldsymbol{w} - \boldsymbol{w}_t \in S_{(t)}$,

$$\begin{aligned}
f(\boldsymbol{w}) &\geq f(\boldsymbol{w}_t) + \langle \nabla f(\boldsymbol{w}_t), \boldsymbol{w} - \boldsymbol{w}_t\rangle + \frac{\lambda\|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2}{2}\\
&= f(\boldsymbol{w}_t) + \left\langle [\nabla f(\boldsymbol{w}_t)]_{S_{(t)}}, \boldsymbol{w} - \boldsymbol{w}_t\right\rangle + \frac{\lambda\|\boldsymbol{w} - \boldsymbol{w}_t\|_2^2}{2}.
\end{aligned}$$

The right hand side, as a function of $\boldsymbol{w}$, is a strongly convex quadratic over $S_{(t)}$, minimized at $\boldsymbol{w}_t - [\nabla f(\boldsymbol{w}_t)]_{S_{(t)}}/\lambda$. Plugging this back in,

$$f(\boldsymbol{w}) \geq f(\boldsymbol{w}_t) - \frac{\|[\nabla f(\boldsymbol{w}_t)]_{S_{(t)}}\|_2^2}{2\lambda},$$

which in particular holds for $\boldsymbol{w} = \boldsymbol{u}_t$ (which satisfies $\boldsymbol{u}_t \in S_{(t)}$), meaning

$$f(\boldsymbol{u}_t) \geq f(\boldsymbol{w}_t) - \frac{\|[\nabla f(\boldsymbol{w}_t)]_{S_{(t)}}\|_2^2}{2\lambda}.$$

Combining this with smoothness and the definition of $\boldsymbol{w}_{t+1}$,

$$\begin{aligned}
f(\boldsymbol{w}_{t+1}) - r_T &\leq f(\boldsymbol{w}_t) - r_T + \left\langle \nabla f(\boldsymbol{w}_t), -\eta_t [\boldsymbol{g}_t]_{S_{(t)}} \right\rangle + \frac{\eta_t^2 \beta \|[\boldsymbol{g}_t]_{S_{(t)}}\|_2^2}{2} \\
&= f(\boldsymbol{w}_t) - r_T + \left\langle [\nabla f(\boldsymbol{w}_t)]_{S_{(t)}}, -\eta_t [\boldsymbol{g}_t]_{S_{(t)}} \right\rangle + \frac{\eta_t^2 \beta \|[\boldsymbol{g}_t]_{S_{(t)}}\|_2^2}{2} \\
&= f(\boldsymbol{w}_t) - r_T - \eta_t \left\langle [\nabla f(\boldsymbol{w}_t)]_{S_{(t)}}, [\nabla f(\boldsymbol{w}_t)]_{S_{(t)}} \right\rangle + \eta_t \varepsilon_t^{(2)} + \frac{\eta_t^2 \beta \|[\boldsymbol{g}_t]_{S_{(t)}}\|_2^2}{2} \\
&\leq f(\boldsymbol{w}_t) - r_T + 2\lambda\eta_t (f(\boldsymbol{u}_t) - r_T + r_T - f(\boldsymbol{w}_t)) + \eta_t \varepsilon_t^{(2)} + \frac{\eta_t^2 \beta G^2}{2} \\
&= (1 - 2\lambda\eta_t)\left(f(\boldsymbol{w}_t) - r_T\right) + \frac{\eta_t^2 \beta G^2}{2} + \eta_t \left(\varepsilon_t^{(1)} + \varepsilon_t^{(2)}\right), \qquad (4)
\end{aligned}$$

thus establishing eq. (3) since $2\lambda\eta_t = 2/(t+1)$.

Next it will be proved by induction that, for any $t \geq 1$,

$$f(\boldsymbol{w}_{t+1}) - r_T \leq \frac{c}{t+1} + \sum_{j=1}^{t} \eta_j (\varepsilon_j^{(1)} + \varepsilon_j^{(2)}) \prod_{l=j+1}^{t} \frac{l}{l+2}, \qquad (5)$$

where the convention $\prod_{l=j+1}^{t} l/(l+2) = 1$ is adopted for $t < j+1$. For the base case $t = 1$, eq. (3) grants

$$\begin{aligned}
f(\boldsymbol{w}_{t+1}) - r_T &\leq \underbrace{\left(\frac{t-1}{t+1}\right)}_{=0} (f(\boldsymbol{w}_t) - r_T) + \frac{\eta_t^2 \beta G^2}{2} + \eta_t \left(\varepsilon_t^{(1)} + \varepsilon_t^{(2)}\right) \\
&\leq \frac{c}{t+1} + \sum_{j=1}^{t} \eta_j (\varepsilon_j^{(1)} + \varepsilon_j^{(2)}) \underbrace{\prod_{l=j+1}^{t} \frac{l}{l+2}}_{=1}.
\end{aligned}$$

On other other hand, in the case $t > 1$, once again starting from eq. (3),

$$\begin{aligned}
f(\boldsymbol{w}_{t+1}) - r_T &\leq \left(\frac{t-1}{t+1}\right) (f(\boldsymbol{w}_t) - r_T) + \frac{\eta_t^2 \beta G^2}{2} + \eta_t \left(\varepsilon_t^{(1)} + \varepsilon_t^{(2)}\right) \\
&\leq \left(\frac{t-1}{t+1}\right) \left(\frac{c}{t} + \sum_{j=1}^{t-1} \eta_j (\varepsilon_j^{(1)} + \varepsilon_j^{(2)}) \prod_{l=j+1}^{t-1} \frac{l}{l+2}\right) + \frac{\eta_t^2 \beta G^2}{2} + \eta_t \left(\varepsilon_t^{(1)} + \varepsilon_t^{(2)}\right) \\
&= \frac{t-1}{t+1} \left(\frac{c}{t} + \frac{\beta G^2}{2\lambda^2 (t-1)(t+1)}\right) + \sum_{j=1}^{t} \eta_j (\varepsilon_j^{(1)} + \varepsilon_j^{(2)}) \prod_{l=j+1}^{t} \frac{l}{l+2},
\end{aligned}$$

thus completing the proof of eq. (5).

To simplify the error term of eq. (5), note

$$\begin{aligned}
j = t \qquad &\implies \qquad \prod_{l=j+1}^{t} \frac{l}{l+2} = 1 = \frac{(j+1)(j+2)}{(t+1)(t+2)}, \\
j = t-1 \qquad &\implies \qquad \prod_{l=j+1}^{t} \frac{l}{l+2} = \frac{t}{t+2} = \frac{(j+1)(j+2)}{(t+1)(t+2)}, \\
j < t-1 \qquad &\implies \qquad \prod_{l=j+1}^{t} \frac{l}{l+2} = \frac{(j+1)(j+2)}{(t+1)(t+2)} \prod_{l=j+3}^{t} \frac{l}{l} = \frac{(j+1)(j+2)}{(t+1)(t+2)};
\end{aligned}$$

thus, for any $t \geq 1$ and $1 \leq j \leq t$,

$$\eta_j \prod_{l=j+1}^{t} \frac{l}{l+2} = \frac{j+2}{\lambda(t+1)(t+2)}.$$

Plugging this simplification back into eq. (5), for $t \geq 1$,

$$\begin{aligned} f(\boldsymbol{w}_{t+1}) - r_T &\leq \frac{c}{t+1} + \sum_{j=1}^{t} \eta_j (\varepsilon_j^{(1)} + \varepsilon_j^{(2)}) \prod_{l=j+1}^{t} \frac{l}{l+2} \\ &= \frac{c}{t+1} + \frac{1}{\lambda(t+1)(t+2)} \sum_{j=1}^{t} (j+2)(\varepsilon_j^{(1)} + \varepsilon_j^{(2)}). \end{aligned} \tag{6}$$

Next, to instantiate the comparator r, consider the choice

$$r_T := \frac{\sum_{j=1}^{T} (j+2) f(\boldsymbol{u}_j)}{\sum_{j=1}^{T} (j+2)}.$$

By construction, this provides

$$\frac{1}{2\lambda} \sum_{j=1}^{T} (j+2)\varepsilon_j^{(1)} = \sum_{j=1}^{T} (j+2) f(\boldsymbol{u}_j) - r_T \sum_{j=1}^{T} (j+2) = \sum_{j=1}^{T} (j+2) f(\boldsymbol{u}_j) - \sum_{j=1}^{T} (j+2) f(\boldsymbol{u}_j) = 0.$$

Consequently, eq. (6) simplifies to

$$f(\boldsymbol{w}_{T+1}) - \frac{\sum_{j=1}^{T} (j+2) f(\boldsymbol{u}_j)}{\sum_{j=1}^{T} (j+2)} \leq \frac{1}{T+1} \left(c + \frac{1}{\lambda(T+2)} \sum_{j=1}^{T} (j+2)\varepsilon_j^{(2)} \right), \tag{7}$$

which is the first part of the desired statement.

For the final desired statement, it remains to control $\varepsilon_j^{(2)}$ within eq. (7). For the expected value, let $\mathcal{F}_j$ be the σ-algebra of information up to time j; then

$$\begin{aligned} \mathbb{E}\left(\sum_{j=1}^{T} \frac{j+2}{T+2} \varepsilon_j^{(2)} \right) &= \mathbb{E}\left(\cdots \mathbb{E}\left(\mathbb{E}\left(\sum_{j=1}^{T} \frac{j+2}{T+2} \varepsilon_j^{(2)} \middle| \mathcal{F}_1 \right) \middle| \mathcal{F}_2 \right) \cdots \middle| \mathcal{F}_T \right) \\ &= \sum_{j=1}^{T} \mathbb{E}\left(\frac{j+2}{T+2} \varepsilon_j^{(2)} \middle| \mathcal{F}_j \right) \\ &= 0. \end{aligned}$$

Here the last equality holds since

$$\begin{aligned} \mathbb{E}\left(\varepsilon_j^{(2)} \middle| \mathcal{F}_j \right) &= \mathbb{E}\left(\left\langle [\nabla f(\boldsymbol{w}_j)]_{S_{(j)}} - [\boldsymbol{g}_j]_{S_{(j)}}, \nabla f(\boldsymbol{w}_j) \right\rangle \middle| \mathcal{F}_j \right) \\ &= \left\langle \mathbb{E}\left([\nabla f(\boldsymbol{w}_j)]_{S_{(j)}} - [\boldsymbol{g}_j]_{S_{(j)}} \middle| \mathcal{F}_j \right), \nabla f(\boldsymbol{w}_j) \right\rangle \\ &= 0, \end{aligned}$$

which used the fact that $\nabla f(\boldsymbol{w}_j)$ is constant in the σ-field $\mathcal{F}_j$. This yields the expectation bound.

For the high probability bound, by Azuma-Hoeffding, with probability at least $1 - \delta$,

$$\begin{aligned} \sum_{j=1}^{T} \frac{j+2}{T+2} \varepsilon_j^{(2)} &= \sum_{j=1}^{T} \frac{j+2}{T+2} \left\langle [\nabla f(\boldsymbol{w}_j)]_{S_{(j)}} - [\boldsymbol{g}_j]_{S_{(j)}}, \nabla f(\boldsymbol{w}_j) \right\rangle \\ &\leq \sqrt{2 \ln(1/\delta) \sum_{j=1}^{T} 4G^4 (j+2)^2/(T+2)^2} \\ &\leq 4G^2 \sqrt{T \ln(1/\delta)}. \end{aligned}$$

□

With the proof of Theorem 2 out of the way, the proof of Theorem 1 follows easily.

Proof of Theorem 1. By the assumptions, $\ell(\langle \boldsymbol{w}, \boldsymbol{x}y \rangle)$ is bounded for any $\|\boldsymbol{w}\| \leq D$ and $\|\boldsymbol{x}y\| \leq X$, and since $\boldsymbol{w} \mapsto \ell(\langle \boldsymbol{w}, \boldsymbol{x}y \rangle)$ is continuous, it follows that $\nabla f(\boldsymbol{w}) = \mathbb{E}[\nabla \ell(\langle \boldsymbol{w}, \boldsymbol{x}y \rangle)]$, meaning in particular that

$$\|\nabla f(\boldsymbol{w})\| \leq \|\mathbb{E}[\boldsymbol{x}y\ell'(\langle \boldsymbol{w}, \boldsymbol{x}y \rangle)]\| + \lambda\|\boldsymbol{w}\| \leq X + \lambda D,$$
$$\|\nabla^2 f(\boldsymbol{w})\| \leq \|\mathbb{E}[\boldsymbol{x}y\ell''(\langle \boldsymbol{w}, \boldsymbol{x}y \rangle)\boldsymbol{x}^\top y]\| + \lambda \leq X^2 + \lambda.$$

Expanding these Lipschitz and smoothness terms in Theorem 2 gives the desired result. □

B Summary of datasets

Below, n is the number of examples, d is the number of base features, and s is the average number of non-zero base features per example.

Dataset	n	s	d	problem
20news	18845	93.8854	101631	binary
a9a	48841	13.8676	123	binary
abalone	4176	8	8	binary
abalone	4177	7.99952	10	regression
activity	165632	18.5489	20	binary
adult	48842	11.9967	105	binary
bio	145750	73.4184	74	binary
cal-housing	20639	8	8	regression
census	299284	32.0072	401	binary
comp-activ-harder	8191	11.5848	12	regression
covtype	581011	11.8789	54	binary
cup98-target	95411	310.982	10825	binary
eeg-eye-state	14980	13.9901	14	binary
ijcnn1	24995	13	22	binary
kdda	8407751	36.349	19306083	binary
kddcup2009	50000	58.4353	71652	binary
letter	20000	15.5807	16	binary
magic04	19020	9.98728	10	binary
maptaskcoref	158546	40.4558	5944	binary
mushroom	8124	22	117	binary
nomao	34465	82.3306	174	binary
poker	946799	10	10	binary
rcv1	781265	75.7171	43001	binary
shuttle	43500	7.04984	9	binary
skin	245057	2.948	3	binary
slice	53500	134.575	384	regression
titanic	2201	3	8	binary
vehv2binary	299254	48.5652	105	binary
w8a	49749	11.6502	300	binary
year	463715	90	90	regression

C Further experimental results

We will show three more sets of results in the appendix. The first set contains a bar plot detailing the performance of the `lin`, `quad` and `cubic` baselines, as well as `apple` with $\alpha \in \{0.125, 0.25, 0.5, 0.75, 1\}$ on all of our 30 datasets. For each method, we present the relative error (2) in Figure 5.

Since the statistical error by itself only tells half the story, we also include a similar plot for relative running times in Figure 6.

Finally, we also want to highlight that despite the competitive statistical performance, our adaptive methods indeed generate a much smaller number of monomials. To this end, we compute the average number of non-zero features per example on all datasets for all methods. These plots are presented in Figure 7, with the same color coding used for algorithms as the previous two plots.

We include tables with all the (non-relative) errors and times in Table 2 and Table 3.

	lin	quad	cubic	apple (1)	apple (0.75)	apple (0.5)	apple (0.25)	apple (0.125)
bio	3.122e-03 2.644e+00	3.122e-03 5.000e+00	3.087e-03 5.965e+02	2.985e-03 3.036e+00	3.053e-03 3.340e+00	2.985e-03 7.516e+00	3.259e-03 1.841e+01	3.156e-03 4.745e+01
a9a	1.510e-01 3.880e-01	1.485e-01 4.320e-01	1.496e-01 4.176e+00	1.488e-01 2.840e-01	1.488e-01 3.000e-01	1.489e-01 4.880e-01	1.478e-01 4.040e-01	1.481e-01 4.080e-01
adult	1.557e-01 3.440e-01	1.529e-01 4.520e-01	1.531e-01 4.252e+00	1.525e-01 2.480e-01	1.525e-01 2.400e-01	1.507e-01 3.680e-01	1.513e-01 3.920e-01	1.474e-01 4.040e-01
titanic	2.182e-01 1.600e-02	2.136e-01 2.000e-02	2.136e-01 9.601e-02	2.136e-01 2.800e-02	2.136e-01 3.600e-02	2.136e-01 4.000e-02	2.136e-01 3.200e-02	2.136e-01 3.600e-02
kdda	1.240e-01 9.492e+01	1.272e-01 3.715e+02	1.253e-01 3.266e+04	1.240e-01 7.629e+01	1.240e-01 6.689e+01	1.240e-01 8.318e+01	1.240e-01 5.768e+01	1.240e-01 8.463e+01
census	4.748e-02 3.068e+00	4.579e-02 7.784e+00	4.651e-02 5.754e+02	4.674e-02 2.200e+00	4.686e-02 2.180e+00	4.698e-02 2.144e+00	4.586e-02 2.532e+00	4.648e-02 3.936e+00
20news	8.119e-02 5.440e-01	8.437e-02 2.303e+01	8.384e-02 5.262e+04	8.437e-02 6.440e-01	9.021e-02 6.480e-01	1.210e-01 8.121e-01	9.525e-02 1.040e+00	7.986e-02 2.112e+00
abalone_bin	2.898e-01 4.400e-02	2.826e-01 4.000e-02	2.719e-01 2.440e-01	2.874e-01 6.000e-02	2.874e-01 4.400e-02	2.766e-01 6.400e-02	2.743e-01 6.800e-02	2.754e-01 1.080e-01
year	1.157e-02 1.261e+01	1.073e-02 6.915e+01	1.113e-02 1.136e+04	1.112e-02 1.529e+01	1.099e-02 2.201e+01	1.083e-02 4.585e+01	1.069e-02 1.189e+02	1.060e-02 4.179e+02
poker	4.555e-01 4.388e+00	4.091e-01 6.736e+00	4.100e-01 2.294e+01	4.119e-01 6.228e+00	4.119e-01 3.836e+00	4.092e-01 6.516e+00	4.099e-01 1.230e+01	4.085e-01 1.657e+01
abalone_reg	8.052e+00 4.000e-02	7.489e+00 4.400e-02	7.107e+00 1.680e-01	7.812e+00 5.600e-02	7.812e+00 5.600e-02	7.690e+00 5.600e-02	7.003e+00 7.200e-02	6.740e+00 8.400e-02
kddcup2009	7.310e-02 7.600e-01	7.310e-02 1.144e+01	7.310e-02 1.213e+03	7.310e-02 9.801e-01	7.310e-02 1.196e+00	7.310e-02 1.208e+00	7.310e-02 1.204e+00	7.310e-02 8.881e-01
covtype	2.450e-01 4.444e+00	2.184e-01 3.116e+00	2.039e-01 2.133e+01	2.331e-01 4.204e+00	2.331e-01 4.488e+00	2.287e-01 2.856e+00	2.261e-01 5.120e+00	2.208e-01 4.192e+00
w8a	1.538e-02 1.320e-01	1.246e-02 4.520e-01	1.337e-02 1.448e+01	1.518e-02 2.760e-01	1.518e-02 3.520e-01	1.417e-02 3.120e-01	1.317e-02 1.480e-01	1.286e-02 1.960e-01
nomao	6.325e-02 5.000e-01	5.063e-02 3.564e+00	5.107e-02 7.038e+02	5.992e-02 6.120e-01	5.701e-02 7.480e-01	5.513e-02 1.280e+00	4.933e-02 2.900e+00	4.904e-02 6.348e+00

Table 2: Errors (first row) and running time in seconds (second row) for each dataset for the baselines and apple variants (first 15 datasets).

	`lin`	`quad`	`cubic`	`apple` (1)	`apple` (0.75)	`apple` (0.5)	`apple` (0.25)	`apple` (0.125)
magic04	2.142e-01 1.040e-01	1.672e-01 1.400e-01	1.638e-01 5.480e-01	1.880e-01 1.960e-01	1.880e-01 1.760e-01	1.801e-01 1.520e-01	1.706e-01 2.600e-01	1.696e-01 3.560e-01
rcv1	4.860e-02 1.627e+01	4.060e-02 1.823e+02	3.701e-02 6.251e+04	4.570e-02 1.895e+01	4.511e-02 1.977e+01	4.438e-02 2.171e+01	4.273e-02 2.401e+01	4.205e-02 3.671e+01
letter	2.273e-01 1.440e-01	1.918e-01 2.040e-01	1.688e-01 1.376e+00	1.872e-01 1.680e-01	1.878e-01 1.920e-01	1.727e-01 3.000e-01	1.670e-01 4.280e-01	1.638e-01 6.800e-01
vehv2binary	3.400e-02 3.364e+00	2.670e-02 9.065e+00	2.505e-02 1.079e+03	8.337e-03 2.880e+00	8.087e-03 4.076e+00	8.772e-03 8.105e+00	1.109e-02 1.825e+01	1.151e-02 4.065e+01
comp	3.409e-03 7.601e-02	2.627e-03 8.000e-02	3.662e-03 3.560e-01	2.587e-03 6.401e-02	2.587e-03 6.400e-02	2.124e-03 1.240e-01	2.038e-03 1.080e-01	2.070e-03 1.560e-01
cal_housing	7.410e-02 9.201e-02	8.651e-02 1.680e-01	1.055e-01 4.000e-01	9.414e-02 1.080e-01	9.414e-02 1.080e-01	9.856e-02 2.120e-01	9.881e-02 2.600e-01	1.183e-01 3.360e-01
cup98	5.697e-02 3.452e+00	3.841e-02 1.342e+02	4.215e-02 1.010e+05	5.646e-02 4.736e+00	5.687e-02 4.180e+00	6.450e-02 4.240e+00	6.780e-02 4.504e+00	5.948e-02 7.084e+00
maptaskcoref	1.087e-01 9.361e-01	7.553e-02 6.412e+00	6.598e-02 7.504e+02	9.997e-02 1.424e+00	9.805e-02 1.424e+00	9.691e-02 1.856e+00	9.338e-02 2.628e+00	8.083e-02 3.240e+00
eeg_eye_state	3.815e-01 1.360e-01	2.573e-01 2.000e-01	2.300e-01 7.960e-01	3.127e-01 1.840e-01	3.127e-01 1.640e-01	2.557e-01 2.040e-01	2.383e-01 2.920e-01	2.136e-01 5.160e-01
activity	1.298e-02 5.560e-01	8.422e-03 9.921e-01	6.762e-03 7.864e+00	8.694e-03 1.116e+00	5.856e-03 1.048e+00	5.977e-03 1.828e+00	4.951e-03 2.304e+00	4.166e-03 4.564e+00
ijcnn1	7.942e-02 2.000e-01	4.501e-02 1.800e-01	3.481e-02 6.960e-01	5.221e-02 2.200e-01	5.221e-02 1.800e-01	4.041e-02 1.840e-01	3.921e-02 3.080e-01	3.901e-02 3.000e-01
shuttle	2.644e-02 1.040e-01	1.391e-02 1.240e-01	9.195e-03 4.160e-01	1.218e-02 2.680e-01	1.218e-02 2.520e-01	5.402e-03 3.000e-01	1.023e-02 3.360e-01	7.931e-03 3.080e-01
slice	7.243e-03 1.600e+00	1.403e-03 3.303e+01	7.797e-04 7.944e+03	5.095e-03 1.972e+00	3.455e-03 2.168e+00	2.302e-03 4.880e+00	1.825e-03 1.609e+01	1.196e-03 3.709e+01
skin	7.421e-02 4.600e-01	1.045e-02 1.148e+00	5.835e-03 1.220e+00	2.165e-02 5.400e-01	2.165e-02 1.220e+00	2.165e-02 1.276e+00	1.585e-02 6.280e-01	7.366e-03 1.384e+00
mushroom	5.723e-02 7.200e-02	5.538e-03 6.800e-02	6.154e-04 2.788e+00	6.646e-02 8.001e-02	9.108e-02 5.600e-02	4.923e-02 6.000e-02	1.538e-02 1.320e-01	1.846e-02 1.760e-01

Table 3: Errors (first row) and running time in seconds (second row) for each dataset for the baselines and `apple` variants (last 15 datasets).

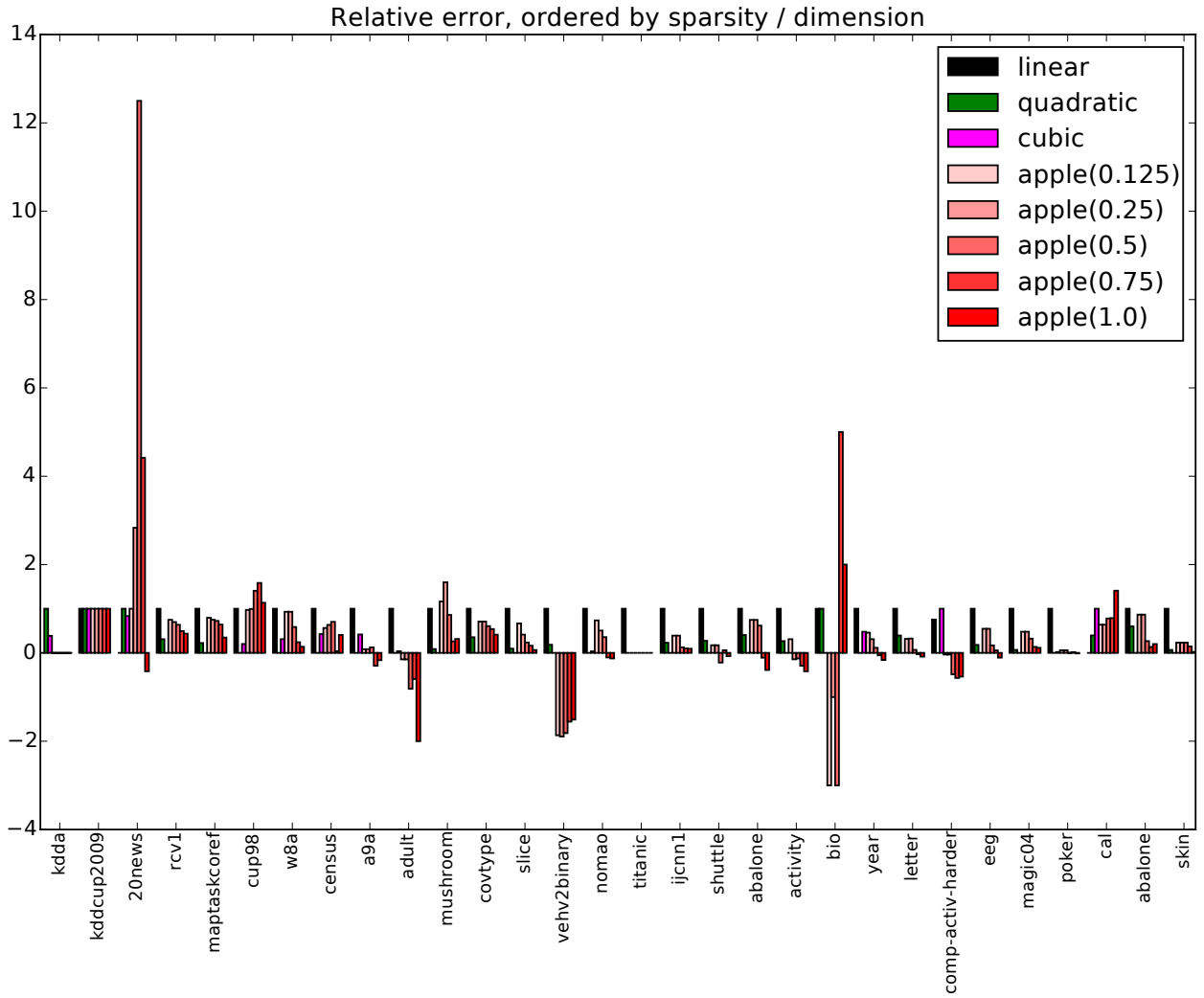

Figure 5: Relative error plots for `apple` and baselines on all 30 datasets. Should be viewed in color.

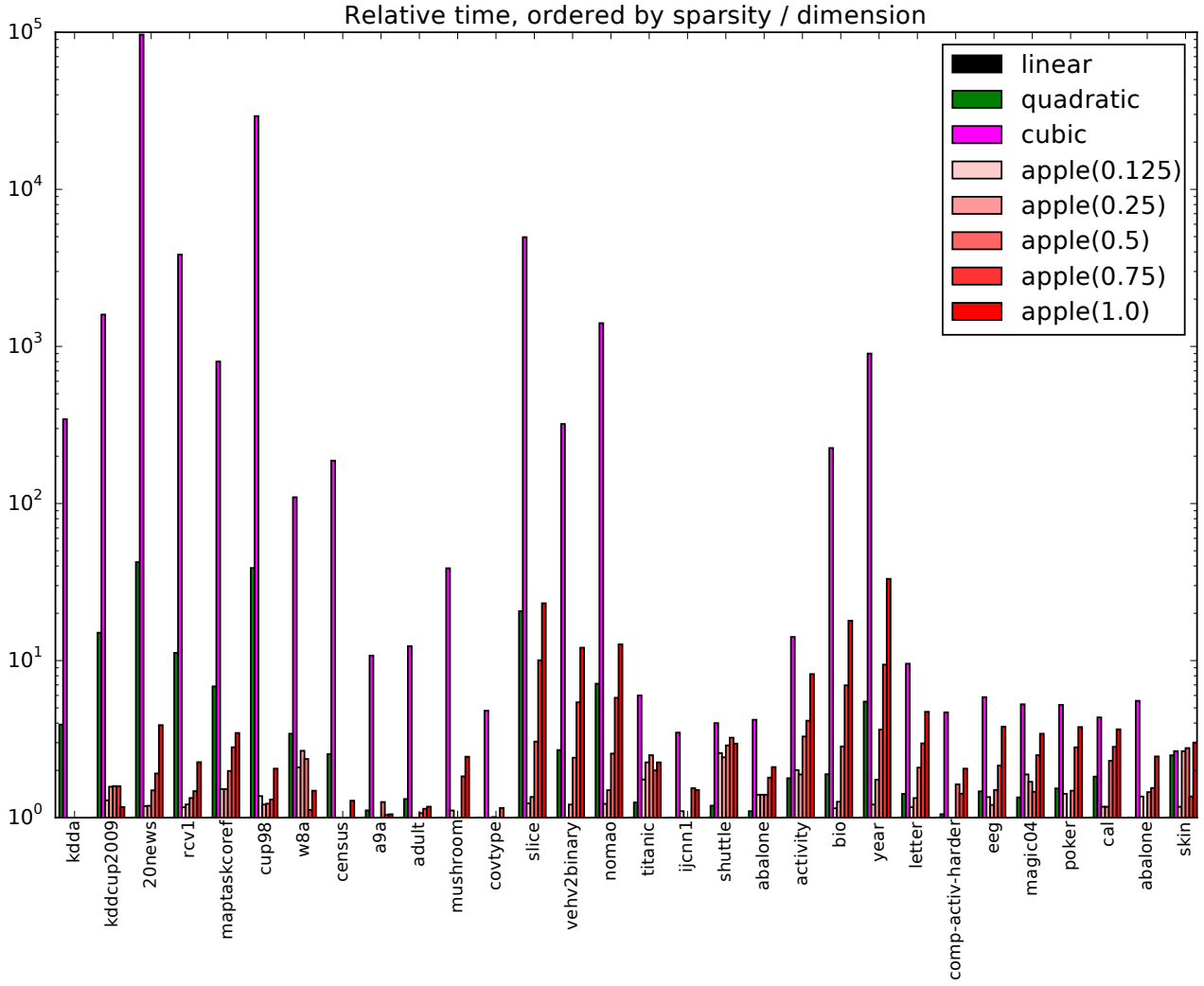

Figure 6: Relative time plots for `apple` and baselines on all 30 datasets. Should be viewed in color.

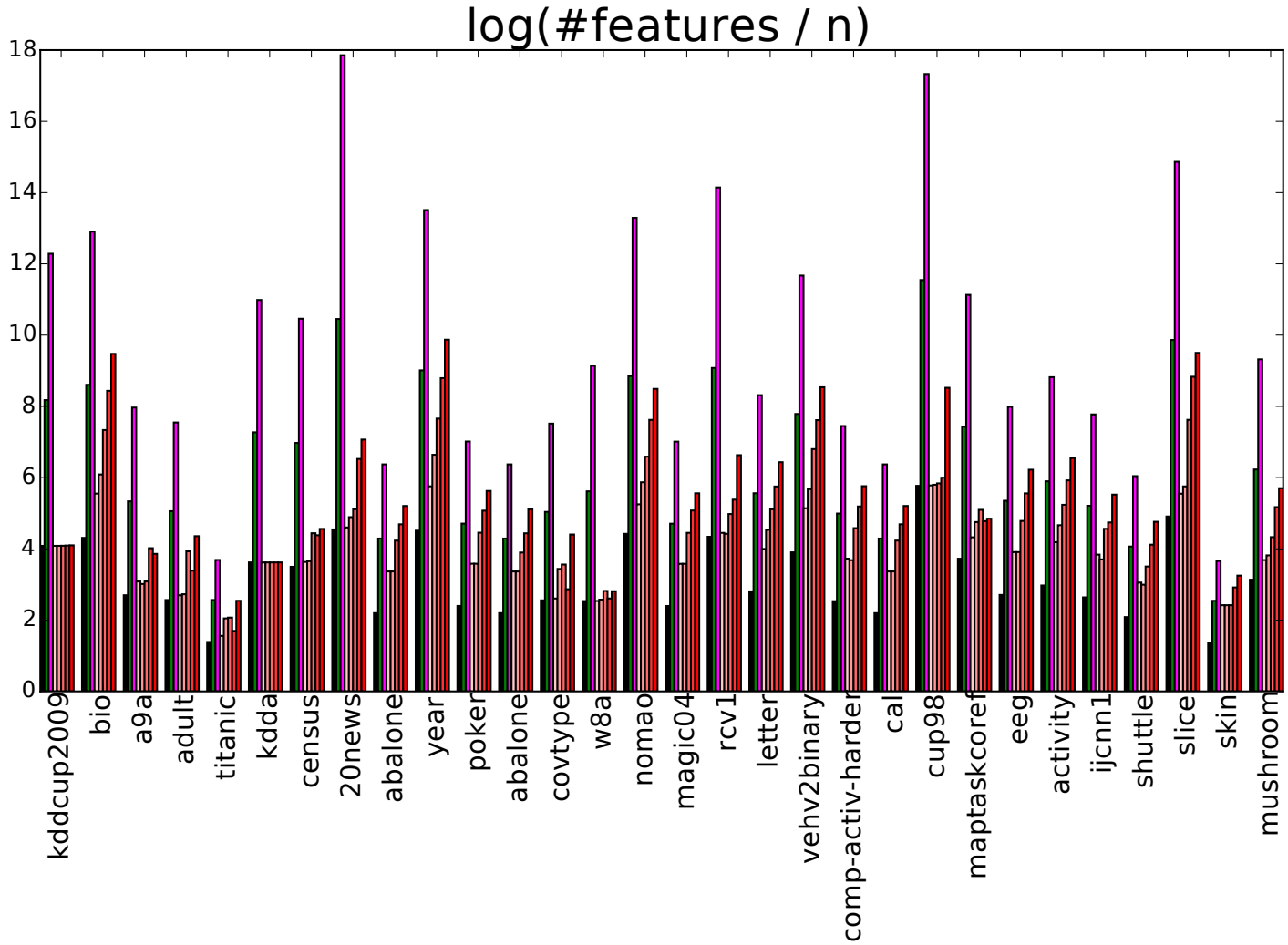

Figure 7: Relative time plots for `apple` and baselines on all 30 datasets. Should be viewed in color.