A Additional Related Works

Two other works [30, 31] are on reducing the dimensionality of the feature space using hashing — note that these works do not consider the combination of gradient descent with dimensionality reduction, and also have a different focus from our work. We note that the work [32] on feature hashing shows that a Countsketch matrix S (with exactly 1 nonzero entry per column) has the JL property when applied to vectors that are flat, i.e., S preserves the norm of x given a bound on $\|x\|_\infty/\|x\|_2$ — we are interested in the opposite case where the weight vector w_* has heavy hitters. [33] studies a more general setting of non-convex optimization/neural networks, but the focus is on compressing the *auxiliary* variables in optimization algorithms such as ADAM, ADAGRAD, etc. [34] is concerned with machine learning problems *directly making use of the compressed weights*, rather than recovering estimates of the optimal weight vectors — in our setting, this corresponds to classification on $\{Rx_t\}_{t\in[T]}$. Though it is not the main focus of our work, in Appendix D we give guarantees on the classification error of our compressed model — our guarantees in Appendix D are somewhat similar to those of [34], though in [34] they also analyze the generalization loss. The work [35] on sketching for logistic regression gives algorithms with poly(d) memory, while our algorithms have memory that is logarithmic in d, and [35] gives $O(1)$- and $O(\log n)$-approximation algorithms for logistic regression, whereas we focus on estimating the coordinates of the optimal weight vector.

B Missing Proofs from Section 2

B.1 Proof of Theorem 2.2

Our proof has the same structure as the batch setting analysis of [1], but we will note the key steps in which it differs. We bound $\|Rw_* - z_*\|_2$ by considering the optimal dual solutions for L and $\widehat{L}$. First, let us compute the dual objective for L. Note that the primal problem can be re-written as $\min_{u,w} \frac{1}{T}\sum_{t=1}^T \ell(u_t) + \frac{\lambda}{2}\|w\|_2^2$ subject to the constraints $u_t = y_t w^T x_t$ for all $t \in [T]$. We can write the Lagrangian as

$$\begin{aligned} F(u,w,\alpha) &= \frac{1}{T}\sum_{t=1}^T \ell(u_t) + \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{T}\sum_{t=1}^T \alpha_t(y_t w^T x_t - u_t) \\ &= \frac{1}{T}\sum_{t=1}^T (\ell(u_t) - \alpha_t u_t) + \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{T}(\widetilde{X}\alpha)^T w \end{aligned} \tag{1}$$

where $\widetilde{X} \in \mathbb{R}^{d\times T}$ such that $\widetilde{X}_t = y_t x_t$ for $t \in [T]$. Thus, the optimal dual solution $\alpha_* \in \mathbb{R}^t$ is the maximizer of

$$\min_{u,w} F(u,w,\alpha) = -\frac{1}{T}\sum_{t=1}^T \ell^*(\alpha_t) - \frac{1}{2\lambda T^2}\alpha^T K \alpha$$

where ℓ^* denotes the Fenchel conjugate of ℓ, and $K = \widetilde{X}^T\widetilde{X}$. Equivalently, α_* is the minimizer of

$$J(\alpha) = \frac{1}{T}\sum_{t=1}^T \ell^*(\alpha_t) + \frac{1}{2\lambda T^2}\alpha^T K \alpha$$

Furthermore, by the Karush-Kuhn-Tucker conditions, $w_* = -\frac{1}{\lambda T}\widetilde{X}\alpha_*$. We can perform a similar analysis for the "sketched" objective: it was shown in [1] that solving the dual of the sketched problem is equivalent to minimizing

$$\widehat{J}(\widehat{\alpha}) = \frac{1}{T}\sum_{t=1}^T \ell^*(\widehat{\alpha}_t) + \frac{1}{2\lambda T^2}\widehat{\alpha}^T \widehat{K}\widehat{\alpha}_*$$

where $\widehat{K} = \widetilde{X}^T R^T R \widetilde{X}$. In addition, if $\widehat{\alpha}_* = \operatorname{argmin} \widehat{J}(\widehat{\alpha})$, then again by the Karush-Kuhn-Tucker conditions, $z_* = -\frac{1}{\lambda T} R\widetilde{X}\widehat{\alpha}_*$. Thus,

$$\|z_* - Rw_*\|_2^2 = \frac{1}{\lambda^2 T^2}\|R\widetilde{X}(\widehat{\alpha}_* - \alpha_*)\|_2^2 = \frac{1}{\lambda^2 T^2}(\widehat{\alpha}_* - \alpha_*)^T \widehat{K}(\widehat{\alpha}_* - \alpha_*) \tag{2}$$

To bound the final quantity in Equation 2, the following lemmas were shown by [1]. Note that they were also shown in the case when R is the Countsketch matrix of [3], and hold regardless of the number of rows R has.

Lemma B.1 ([1])**.**

$$\frac{1}{\lambda T^2}(\widehat{\alpha}_* - \alpha_*)^T \widehat{K}(\widehat{\alpha}_* - \alpha_*) \leq \frac{1}{\lambda T^2}(\alpha_* - \widehat{\alpha}_*)^T(\widehat{K} - K)\alpha_* \leq \frac{1}{T}\|\widehat{\alpha}_* - \alpha_*\|_1 \|\Delta\|_\infty \tag{3}$$

where $\Delta := \frac{1}{\lambda T}(\widehat{K} - K)\alpha$.

Lemma B.2 ([1])**.** $\|\widehat{\alpha}_* - \alpha_*\|_1 \leq 2T\beta\|\Delta\|_\infty$

Combining the above two lemmas with Equation 2, we can conclude that

$$\|z_* - Rw_*\|_2^2 \leq \frac{2\beta}{\lambda}\|\Delta\|_\infty^2 \tag{4}$$

Now, to obtain the ℓ_2 guarantee, we bound $\|\Delta\|_\infty$ more carefully than [1], using Theorem 2.1, as follows (see Remark B.5 for a comparison of our argument to that of [1]):

Lemma B.3. *For $F \in (0,1)$, if R is a sparse JL matrix with $k = \Theta(\frac{\log(dT/\delta)}{F^2})$ rows, then with probability $1-\delta$, the following hold simultaneously for all $i, j \in [d]$ and $t \in [T]$:*

1. $|\langle e_i, w_*\rangle - \langle R_i, Rw_*\rangle| \leq F\|w_*\|_2$
2. $|\langle Rx_t, Rw_*\rangle - \langle x_t, w_*\rangle| \leq F\|x_t\|_2\|w_*\|_2$
3. $|\langle R_i, R_j\rangle - \langle e_i, e_j\rangle| \leq F$
4. $\|Rx_t\|_2 = (1 \pm F)\|x_t\|_2 \leq 1 + F$
5. $\|Rw_*\|_2 = (1 \pm F)\|w_*\|_2$

Proof. First, let $u, v \in \mathbb{R}^d$ with $\|u\|_2 = \|v\|_2 = 1$. If R has $\Theta(\frac{1}{F^2}\log(\frac{1}{\delta'}))$ rows, then with probability $1-\delta'$, R preserves the ℓ_2 norms of $\mathbf{u}+\mathbf{v}$ and $\mathbf{u}-\mathbf{v}$ up to a factor of $1 \pm F$, by Theorem 2.1. Thus,

$$\begin{aligned}
\langle R\mathbf{u}, R\mathbf{v}\rangle &= \frac{\|R(\mathbf{u}+\mathbf{v})\|_2^2 - \|R(\mathbf{u}-\mathbf{v})\|_2^2}{4} \\
&= \frac{(1\pm F)\|\mathbf{u}+\mathbf{v}\|_2^2 - (1\pm F)\|\mathbf{u}-\mathbf{v}\|_2^2}{4} \\
&= \frac{\|\mathbf{u}+\mathbf{v}\|_2^2 - \|\mathbf{u}-\mathbf{v}\|_2^2}{4} \pm F \cdot \frac{\|\mathbf{u}+\mathbf{v}\|_2^2 + \|\mathbf{u}-\mathbf{v}\|_2^2}{4} \\
&= \langle \mathbf{u}, \mathbf{v}\rangle \pm O(F)
\end{aligned} \tag{5}$$

where the last inequality is because $\|\mathbf{u}+\mathbf{v}\|_2, \|\mathbf{u}-\mathbf{v}\|_2 = O(1)$. Thus, in general, if $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ and R has $\Theta(\frac{1}{F^2}\log(\frac{1}{\delta'}))$ rows, then with probability $1-\delta'$, $\langle R\mathbf{u}, R\mathbf{v}\rangle = \langle \mathbf{u}, \mathbf{v}\rangle \pm O(F\|\mathbf{u}\|_2\|\mathbf{v}\|_2)$. Finally, in order for the conclusions of Lemma B.3 to hold, it suffices to have $\delta' = \frac{\delta}{d+T+d^2+T+1} = \Theta(\frac{\delta}{T+d^2})$, since we can then union bound over the relevant pairs of points in $\mathbb{R}^d$. □

Lemma B.3 allows us to show the following bound on $\|\Delta\|_\infty$:

Lemma B.4. $\|\Delta\|_\infty \leq F\|w_*\|_2$

Proof. We can write

$$\Delta = \frac{1}{\lambda T}(\widehat{K} - K)\alpha_* = \frac{1}{\lambda T}(\widetilde{X}^T R^T R\widetilde{X} - \widetilde{X}^T\widetilde{X})\alpha_* = -\widetilde{X}^T(R^T R - I)w_* \tag{6}$$

since $w_* = -\frac{1}{\lambda T}\widetilde{X}\alpha_*$. Thus,

$$\begin{aligned}
\|\Delta\|_\infty &= \max_{t\in[T]} |y_t x_t^T(R^T R - I)w_*| \\
&= \max_{t\in[T]} |\langle Rx_t, Rw_*\rangle - \langle x_t, w_*\rangle| \\
&\leq \max_{t\in[T]} F\|x_t\|_2\|w_*\|_2 \\
&\leq F\|w_*\|_2
\end{aligned} \tag{7}$$

where the first inequality is by Property (3) of Lemma B.3. □

Remark B.5. *Instead of Lemma B.3, [1] only used the fact that $|\langle R_i, R_j\rangle - \langle e_i, e_j\rangle| \le F$, as long as R has enough rows. Once that fact is shown, then [1] used it to show that $\|\Delta\|_\infty \le O(F\|w_*\|_1)$ instead, using an argument which the proof of our Lemma B.4 is based on.*

Thus, by Equation 4 and Lemma B.4,

$$\|z_* - Rw_*\|_2^2 \le \frac{2\beta}{\lambda}\|\Delta\|_\infty^2 \le \frac{2\beta F^2}{\lambda}\|w_*\|_2^2 \tag{8}$$

and taking square roots gives

$$\|z_* - Rw_*\|_2 \le \sqrt{\frac{2\beta}{\lambda}}F\|w_*\|_2$$

In particular, if we let $F = \varepsilon\sqrt{\frac{\lambda}{2\beta}}$, then we get the desired result. This completes the proof.

B.2 Bound on $\|w_*\|_2$ and $\|z_*\|_2$

In this section we give an auxiliary lemma which we use in later proofs:

Lemma B.6. *$\|w_*\|_2 \le H/\lambda$. In addition, if R is a sketching matrix for which the conclusion of Lemma B.3 holds (e.g. if R is a sparse JL matrix with $k = \Theta(\frac{\log(dT/\delta)}{F^2})$ rows for $F \in (0,1)$), then $\|z_*\|_2 \le O(H/\lambda)$.*

Proof. Recall that

$$w_* = \mathrm{argmin}_{w\in\mathbb{R}^d}\frac{1}{T}\sum_{t=1}^{T}\ell(y_t w^T x_t) + \frac{\lambda}{2}\|w\|_2^2 =: L(w)$$

Note that L is $\beta+\lambda$-smooth, since ℓ is β-smooth and $\|x_t\|_2 \le 1$. Thus, by Theorem 3.3 of [36], if w_k' is the k^{th} iterate of gradient descent on L (at an arbitrary initialization), then $L(w_k') \to L(w_*)$ as $k \to \infty$. Moreover, since L is λ-strongly convex, this implies that $w_k' \to w_*$ as $k \to \infty$. However, the gradient descent update to w_k' can be written as follows:

$$w_{k+1}' \leftarrow (1-\lambda\eta_k)w_k' - \eta_k \cdot \frac{1}{T}\sum_{t=1}^{T}\ell'(y_t(w_k')^T x_t)\cdot y_t x_t$$

Since ℓ is H-Lipschitz and $\|x_t\|_2 \le 1$ for all $t \in [T]$, we can show by induction that $\|w_{k+1}'\|_2 \le \frac{H}{\lambda}$, and taking the limit as $k \to \infty$ gives $\|w_*\|_2 \le \frac{H}{\lambda}$. We can similarly show that $\|z_*\|_2 \le O(\frac{H}{\lambda})$, with the only change in the proof being the observation that $\|Rx_t\|_2 \le O(1)$ by Lemma B.3. □

B.3 Proof of Theorem 2.6

Throughout the proof of this theorem, R is a sparse JL matrix with $O(\frac{\log(dT/\delta)}{F^2})$ rows — we will then choose $F \in (0,1)$ appropriately at the end of the proof. As before, $w_* = \mathrm{argmin}_{w\in\mathbb{R}^d}\frac{1}{T}\sum_{t=1}^{T}\ell(y_t w^T x_t) + \frac{\lambda}{2}\|w\|_2^2$, and z_* is defined analogously for the corresponding sketched problem.

Lemma B.7. *For all $t \ge 1$,*

$$\begin{aligned}\|w_{t+1} - \widehat{w_{t+1}}\|_2 \le{}& (1-\lambda\eta_t)\|w_t - \widehat{w_t}\|_2 \\ &+ \eta_t\beta\Big(\|w_t - w_*\|_2 + O(1)\|z_t - z_*\|_2 + C_{\beta,\lambda}F\|w_*\|_2\Big)\end{aligned} \tag{9}$$

Proof. Recall that w_t and $\widehat{w_t}$ are updated respectively according to

$$w_{t+1} \leftarrow (1-\lambda\eta_t)w_t - \eta_t y_t \ell'(y_t w_t^T x_t)x_t$$

and

$$\widehat{w_{t+1}} \leftarrow (1-\lambda\eta_t)\widehat{w_t} - \eta_t y_t \ell'(y_t z_t^T R x_t) x_t$$

Therefore, by the triangle inequality,

$$\begin{aligned} \|w_{t+1} - \widehat{w_{t+1}}\|_2 &\le (1-\lambda\eta_t)\|w_t - \widehat{w_t}\|_2 + \eta_t |\ell'(y_t w_t^T x_t) - \ell'(y_t z_t^T R x_t)| \|x_t\|_2 \\ &\le (1-\lambda\eta_t)\|w_t - \widehat{w_t}\|_2 + \eta_t |\ell'(y_t w_t^T x_t) - \ell'(y_t z_t^T R x_t)| \end{aligned} \tag{10}$$

where the second inequality is because $\|x_t\|_2 \le 1$. Now, let us bound the difference of $\ell'(y_t w_t^T x_t)$ and $\ell'(y_t z_t^T R x_t)$: since ℓ is β-smooth,

$$|\ell'(y_t w_t^T x_t) - \ell'(y_t z_t^T R x_t)| \le \beta |\langle w_t, x_t\rangle - \langle z_t, R x_t\rangle| \tag{11}$$

We can bound the difference of inner products on the right-hand side using the fact that R is a JL matrix:

$$\begin{aligned} |\langle w_t, x_t\rangle - \langle z_t, R x_t\rangle| \le{}& |\langle w_t, x_t\rangle - \langle w_*, x_t\rangle| + |\langle w_*, x_t\rangle - \langle z_*, R x_t\rangle| \\ &+ |\langle z_*, R x_t\rangle - \langle z_t, R x_t\rangle| \end{aligned} \tag{12}$$

The first term on the right-hand side of (12) can be bounded as

$$|\langle w_t, x_t\rangle - \langle w_*, x_t\rangle| = |\langle w_t - w_*, x_t\rangle| \le \|w_t - w_*\|_2 \|x_t\|_2 \le \|w_t - w_*\|_2$$

where the last inequality is because $\|x_t\|_2 \le 1$. Similarly, the third term can be bounded as

$$|\langle z_*, R x_t\rangle - \langle z_t, R x_t\rangle| = |\langle z_* - z_t, R x_t\rangle| \le \|z_t - z_*\|_2 \|R x_t\|_2 \le O(1)\|z_t - z_*\|_2$$

where the last inequality is by Lemma B.3. Finally, we can bound the second term as follows:

$$\begin{aligned} |\langle w_*, x_t\rangle - \langle z_*, R x_t\rangle| &\le |\langle w_*, x_t\rangle - \langle R w_*, R x_t\rangle| + |\langle R w_*, R x_t\rangle - \langle z_*, R x_t\rangle| \\ &\le F\|x_t\|_2 \|w_*\|_2 + O(1)\|z_* - R w_*\|_2 \\ &\le F\|w_*\|_2 + O(1)\|z_* - R w_*\|_2 \\ &\le O\Big(1 + \sqrt{\frac{\beta}{\lambda}}\Big) F \|w_*\|_2 \end{aligned} \tag{13}$$

Here in the second inequality, we bounded the first term according to property (2) of Lemma B.3, and the second term according to property (4) of Lemma B.3, together with the Cauchy-Schwarz inequality. The third inequality is because $\|x_t\|_2 \le 1$. Finally, the fourth inequality is because, by the proof of Theorem 2.2 in the previous section, $\|z_* - R w_*\|_2 \le \sqrt{\frac{2\beta}{\lambda}} F \|w_*\|_2$ when R has $O(\frac{\log(dT/\delta)}{F^2})$ rows. Thus, for convenience, if we define $C_{\beta,\lambda} = O\Big(1 + \sqrt{\frac{\beta}{\lambda}}\Big)$, then

$$|\langle w_t, x_t\rangle - \langle z_t, R x_t\rangle| \le \|w_t - w_*\|_2 + O(1)\|z_t - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2$$

In summary,

$$\begin{aligned} \|w_{t+1} - \widehat{w_{t+1}}\|_2 \le{}& (1-\lambda\eta_t)\|w_t - \widehat{w_t}\|_2 \\ &+ \eta_t \beta \Big(\|w_t - w_*\|_2 + O(1)\|z_t - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2 \Big) \end{aligned} \tag{14}$$

□

From the above lemma, we can obtain the following non-recursive bound on $\|w_t - \widehat{w_t}\|_2$:

Lemma B.8. *For all $t \in \mathbb{N}$,*

$$\|w_t - \widehat{w_t}\|_2 \le \sum_{s=1}^{t-1} \eta_s \beta \Big(\|w_s - w_*\|_2 + O(1)\|z_s - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2 \Big) \prod_{j=s+1}^{t-1} (1 - \lambda\eta_j)$$

where the product $\prod_{j=s+1}^{t-1}(1-\lambda\eta_j)$ is defined to be 1 if $s+1 > t-1$.

Proof. We proceed by induction on t. For $t = 1$, this trivially holds because $w_t = \widehat{w_t} = 0$. Now, suppose that for some $t \in \mathbb{N}$,

$$\|w_t - \widehat{w_t}\|_2 \le \sum_{s=1}^{t-1} \eta_s \beta \Big(\|w_s - w_*\|_2 + O(1)\|z_s - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2 \Big) \prod_{j=s+1}^{t-1} (1 - \lambda \eta_j)$$

Then,

$$\begin{aligned}
&\|w_{t+1} - \widehat{w_{t+1}}\|_2 \\
&\quad \le (1 - \lambda\eta_t)\|w_t - \widehat{w_t}\|_2 + \eta_t \beta \Big(\|w_t - w_*\|_2 + O(1)\|z_t - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2 \Big) \\
&\quad \le (1 - \lambda\eta_t) \sum_{s=1}^{t-1} \eta_s \beta \Big(\|w_s - w_*\|_2 + O(1)\|z_s - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2 \Big) \prod_{j=s+1}^{t-1} (1 - \lambda\eta_j) \\
&\qquad + \eta_t \beta \Big(\|w_t - w_*\|_2 + O(1)\|z_t - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2 \Big) \\
&\quad = \sum_{s=1}^{t-1} \eta_s \beta \Big(\|w_s - w_*\|_2 + O(1)\|z_s - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2 \Big) \prod_{j=s+1}^{t} (1 - \lambda\eta_j) \\
&\qquad + \eta_t \beta \Big(\|w_t - w_*\|_2 + O(1)\|z_t - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2 \Big) \\
&\quad = \sum_{s=1}^{t} \eta_s \beta \Big(\|w_s - w_*\|_2 + O(1)\|z_s - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2 \Big) \prod_{j=s+1}^{t} (1 - \lambda\eta_j)
\end{aligned} \tag{15}$$

Here, the first inequality is by Lemma B.7, and the second inequality is by the inductive hypothesis. The last equality is because for $s = t$, $\prod_{j=s+1}^{t}(1 - \lambda\eta_j)$ is simply equal to 1. This completes the induction. □

We can use the above lemma to bound $\|\overline{w_T} - \overline{\widehat{w_T}}\|_2$, as follows. In the following let $S_{a,b} = \sum_{r=a}^{b} -\lambda\eta_r$.

$$\begin{aligned}
&\|\overline{w_T} - \overline{\widehat{w_T}}\|_2 \\
&\quad \le \frac{1}{T} \sum_{t=1}^{T} \|w_t - \widehat{w_t}\|_2 \\
&\quad \le \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{t-1} \eta_s \beta \Big(\|w_s - w_*\|_2 + O(1)\|z_s - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2 \Big) \prod_{j=s+1}^{t-1} (1 - \lambda\eta_j) \\
&\quad \le \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{t-1} \eta_s \beta \Big(\|w_s - w_*\|_2 + O(1)\|z_s - z_*\|_2 + C_{\beta,\lambda} F \|w_*\|_2 \Big) e^{S_{s+1,t-1}}
\end{aligned} \tag{16}$$

where the first inequality is by the triangle inequality and the second is by Lemma B.8. The third inequality is by $1 + x \le e^x$.

Let us now break up the above sum, into the following three sums: define

$$S_1 = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{t-1} \eta_s \beta \|w_s - w_*\|_2 e^{S_{s+1,t-1}}$$

and

$$S_2 = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{t-1} \eta_s \beta \|z_s - z_*\|_2 e^{S_{s+1,t-1}}$$

and

$$S_3 = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{t-1} \eta_s \beta C_{\beta,\lambda} F \|w_*\|_2 e^{S_{s+1,t-1}}$$

Then, clearly

$$\|\overline{w_T} - \overline{\widehat{w_T}}\|_2 \leq O(1)(S_1 + S_2 + S_3)$$

Let us first bound S_1. Recall that $\eta_j = D/(G\sqrt{j})$. Observe that we can take the diameter D of the constraint set to be $O(\frac{H}{\lambda})$, since by Lemma B.6, $\|z_*\|_2, \|w_*\|_2 \leq O(H/\lambda)$. Moreover, we do not need to apply the projection step in online gradient descent to w_t or z_t, since by induction on t, $\|w_t\|_2 \leq H/\lambda$ and $\|z_t\|_2 \leq O(H/\lambda)$ even without a projection step (by the gradient update, we have the inequality $\|w_{t+1}\|_2 \leq (1-\lambda\eta_t)\|w_t\|_2 + \eta_t|\ell'(y_t w_t^T x_t)|\|x_t\|_2 \leq (1-\lambda\eta_t)\|w_t\|_2 + \eta_t H$, and a similar inequality holds for z_t). In addition, for any $w \in \mathbb{R}^n$, since the loss function on the t^{th} iteration is $f_t(w) = \ell(y_t w^T x_t) + \frac{\lambda}{2}\|w\|_2^2$,

$$\nabla f_t(w) = \ell'(y_t w^T x_t) \cdot (y_t x_t) + \lambda w$$

meaning that

$$\|\nabla f_t(w)\|_2 \leq H + \lambda\|w\|_2 \leq 2H$$

and thus, we can take $G = 2H$. Therefore, for $s < t$,

$$\begin{aligned} S_{s+1,t-1} &= \sum_{r=s+1}^{t-1} -\lambda\eta_r \\ &= \sum_{r=s+1}^{t-1} -\lambda \cdot \frac{D}{G\sqrt{r}} \\ &\leq \sum_{r=s+1}^{t-1} -\lambda \cdot \frac{H}{(2H)\cdot\lambda\sqrt{r}} \\ &= -\frac{1}{2}\sum_{r=s+1}^{t-1} \frac{1}{\sqrt{r}} \\ &\leq \sqrt{s} - \sqrt{t} + 1 \end{aligned} \tag{17}$$

Here, for the last inequality, note that $\sum_{r=s+1}^{t-1} \frac{1}{\sqrt{r}} \geq \int_{s+1}^{t} \frac{1}{\sqrt{x}} dx = 2(\sqrt{t} - \sqrt{s+1})$. In summary,

$$e^{S_{s+1,t-1}} \leq Ce^{\sqrt{s}-\sqrt{t}}$$

for some absolute constant C, meaning that

$$S_1 \leq \frac{C\beta}{T}\sum_{t=1}^{T}\sum_{s=1}^{t-1} \eta_s \|w_s - w_*\|_2 e^{\sqrt{s}-\sqrt{t}} \leq \frac{C\beta}{\lambda T}\sum_{t=1}^{T}\sum_{s=1}^{t-1} \frac{1}{\sqrt{s}}\|w_s - w_*\|_2 e^{\sqrt{s}-\sqrt{t}}$$

Next, we switch the order of summation, to obtain

$$\begin{aligned} S_1 &\leq \frac{C\beta}{\lambda T}\sum_{t=1}^{T}\sum_{s=1}^{t-1} \frac{1}{\sqrt{s}} e^{\sqrt{s}-\sqrt{t}}\|w_s - w_*\|_2 \\ &= \frac{C\beta}{\lambda T}\sum_{s=1}^{T}\sum_{t=s+1}^{T} \frac{1}{\sqrt{s}} e^{\sqrt{s}-\sqrt{t}}\|w_s - w_*\|_2 \\ &= \frac{C\beta}{\lambda T}\sum_{s=1}^{T} \frac{1}{\sqrt{s}} e^{\sqrt{s}}\|w_s - w_*\|_2 \sum_{t=s+1}^{T} e^{-\sqrt{t}} \end{aligned} \tag{18}$$

where in the first equality above, we switched the order of summation. We bound the innermost sum above:

$$\begin{aligned}\sum_{t=s+1}^{T} e^{-\sqrt{t}} &\le \int_{s}^{\infty} e^{-\sqrt{x}} dx \\ &= \int_{\sqrt{s}}^{\infty} 2ye^{-y} dy \\ &= \Big(-2ye^{-y} - 2e^{-y} \Big)\Big|_{\sqrt{s}}^{\infty} \\ &= 2\sqrt{s}e^{-\sqrt{s}} + 2e^{-\sqrt{s}} \\ &\le O(\sqrt{s}e^{-\sqrt{s}})\end{aligned} \tag{19}$$

Here in the first equality, we used the change of variables $x = y^2$. For the second equality, note that the antiderivative of $2ye^{-y}$ is $-2ye^{-y} - 2e^{-y}$. Thus,

$$S_1 \le \frac{C\beta}{\lambda T} \sum_{s=1}^{T} \frac{e^{\sqrt{s}}}{\sqrt{s}} \|w_s - w_*\|_2 \cdot O(\sqrt{s}e^{-\sqrt{s}}) = O\Big(\frac{\beta}{\lambda T}\Big) \sum_{s=1}^{T} \|w_s - w_*\|_2$$

Finally, the right-hand side expression can be bounded as follows:

Lemma B.9.

$$\frac{1}{T} \sum_{s=1}^{T} \|w_s - w_*\|_2 \le O\Big(\frac{H}{\lambda T^{1/4}}\Big)$$

Proof. Observe that by online gradient descent regret bounds (e.g. [22], Theorem 3.1),

$$\frac{3}{2} GD\sqrt{T} \ge \sum_{s=1}^{T} (f_t(w_s) - f_t(w_*)) \ge \sum_{s=1}^{T} \frac{\lambda}{2} \|w_s - w_*\|_2^2$$

where the first inequality is Theorem 3.1 of [22], and the second is because the f_t are λ-strongly convex. (Note that we can apply the online gradient descent regret bounds since $\|w_*\|_2 \le H/\lambda$ and $\|w_t\|_2 \le H/\lambda$, meaning we can simply let $D = H/\lambda$ as discussed above, and do not need to apply the projection step to w_t.) Thus, by the Cauchy-Schwarz inequality,

$$\Big(\sum_{s=1}^{T} \|w_s - w_*\|_2 \Big)^2 \le T \Big(\sum_{s=1}^{T} \|w_s - w_*\|_2^2 \Big) \le T \cdot \frac{2}{\lambda} \cdot \frac{3}{2} GD\sqrt{T} = \frac{3GDT^{3/2}}{\lambda}$$

Taking square roots gives

$$\sum_{s=1}^{T} \|w_s - w_*\|_2 \le \sqrt{\frac{3GD}{\lambda}} T^{3/4}$$

and dividing by T gives

$$\frac{1}{T} \sum_{s=1}^{T} \|w_s - w_*\|_2 \le \sqrt{\frac{3GD}{\lambda}} \cdot \frac{1}{T^{1/4}} \le O\Big(\frac{H}{\lambda T^{1/4}}\Big)$$

where the last inequality is because $D \le \frac{H}{\lambda}$ and $G \le O(H)$. This proves the lemma. □

Thus, in summary,

$$S_1 \le O\Big(\frac{\beta}{\lambda T}\Big) \sum_{s=1}^{T} \|w_s - w_*\|_2 \le O\Big(\frac{\beta H}{\lambda^2 T^{1/4}}\Big)$$

By a similar argument, we can show that

$$S_2 \le O\Big(\frac{\beta H}{\lambda^2 T^{1/4}}\Big)$$

with the only differences being that Lemma B.9 holds for $\frac{1}{T}\sum_{s=0}^{T}\|z_s - z_*\|_2$ as well, and we now have $D = O(H/\lambda)$ and $G = O(H)$.

Finally, let us bound S_3:

$$\begin{aligned}
S_3 &= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{t-1}\eta_s\beta C_{\beta,\lambda}F\|w_*\|_2 e^{S_{s+1,t-1}} \\
&= \frac{\beta C_{\beta,\lambda}F\|w_*\|_2}{T}\sum_{t=1}^{T}\sum_{s=1}^{t-1}\eta_s e^{S_{s+1,t-1}} \\
&= O\Big(\frac{\beta C_{\beta,\lambda}F\|w_*\|_2}{T}\Big)\sum_{t=1}^{T}\sum_{s=1}^{t-1}\frac{1}{\lambda\sqrt{s}}e^{\sqrt{s}-\sqrt{t}} \\
&= O\Big(\frac{\beta C_{\beta,\lambda}F\|w_*\|_2}{\lambda T}\Big)\sum_{t=1}^{T}\sum_{s=1}^{t-1}\frac{1}{\sqrt{s}}e^{\sqrt{s}-\sqrt{t}} \\
&= O\Big(\frac{\beta C_{\beta,\lambda}F\|w_*\|_2}{\lambda T}\Big)\sum_{s=1}^{T}\sum_{t=s+1}^{T}\frac{1}{\sqrt{s}}e^{\sqrt{s}-\sqrt{t}} \\
&= O\Big(\frac{\beta C_{\beta,\lambda}F\|w_*\|_2}{\lambda T}\Big)\sum_{s=1}^{T}\frac{1}{\sqrt{s}}e^{\sqrt{s}}\sum_{t=s+1}^{T}e^{-\sqrt{t}} \\
&= O\Big(\frac{\beta C_{\beta,\lambda}F\|w_*\|_2}{\lambda T}\Big)\sum_{s=1}^{T}\frac{1}{\sqrt{s}}e^{\sqrt{s}}O(\sqrt{s}e^{-\sqrt{s}}) \\
&= O\Big(\frac{\beta C_{\beta,\lambda}F\|w_*\|_2}{\lambda}\Big)
\end{aligned} \tag{20}$$

In summary,

$$\|\overline{w_T} - \overline{\widehat{w_T}}\|_2 \le O(\beta C_{\beta,\lambda}F\|w_*\|_2/\lambda) + O\Big(\frac{\beta H}{\lambda^2 T^{1/4}}\Big)$$

To complete the proof, we bound $\|\overline{w_T} - w_*\|_2$, again using regret bounds:

$$\|\overline{w_T} - w_*\|_2 = \Big\|\frac{1}{T}\sum_{t=1}^{T}w_t - w_*\Big\|_2 \le \frac{1}{T}\sum_{t=1}^{T}\|w_t - w_*\|_2 \le O\Big(\frac{H}{\lambda T^{1/4}}\Big)$$

where the first inequality is by the triangle inequality, and the second is by Lemma B.9. Thus,

$$\|\overline{\widehat{w_T}} - w_*\|_2 \le O(\beta C_{\beta,\lambda}F\|w_*\|_2/\lambda) + O\Big(\frac{\beta H}{\lambda^2 T^{1/4}}\Big) + O\Big(\frac{H}{\lambda T^{1/4}}\Big)$$

Now, we must determine what F and T must be for the right-hand side to be at most $\varepsilon\|w_*\|_2$. For the first term to be at most $\varepsilon\|w_*\|_2$, we must have $\beta C_{\beta,\lambda}F/\lambda \le O(\varepsilon)$ — in other words, $F \le O(\frac{\lambda\varepsilon}{\beta}\cdot\frac{1}{C_{\beta,\lambda}}) = \Theta(\frac{\lambda\varepsilon}{\beta}\cdot\min(1,\sqrt{\lambda/\beta}))$. For the second term to be at most $\varepsilon\|w_*\|_2$, we must have $\frac{\beta H}{\lambda^2 T^{1/4}} \le \varepsilon\|w_*\|_2$, and for this to occur, it is sufficient to have $T \ge \Omega((\beta^4 H^4)/(\lambda^8\varepsilon^4\tau^4))$. Finally, for the third term to be at most $\varepsilon\|w_*\|_2$, it suffices to have $T \ge \Omega(H^4/(\lambda^4\varepsilon^4\tau^4))$. This completes the proof.

B.4 Proof of Theorem 2.7

This is a corollary of Theorem 2.6. First, note that by Theorem 2.6, if R is a JL matrix with $O(\lambda^{-2}\varepsilon^{-2}\beta^2\log(dT/\delta)\max(1,\beta/\lambda))$ rows, then $\|\overline{\widehat{w_T}} - w_*\|_2 \le \varepsilon\|w_*\|_2$ with probability $1-\delta$. In particular, this means that $\|\overline{\widehat{w_T}} - w_*\|_\infty \le \varepsilon\|w_*\|_2$. Thus, first of all, if we have an ℓ_2 point query algorithm that gives an estimate $v \in \mathbb{R}^d$ such that $\|v - \overline{\widehat{w_T}}\|_\infty \le O(\varepsilon\|\overline{\widehat{w_T}}\|_2) \le O(\varepsilon\|w_*\|_2)$, then $\|v - w_*\|_\infty \le O(\varepsilon\|w_*\|_2)$. This also implies that $i \in [d]$ is an ℓ_2 heavy hitter of $\overline{\widehat{w_T}}$ if and only if it is an ℓ_2 heavy hitter of w_*, since $|\overline{\widehat{w_T}}_i| \ge \Theta(\varepsilon\|\overline{\widehat{w_T}}\|_2)$ if and only if $|w_{*,i}| \ge \Theta(\varepsilon\|w_*\|_2)$ — this is

because, if $\|\widehat{w_T} - w_*\|_2 \leq c\varepsilon\|w_*\|_2$ for a sufficiently small constant $c > 0$ (with the number of rows in R being increased by $1/c^2$), then $|\widehat{w_T}_i - w_{*,i}| \leq c\varepsilon\|w_*\|_2$, and moreover, $\|\widehat{w_T}\|_2 = \Theta(\|w_*\|_2)$.

Thus, the space complexities follow by adding the turnstile space complexities with the number of rows in R, and the update times follows from adding the time required to compute z_t with the time required to update the point query/heavy hitters data structures (which is $\text{nnz}(x_t)$ multiplied by the the turnstile update time of the point query/heavy hitters data structures). This is because, by the discussion in the previous paragraph, we can simply apply the existing data structures mentioned in Theorems 2.4 and 2.5, getting an additional failure probability of δ for point query and $\frac{1}{\text{poly}(d)}$ for heavy hitters.

C Missing Details and Proofs from Section 3

C.1 Theorem 3.1 - Generating Columns of R on Demand

Let A be an ε-incoherent matrix — then A can be constructed efficiently. Moreover, R does not have to be stored explicitly — instead, each column can be generated on demand. If $m = O(\frac{1}{\varepsilon^2} \cdot (\frac{\log n}{\log\log n + \log 1/\varepsilon})^2)$ is desired, then A can be constructed using codes, as described on page 6 of [6] [5]. For $m = O(\frac{\log n}{\varepsilon^2})$, A is instead a derandomized JL matrix, for which the entries are in $\{\pm 1/\sqrt{m}\}$ [37]. In both cases, the sketching matrices need not be stored, and any single column can be accessed whenever desired. Indeed, in the case of Reed-Solomon codes, one can use fast multipoint evaluation as described in [6] to generate individual entries of a column quickly, compute the product with the input, update the accumulated matrix-vector product, and then reuse the memory for that column. For the derandomized JL matrix given in [37], the argument in [37] gives a small number of seeds to enumerate over. Once you have found the seed, which can be found in a preprocessing step applied to the standard basis vectors $e_1, \ldots, e_d$, one can generate each entry of any desired column of the sketching matrix using Nisan's pseudorandom generator in $\text{poly}(\log d)$ time, and then generate the entire column, compute the product with the input, update the accumulated matrix-vector product, and then re-use the memory for that column.

C.2 Proof of Theorem 3.2

We first show that a good estimate to each of the coordinates to w_* can be obtained by applying the recovery procedure in Algorithm 2 to z_* (i.e. we analyze our algorithm in the batch setting as in [1]):

Lemma C.1. *Suppose all of the assumptions in Definition 1.1 hold. In addition, assume that $\|x_t\|_1 \leq \gamma$ for all $t \in [T]$. If R is defined as in Algorithm 2, with R being an incoherent matrix with $O(\varepsilon^{-2}\log d \cdot \max(1, 2\gamma^2\beta/\lambda))$ rows, then $\|R^T z_* - w_*\|_\infty \leq \varepsilon\|w_*\|_1$.*

Proof. The proof is nearly the same as that of Theorem 2.2 — we let R be an F-incoherent matrix with $O(\log d/F^2)$ rows where F is selected appropriately later. In the place of Lemma B.3, we use the following key property of R:

Lemma C.2 (Lemma 2 of [6])**.** *For $v_1, v_2 \in \mathbb{R}^d$, $|v_1^T v_2 - (Rv_1)^T(Rv_2)| \leq F\|v_1\|_1\|v_2\|_1$.*

This was previously observed in [6], and a similar property was used in Lemma 4 of [1], though there this argument was applied when R was a Countsketch matrix, and therefore a JL matrix.

Lemma C.3. *$\|\Delta\|_\infty \leq \gamma F\|w_*\|_1$, where Δ is defined as in the proof of Theorem 2.2.*

Proof. This is essentially Lemma 5 of [1], and the proof is unchanged — we include it here for completeness. First, note that

$$\Delta = \frac{1}{\lambda T}(\widehat{K} - K)\alpha_* = \frac{1}{\lambda T}(\widetilde{X}^T R^T R\widetilde{X} - \widetilde{X}^T\widetilde{X})\alpha_* = -\widetilde{X}^T(R^T R - I)w_* \tag{21}$$

[5] `https://arxiv.org/pdf/1206.5725.pdf`

where the third equality is because $w_* = -\frac{1}{\lambda T}\widetilde{X}\alpha_*$. Thus,

$$\begin{aligned}\|\Delta\|_\infty &= \max_{t\in[T]} |y_t x_t^T (R^T R - I) w_*| \\ &= \max_{t\in[T]} |x_t^T R^T R w_* - x_t^T w_*| \\ &= \max_{t\in[T]} |(Rx_t)^T (Rw_*) - x_t^T w_*| \\ &\le F\|x_t\|_1 \|w_*\|_1 \\ &\le F\gamma \|w_*\|_1\end{aligned} \tag{22}$$

where the first inequality above is by Lemma C.2, and the second is because $\gamma = \max_{t\in[T]} \|x_t\|_1$. □

Now, arguing as in [1] and in the proof of Theorem 2.2 (with the argument being identical except for the fact that we use $\|\Delta\|_\infty \le \gamma F \|w_*\|_1$ instead of $\|\Delta\|_\infty \le F\|w_*\|_2$), we obtain

$$\|z_* - Rw_*\|_2^2 \le \frac{2\beta}{\lambda}\gamma^2 F^2 \|w_*\|_1^2 \tag{23}$$

To finish the proof, we now bound $\|R^T z_* - w_*\|_\infty$ using Equation 23, together with properties of ε-incoherent matrices. By the triangle inequality,

$$\|R^T z_* - w_*\|_\infty \le \|R^T z_* - R^T R w_*\|_\infty + \|R^T R w_* - w_*\|_\infty \tag{24}$$

First we bound the first term:

$$\begin{aligned}\|R^T z_* - R^T R w_*\|_\infty &= \max_{i\in[d]} |\langle R_i, z_* - Rw_*\rangle| \\ &\le \max_{i\in[d]} \|R_i\|_2 \|z_* - Rw_*\|_2 \\ &= \|z_* - Rw_*\|_2 \\ &\le \sqrt{\frac{2\beta}{\lambda}}\gamma F \|w_*\|_1\end{aligned} \tag{25}$$

where the first inequality is by the Cauchy-Schwarz inequality, and the second inequality is by Equation 23. In addition,

Lemma C.4. $\|R^T R w_* - w_*\|_\infty \le F\|w_*\|_1$

Proof. For $i \in [d]$,

$$\begin{aligned}|\langle R_i, Rw_*\rangle - w_{*,i}| &= |\langle Re_i, Rw_*\rangle - \langle e_i, w_*\rangle| \\ &\le F\|e_i\|_1 \|w_*\|_1 \\ &= F\|w_*\|_1\end{aligned} \tag{26}$$

where the first inequality is due to Lemma C.2. □

Combining Equations 24 and 25 and Lemma C.4, we find that

$$\|R^T z_* - w_*\|_\infty \le \Big(\sqrt{\frac{2\beta}{\lambda}}\gamma F + F\Big)\|w_*\|_1$$

For the right-hand side to be at most $\varepsilon\|w_*\|_1$, it suffices to choose $F = \frac{\varepsilon}{2}\cdot\min(1, \frac{1}{\gamma}\cdot\sqrt{\frac{\lambda}{2\beta}})$. Thus, by Theorem 2.2, it suffices for R to have $O\Big(\frac{\log d}{\varepsilon^2}\cdot\max\Big(1, \frac{\gamma^2\beta}{\lambda}\Big)\Big)$ rows. □

We now complete the analysis of Algorithm 2, in the online setting. First, let us prove the following bound on $\|z_* - \overline{z}\|_2$:

Lemma C.5. $\|\overline{z} - z_*\|_2 \le O\Big(\frac{H(1+\sqrt{\varepsilon}\gamma)}{\lambda T^{1/4}}\Big)$.

Proof. The proof is similar to that of Lemma B.9. Observe that by online gradient descent regret bounds (such as Theorem 3.1 of [22]),

$$\frac{3}{2}GD\sqrt{T} \geq \sum_{t=1}^{T}(f_t(z_t) - f_t(z_*)) \geq \sum_{t=1}^{T}\frac{\lambda}{2}\|z_t - z_*\|_2^2$$

where the second inequality is because f_t is λ-strongly convex. Thus, dividing by T gives

$$\frac{3GD}{\lambda\sqrt{T}} \geq \frac{1}{T}\sum_{t=1}^{T}\|z_t - z_*\|_2^2 \geq \|\overline{z} - z_*\|_2^2$$

where the last inequality is because the function $g(x) = \|x - z_*\|_2^2$ is convex. Thus, taking square roots gives

$$\|\overline{z} - z_*\|_2 \leq \sqrt{\frac{3GD}{\lambda}} \cdot \frac{1}{T^{1/4}}$$

Let us next bound D and G — our proof of these bounds is similar to that in [1], though we do not assume any bounds on $\|w_*\|_2$ and $\|w_*\|_1$. First, by Equation 23,

$$\|z_*\|_2 \leq \|z_* - Rw_*\|_2 + \|Rw_*\|_2 \leq \frac{2\beta}{\lambda}\gamma F\|w_*\|_1 + \|Rw_*\|_2 \tag{27}$$

In addition, by Lemma C.2

$$|\|Rw_*\|_2^2 - \|w_*\|_2^2| \leq F\|w_*\|_1^2 \tag{28}$$

that is,

$$\|Rw_*\|_2 \leq \sqrt{\|w_*\|_2^2 + F\|w_*\|_1^2} \leq \|w_*\|_2 + \sqrt{F}\|w_*\|_1$$

Thus,

$$\|z_*\|_2 \leq \|w_*\|_2 + \sqrt{\varepsilon}\|w_*\|_1$$

We can bound $\|w_*\|_1$ as follows:

Lemma C.6. $\|w_*\|_1 \leq H\gamma/\lambda$.

Proof. The proof is similar to Lemma B.6. Recall that

$$w_* = \mathrm{argmin}_{w\in\mathbb{R}^d}\frac{1}{T}\sum_{t=1}^{T}\ell(y_t w^T x_t) + \frac{\lambda}{2}\|w\|_2^2 = \mathrm{argmin}_{w\in\mathbb{R}^d} L(w)$$

and if we let w'_k be the k^{th} iterate of gradient descent on L (from any initialization, e.g. 0), then $w'_k \to w_*$. However,

$$w'_{k+1} = (1 - \lambda\eta_t)w'_k - \eta_t\frac{1}{T}\sum_{t=1}^{T}y_t\ell'(y_t w^T x_t)x_t$$

meaning that

$$\begin{aligned}
\|w'_{k+1}\|_1 &\leq (1 - \lambda\eta_t)\|w'_k\|_1 + \eta_t \cdot \left\|\frac{1}{T}\sum_{t=1}^{T}y_t\ell'(y_t w^T x_t)x_t\right\|_1 \\
&\leq (1 - \lambda\eta_t)\|w'_k\|_1 + \lambda\eta_t \cdot \frac{1}{\lambda T}\sum_{t=1}^{T}H\gamma \\
&\leq (1 - \lambda\eta_t)\|w'_k\|_1 + \lambda\eta_t \cdot \frac{H\gamma}{\lambda}
\end{aligned} \tag{29}$$

where the second equality is because $|\ell'(y_t w^T x_t| \leq H$ and $\|x_t\|_1 \leq \gamma$. Thus, we can show by induction on k that $\|w'_k\|_1 \leq \frac{H\gamma}{\lambda}$, and taking the limit as $k \to \infty$ gives $\|w_*\|_1 \leq \frac{H\gamma}{\lambda}$. □

Therefore, since $\|w_*\|_1 \le H\gamma/\lambda$ and $\|w_*\|_2 \le H/\lambda$ by Lemma B.6,

$$\|z_*\|_2 \le \frac{H(1+\sqrt{\varepsilon}\gamma)}{\lambda}$$

and therefore we can define $D = \frac{H(1+\sqrt{\varepsilon}\gamma)}{\lambda}$.

Remark C.7. *The projection step in online gradient descent is unnecessary here, i.e. it is not necessary to project z_t onto the ℓ_2 ball of radius D. To see this, if we use the following online gradient descent step:*

$$z_{t+1} \leftarrow (1-\lambda\eta_t)z_t - \eta_t \ell'(y_t z_t^T R x_t) R x_t$$

then

$$\|z_{t+1}\|_2 \le (1-\lambda\eta_t)\|z_t\|_2 + (\lambda\eta_t)\cdot\frac{H}{\lambda}\|Rx_t\|_2$$

Since R is an ε-incoherent matrix, by Lemma C.2,

$$|\langle Rx_t, Rx_t\rangle - \langle x_t, x_t\rangle| \le \varepsilon\|x_t\|_1^2 \le \varepsilon\gamma^2$$

and therefore,

$$\|Rx_t\|_2 \le \sqrt{\|x_t\|_2^2 + \varepsilon\gamma^2} \le \sqrt{1+\varepsilon\gamma^2} \le 1+\sqrt{\varepsilon}\gamma$$

Thus,

$$\|z_{t+1}\|_2 \le (1-\lambda\eta_t)\|z_t\|_2 + (\lambda\eta_t)\cdot\frac{H(1+\sqrt{\varepsilon}\gamma)}{\lambda}$$

and by induction on t, we can show that $\|z_t\|_2 \le \frac{H(1+\sqrt{\varepsilon}\gamma)}{\lambda}$ even without the projection step.

Finally, let us bound G. For all $t \in [T]$ and $\|z\|_2 \le D$,

$$\begin{aligned}
\|\nabla f_t(z)\|_2 &= \|y_t\ell'(y_t z^T R x_t)Rx_t + \lambda z\|_2 \\
&\le |\ell'(y_t z^T R x_t)|\|Rx_t\|_2 + \lambda\|z\|_2 \\
&\le H\cdot(1+\sqrt{\varepsilon}\gamma) + \lambda D \\
&\le H\cdot(1+\sqrt{\varepsilon}\gamma) + \lambda\cdot\frac{H(1+\sqrt{\varepsilon}\gamma)}{\lambda} \\
&= O(H(1+\sqrt{\varepsilon}\gamma))
\end{aligned} \tag{30}$$

Therefore,

$$\|\overline{z} - z_*\|_2 \le \sqrt{\frac{3GD}{\lambda}}\cdot\frac{1}{T^{1/4}} \le O\Big(\frac{H(1+\sqrt{\varepsilon}\gamma)}{\lambda T^{1/4}}\Big)$$

where the second inequality is because $G = O(H(1+\sqrt{\varepsilon}\gamma)/\lambda)$ and $D = O(H(1+\sqrt{\varepsilon}\gamma))$, meaning $\sqrt{GD/\lambda} = H(1+\sqrt{\varepsilon}\gamma)/\lambda$. This proves Lemma C.5. □

Therefore,

$$\begin{aligned}
\|R^T\overline{z} - w_*\|_\infty &\le \|R^T\overline{z} - R^T z_*\|_\infty + \|R^T z_* - w_*\|_\infty \\
&\le \|R^T\overline{z} - R^T z_*\|_\infty + O(\varepsilon\|w_*\|_1) \\
&\le \sup_{i\in[d]} |\langle R_i, \overline{z} - z_*\rangle| + O(\varepsilon\|w_*\|_1) \\
&\le \|R_i\|_2\|\overline{z} - z_*\|_2 + O(\varepsilon\|w_*\|_1) \\
&= \|\overline{z} - z_*\|_2 + O(\varepsilon\|w_*\|_1) \\
&\le O\Big(\frac{H(1+\sqrt{\varepsilon}\gamma)}{\lambda T^{1/4}}\Big) + O(\varepsilon\|w_*\|_1)
\end{aligned} \tag{31}$$

Here the second inequality is by Lemma C.1. The fourth inequality is by the Cauchy-Schwarz inequality. The first equality is because R is an incoherent matrix. The fifth inequality is by Lemma C.5. Thus, for the right-hand side above to be most $O(\varepsilon\|w_*\|_1)$, it suffices to have

$$O\Big(\frac{H(1+\sqrt{\varepsilon}\gamma)}{\lambda T^{1/4}}\Big) \le \varepsilon\theta \le \varepsilon\|w_*\|_1$$

Rearranging gives

$$T^{1/4} \geq O\Big(\frac{H(1+\sqrt{\varepsilon}\gamma)}{\lambda\varepsilon\theta}\Big)$$

or

$$T \geq O\Big(\frac{H^4(1+\sqrt{\varepsilon}\gamma)^4}{\lambda^4\varepsilon^4\theta^4}\Big)$$

This completes the proof.

C.3 Proof of Theorem 3.3

By Theorem 2.2, we know that $\|z_* - Rw_*\|_2 \leq \sqrt{\frac{2\beta}{\lambda}}F\|w_*\|_2$ (where R has $O(\frac{\log(dT/\delta)}{F^2})$ rows and we will select F appropriately). Now, we bound $\|R^T z_* - w_*\|_\infty$. First, note that by the triangle inequality,

$$\|R^T z_* - w_*\|_\infty \leq \|R^T z_* - R^T R w_*\|_\infty + \|R^T R w_* - w_*\|_\infty \tag{32}$$

The first term in Equation 32 can be bounded using Theorem 2.2:

$$\begin{aligned}\|R^T z_* - R^T R w_*\|_\infty &= \max_{i\in[d]} |\langle R_i, z_*\rangle - \langle R_i, Rw_*\rangle| \\ &\leq \max_{i\in[d]} \|R_i\|_2 \|z_* - Rw_*\|_2 \\ &\leq O(\|z_* - Rw_*\|_2) \\ &\leq O\Big(\frac{F\sqrt{\beta}}{\sqrt{\lambda}}\|w_*\|_2\Big)\end{aligned} \tag{33}$$

where the first inequality is by the Cauchy-Schwarz inequality. In addition, the second term in Equation 32 can be bounded using Property (1) of Lemma B.3:

$$\|R^T R w_* - w_*\|_\infty = \max_{i\in[d]} |\langle R_i, Rw_*\rangle - \langle e_i, w_*\rangle| \leq F\|w_*\|_2 \tag{34}$$

In summary, by Equations 32, 33 and 34, $\|R^T z_* - w_*\|_\infty \leq F\|w_*\|_2 \cdot O\Big(1 + \sqrt{\frac{\beta}{\lambda}}\Big)$ and to have $\|R^T z_* - w_*\|_\infty$ be at most $\varepsilon\|w_*\|_2$, it suffices to let $F = O(\varepsilon \cdot \min(1, \sqrt{\frac{\lambda}{\beta}}))$.

Finally, to extend to the online setting (i.e., to show that $\|R^T\overline{z} - w_*\|_\infty \leq \varepsilon\|w_*\|_2$ for sufficiently large T), we note that we can show that $\frac{1}{T}\sum_{s=1}^T \|z_s - z_*\|_2 \leq O(\frac{H}{\lambda T^{1/4}})$ using the same argument as in Lemma B.9, and thus, by convexity of $\|\cdot\|_2$, $\|\overline{z} - z_*\|_2 \leq O(\frac{H}{\lambda T^{1/4}})$. Therefore,

$$\begin{aligned}\|R^T\overline{z} - w_*\|_\infty &\leq \|R^T\overline{z} - R^T z_*\|_\infty + \|R^T z_* - w_*\|_\infty \\ &\leq \|R^T\overline{z} - R^T z_*\|_\infty + \varepsilon\|w_*\|_2 \\ &\leq \max_{i\in[d]} |\langle R_i, \overline{z} - z_*\rangle| + \varepsilon\|w_*\|_2 \\ &\leq O(\|\overline{z} - z_*\|_2) + \varepsilon\|w_*\|_2 \\ &\leq O\Big(\frac{H}{\lambda T^{1/4}}\Big) + \varepsilon\|w_*\|_2\end{aligned} \tag{35}$$

Here the second inequality holds as long as $F = O(\varepsilon \cdot \min(1, \sqrt{\lambda/\beta}))$. The fourth inequality is by Cauchy-Schwarz. Finally, the sixth inequality holds as long as $T \geq \Omega(\frac{H^4}{\lambda^4\varepsilon^4\tau^4})$.

D Missing Pseudocode, Proofs, and Kernel Logistic Regression Results from Section 4

D.1 Pseudocode for Tensor Classification Point Query

Our algorithm for ℓ_2 point query on $w_* \in \mathbb{R}^{d^p}$ is shown in Algorithm 4.

Algorithm 4 Algorithm for ℓ_2 point query on w_* with low-rank tensor inputs. Note that Rx_t can be computed in time $\text{poly}(\varepsilon^{-1}p\log(dT/\delta)\max(1,\beta/\lambda))\sum_{i=1}^{p}\text{nnz}(x_t^{(i)})$. It can also be computed in one pass over the nonzero entries of $x_t^{(1)},\ldots,x_t^{(p)}$. Note that computing R_i for $i=(i_1,i_2,\ldots,i_p)$ can also be done in $\text{poly}(\varepsilon^{-1}p\log(dT/\delta)\max(1,\beta/\lambda))$ time, since $R_i = R(e_{i_1}\otimes\ldots\otimes e_{i_p})$.

function INITIALIZATION()
    $R\in\mathbb{R}^{k\times d^p}$ is the sketch described in Theorem 5.1, with $\text{poly}(\varepsilon^{-1}p\log(dT/\delta)\max(1,\beta/\lambda))$ rows.
    $z_1\in\mathbb{R}^k$ is initially set to 0.
end function

// Here x_t is a rank-k tensor, and R is the sketching matrix of Theorem 4.1, meaning the procedure
// for computing Rx_t is different from Algorithm 2. The algorithm is the same otherwise.
function UPDATE($x_t=\sum_{i=1}^{k}x_t^{(i,1)}\otimes\ldots\otimes x_t^{(i,p)},y_t$)
    $z_{t+1}\leftarrow(1-\lambda\eta_t)z_t-\eta_t\ell'(y_t z_t^T Rx_t)Rx_t$
end function

function ESTIMATE-WEIGHTS($i=(i_1,i_2,\ldots,i_p)$)
    $\overline{z_T}\leftarrow\frac{1}{T}\sum_{t=1}^{T}z_t$
    $R_i\leftarrow R(e_{i_1}\otimes\ldots\otimes e_{i_p})$
    return $R_i^T\overline{z_T}$
end function

D.2 Proof of Theorem 4.2

This follows from Theorem 4.1, using the same arguments as in the proof of Theorem 2.2 and Theorem 3.3. This is because if R is the sketching matrix described in Theorem 4.1, then it has all of the properties of the sparse JL matrix described in Theorem 2.1 — thus, as argued in the proof of Theorem 3.3, if F is the accuracy parameter in Theorem 4.1 and ε is our desired accuracy parameter for point query, it suffices to let $F=O(\varepsilon\cdot\min(1,\sqrt{\lambda/\beta}))$. Furthermore, if δ' is the failure probability with which we apply Theorem 4.1, then it suffices to let $\delta'=\frac{\delta}{\text{poly}(dT)}$ as in the proof of Lemma B.3. Thus, if R is the sketching matrix of Theorem 4.1 used in Algorithm 4, then it suffices for R to have $\text{poly}(\frac{p\log(dT/\delta)}{\varepsilon}\cdot\max(1,\sqrt{\beta/\lambda}))$ rows. By Theorem 4.1, for any rank-1 tensor $x_1\otimes\ldots\otimes x_p$, $R(x_1\otimes\ldots\otimes x_p)$ can be computed in $\text{poly}(\varepsilon^{-1}p\log(dT/\delta)\max(1,\beta/\lambda))\cdot\sum_{i=1}^{p}\text{nnz}(x_i)$ time, giving the desired update time. Finally, the desired query time follows from the fact that for $i=(i_1,\ldots,i_p)$, $R_i=R(e_{i_1}\otimes\ldots\otimes e_{i_p})$ can be computed in $\text{poly}(\varepsilon^{-1}p\log(dT/\delta)\cdot\max(1,\beta/\lambda))$ time, together with the fact that R_i and $\overline{z}$ have $\text{poly}(\varepsilon^{-1}p\log(dT/\delta)\cdot\max(1,\beta/\lambda))$ entries. The minimum required value of T is also unchanged — note that it follows from the same argument as the one used in Lemma B.9 (only needing the additional properties that $\|Rx_t\|_2\le O(1)$, which follows from the same argument as in Lemma B.3, and $\|z_*\|_2\le O(H/\lambda)$, which follows from $\|Rx_t\|_2\le O(1)$ together with an argument based on gradient descent as in Lemma B.6). This completes the proof.

D.3 Proof of Theorem 4.3

For convenience, we denote the t^{th} update to v by $x_t=x_t^{(1)}\otimes\ldots\otimes x_t^{(p)}\in\mathbb{R}^{d^p}$, meaning that $v=\sum_{t\in[T]}x_t^{(1)}\otimes\ldots\otimes x_t^{(p)}$. We show that Algorithm 3 returns a list L containing all indices $(i_1,\ldots,i_p)$ which are heavy hitters. Note that if $(i_1,\ldots,i_p)$ is an ε ℓ_2 heavy hitter, meaning that $|v(i_1,\ldots,i_p)|\ge\varepsilon\|v\|_2$, then in particular, for each $j\in[p]$,

$$\sum_{(k_1,\ldots,k_p)\in[d]^p|k_j=i_j}|v(k_1,\ldots,k_p)|^2\ge\varepsilon^2\|v\|_2^2$$

Thus, for each mode $j\in[p]$, we first wish to find the indices $i_j\in[d]$ for which the above holds. This is what is achieved by the first two steps of the QUERY function. To see why, let us consider the case $j=p$ — the other cases are similar. The first step of the QUERY function finds all $\frac{\varepsilon}{\text{poly}(p\log(d/\delta))}$-heavy hitters of $\widehat{v}=\sum_t x_t^{(p)}\otimes\text{COMPRESSOTHERMODES}^{(p)}$, by applying the

standard heavy hitters data structure ONEMODESKETCH$^{(p)}$ to $\widehat{v}$. This is what is achieved by the first step of the QUERY function. To see why, let us consider the case $j = p$ — the other cases are similar. This step finds all $\frac{\varepsilon}{\text{poly}(p\log(d/\delta))}$-heavy hitters of $\sum_t x_t^{(p)} \otimes$ COMPRESSOTHERMODES$^{(j)}$, by applying the standard heavy hitters data structure ONEMODESKETCH$^{(p)}$ to the vector $\widehat{v} = \sum_t x_t^{(p)} \otimes$ COMPRESSOTHERMODES$^{(p)}$. Note that for each $i \in [d]$,

$$\widehat{v(i,\cdot)} = \sum_t x_t^{(p)}(i)\text{COMPRESSOTHERMODES}^{(p)}$$

where $\widehat{v(i,\cdot)}$ is the slice of $\widehat{v}$ whose index in the first mode is i, and $x_t^{(p)}(i)$ is the i^{th} coordinate of $x_t^{(p)}$. Since COMPRESSOTHERMODES$^{(p)}$ is itself given by a JL matrix according to Theorem 5.1, this means that

$$\|\widehat{v}(i,\cdot)\|_2 = \Theta(1)\Big\|\sum_t x_t^{(p)}(i)\cdot x_t^{(1)} \otimes \ldots \otimes x_t^{(p-1)}\Big\|_2 = \Theta(1)\|v(\cdot,\ldots,\cdot,i)\|_2$$

In summary, if i is a heavy coordinate of v on the p^{th} mode, then i is also a heavy coordinate of $\widehat{v} = \sum_t x_t^{(p)} \otimes$ COMPRESSOTHERMODES$^{(j)}$ in the first mode. We can find all heavy coordinates of $\sum_t x_t^{(p)} \otimes$ COMPRESSOTHERMODES$^{(j)}$ in the first mode using a standard heavy hitters data structure with accuracy $\frac{\varepsilon}{\text{poly}(p\log(d/\delta))}$ instead of ε (to account for the weight of i being evenly spread across all poly$(\frac{p\log(d/\delta)}{\varepsilon})$ coordinates in the second mode). The same is true for all modes $i \in [p]$.

Now, using ONEMODESKETCH$^{(i)}$ we obtain a list L_i of heavy coordinates on the i^{th} mode, for all modes $i \in [p]$. We next iterate over all the modes $i \in [p]$, iteratively building a list of all the heavy prefixes of length i. Suppose at the $(i+1)^{th}$ iteration, we have a list L of prefixes of length i, with $|L| \le \frac{C}{\varepsilon^2}$ and containing all prefixes $(j_1,\ldots,j_i)$ such that $|v(j_1,\ldots,j_i,\cdot)| \ge c\varepsilon\|v\|_2$ for some constant $C, c > 0$. Then, clearly all ε-heavy prefixes of length $(i+1)$ must be among the prefixes of length $(i+1)$ in $L \times L_i$. Now, to update L, we take the top $\frac{C}{\varepsilon^2}$ elements of $L \times L_i$ according to PREFIXPOINTQUERY$^{(i+1)}$. It suffices to argue that PREFIXPOINTQUERY$^{(i+1)}$ can estimate the ℓ_2 norm of all prefixes $(j_1,\ldots,j_i,j_{i+1})$ up to an $\varepsilon\|v\|_2$ additive error. Since PREFIXPOINTQUERY$^{(i+1)}$ is a JL matrix, it gives an estimate $N(j_1,\ldots,j_i,j_{i+1})$ of the weight of the prefix $(j_1,\ldots,j_i,j_{i+1})$ that satisfies

$$\begin{aligned} N(j_1,\ldots,j_i,j_{i+1}) = &\Big\|\sum_t x_t^{(1)}(j_1)\ldots x_t^{(i+1)}(j_{i+1})\text{COMPRESSSUFFIX}(x_t^{(i+1)}\otimes\ldots\otimes x_t^{(p)})\Big\|_2 \\ &\pm O(\varepsilon)\Big\|\sum_t x_t^{(1)}\otimes\ldots\otimes x_t^{(i+1)}\otimes\text{COMPRESSSUFFIX}(x_t^{(i+1)}\otimes\ldots\otimes x_t^{(p)})\Big\|_2 \end{aligned} \tag{36}$$

This is because it can give an estimate of an individual coordinate of $\sum_t x_t^{(1)}\otimes\ldots\otimes x_t^{(i+1)}\otimes$ COMPRESSSUFFIX$(x_t^{(i+1)}\otimes\ldots\otimes x_t^{(p)})$, up to additive error $O(\frac{\varepsilon}{\text{poly}(\frac{p\log(d/\delta)}{\varepsilon})})\Big\|\sum_t x_t^{(1)}\otimes\ldots\otimes x_t^{(i+1)}\otimes\text{COMPRESSSUFFIX}(x_t^{(i+1)}\otimes\ldots\otimes x_t^{(p)})\Big\|_2$ — this simply follows from the arguments in Section 3 on using JL matrices for point query, as well as Theorem 4.1. Since COMPRESSSUFFIX$^{(i+1)}$ is itself a JL matrix, $\Big\|\sum_t x_t^{(1)}(j_1)\ldots x_t^{(i+1)}(j_{i+1})\text{COMPRESSSUFFIX}(x_t^{(i+1)}\otimes\ldots\otimes x_t^{(p)})\Big\|_2$ is equal to the weight of the prefix $(j_1,\ldots,j_{i+1})$ up to a constant factor. By Lemma 14 of [11] [6], since COMPRESSSUFFIX$^{(i)}$ has the JL property, it is also true that $\Big\|\sum_t x_t^{(1)}\otimes\ldots\otimes x_t^{(i+1)}\otimes\text{COMPRESSSUFFIX}(x_t^{(i+1)}\otimes\ldots\otimes x_t^{(p)})\Big\|_2$ is equal to $\|v\|_2$ up to a constant factor. Thus, $N(j_1,\ldots,j_i,j_{i+1})$ is equal to $\|v(j_1,\ldots,j_i,j_{i+1},\cdot,\ldots,\cdot)\|_2$ up to a constant factor and additive error $O(\varepsilon\|v\|_2)$ — this shows the correctness of selecting the top $\frac{C}{\varepsilon^2}$ prefixes according to PREFIXPOINTQUERY and COMPRESSSUFFIX as done in our algorithm, and thus completes the proof of correctness. The bounds on the space complexity follow from Theorem 4.1.

[6] See page 23 of the arxiv version.

D.4 Proof of Theorem 4.4

This is a corollary of Theorem 4.3 and Theorem 2.6 — the only change we make to Theorem 2.6 is that R is now the sketching matrix of Theorem 4.1. If F is the accuracy parameter of R, then as argued in the proof of Theorem 2.6, it suffices to let $F \leq O(\frac{\lambda\varepsilon}{\beta} \cdot \min(1, \sqrt{\lambda/\beta}))$, and therefore it suffices to let R have $\text{poly}(F^{-1}p\log(dT/\delta)) = \text{poly}(\beta\lambda^{-1}\varepsilon^{-1}p\log(dT/\delta)\max(1,\beta/\lambda))$ rows. Therefore, we can obtain the space complexity, update time, and query time as follows:

- Since R has $\text{poly}(\beta\lambda^{-1}\varepsilon^{-1}p\log(dT/\delta)\max(1,\beta/\lambda))$ rows and the space complexity of Algorithm 3 is $\text{poly}(\varepsilon^{-1}p\log(d/\delta))$, the overall space complexity for classification is also $\text{poly}(\varepsilon^{-1}p\log(dT/\delta)\max(1,\beta/\lambda))$.
- $\text{poly}(\varepsilon^{-1}p\log(dT/\delta)\max(1,\beta/\lambda))\sum_{i=1}^{k}\sum_{j=1}^{p}\text{nnz}(x_t^{(i,j)})$ time is needed to compute Rx_t where $x_t = \sum_{i=1}^{k} x_t^{(i,1)} \otimes \ldots \otimes x_t^{(i,p)}$, and the update time of Algorithm 3 in this case is $\text{poly}(\varepsilon^{-1}p\log(d/\delta))\sum_{i=1}^{k}\sum_{j=1}^{p}\text{nnz}(x_t^{(i,j)})$. Thus, the overall update time is as desired.
- The query time is that of Algorithm 3, which is $\text{poly}(\varepsilon^{-1}p\log(d/\delta))$.

Finally, the required minimum values of T are exactly as in Theorem 2.6. This completes the proof.

D.5 Polynomial Kernel and Gaussian Kernel — Corollaries of Theorems 4.2, 4.3, and 4.4

We note that the *polynomial kernel* of degree q, given a point $x_t \in \mathbb{R}^d$, corresponds to the explicit feature mapping into $\mathbb{R}^{d^q}$ given by the self tensoring $x_t^{\otimes q}$. Thus, given a stream of points x_t in $\mathbb{R}^d$, we can apply the above theorems to $x_t^{\otimes q}$ without having to explicitly perform this self-tensoring. The polynomial kernel is not only useful by itself, with $q = 2$ and $q = 3$ common in natural language processing applications, but it is also used for approximating other kernels via Taylor series, e.g., the Gaussian or RBF kernel, where typically q is logarithmic and q also depends on the radius of the input point set. We refer the reader to Theorem 5 of [11] for more background, but just as in the case of the polynomial kernel, given a stream of points x_t, we can implicitly form a polynomial kernel of the appropriate degree, and use this to approximate the Gaussian kernel. Thus, we obtain ℓ_2 heavy hitter and point query algorithms for classification for this important class of kernels as well.

We now formally define the problem of heavy hitters for polynomial kernel classification.

Definition D.1 (ℓ_2 Heavy Hitters for Kernel Classification)**.** *Let $T \in \mathbb{N}$, and let $x_t \in \mathbb{R}^d$, $y_t \in \{-1, 1\}$ for $t \in [T]$. In addition, define*

$$L(w) = \frac{1}{T}\sum_{t=1}^{T}\ell(y_t w^T x_t^{\otimes p}) + \frac{\lambda}{2}\|w\|_2^2$$

*and $w_* = \text{argmin}_{w\in\mathbb{R}^d} L(w)$. For $\varepsilon > 0$, we say $i = (i_1, \ldots, i_p) \in [d]^p$ is an ε **polynomial kernel ℓ_2 heavy hitter** for w_* if*

$$\sum_{\substack{(j_1,\ldots,j_p)\in[d]^p \\ M(j_1,\ldots,j_p)=M(i_1,\ldots,i_p)}} |w(j_1, \ldots, j_p)|^2 \geq \varepsilon^2\|w_*\|_2^2$$

where $M(j_1, \ldots, j_p)$ denotes the multiset which is formed by $j_1, \ldots, j_p$.

Note that $M(j_1, \ldots, j_p) = M(i_1, \ldots, i_p)$ if and only if there is a permutation σ on $[d]$ such that $\sigma(i_k) = \sigma(j_k)$. In other words, we wish to consider $(i_1, \ldots, i_p)$ as a polynomial kernel ℓ_2 heavy hitter if all permutations of $(i_1, \ldots, i_p)$ contribute an ε fraction of the ℓ_2 norm of w_*. In this setting, we obtain the following result:

Theorem D.2 (Algorithm for Polynomial Kernel ℓ_2 Heavy Hitters)**.** *Let $\varepsilon, \delta \in (0, 1)$, and suppose all the assumptions in Definition 1.1 hold (in particular, $\|x_t\|_2 \leq 1$ for all $t \in [T]$, meaning $\|x_t^{\otimes p}\|_2 \leq 1$). In addition, let $c \in [p]$. Then, there is an algorithm to find all ε polynomial kernel ℓ_2 heavy hitters $(i_1, \ldots, i_p) \in [d]^p$ such that the Hamming distance of $(i_1, \ldots, i_p)$ from $\{(i, \ldots, i) \mid i \in [p]\}$ is at most c. The space complexity of this algorithm is $p^{O(C)}\text{poly}(\varepsilon^{-1}\log(dT/\delta)(1+\beta/\lambda))$, the query time of this algorithm is $p^{O(c)}\text{poly}(\varepsilon^{-1}\log(d/\delta))$, and the update time is $p^{O(c)}\text{poly}(\varepsilon^{-1}\log(dT/\delta)(1 + \beta/\lambda))\text{nnz}(x_t)$.*

Proof. This follows from applying the algorithm described in Theorem 4.4, with $\frac{\varepsilon}{p^C}$ in place of ε. To see this, first note that w_* is a symmetric tensor, since if gradient descent is performed on

$$L(w) = \frac{1}{T}\sum_{t=1}^{T} \ell(y_t w^T x_t^{\otimes p}) + \frac{\lambda}{2}\|w\|_2^2$$

then each gradient update is a symmetric tensor, and the iterates of gradient descent on L converge to w_* (see the proof of Lemma B.6). Now, suppose $(i_1, \ldots, i_p)$ is an ε polynomial kernel ℓ_2 heavy hitter of w_*, such that $(i_1, \ldots, i_p)$ has Hamming distance c from $\{(i, \ldots, i) \mid i \in [p]\}$. This means that $(i_1, \ldots, i_p)$ has at most $c+1$ distinct coordinates, and one of those coordinates occurs $p-c$ times. Thus, if we let $p-c, f_1, \ldots, f_c$ be the number of times these coordinates occur (with some of the f_j being 0 if there are less than $c+1$ distinct coordinates) then the number of distinct permutations of $(i_1, \ldots, i_p)$ is

$$K := \binom{p}{p-c, f_1, \ldots, f_c} = \frac{p!}{(p-c)!f_1! \ldots f_c!} \leq \frac{p!}{(p-c)!} = p \cdot (p-1) \cdot \ldots \cdot (p-c+1) \leq p^c$$

Since $(i_1, \ldots, i_p)$ is a polynomial kernel ℓ_2 heavy hitter, this means that

$$K|w(i_1, \ldots, i_p)|^2 \geq \varepsilon^2 \|w_*\|_2^2$$

and therefore, $(i_1, \ldots, i_p)$ is an $\frac{\varepsilon}{\sqrt{K}}$ ℓ_2 heavy hitter in the usual sense. In particular, since $K \leq p^c$, $(i_1, \ldots, i_p)$ is an $\frac{\varepsilon}{p^{C/2}}$ ℓ_2 heavy hitter in the usual sense. Thus, we could do the following:

- Apply the algorithm described in Theorem 4.4 to obtain a list L of $\frac{\varepsilon}{p^{C/2}}$ ℓ_2 heavy hitters, with $|L| \leq O(p^C/\varepsilon^2)$.
- Then, we could incur an additional running time of $p^{O(c)}/\varepsilon^{O(1)}$ to iterate through the elements of L and perform point query for each element $(i_1, \ldots, i_p)$ to determine if it is an $\frac{\varepsilon}{\sqrt{K}}$ heavy hitter in the usual sense, where K is the number of distinct permutations of $(i_1, \ldots, i_p)$. We would also use this step to remove duplicates in L (treating indices which are permutations of each other as the same).

Thus, the time and space complexities are as follows:

- The space complexity is $p^{O(c)}\text{poly}(\varepsilon^{-1}\log(dT/\delta)(1+\beta/\lambda))$.
- The update time is $p^{O(c)}\text{poly}(\varepsilon^{-1}\log(dT/\delta)(1+\beta/\lambda))\text{nnz}(x_t)$.
- The query time is $p^{O(c)}\text{poly}(\varepsilon^{-1}\log(d/\delta))$ (note that the additional processing on the list of heavy hitters output by Algorithm 3 is also $p^{O(c)}/\varepsilon^{O(1)}$).

□

E Training Loss Using $\overline{z}$

Here, we show that when using $\overline{z}$ rather than w_*, the loss function does not significantly increase:

Theorem E.1. *Suppose all of the assumptions in Definition 1.1 hold. Let R be a sparse JL matrix and suppose z_t is updated according to*

$$z_{t+1} \leftarrow (1-\lambda\eta_t)z_t - \eta_t y_t \ell'(y_t z_t^T R x_t) R x_t$$

In addition, define $\overline{z} = \frac{1}{T}\sum_{t=1}^T z_t$, and define w_* as before. Finally, define

$$L_* = \frac{1}{T}\sum_{t=1}^T \ell(y_t w_*^T x_t) + \frac{\lambda}{2}\|w_*\|_2^2$$

and

$$\widehat{L} = \frac{1}{T}\sum_{t=1}^T \ell(y_t \overline{z}^T R x_t) + \frac{\lambda}{2}\|\overline{z}\|_2^2$$

Then, $|L_* - \widehat{L}| \le \varepsilon\|w_*\|_2$ with probability $1-\delta$ as long as R has at least $O(\frac{H^2\log(dT/\delta)}{\varepsilon^2}\cdot \max(1,1/\lambda)\max(1,O(\sqrt{2\beta/\lambda})))$ rows, and $T \ge \max(H^8/(\lambda^4\varepsilon^4\tau^4), H^8/(\lambda^8\varepsilon^4\tau^4))$. In addition,

$$\Big|\frac{1}{T}\sum_{t=1}^T \ell(y_t w_*^T x_t) - \frac{1}{T}\sum_{t=1}^T \ell(y_t \overline{z}^T R x_t)\Big| \le \varepsilon\|w_*\|_2$$

with probability $1-\delta$ as long as R has at least $O(\frac{H^2\log(dT/\delta)}{\varepsilon^2}\cdot\max(1,\sqrt{\frac{2\beta}{\lambda}}))$ rows and $T \ge \Omega(H^8/(\lambda^4\varepsilon^4\tau^4))$.

Proof. Let R be a sparse JL matrix with $\Theta(\frac{\log(dT/\delta)}{F^2})$ rows, where F will be determined later. Let

$$L_* = \frac{1}{T}\sum_{t=1}^T \ell(y_t w_*^T x_t) + \frac{\lambda}{2}\|w_*\|_2^2$$

and

$$\widehat{L} = \frac{1}{T}\sum_{t=1}^T \ell(y_t \overline{z}^T R x_t) + \frac{\lambda}{2}\|\overline{z}\|_2^2$$

Then,

$$\begin{aligned}
|\widehat{L} - L_*| &\le \frac{1}{T}\sum_{t=1}^T |\ell(y_t w_*^T x_t) - \ell(y_t \overline{z}^T R x_t)| + \frac{\lambda}{2}|\|w_*\|_2^2 - \|\overline{z}\|_2^2| \\
&\le \frac{H}{T}\sum_{t=1}^T |\langle w_*, x_t\rangle - \langle \overline{z}, Rx_t\rangle| \\
&\quad + \frac{\lambda}{2}\Big(\Big|\|w_*\|_2^2 - \|Rw_*\|_2^2\Big| + \Big|\|Rw_*\|_2^2 - \|z_*\|_2^2\Big| + \Big|\|z_*\|_2^2 - \|\overline{z}\|_2^2\Big|\Big)
\end{aligned} \tag{37}$$

Here, the second inequality is because ℓ is H-Lipschitz, and by the triangle inequality.

Now, for $t \in [T]$, observe that

$$\begin{aligned}
|\langle w_*, x_t\rangle - \langle \overline{z}, Rx_t\rangle| &\le |\langle w_*, x_t\rangle - \langle Rw_*, Rx_t\rangle| + |\langle Rw_*, Rx_t\rangle - \langle z_*, Rx_t\rangle| \\
&\quad + |\langle z_*, Rx_t\rangle - \langle \overline{z}, Rx_t\rangle| \\
&\le F\|w_*\|_2 + \|z_* - Rw_*\|_2\|Rx_t\|_2 + \|z_* - \overline{z}\|_2\|Rx_t\|_2 \\
&\le F\|w_*\|_2 + O(1)\|z_* - Rw_*\|_2 + O(1)\|z_* - \overline{z}\|_2 \\
&\le F\|w_*\|_2 + O(1)\cdot\sqrt{\frac{2\beta}{\lambda}}F\|w_*\|_2 + O(1)\|z_* - \overline{z}\|_2 \\
&\le F\|w_*\|_2 + O(1)\cdot\sqrt{\frac{2\beta}{\lambda}}F\|w_*\|_2 + O\Big(\frac{H}{\lambda T^{1/4}}\Big)
\end{aligned} \tag{38}$$

Here, the second inequality holds by Lemma B.3, and the Cauchy-Schwarz inequality, and the third inequality also holds by Lemma B.3 as long as $F \le \varepsilon$. The fourth inequality is due to Equation 8. Finally, for the fifth inequality, observe that as in Lemma B.9, we can show that $\frac{1}{T}\sum_{s=1}^{T}\|z_s - z_*\|_2 \le O(\frac{H}{\lambda T^{1/4}})$ (since D is still $O(\frac{H}{\lambda})$ since $\|Rx_t\|_2 \le 1 + F \le O(1)$, and G is at most $O(H)$ for the same reason) — by the convexity of the ℓ_2 norm, this implies that $\|\overline{z} - z_*\|_2 \le O(\frac{H}{\lambda T^{1/4}})$. In summary,

$$\frac{1}{T}\sum_{t=1}^{T}|\ell(y_t w_*^T x_t) - \ell(y_t \overline{z}^T R x_t)| \le HF\|w_*\|_2 + O(1)\cdot\sqrt{\frac{2\beta}{\lambda}}HF\|w_*\|_2 + O\Big(\frac{H^2}{\lambda T^{1/4}}\Big)$$

Therefore, as long as $F \le O(\varepsilon/H \cdot \min(1, \sqrt{\frac{\lambda}{2\beta}}))$ (i.e. R has at least $O(\frac{H^2\log(dT/\delta)}{\varepsilon^2}\cdot\max(1,\sqrt{\frac{2\beta}{\lambda}}))$ rows) and $T \ge \Omega(H^8/(\lambda^4\varepsilon^4\tau^4))$, the second statement of the theorem holds.

Now, we bound the remaining terms in the last expression in Equation 37. By Lemma B.6, the function $\|\cdot\|_2^2$ is $O(\frac{H}{\lambda})$-Lipschitz on a region containing Rw_*, z_* and $\overline{z}$, since these vectors all have ℓ_2 norm at most $O(\frac{H}{\lambda})$. We can thus bound the remaining terms as follows:

$$\begin{aligned}
&\Big|\|w_*\|_2^2 - \|Rw_*\|_2^2\Big| + \Big|\|Rw_*\|_2^2 - \|z_*\|_2^2\Big| + \Big|\|z_*\|_2^2 - \|\overline{z}\|_2^2\Big| \\
&\qquad \le O(F)\cdot\frac{H}{\lambda}\|w_*\|_2 + \Big|\|Rw_*\|_2^2 - \|z_*\|_2^2\Big| + \Big|\|z_*\|_2^2 - \|\overline{z}\|_2^2\Big| \\
&\qquad \le \frac{FH}{\lambda}\|w_*\|_2 + O\Big(\frac{H}{\lambda}\Big)\|z_* - Rw_*\|_2 + O\Big(\frac{H}{\lambda}\Big)\|z_* - \overline{z}\|_2 \\
&\qquad \le \frac{FH}{\lambda}\|w_*\|_2 + O\Big(\frac{H}{\lambda}\Big)\cdot\sqrt{\frac{2\beta}{\lambda}}F\|w_*\|_2 + O\Big(\frac{H^2}{\lambda^2 T^{1/4}}\Big)
\end{aligned} \tag{39}$$

The first inequality is by Lemma B.3 and because $\|w_*\|_2 \le \frac{H}{\lambda}$. The second inequality is because $\|\cdot\|_2^2$ is $O(\frac{H}{\lambda})$-Lipschitz on a region containing Rw_*, z_* and $\overline{z}$. Finally, the third inequality is by applying the bounds from above on $\|z_* - Rw_*\|_2$ and $\|z_* - \overline{z}\|_2$.

In summary,

$$\begin{aligned}
|\widehat{L} - L_*| &\le HF\|w_*\|_2 + O(1)\cdot\sqrt{\frac{2\beta}{\lambda}}HF\|w_*\|_2 + O\Big(\frac{H^2}{\lambda T^{1/4}}\Big) \\
&\quad + \frac{FH}{\lambda}\|w_*\|_2 + O\Big(\frac{H}{\lambda}\Big)\cdot\sqrt{\frac{2\beta}{\lambda}}\cdot F\|w_*\|_2 + O\Big(\frac{H^2}{\lambda^2 T^{1/4}}\Big) \\
&\le HF\|w_*\|_2\cdot\Big(1 + O(1)\cdot\sqrt{\frac{2\beta}{\lambda}} + \frac{1}{\lambda} + \frac{1}{\lambda}\cdot O(1)\cdot\sqrt{\frac{2\beta}{\lambda}}\Big) \\
&\quad + O\Big(\frac{H^2}{\lambda T^{1/4}} + \frac{H^2}{\lambda^2 T^{1/4}}\Big) \\
&\le HF\|w_*\|_2\cdot(1 + 1/\lambda)\cdot(1 + O(\sqrt{2\beta/\lambda})) + O\Big(\frac{H^2}{\lambda T^{1/4}} + \frac{H^2}{\lambda^2 T^{1/4}}\Big) \\
&\le HF\|w_*\|_2\cdot\max(1, 1/\lambda)\cdot\max(1, O(\sqrt{2\beta/\lambda})) + O\Big(\frac{H^2}{\lambda T^{1/4}} + \frac{H^2}{\lambda^2 T^{1/4}}\Big)
\end{aligned} \tag{40}$$

Thus, as long as $T \ge \max(H^8/(\lambda^4\varepsilon^4\tau^4), H^8/(\lambda^8\varepsilon^4\tau^4))$, and $F \le O(\varepsilon/H)\cdot\min(1,\lambda), \min(1, O(\sqrt{\lambda/2\beta}))$ (i.e. R has at least $O(\frac{H^2\log(dT/\delta)}{\varepsilon^2}\cdot\max(1,1/\lambda)\max(1, O(\sqrt{2\beta/\lambda})))$ rows), $|\widehat{L} - L_*| \le \varepsilon\|w_*\|_2$. □

F Classification Error using Top K Weights

Number of Weights Used (K)	Test Error
Full weights	0.9573693534100974
50	0.8555949217596693
60	0.8679362267493357
70	0.8767581930912312
80	0.8782285208148805
90	0.8811662237968704
100	0.8952583407144966
200	0.9232122822556835
300	0.9370711544139356
400	0.9396220844405079
500	0.942849129022734
600	0.9442810746973723
700	0.9470386772955418
800	0.9481842338352524
900	0.9491378801299085
1000	0.9501505757307351
10000	0.957472689695896
20000	0.9573516386182462
40000	0.9573693534100974

Table 1: The number of weights K that we use, together with the test error achieved when the top K weights are used. Note that the number of nonzero weights in the original trained linear classifier is 41130. With only 400 weights, the test accuracy is 93.9%, while the test accuracy with all 41130 weights is 95.7%.

We performed experiments with the RCV1 dataset [13] to determine the effect of using only the top K weights on classification performance, for K much smaller than the total number of nonzero weights. First, we split the RCV1 dataset into two halves, one for training and one for testing. We obtain a vector $w \in \mathbb{R}^d$ (where $d = 41130$ is the number of nonzero features) using online logistic regression on the training half. Then, we calculate the test accuracy on the testing half when using w, as well as when using w_K, for $K \ll d$ — here, given a data point (x_t, y_t) where $x_t \in \mathbb{R}^d$ and $y_t \in \{\pm 1\}$, a vector v classifies x_t properly if $v^T x_t$ has the same sign as y_t. The results are shown in Table 1 — note that good test accuracy is obtained even with w_K for $K \ll d$.

Our code used to obtain these results is available in the supplementary material, in the folder "Appendix F Experiments." Some of the files in this folder are also based on files due to the authors of [1] at `https://github.com/stanford-futuredata/wmsketch`.

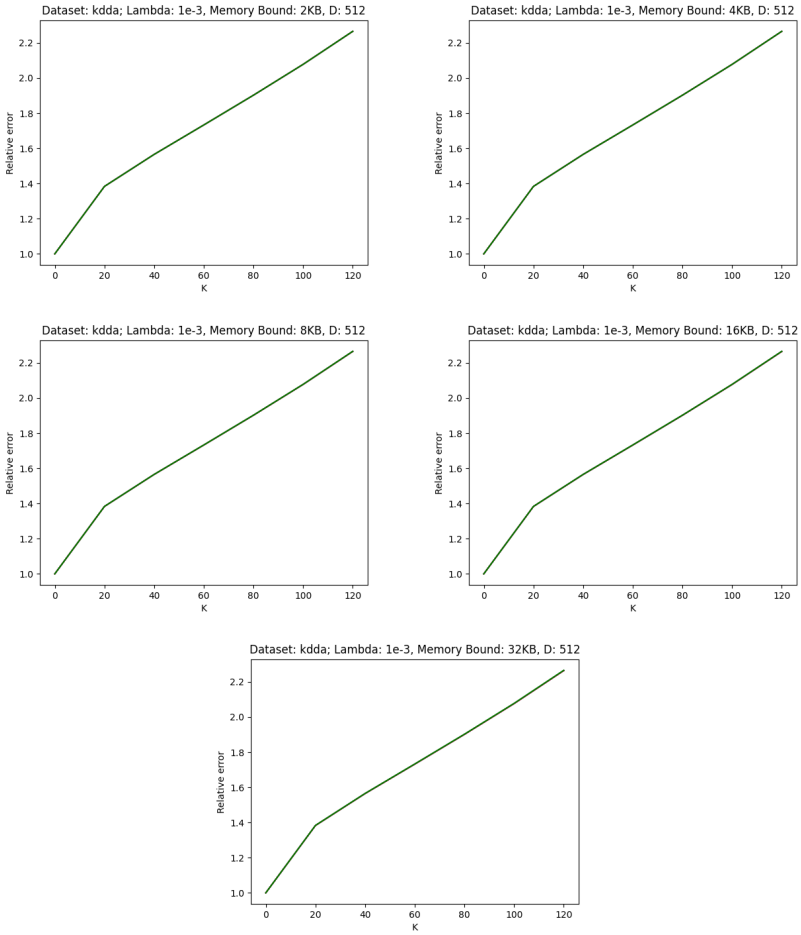

Figure 2: Results on KDD CUP 2010 dataset with $\lambda = 10^{-3}$. Blue denotes the performance of our algorithm from Section 4. Green denotes the performance of our black-box reduction based algorithm from Section 2, and red is the algorithm of [1].

G All Plots for Experiments

Complete results for our experiments are shown in Figures 2, 3, 4, 5, 6, 7, 8, 9, 10.

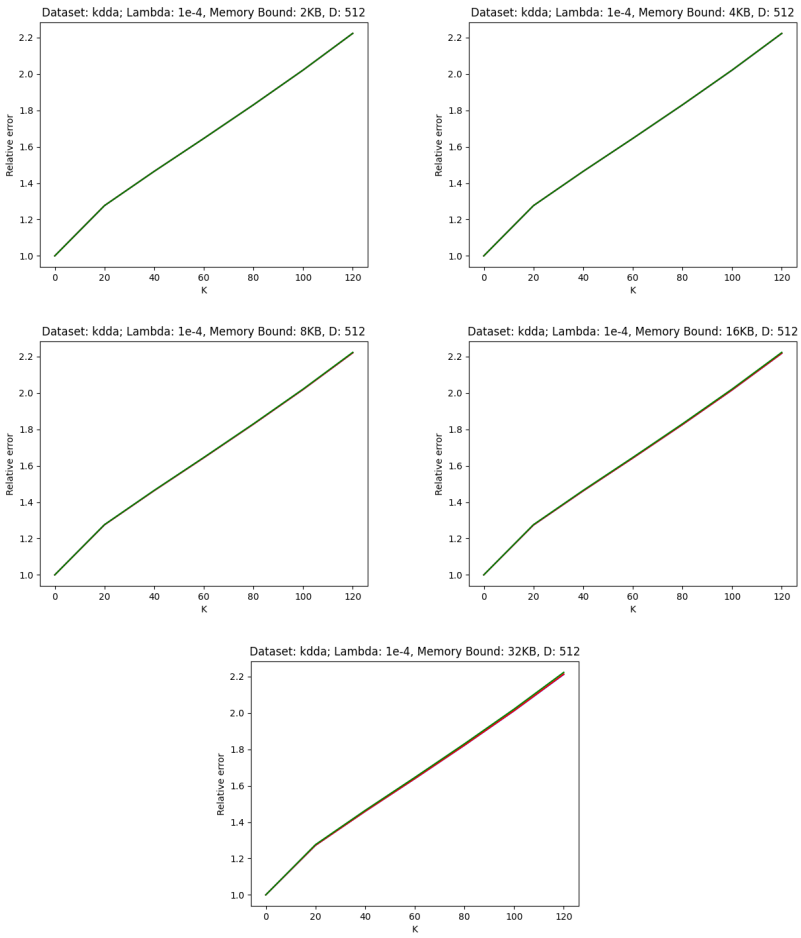

Figure 3: Results on KDD CUP 2010 dataset with $\lambda = 10^{-4}$. Blue denotes the performance of our algorithm from Section 4. Green denotes the performance of our black-box reduction based algorithm from Section 2, and red is the algorithm of [1].

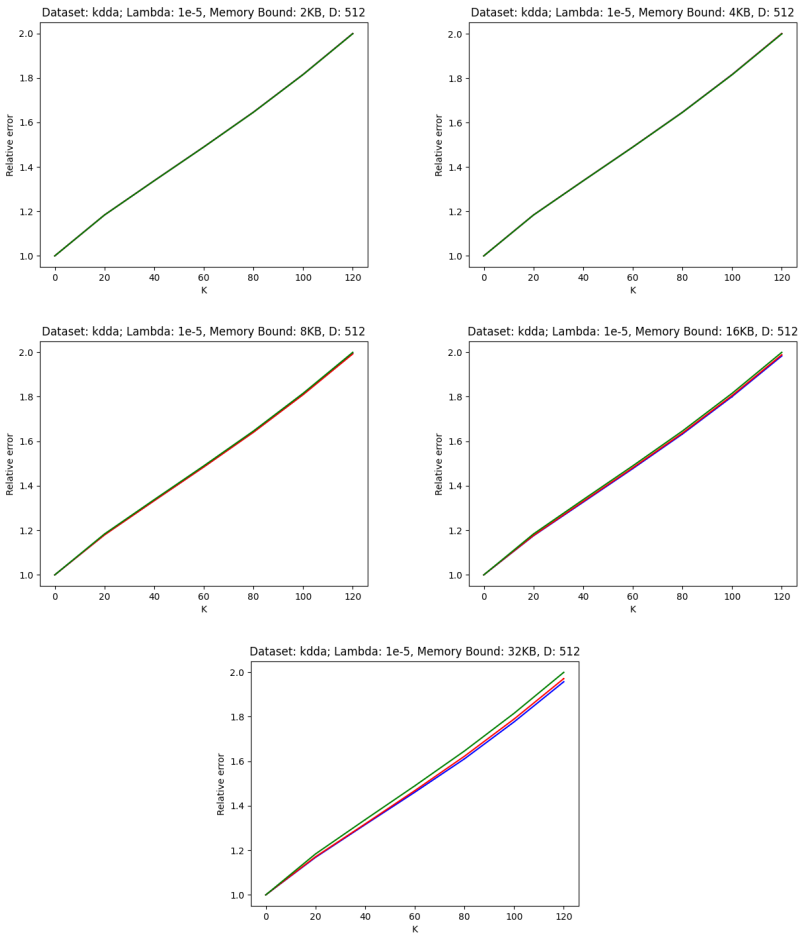

Figure 4: Results on KDD CUP 2010 dataset with $\lambda = 10^{-5}$. Blue denotes the performance of our algorithm from Section 4. Green denotes the performance of our black-box reduction based algorithm from Section 2, and red is the algorithm of [1].

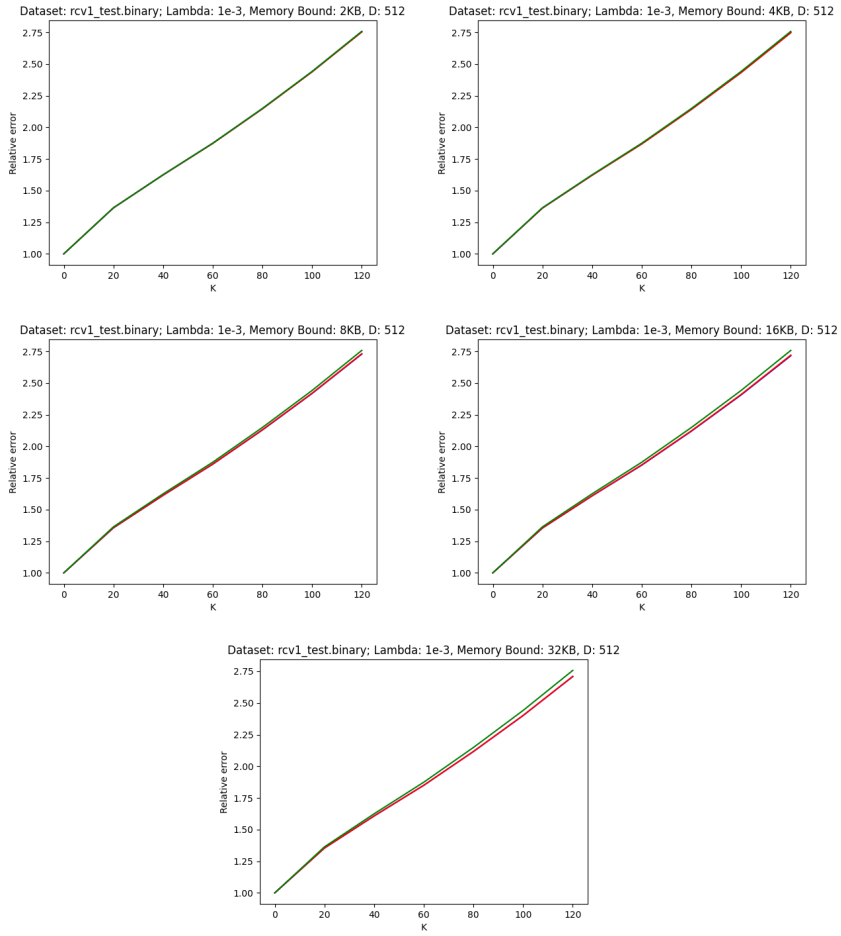

Figure 5: Results on RCV1 dataset with $\lambda = 10^{-3}$. Blue denotes the performance of our algorithm from Section 4. Green denotes the performance of our black-box reduction based algorithm from Section 2, and red is the algorithm of [1].

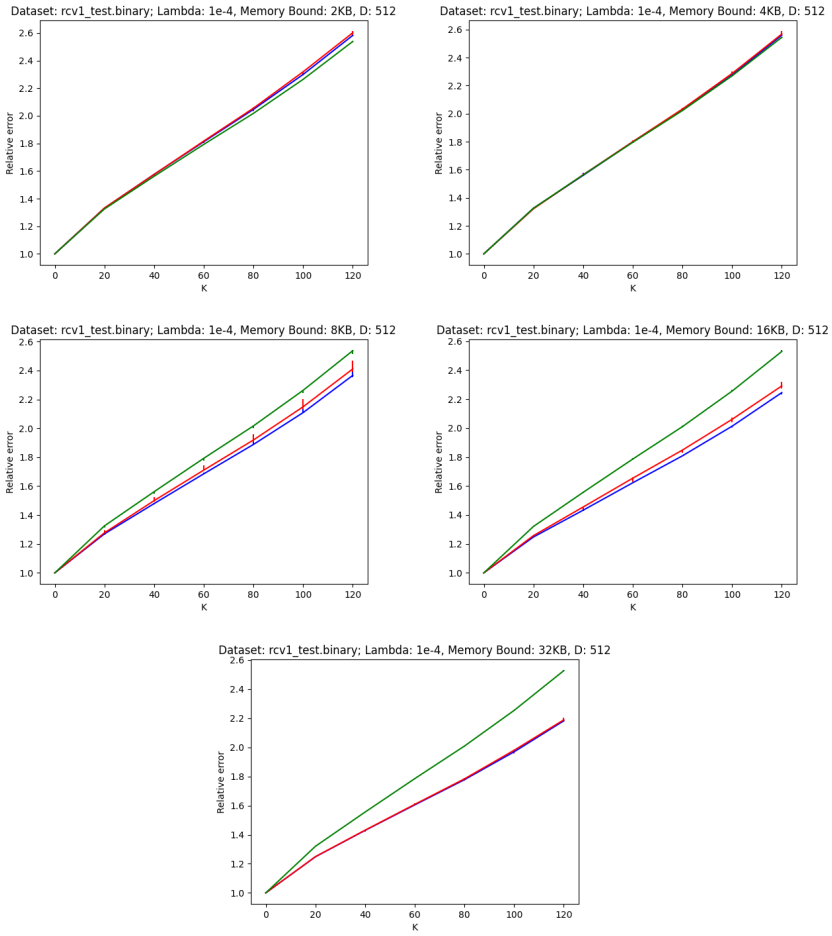

Figure 6: Results on RCV1 dataset with $\lambda = 10^{-4}$. Blue denotes the performance of our algorithm from Section 4. Green denotes the performance of our black-box reduction based algorithm from Section 2, and red is the algorithm of [1].

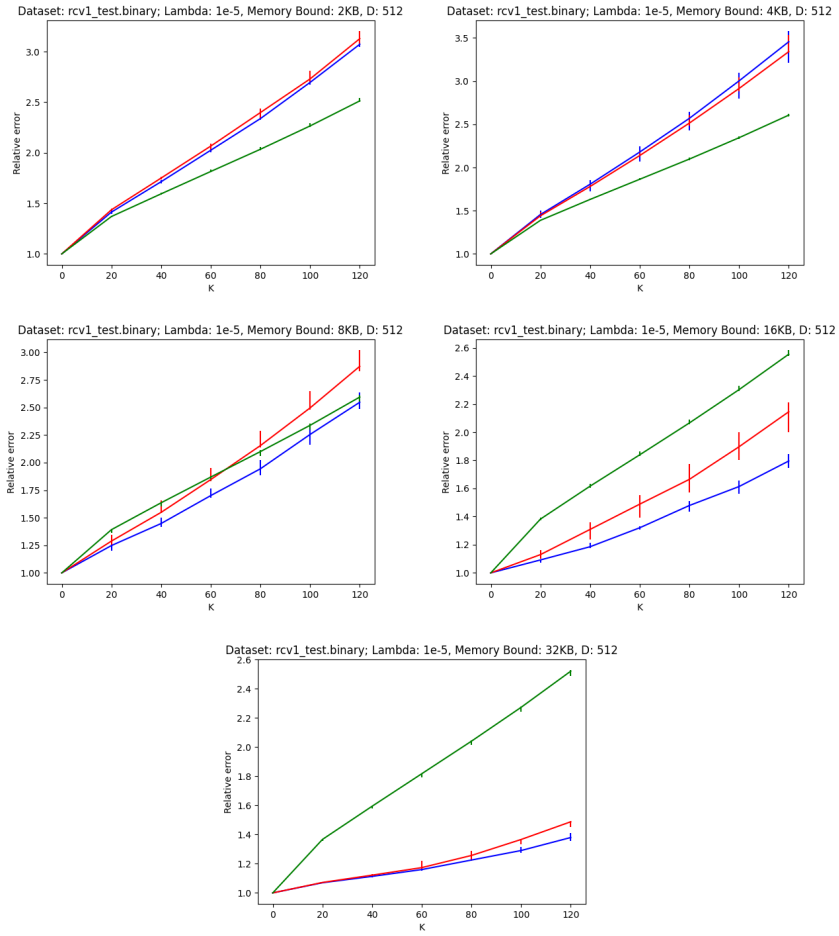

Figure 7: Results on RCV1 dataset with $\lambda = 10^{-5}$. Blue denotes the performance of our algorithm from Section 4. Green denotes the performance of our black-box reduction based algorithm from Section 2, and red is the algorithm of [1].

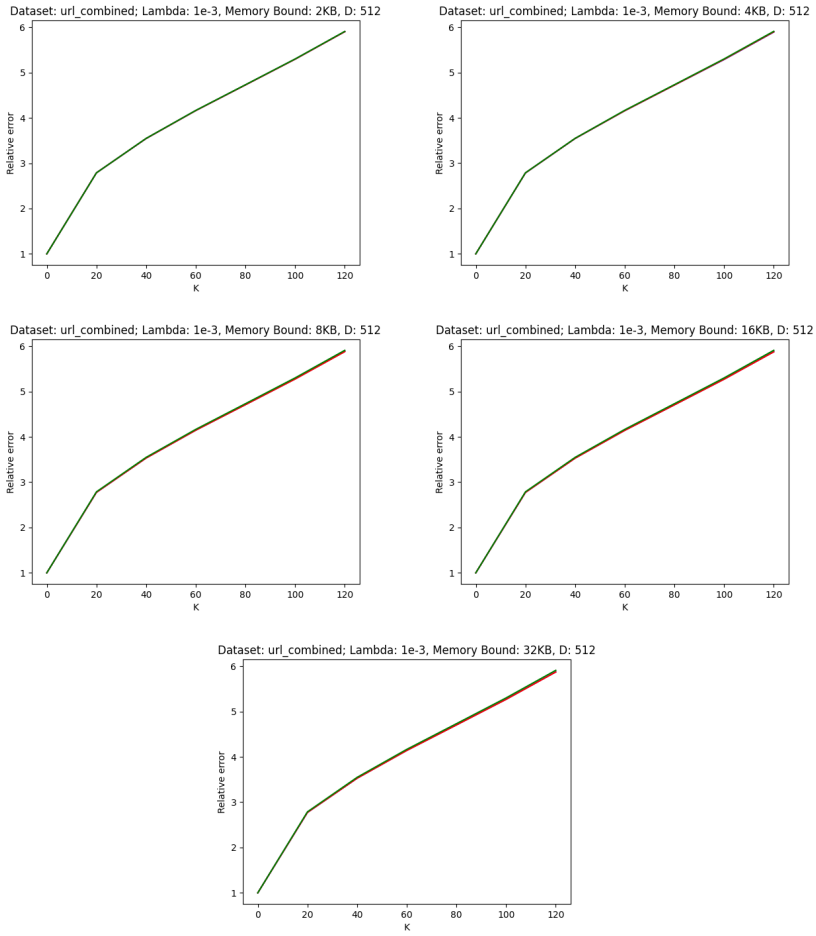

Figure 8: Results on URL dataset with $\lambda = 10^{-3}$. Blue denotes the performance of our algorithm from Section 4. Green denotes the performance of our black-box reduction based algorithm from Section 2, and red is the algorithm of [1].

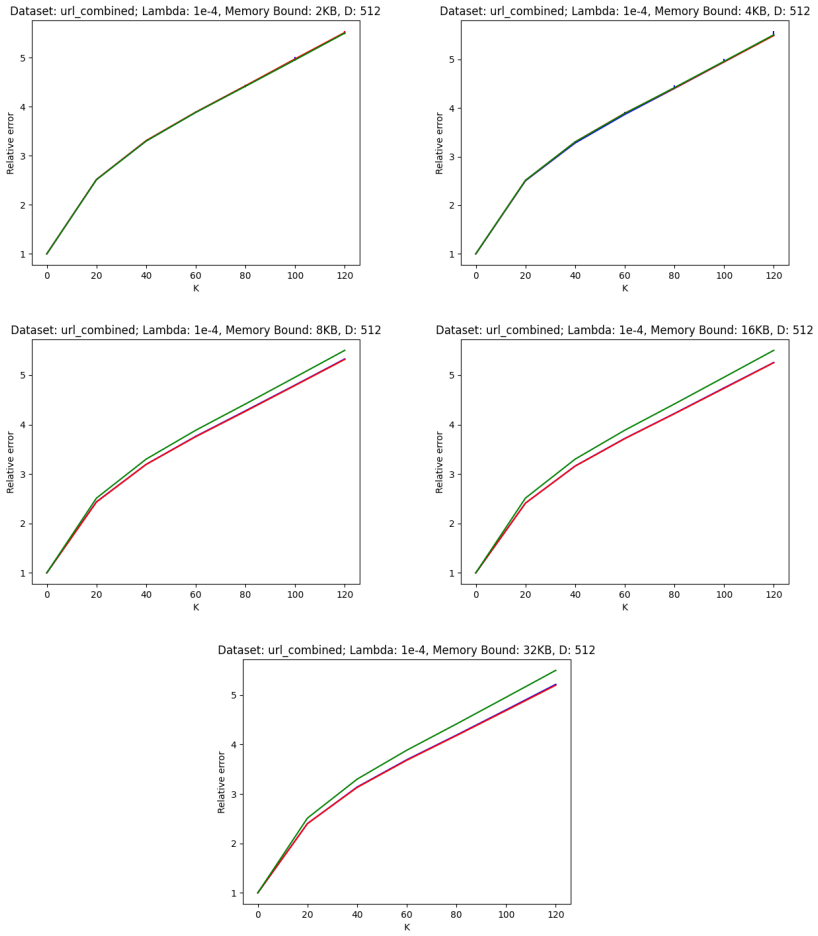

Figure 9: Results on URL dataset with $\lambda = 10^{-4}$. Blue denotes the performance of our algorithm from Section 4. Green denotes the performance of our black-box reduction based algorithm from Section 2, and red is the algorithm of [1].

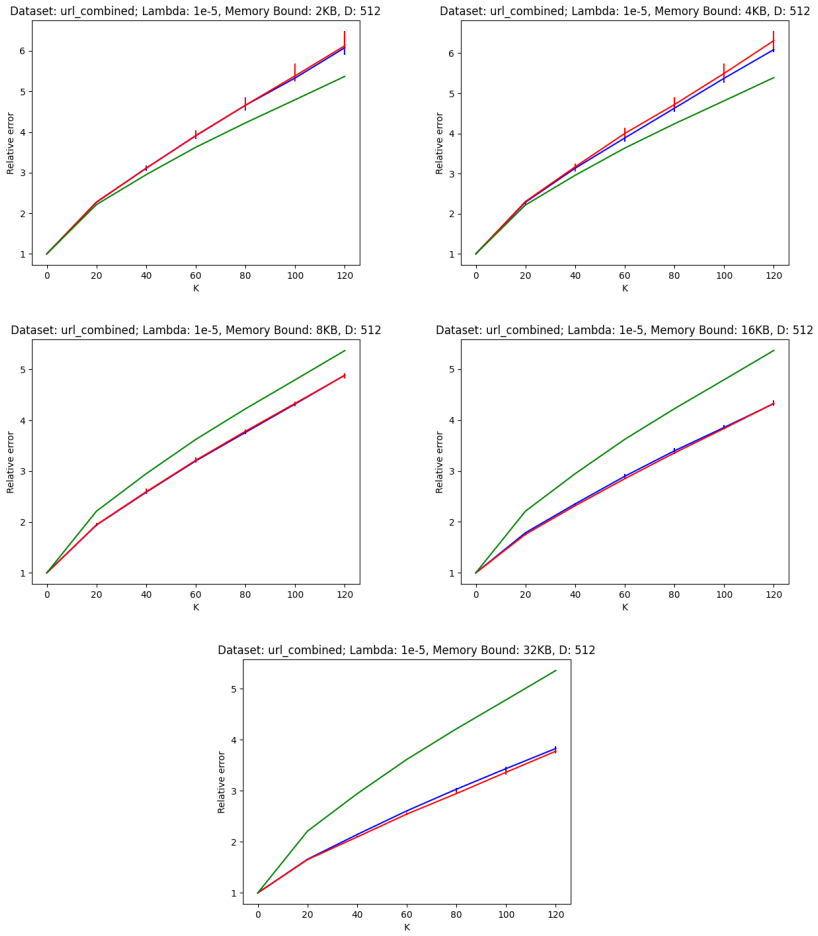

Figure 10: Results on URL dataset with $\lambda = 10^{-5}$. Blue denotes the performance of our algorithm from Section 4. Green denotes the performance of our black-box reduction based algorithm from Section 2, and red is the algorithm of [1].