A. Algorithms

A.1. Two-Phased DDAERR

Algorithm 2 Two-Phased DDAERR

```
Parameters: $m_1, m_2, \delta, B, \eta > 0$
Input: training set $S = \{(\mathbf{x}_t, y_t)\}_{t\in[m_1+m_2]}$ and $k > 0$
Output: regressor $\bar{\mathbf{w}}$ with $\|\bar{\mathbf{w}}\|_2 \le B$
Initialize $\mathbf{w}_1 \neq 0$, $\|\mathbf{w}_1\|_2 \le B$ arbitrarily
Initialize $\mathbf{A}$, $counts$ and $square\_sums$ - arrays of size $d$ with zeros
for $t = 1$ to $m_1$ do
    for $r = 1$ to $k + 1$ do
        Pick $i_{t,r} \in [d]$ uniformly at random
        $counts[i_{t,r}] \leftarrow counts[i_{t,r}] + 1$
        $square\_sums[i_{t,r}] \leftarrow square\_sums[i_{t,r}] + \mathbf{x}_t[i_{t,r}]^2$
    end for
end for
for $i = 1$ to $d$ do
    $\mathbf{A}[i] \leftarrow \frac{square\_sums[i]}{counts[i]}$
end for
$\epsilon \leftarrow \frac{d \log \frac{2d}{\delta}}{(k+1)m_1}$
Run GAERR with $q_i = \frac{\sqrt{\mathbf{A}[i]+\frac{13}{6}\epsilon}}{\sum_{j=1}^{d}\sqrt{\mathbf{A}[j]+\frac{13}{6}\epsilon}}$ on the following $m_2$ examples and return its output
```

A.2. GAELR

The GAELR algorithm is based in the EG algorithm with gradient estimates. The EG algorithm goes over the training set, and for each example builds an unbiased estimator of the gradient and clips it (where the *clip* operation is defined as $clip(x, c) = \max\{\min\{x, c\}, -c\}$) to make the updates more robust. Afterwards, the algorithm updates $\mathbf{w}_t$ by performing multiplicative updates of size η. The result is projected over the L_1 ball of size B, yielding $\mathbf{w}_{t+1}$. At the end, the algorithm outputs the average of all $\mathbf{w}_t$.

The gradient estimate is done here similarly to the GAERR algorithm: we use k attributes to estimate the data point $\mathbf{x}_t$ and 1 attribute to estimate the inner product. The only difference here is that here we use $p_{j_t} = |w_{t,j_t}| / \|\mathbf{w}_t\|_1$ when estimating the inner product, instead of $p_{j_t} = w_{t,j_t}^2 / \|\mathbf{w}_t\|_2^2$ as in the GAERR algorithm.

Algorithm 3 GAELR

```
Parameters: $B, \eta > 0$ and $q_i$ for $i \in [d]$
Input: training set $S = \{(\mathbf{x}_t, y_t)\}_{t\in[m]}$ and $k > 0$
Output: regressor $\bar{\mathbf{w}}$ with $\|\bar{\mathbf{w}}\|_1 \le B$
Initialize $\mathbf{z}_1^+ \leftarrow \mathbf{1}_d, \mathbf{z}_1^- \leftarrow \mathbf{1}_d$
for $t = 1$ to $m$ do
    $\mathbf{w}_t \leftarrow (\mathbf{z}_t^+ - \mathbf{z}_t^-) \cdot B / (\|\mathbf{z}_t^+\|_1 + \|\mathbf{z}_t^-\|_1)$
    for $r = 1$ to $k$ do
        Pick $i_{t,r} \in [d]$ with probability $q_{i_{t,r}}$ and observe $\mathbf{x}_t[i_{t,r}]$
        $\widetilde{\mathbf{x}}_{t,r} \leftarrow \frac{1}{q_{i_{t,r}}}\mathbf{x}_t[i_{t,r}] \cdot \mathbf{e}_{i_{t,r}}$
    end for
    $\widetilde{\mathbf{x}}_t \leftarrow \frac{1}{k}\sum_{r=1}^{k}\widetilde{\mathbf{x}}_{t,r}$
    Choose $j_t \in [d]$ with probability $p_{j_t} = \frac{|\mathbf{w}_t[j_t]|}{\|\mathbf{w}_t\|_1}$ and observe $\mathbf{x}_t[j_t]$
    $\widetilde{\phi}_t \leftarrow \frac{w_{t,j}}{p_j}\mathbf{x}_t[j_t] - y_t$
    $\widetilde{\mathbf{g}}_t \leftarrow \widetilde{\phi}_t \cdot \widetilde{\mathbf{x}}_t$
    for $i = 1$ to $d$ do
        $\bar{\mathbf{g}}_t[i] = clip(\widetilde{\mathbf{g}}_t[i], 1/\eta)$
        $\mathbf{z}_{t+1}^+[i] \leftarrow \mathbf{z}_t^+[i] \cdot \exp(-\eta\bar{\mathbf{g}}_t[i])$
        $\mathbf{z}_{t+1}^-[i] \leftarrow \mathbf{z}_t^-[i] \cdot \exp(+\eta\bar{\mathbf{g}}_t[i])$
    end for
end for
$\bar{\mathbf{w}} \leftarrow \frac{1}{m}\sum_{t=1}^{m}\mathbf{w}_t$
```

A.3. Two-Phased DDAELR

Algorithm 4 Two-Phased DDAELR

```
Parameters: $m_1, m_2, \delta, B, \eta > 0$
Input: training set $S = \{(\mathbf{x}_t, y_t)\}_{t\in[m_1+m_2]}$ and $k > 0$
Output: regressor $\bar{\mathbf{w}}$ with $\|\bar{\mathbf{w}}\| \le B$
Initialize $\mathbf{w_1} \neq \mathbf{0}$, $\|\mathbf{w_1}\|_2 \le B$ arbitrarily
Initialize $\mathbf{A}$, $counts$ and $square\_sums$ - arrays of size $d$ with zeros
for $t = 1$ to $m_1$ do
    for $r = 1$ to $k + 1$ do
        Pick $i_{t,r} \in [d]$ uniformly at random
        $counts[i_{t,r}] \leftarrow counts[i_{t,r}] + 1$
        $square\_sums[i_{t,r}] \leftarrow square\_sums[i_{t,r}] + \mathbf{x}_t[i_{t,r}]^2$
    end for
end for
for $i = 1$ to $d$ do
    $\mathbf{A}[i] \leftarrow \frac{square\_sums[i]}{counts[i]}$
end for
$\epsilon \leftarrow \min\left(\frac{d \log \frac{2d}{\delta}}{(k+1)m_1}, 1\right)$
Run GAELR with $q_i = \frac{\mathbf{A}[i]+\frac{13}{6}\epsilon}{\sum_{j=1}^{d}(\mathbf{A}[j]+\frac{13}{6}\epsilon)}$ on the following $m_2$ examples and return its output
```

B. Additional Experiment - Covertype

In this experiment we used the Covertype (Blackard & Dean, 1999) data set which aims to predict the forest cover type i.e. the dominant species of tree, from cartographic variables. This data set is designed for multi class classification, but we reduce it to binary classification by choosing one of the tree species and address the problem by regressing the -1 and $+1$ labels. For both the Ridge and Lasso scenarios, we used a budget of $k+1=5$. For this data set we have $d=54$, $\rho_{\text{Ridge}}=0.49$ and $\rho_{\text{Lasso}}=0.08$.

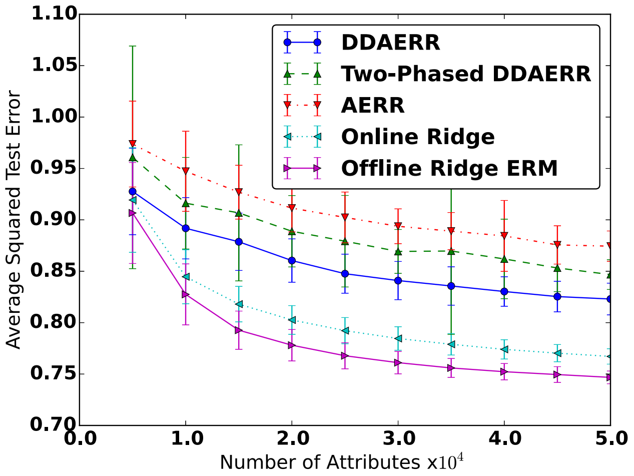

Figure 5. Test error for the algorithms with $k+1=5$ in the Ridge scenario over the classification task in the Cover Type data set.

The results for the Ridge scenario appear in figure 5: Again, our DDAERR algorithm performs considerably better than the AERR algorithm. Also, the DDAERR algorithm performs similarly to the online Ridge algorithm for a small number of examined attributes. The performance of the Two-Phased DDAERR is between those of the AERR algorithm and the DDAERR algorithm, and given a larger training set will probably converge towards the DDAERR algorithm as the number of observed attributes grow. This time, however, the full-information Ridge algorithms outperform the attribute efficient ones.

The results for the Lasso scenario in figure 6 are similar: The DDAELR algorithm performs better than the AELR algorithm. Also, the performance of the Two-Phased DDAELR is between those of the AELR algorithm and the DDAELR algorithm and converges towards the DDAELR algorithm, as the number of attributes grows. For a small number of examined attributes, the DDAELR algorithm performs similarly to the online Lasso algorithm but as the number of examined attributes grow, the algorithms drift apart.

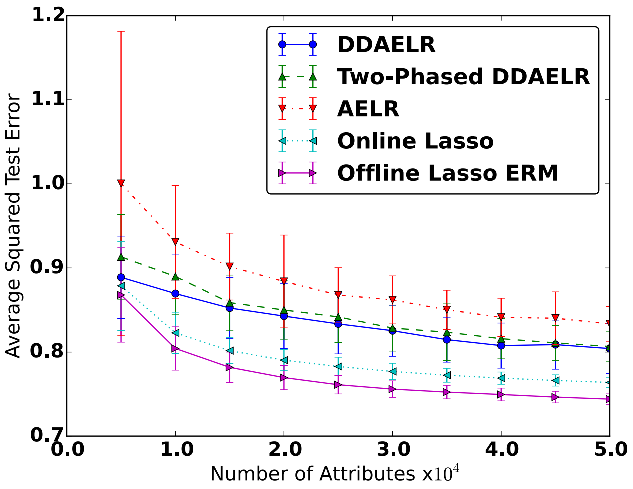

Figure 6. Test error for the algorithms with $k+1=5$ in the Lasso scenario over the classification task in the Cover Type data set.

C. Proofs

C.1. Proof of Theorem 3.1

We follow the path of the proof of Theorem 3.3 in (Hazan & Koren, 2012) by using the standard analysis of the OGD algorithm. Its expected excess risk bound is stated in the following lemma.

Lemma C.1 (Zinkevich, 2003)**.** *For any $\|\mathbf{w}^*\| \le B$, we have*

$$\sum_{t=1}^{m} \tilde{\mathbf{g}}_t^T \left(\mathbf{w}_t - \mathbf{w}^*\right) \le \frac{2B^2}{\eta} + \frac{\eta}{2} \sum_{t=1}^{m} \|\tilde{\mathbf{g}}_t\|_2^2 . \qquad (4)$$

To use this lemma, first we need to prove that the GAERR algorithm actually corresponds to OGD with unbiased gradient estimates, as implied by the following lemma:

Lemma C.2. *The vector $\tilde{\mathbf{g}}_t$ is an unbiased estimator of the gradient $\mathbf{g}_t = \left(\mathbf{w}_t^T \mathbf{x}_t - y_t\right) \mathbf{x}_t$, that is $\mathbb{E}_A\left[\tilde{\mathbf{g}}_t\right] = \mathbf{g}_t$.*

Now, we can take the expectation of equation (4) with respect to the randomization of the algorithm and the data distribution, and using Lemma C.2 we have

$$\mathbb{E}_{D,A}\left[\sum_{t=1}^{m} \mathbf{g}_t^T \left(\mathbf{w}_t - \mathbf{w}^*\right)\right] \le \frac{2B^2}{\eta} + \frac{\eta}{2} G^2 m.$$

On the other hand, the convexity of ℓ gives $\ell_t\left(\mathbf{w}_t\right) - \ell_t\left(\mathbf{w}^*\right) \le \mathbf{g}^T\left(\mathbf{w}_t - \mathbf{w}^*\right)$. Together with the above we have

$$\mathbb{E}_{D,A}\left[\frac{1}{m}\sum_{t=1}^{m} \ell_t\left(\mathbf{w}_t\right)\right] \le \mathbb{E}_{D,A}\left[\frac{1}{m}\sum_{t=1}^{m} \ell_t\left(\mathbf{w}^*\right)\right] + \frac{2B^2}{\eta m} + \frac{\eta}{2} G^2,$$

or

$$\mathbb{E}_{D,A}\left[\frac{1}{m}\sum_{t=1}^{m} L_{\mathcal{D}}\left(\mathbf{w}_t\right)\right] \le L_{\mathcal{D}}\left(\mathbf{w}^*\right) + \frac{2B^2}{\eta m} + \frac{\eta}{2}G^2,$$

Using the convexity of $L_{\mathcal{D}}$ and Jensen's inequality, the theorem follows.

Proof of Lemma C.2. First, it is straightforward to see $\mathbb{E}_A\left[\widetilde{\mathbf{x}}_{t,r}\right] = \mathbf{x}_t$ for all r thus also $\mathbb{E}_A\left[\widetilde{\mathbf{x}}_t\right] = \mathbf{x}_t$. Also, a simple calculation reveals that

$$\mathbb{E}_A\left[\widetilde{\phi}_t\right] = \sum_{j=1}^{d} p_j\left(\frac{w_{t,j}}{p_j}\mathbf{x}_t\left[j\right] - y_t\right) = \mathbf{w}_t^T\mathbf{x}_t - y_t.$$

Since $\widetilde{\mathbf{x}}_t$ and $\widetilde{\phi}_t$ are independent given $\mathbf{x}_t$, we obtain that $\mathbb{E}_A\left[\widetilde{\mathbf{g}}_t\right] = \left(\mathbf{w}_t^T\mathbf{x}_t - y_t\right)\cdot\mathbf{x}_t$, which is the required gradient. □

C.2. Proof of Lemma 3.2

We will use two auxiliary lemmas. The first will help us bound the 2-norm of the data point estimator.

Lemma C.3. *For every distribution* $(q_1,..,q_d)$ *where* $q_i \ge 0$ *and* $\sum_{i=1}^{d} q_i = 1$, *we have* $\mathbb{E}_{D,A}\left[\|\widetilde{\mathbf{x}}_t\|_2^2\right] \le \frac{1}{k}\mathbb{E}_{D,A}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right] + \frac{k-1}{k}\mathbb{E}_D\left[\|\mathbf{x}\|_2^2\right].$

The second will help us bound the square of the estimator of the inner product (minus the label).

Lemma C.4. *Using our sampling method we have* $\mathbb{E}_{D,A}\left[\widetilde{\phi}_t^{\,2}\right] \le 4B^2$.

Now, the lemma follows directly from Lemmas C.3 and C.4, using the independence of $\widetilde{\mathbf{x}}_t$ and $\widetilde{\phi}_t$ given $\mathbf{x}_t$ and $\|\mathbf{x}\|_2 \le 1$.

Proof of Lemma C.3. From the definition of $\widetilde{\mathbf{x}}_t$,

$$\begin{aligned}\mathbb{E}_{D,A}\left[\|\widetilde{\mathbf{x}}_t\|_2^2\right] &= \frac{1}{k^2}\mathbb{E}_{D,A}\left[\left\|\sum_{r=1}^{k}\widetilde{\mathbf{x}}_{t,r}\right\|_2^2\right]\\ &= \frac{1}{k^2}\sum_{r=1}^{k}\mathbb{E}_{D,A}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right]+\\ &\qquad \frac{1}{k^2}\sum_{r=1}^{k}\sum_{s\neq r}^{k}\mathbb{E}_{D,A}\left[\langle\widetilde{\mathbf{x}}_{t,r},\widetilde{\mathbf{x}}_{t,s}\rangle\right].\end{aligned}$$

Since $\widetilde{\mathbf{x}}_{t,r}$ and $\widetilde{\mathbf{x}}_{t,s}$ are independent of each other and $\mathbb{E}_{D,A}\left[\widetilde{\mathbf{x}}_{t,r}\right] = \mathbb{E}_D\left[\mathbf{x}\right]$, we finally have

$$\begin{aligned}\mathbb{E}_{D,A}\left[\|\widetilde{\mathbf{x}}_t\|_2^2\right] &= \frac{1}{k}\mathbb{E}_{D,A}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right] + \frac{k^2-k}{k^2}\left\|\mathbb{E}_D\left[\mathbf{x}\right]\right\|_2^2\\ &= \frac{1}{k}\mathbb{E}_{D,A}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right] + \frac{k-1}{k}\left\|\mathbb{E}_D\left[\mathbf{x}\right]\right\|_2^2.\end{aligned}$$

Using the convexity of the 2-norm and Jensen's inequality, the lemma follows. □

Proof of Lemma C.4. Recalling $|y_t| \le B$ and using the inequality $(a-b)^2 \le 2\left(a^2+b^2\right)$, by a straightforward calculation we obtain

$$\begin{aligned}\mathbb{E}_{D,A}\left[\widetilde{\phi}_t^{\,2}\right] &= \mathbb{E}_{D,A}\left[\left(\frac{w_{t,j}}{p_j}\mathbf{x}_t\left[j_t\right] - y_t\right)^2\right]\\ &\le 2\mathbb{E}_{D,A}\left[\left(\frac{w_{t,j}}{p_j}\mathbf{x}_t\left[j_t\right]\right)^2 + y_t^2\right]\\ &\le 2\sum_{j=1}^{d}\frac{1}{p_j}w_{t,j}^2\mathbb{E}_D\left[x_j^2\right] + 2B^2\\ &= 2\left\|\mathbf{w}_t\right\|_2^2\mathbb{E}_D\left[\|\mathbf{x}\|_2^2\right] + 2B^2\\ &\le 4B^2.\end{aligned}$$

□

C.3. Proof of Theorem 3.3

The theorem follows directly from Theorem 3.1, Lemma 3.2, equation (1) and the calculated q_i-s in equation (2).

C.4. Proof of Theorem 3.4

The main goal of the proof is to bound the expected squared 2-norm of the gradient estimator from above. By using Lemma 3.2, all that remains is to upper bound $\mathbb{E}_{D,A_2}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right]$. In the next lemma we show two different upper bounds on $\mathbb{E}_{D,A_2}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right]$. The first states that with probability 1 over the first phase $\mathbb{E}_{D,A_2}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right] \le 5d$, meaning that up to a constant factor the bound is the same as in the AERR algorithm. The second bound decreases in ϵ, and will help up to analyze the convergence rate of the algorithm.

Lemma C.5. *For all* m_1 *and* $t > m_1$, *with probability* 1 *over the first phase, we have*

$$\mathbb{E}_{D,A_2}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right] \le 5d,$$

and with probability $\ge 1-\delta$ *over the first phase, we have*

$$\mathbb{E}_{D,A_2}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right] \le 2\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} + 2\sqrt{\frac{5\epsilon}{3}}d\sqrt{\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}}}.$$

The proof can be found in Appendix C.5.

We will treat each bound separately, and later join the results into a single lemma. First, we prove that with a proper choice of η, the bound of our Two-Phased DDAERR algorithm is with probability 1 over the first phase equal to the bound of the AERR algorithm, up to a constant factor.

Lemma C.6. *Let $\bar{\mathbf{w}}$ be the output of Two-Phased DDAERR when run with $\eta = \sqrt{\frac{k}{6dm_2}}$. Then with probability 1 over the first phase, we have for all m_1 and for any $\mathbf{w}^* \in \mathbb{R}^d$ with $\|\mathbf{w}^*\|_2 \leq B$,*

$$\mathbb{E}_{D,A_2}\left[L_{\mathcal{D}}\left(\bar{\mathbf{w}}\right)\right] - L_{\mathcal{D}}\left(\mathbf{w}^*\right) \leq 4B^2\sqrt{\frac{6d}{km_2}}.$$

The proof can be found in Appendix C.6.

Assume for simplicity that we have an estimator for $\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}}$ that satisfies $H \geq \left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}}$. We can use it to calculate an appropriate step size and to bound the risk, as shown in the next lemma.

Lemma C.7. *Assume we have a value H that satisfies $H \geq \left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}}$. Let $\bar{\mathbf{w}}$ be the output of Two-Phased DDAERR when run with $\eta = \frac{1}{\sqrt{m_2\left(\frac{2}{k}H+\frac{2}{k}\sqrt{\frac{5}{3}}d\sqrt{H}\sqrt{\epsilon}+1\right)}}$. Then with probability $\geq 1-\delta$ over the first phase, we have for all m_1 and for any $\mathbf{w}^* \in \mathbb{R}^d$ with $\|\mathbf{w}^*\|_2 \leq B$,*

$$\mathbb{E}_{D,A_2}\left[L_{\mathcal{D}}\left(\bar{\mathbf{w}}\right)\right] - L_{\mathcal{D}}\left(\mathbf{w}^*\right) \leq \frac{4B^2}{\sqrt{m_2}}\sqrt{\frac{2}{k}H + \frac{2}{k}\sqrt{\frac{5}{3}}d\sqrt{H}\sqrt{\epsilon} + 1}.$$

The proof can be found in Appendix C.7.

This lemma gives a non-trivial expected excess risk bound only if ϵ is small enough, but when m_1 is small, this is not necessarily the case. Therefore, we would like to unite these two lemmas to ensure that even in the worst case, we will not have a worse bound than the AERR algorithm.

Lemma C.8. *Assume we have a value H that satisfies $H \geq \left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}}$. Let $\bar{\mathbf{w}}$ be the output of Two-Phased DDAERR when run with $\eta = \max\left(\eta_1, \eta_2\right)$ where $\eta_1 = \sqrt{\frac{k}{6dm_2}}$ and $\eta_2 = \sqrt{\frac{k}{m_2\left(2H+2\sqrt{\frac{5}{3}}d\sqrt{H}\sqrt{\frac{d\log\frac{2d}{\delta}}{(k+1)m_1}}+k\right)}}$.*

Then for all m_1 and for any $\mathbf{w}^* \in \mathbb{R}^d$ with $\|\mathbf{w}^*\|_2 \leq B$, with probability 1 over the first phase, we have

$$\mathbb{E}_{D,A_2}\left[L_{\mathcal{D}}\left(\bar{\mathbf{w}}\right)\right] - L_{\mathcal{D}}\left(\mathbf{w}^*\right) \leq \frac{4B^2}{\sqrt{m_2}}\sqrt{\frac{6d}{k}}.$$

Also, with probability $\geq 1-\delta$ over the first phase, we have

$$\mathbb{E}_{D,A_2}\left[L_{\mathcal{D}}\left(\bar{\mathbf{w}}\right)\right] - L_{\mathcal{D}}\left(\mathbf{w}^*\right) \leq \frac{4B^2}{\sqrt{m_2}}\sqrt{\frac{2}{k}H + \frac{2}{k}\sqrt{\frac{5}{3}}d\sqrt{H}\sqrt{\frac{d\log\frac{2d}{\delta}}{(k+1)\,m_1}} + 1}.$$

The proof can be found in Appendix C.8.

The last thing we require is an estimator for $\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}}$. We could always naively bound $\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}}$ from above by d, but then, even if m_1 tends to infinity, the bound of the algorithm will not be better than the bound of the AERR algorithm. A better estimator is stated in the next lemma:

Lemma C.9. *The estimator $H = \left\|2\mathbf{A} + \frac{10}{3}\epsilon\right\|_{\frac{1}{2}}$, satisfies, with probability $\geq 1-\delta$ over the first phase, $\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} \leq H \leq 8\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} + \frac{34}{3}d^2\epsilon$.*

The proof can be found in Appendix C.9.

Finally, the proof of the main theorem is straightforward, using Lemma C.8, Lemma C.9 and some algebraic manipulations.

C.5. Proof of Lemma C.5

First, we state a simple probabilistic lemma that will be used to bound our estimates for the second moment of the attributes. The proof appears a bit later in the section.

Lemma C.10. *Let $Z_1, Z_2, ..., Z_n$ be i.i.d random variables. $Z_i \in [0,1]$. Let $\hat{\mathbb{E}}[Z] = \frac{1}{n}\sum_{i=1}^{n} Z_i$ be their average. Then, with probability $\geq 1-\delta$*

$$\hat{\mathbb{E}}\left[Z\right] \leq 2\mathbb{E}\left[Z\right] + \frac{7\log\frac{1}{\delta}}{6n}.$$

Also, with probability $\geq 1-\delta$

$$\hat{\mathbb{E}}\left[Z\right] \geq \frac{1}{2}\mathbb{E}\left[Z\right] - \frac{5\log\frac{1}{\delta}}{3n}.$$

We prefer to use this lemma rather than a direct application of the more standard Hoeffding or Bernstein inequality, because we are interested in a fast convergence rate of $\frac{1}{n}$, and are willing to pay the price of an additional constant factor in front of the expectation.

To prove our lemma, we use the definition of $\|\widetilde{\mathbf{x}}_{t,r}\|_2^2$,

$$\begin{aligned}\mathbb{E}_{D,A_2}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right] &= \mathbb{E}_{D,A_2}\left[\widetilde{\mathbf{x}}_{t,r}\left[i_{t,r}\right]^2\right] \\ &= \sum_{i=1}^{d}\frac{1}{q_i}\mathbb{E}_D\left[x_i^2\right] \\ &= \sum_{j=1}^{d}\sqrt{A\left[j\right] + \frac{13}{6}\epsilon}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{A\left[i\right] + \frac{13}{6}\epsilon}}.\end{aligned}$$

For all $i \in [d]$ let T_i be a random variable describing the amount of times the algorithm sampled the i-th attribute during the first phase. For every realization t_i of T_i, since T_i and the samples themselves are independent, we can use Lemma C.10 and by the union bound have that with probability larger than $1-\delta$, $\mathbf{A}\left[i\right] \leq 2\mathbb{E}_D\left[x_i^2\right] + \frac{7}{6}\mathbb{E}_{A_1}\left[\epsilon_i\right]$, and

$\mathbf{A}\left[i\right] \geq \frac{1}{2}\mathbb{E}_D\left[x_i^2\right] - \frac{5}{3}\mathbb{E}_{A_1}\left[\epsilon_i\right]$ where $\epsilon_i = \frac{\log\frac{2d}{\delta}}{t_i}$. Clearly, $\mathbb{E}_{A_1}\left[T_i\right] = \frac{(k+1)m_1}{d}$, and using the convexity of $f\left(x\right) = \frac{1}{x}$ we have $\mathbb{E}_{A_1}\left[\epsilon_i\right] \geq \frac{d\log\frac{2d}{\delta}}{(k+1)m_1} = \epsilon$. Therefore, with probability $\geq 1-\delta$ over the first phase, we have

$$\begin{cases} \mathbf{A}\left[i\right] \leq 2\mathbb{E}_D\left[x_i^2\right] + \frac{7}{6}\epsilon \\ \mathbf{A}\left[i\right] \geq \frac{1}{2}\mathbb{E}_D\left[x_i^2\right] - \frac{5}{3}\epsilon. \end{cases} \tag{5}$$

Note that these equations also hold trivially for any $\epsilon \geq 1$ as with probability 1 we have $x_i^2 \leq 1$ for all $i \in [d]$.

Now we can continue and get that,

$$\begin{aligned}
&\mathbb{E}_{D,A_2}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right] \\
&\leq \sum_{j=1}^{d}\sqrt{2\mathbb{E}_D\left[x_i^2\right] + \frac{7}{6}\epsilon + \frac{13}{6}\epsilon}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\frac{1}{2}\mathbb{E}_D\left[x_i^2\right] - \frac{5}{3}\epsilon + \frac{13}{6}\epsilon}} \\
&= \sum_{j=1}^{d}\sqrt{2\left(\mathbb{E}_D\left[x_i^2\right] + \frac{5}{3}\epsilon\right)}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\frac{1}{2}\left(\mathbb{E}_D\left[x_i^2\right] + \epsilon\right)}} \\
&= 2\sum_{j=1}^{d}\sqrt{\mathbb{E}_D\left[x_i^2\right] + \frac{5}{3}\epsilon}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\mathbb{E}_D\left[x_i^2\right] + \epsilon}}.
\end{aligned}$$

We shall bound this value in two ways. For the first part of the lemma, we have

$$\begin{aligned}
&\mathbb{E}_{D,A_2}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right] \\
&\leq 2\sum_{j=1}^{d}\sqrt{\mathbb{E}_D\left[x_j^2\right] + \frac{5}{3}\epsilon}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\mathbb{E}_D\left[x_i^2\right] + \epsilon}} \\
&\leq 2\sum_{j=1}^{d}\sqrt{\mathbb{E}_D\left[x_j^2\right]}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\mathbb{E}_D\left[x_i^2\right] + \epsilon}} \\
&\qquad + 2\sum_{j=1}^{d}\sqrt{\frac{5}{3}\epsilon}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\mathbb{E}_D\left[x_i^2\right] + \epsilon}}.
\end{aligned}$$

Continuing, we upper bound the above by

$$\begin{aligned}
&\leq 2\sum_{j=1}^{d}\sqrt{\mathbb{E}_D\left[x_j^2\right]}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\mathbb{E}_D\left[x_i^2\right]}} \\
&\qquad + 2\sum_{j=1}^{d}\sqrt{\frac{5}{3}\epsilon}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\epsilon}} \\
&\leq 2\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} + 2d\sqrt{\frac{5}{3}}\sum_{i=1}^{d}\mathbb{E}_D\left[x_i^2\right] \\
&\leq 2\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} + 2\sqrt{\frac{5}{3}}d \\
&\leq 5d.
\end{aligned}$$

As this bound is independent of ϵ, it holds with probability 1 over the first phase.

For the second part of the lemma, we have with probability $\geq 1-\delta$,

$$\begin{aligned}
&\mathbb{E}_{D,A_2}\left[\|\widetilde{\mathbf{x}}_{t,r}\|_2^2\right] \\
&\leq 2\sum_{j=1}^{d}\sqrt{\mathbb{E}_D\left[x_j^2\right] + \frac{5}{3}\epsilon}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\mathbb{E}_D\left[x_i^2\right] + \epsilon}} \\
&\leq 2\sum_{j=1}^{d}\sqrt{\mathbb{E}_D\left[x_j^2\right] + \frac{5}{3}\epsilon}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\mathbb{E}_D\left[x_i^2\right]}} \\
&\leq 2\sum_{j=1}^{d}\sqrt{\mathbb{E}_D\left[x_j^2\right]}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\mathbb{E}_D\left[x_i^2\right]}} \\
&\qquad + 2\sum_{j=1}^{d}\sqrt{\frac{5}{3}\epsilon}\sum_{i=1}^{d}\frac{\mathbb{E}_D\left[x_i^2\right]}{\sqrt{\mathbb{E}_D\left[x_i^2\right]}} \\
&\leq 2\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} + 2\sqrt{\frac{5}{3}}d\sqrt{\|\mathbb{E}_D\left[\mathbf{x}^2\right]\|_{\frac{1}{2}}}\sqrt{\epsilon},
\end{aligned}$$

which concludes the proof.

Proof of Lemma C.10. Let us denote the variance of Z by $\sigma^2 = \mathbb{E}\left[Z^2\right] - \mathbb{E}\left[Z\right]^2$. By Bernstein's inequality, with probability $\geq 1-\delta$, we have

$$\hat{\mathbb{E}}\left[Z\right] \leq \mathbb{E}\left[Z\right] + \sqrt{\frac{2\sigma^2\log\frac{1}{\delta}}{n}} + \frac{2\log\frac{1}{\delta}}{3n}.$$

Using $Z_i \in [0,1]$, we obtain $\sigma^2 = \mathbb{E}\left[Z^2\right] - \mathbb{E}\left[Z\right]^2 \leq \mathbb{E}\left[Z^2\right] \leq \mathbb{E}\left[Z\right]$. Plugging back in the expression for $\hat{\mathbb{E}}\left[Z\right]$,

$$\hat{\mathbb{E}}\left[Z\right] \leq \mathbb{E}\left[Z\right] + \sqrt{\frac{2\mathbb{E}\left[Z\right]\log\frac{1}{\delta}}{n}} + \frac{2\log\frac{1}{\delta}}{3n}.$$

Using the fact that the geometric mean is smaller or equal to the arithmetic mean, we have

$$\hat{\mathbb{E}}\left[Z\right] \leq \mathbb{E}\left[Z\right] + \frac{2\mathbb{E}\left[Z\right]}{2} + \frac{\log\frac{1}{\delta}}{2n} + \frac{2\log\frac{1}{\delta}}{3n}$$

or,

$$\hat{\mathbb{E}}\left[Z\right] \leq 2\mathbb{E}\left[Z\right] + \frac{7\log\frac{1}{\delta}}{6n},$$

which concludes the first part of the proof.

Similarly, by Bernstein's inequality again, with probability $\geq 1-\delta$, we have

$$\hat{\mathbb{E}}\left[Z\right] \geq \mathbb{E}\left[Z\right] - \sqrt{\frac{2\sigma^2\log\frac{1}{\delta}}{n}} - \frac{2\log\frac{1}{\delta}}{3n}.$$

Using $\sigma^2 \leq \mathbb{E}\left[Z\right]$, this turns to

$$\hat{\mathbb{E}}\left[Z\right] \geq \mathbb{E}\left[Z\right] - \sqrt{\frac{2\mathbb{E}\left[Z\right] \log \frac{1}{\delta}}{n}} - \frac{2 \log \frac{1}{\delta}}{3n}.$$

Again using the fact that the geometric mean is smaller or equal to the arithmetic mean, we have

$$\hat{\mathbb{E}}\left[Z\right] \geq \mathbb{E}\left[Z\right] - \frac{\mathbb{E}\left[Z\right]}{2} - \frac{2 \log \frac{1}{\delta}}{2n} - \frac{2 \log \frac{1}{\delta}}{3n}$$

or,

$$\hat{\mathbb{E}}\left[Z\right] \geq \frac{1}{2}\mathbb{E}\left[Z\right] - \frac{5 \log \frac{1}{\delta}}{3n},$$

which concludes the proof. □

C.6. Proof of Lemma C.6

First, using Theorem 3.1 on the second phase of the algorithm, we have

$$\mathbb{E}_{D,A_2}\left[L_{\mathcal{D}}\left(\bar{\mathbf{w}}\right)\right] - L_{\mathcal{D}}\left(\mathbf{w}^*\right) \leq \frac{2B^2}{\eta m_2} + \frac{\eta}{2}G^2. \tag{6}$$

Now we use the first part of Lemma C.5, plug it into Lemma 3.2 and obtain that with probability 1, we have $G^2 \leq 4B^2\left(\frac{5d}{k}+1\right) \leq 24B^2\frac{d}{k}$. Plugging $\eta = \sqrt{\frac{k}{6dm_2}}$ into equation (6) finishes the proof.

C.7. Proof of Lemma C.7

We use the second part of Lemma C.5, plug it into Lemma 3.2 and obtain that with probability $\geq 1-\delta$, we have $G^2 \leq 4B^2\left(\frac{2}{k}\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} + \frac{2}{k}\sqrt{\frac{5}{3}}d\sqrt{\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}}}\sqrt{\epsilon}+1\right)$. We denote $\widehat{G^2} = 4B^2\left(\frac{2}{k}H + \frac{2}{k}\sqrt{\frac{5}{3}}d\sqrt{H}\sqrt{\epsilon}+1\right)$. Since $H \geq \left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}}$ we have $G^2 \leq \widehat{G^2}$. Plugging $\eta = \frac{2B}{\sqrt{\widehat{G^2}m_2}} = \frac{1}{\sqrt{m_2\left(\frac{2}{k}H+\frac{2}{k}\sqrt{\frac{5}{3}}d\sqrt{H}\sqrt{\epsilon}+1\right)}}$ into equation (6), we get

$$\begin{aligned}\mathbb{E}_{D,A_2}\left[L_{\mathcal{D}}\left(\bar{\mathbf{w}}\right)\right] - L_{\mathcal{D}}\left(\mathbf{w}_*\right) & \\ &\leq \frac{2B^2}{\eta m_2} + \frac{\eta}{2}G^2 \\ &\leq \frac{2B^2}{\eta m_2} + \frac{\eta}{2}\widehat{G^2} \\ &\leq \frac{2B}{\sqrt{m_2}}\sqrt{\widehat{G^2}} \\ &= \frac{4B^2}{\sqrt{m_2}}\sqrt{\frac{2}{k}H + \frac{2}{k}\sqrt{\frac{5}{3}}d\sqrt{H}\sqrt{\epsilon}+1}.\end{aligned}$$

C.8. Proof of Lemma C.8

First, we state a simple lemma that will allow us to combine two risk bounds, each is achieved by a different value of η.

Lemma C.11. *Let $f\left(\eta\right) = \frac{A}{\eta} + \eta BG^2$ for some positive constants A, B, G, where $G \leq \min\left(G_1, G_2\right)$. Let $\eta_i = \frac{1}{G_i}\sqrt{\frac{A}{B}}$ for $i = 1, 2$. Then $f\left(\max\left(\eta_1, \eta_2\right)\right) \leq \min\left(f\left(\eta_1\right), f\left(\eta_2\right)\right)$.*

By Lemma C.6 , using $\eta = \sqrt{\frac{k}{12dm_2}}$, we have with probability 1,

$$\mathbb{E}_{D,A_2}\left[L_{\mathcal{D}}\left(\bar{\mathbf{w}}\right)\right] - L_{\mathcal{D}}\left(\mathbf{w}^*\right) \leq 4B^2\sqrt{\frac{6d}{km_2}}.$$

Similarly, by Lemma C.7, using $\eta = \frac{1}{\sqrt{m_2\left(\frac{2}{k}H+\frac{2}{k}\sqrt{\frac{5}{3}}d\sqrt{H}\sqrt{\frac{d\log\frac{2d}{\delta}}{(k+1)m_1}}+1\right)}}$, we have with probability $\geq 1-\delta$,

$$\begin{aligned}&\mathbb{E}_{D,A_2}\left[L_{\mathcal{D}}\left(\bar{\mathbf{w}}\right)\right] - L_{\mathcal{D}}\left(\mathbf{w}^*\right) \\ &\leq \frac{4B^2}{\sqrt{m_2}}\sqrt{\frac{2}{k}H + \frac{2}{k}\sqrt{\frac{5}{3}}d\sqrt{H}\sqrt{\frac{d\log\frac{2d}{\delta}}{\left(k+1\right)m_1}}+1}.\end{aligned}$$

Using Theorem 3.1, the expected excess risk bound has the form of the function in Lemma C.11, and the theorem follows directly.

Proof of Lemma C.11. Assume without loss of generality that $G_1 \geq G_2$, therefore we also have $\eta_2 \geq \eta_1$. It is enough to prove $f\left(\eta_2\right) \leq f\left(\eta_1\right)$ which follows directly by simple algebraic manipulations. □

C.9. Proof of Lemma C.9

First, using the second inequality in equation (5) we have with probability $\geq 1-\delta$, that $\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} \leq \left\|2\mathbf{A} + \frac{10}{3}\epsilon\right\|_{\frac{1}{2}}$. Using the first inequality in equation (5) and the identity $\left\|\mathbf{a}+\mathbf{b}\right\|_{\frac{1}{2}} \leq 2\left\|\mathbf{a}\right\|_{\frac{1}{2}} + 2\left\|\mathbf{b}\right\|_{\frac{1}{2}}$ we can see that with probability $\geq 1-\delta$,

$$\begin{aligned}\left\|2\mathbf{A} + \frac{10}{3}\epsilon\right\|_{\frac{1}{2}} \leq &\left\|4\mathbb{E}_D\left[\mathbf{x}^2\right] + \frac{14}{6}\epsilon + \frac{10}{3}\epsilon\right\|_{\frac{1}{2}} \\ \leq &8\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_{\frac{1}{2}} + \frac{34}{3}d^2\epsilon.\end{aligned} \tag{7}$$

C.10. Proof of Theorem 4.1

Our analysis is based on the analysis in (Hazan & Koren, 2012) and brought here for completeness. First, we state the second-order bound for the EG algorithm.

Lemma C.12 (simplified version of Lemma II.3 of (Clarkson et al., 2012))**.** *Let $\eta > 0$, and let $\mathbf{c}_1, .., \mathbf{c}_t$ be an arbitrary sequence of vectors in $\mathbb{R}^n$, with $\mathbf{c}_t[i] \geq -\frac{1}{\eta}$ for all t and all $i \in [n]$. Define a sequence $\mathbf{z}_1, .., .\mathbf{z}_T$ by letting $\mathbf{z}_1 = \mathbf{1}_n$ and for $t \geq 1$,*

$$\mathbf{z}_{t+1}[i] = \mathbf{z}_t[i] \cdot \exp\left(-\eta \mathbf{c}_t[i]\right) \qquad i = 1, .., n.$$

Then, for the vectors $\mathbf{p}_t = \frac{\mathbf{z}'_t}{\|\mathbf{z}'_t\|_1}$ we have

$$\sum_{t=1}^{m} \mathbf{p}_t^T \mathbf{c}_t \leq \min_{i \in [n]} \sum_{t=1}^{m} \mathbf{c}_t[i] + \frac{\log n}{\eta} + \eta \sum_{t=1}^{m} \mathbf{p}_t^T \mathbf{c}_t^2.$$

Now we examine the vectors $\mathbf{z}' = \left(\mathbf{z}_t^+, \mathbf{z}_t^-\right) \in \mathbb{R}^{2d}$ and $\bar{\mathbf{g}}'_t = (\bar{\mathbf{g}}, -\bar{\mathbf{g}}) \in \mathbb{R}^{2d}$, and setting $\mathbf{p}_t = \frac{\mathbf{z}'_t}{\|\mathbf{z}'_t\|_1}$. We have the following lemma:

Lemma C.13 (Lemma 3.5 of (Hazan & Koren, 2012))**.**

$$\sum_{t=1}^{m} \mathbf{p}_t^T \bar{\mathbf{g}}'_t \leq \min_{i \in [2d]} \sum_{t=1}^{m} \bar{\mathbf{g}}'_t[i] + \frac{\log 2d}{\eta} + \eta \sum_{t=1}^{m} \mathbf{p}_t^T \left(\bar{\mathbf{g}}'_t\right)^2.$$

Using this lemma, we establish an expected excess risk bound with respect to the clipped linear functions $\bar{\mathbf{g}}_t^T \mathbf{w}$:

Lemma C.14 (Lemma 3.6 of (Hazan & Koren, 2012))**.** *Assume that $\left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{g}}_t^2\right]\right\|_\infty \leq G^2$ for all t, for some $G \geq 0$. Then, for any $\|\mathbf{w}^*\|_1 \leq B$, we have*

$$\mathbb{E}_{D,A}\left[\sum_{t=1}^{m} \bar{\mathbf{g}}_t^T \mathbf{w}_t\right] \leq \mathbb{E}_{D,A}\left[\sum_{t=1}^{m} \bar{\mathbf{g}}_t^T \mathbf{w}^*\right] + B\left(\frac{\log 2d}{\eta} + \eta G^2 m\right).$$

For the proof of Lemma C.16 we will need a simple lemma, that allows us to bound the deviation of the expected value of a clipped random variable from that of the original variable, in terms of its variance.

Lemma C.15. *Let X be a random variable with $|\mathbb{E}[X]| \leq \frac{C}{2}$ for some $C > 0$. Then for the clipped variable $\bar{X} = \operatorname{clip}(X, C) = \max\{\min\{X, C\}, -C\}$ we have*

$$\left|\mathbb{E}\left[\bar{X}\right] - \mathbb{E}[X]\right| \leq 2\frac{\operatorname{Var}[X]}{C}.$$

The next step is to relate the risk generated by the linear functions $\widetilde{\mathbf{g}}_t^T \mathbf{w}$, to that generated by the clipped functions, $\bar{\mathbf{g}}_t^T \mathbf{w}$.

Lemma C.16 (A correction of Lemma 3.7 of (Hazan & Koren, 2012))**.** *Assume that $\left\|\mathbb{E}\left[\widetilde{\mathbf{g}}_t^2\right]\right\|_\infty \leq G^2$ for all t, for some $G \geq 0$. Then, for $0 \leq \eta \leq \frac{1}{2G}$, we have*

$$\mathbb{E}_{D,A}\left[\sum_{t=1}^{m} \widetilde{\mathbf{g}}_t^T \left(\mathbf{w}_t - \mathbf{w}^*\right)\right] \leq \mathbb{E}_{D,A}\left[\sum_{t=1}^{m} \bar{\mathbf{g}}_t^T \left(\mathbf{w}_t - \mathbf{w}^*\right)\right] + 4B\eta G^2 m.$$

Using these lemmas, we proceed to the proof of the theorem. First, from Lemma C.2, as the GAERR and GAELR algorithm build the gradient estimator using the same method, we have $\mathbb{E}_A\left[\widetilde{\mathbf{g}}_t\right] = \mathbf{g_t}$. From this follows that $\mathbb{E}_A\left[\sum_{t=1}^m \widetilde{\mathbf{g}}_t^T \left(\mathbf{w}_t - \mathbf{w}^*\right)\right] = \mathbb{E}_A\left[\sum_{t=1}^m \mathbf{g}_t^T \left(\mathbf{w}_t - \mathbf{w}^*\right)\right]$. Combining this with Lemmas C.14 and C.16, for $\eta \leq \frac{1}{2G}$, we have

$$\mathbb{E}_{D,A}\left[\sum_{t=1}^{m} \mathbf{g}_t^T \left(\mathbf{w}_t - \mathbf{w}^*\right)\right] \leq \frac{B \log 2d}{\eta} + 5B\eta G^2 m.$$

Proceeding as in the proof of Theorem 3.1 finishes the proof of Theorem 4.1.

Proof of Lemma C.12. Using the fact that $e^x \leq 1 + x + x^2$, for $x \leq 1$, we have

$$\begin{aligned}\|\mathbf{z}_{t+1}\|_1 &= \sum_{i=1}^{n} \mathbf{z}_t[i] \cdot e^{-\eta \mathbf{c}_t[i]} \\ &\leq \sum_{i=1}^{n} \mathbf{z}_t[i] \cdot \left(1 - \eta \mathbf{c}_t[i] + \eta^2 \mathbf{c}_t[i]^2\right) \\ &= \|\mathbf{z}_t\|_1 \cdot \left(1 - \eta \mathbf{p}_t^T \mathbf{c}_t + \eta^2 \mathbf{p}_t^T \mathbf{c}_t^2\right),\end{aligned}$$

and since $e^z \geq 1 + z$ for $z \in \mathbb{R}$, this implies by induction that

$$\begin{aligned}\log \|\mathbf{z}_{T+1}\|_1 &= \log n + \sum_{t=1}^{T} \log\left(1 - \eta \mathbf{p}_t^T \mathbf{c}_t + \eta^2 \mathbf{p}_t^T \mathbf{c}_t^2\right) \\ &\leq \log n - \eta \sum_{t=1}^{T} \mathbf{p}_t^T \mathbf{c}_t + \eta^2 \sum_{t=1}^{T} \mathbf{p}_t^T \mathbf{c}_t^2.\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\log \|\mathbf{z}_{T+1}\|_1 &= \log \sum_{i=1}^{n} \prod_{t=1}^{T} e^{\eta \mathbf{c}_t[i]} \\ &\geq \log \prod_{t=1}^{T} e^{\eta \mathbf{c}_t[i^*]} \\ &= -\eta \sum_{t=1}^{T} \mathbf{c}_t[i^*].\end{aligned}$$

Combining these two and rearranging, we obtain

$$\sum_{t=1}^{m} \mathbf{p}_t^T \mathbf{c}_t \leq \sum_{t=1}^{m} \mathbf{c}_t \left[i^*\right] + \frac{\log n}{\eta} + \eta \sum_{t=1}^{m} \mathbf{p}_t^T \mathbf{c}_t^2$$

for any i^*, which completes the proof. □

Proof of Lemma C.13. To see how Lemma C.13 follows from Lemma C.12, note that we can write the update rule of the GAELR algorithm in the terms of the augmented vectors, $\mathbf{z}_t$ and $\bar{\mathbf{g}}_t'$ as follows

$$\mathbf{z}_{t+1}\left[i\right] = \mathbf{z}_t\left[i\right] \cdot \exp\left(-\eta \bar{\mathbf{g}}_t'\left[i\right]\right) \qquad i = 1, .., 2d.$$

That is, $\mathbf{z}_{t+1}$ is obtained from $\mathbf{z}_t$ by a multiplicative update based on the vector $\bar{\mathbf{g}}_t'$. Noticing that $\|\bar{\mathbf{g}}_t'\|_\infty = \|\bar{\mathbf{g}}_t\|_\infty \leq \frac{1}{\eta}$, we see from Lemma C.12 that for any i^*,

$$\sum_{t=1}^{m} \mathbf{p}_t^T \bar{\mathbf{g}}_t' \leq \sum_{t=1}^{m} \bar{\mathbf{g}}_t'\left[i^*\right] + \frac{\log 2d}{\eta} + \eta \sum_{t=1}^{m} \mathbf{p}_t^T \left(\bar{\mathbf{g}}_t'\right)^2,$$

where $\mathbf{p}_t = \frac{\mathbf{z}_t'}{\|\mathbf{z}_t'\|_1}$, which gives the lemma. □

Proof of Lemma C.14. Notice that by our notation,

$$\sum_{t=1}^{m} \mathbf{p}_t^T \bar{\mathbf{g}}_t' = \sum_{t=1}^{m} \frac{\left(\mathbf{z}_t^+, \mathbf{z}_t^-\right)^T \left(\bar{\mathbf{g}}_t, -\bar{\mathbf{g}}_t\right)}{\left\|\mathbf{z}_t^+\right\|_1 + \left\|\mathbf{z}_t^-\right\|_1} = \frac{1}{B} \sum_{t=1}^{m} \mathbf{w}_t^T \bar{\mathbf{g}}_t$$

and

$$\min_i \sum_{t=1}^{m} \bar{\mathbf{g}}_t'\left[i\right] = \min_{\|\mathbf{w}\|_1 \leq B} \frac{1}{B} \sum_{t=1}^{m} \mathbf{w}^T \bar{\mathbf{g}}_t \leq \frac{1}{B} \sum_{t=1}^{m} \mathbf{w}^{*T} \bar{\mathbf{g}}_t$$

for any $\mathbf{w}^*$ with $\|\mathbf{w}^*\|_1 \leq B$. Plugging into the bound of Lemma C.13, we get

$$\sum_{t=1}^{m} \bar{\mathbf{g}}_t \left(\mathbf{w}_t - \mathbf{w}^*\right) \leq B \left(\frac{\log 2d}{\eta} + \eta \sum_{t=1}^{m} \mathbf{p}_t^T \left(\bar{\mathbf{g}}_t'\right)^2 \right).$$

Finally, taking the expectation with respect to the randomization of the algorithm and the data distribution, and noticing that $\left\|\mathbb{E}_{D,A}\left[\left(\bar{\mathbf{g}}_t'\right)^2\right]\right\|_\infty \leq \left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{g}}_t^2\right]\right\|_\infty \leq G^2$, the proof is complete. □

Proof of Lemma C.15. As a first step, note that for $x > C$ we have $x - \mathbb{E}\left[X\right] \geq C/2$, so that

$$C\left(x - C\right) \leq 2\left(x - \mathbb{E}\left[X\right]\right)\left(x - C\right) \leq 2\left(x - \mathbb{E}\left[X\right]\right)^2.$$

Hence, denoting by μ the probability measure of X, we obtain

$$\begin{aligned} \left|\mathbb{E}\left[\bar{X}\right] - \mathbb{E}\left[X\right]\right| &\leq \int_{x<-C} \left(x + C\right) d\mu + \int_{x>C} \left(x - C\right) d\mu \\ &\leq \int_{x>C} \left(x - C\right) d\mu \\ &\leq \frac{2}{C} \int_{x>C} \left(x - \mathbb{E}\left[X\right]\right)^2 d\mu \\ &\leq 2\frac{\mathrm{Var}\left[X\right]}{C}. \end{aligned}$$

Similarly one can prove that $\mathbb{E}\left[\bar{X}\right] - \mathbb{E}\left[X\right] \geq -2\mathrm{Var}\left[X\right]/C$, and the result follows. □

Proof of Lemma C.16. Notice that $\left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{g}}_t^2\right]\right\|_\infty \leq G^2$ implies $\left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{g}}_t\right]\right\|_\infty \leq G$ as

$$\left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{g}}_t\right]\right\|_\infty^2 = \left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{g}}_t\right]^2\right\|_\infty \leq \left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{g}}_t^2\right]\right\|_\infty.$$

Since $\bar{\mathbf{g}}\left[i\right] = \mathrm{clip}\left(\widetilde{\mathbf{g}}\left[i\right], 1/\eta\right)$ and $\left|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{g}}_t\left[i\right]\right]\right| \leq G \leq 1/2\eta$ the above lemma implies that

$$\left|\mathbb{E}_{D,A}\left[\bar{\mathbf{g}}_t\left[i\right]\right] - \mathbb{E}_{D,A}\left[\widetilde{\mathbf{g}}_t\left[i\right]\right]\right| \leq 2\eta \mathbb{E}_{D,A}\left[\widetilde{\mathbf{g}}_t\left[i\right]^2\right] \leq 2\eta G^2$$

for all i, which means $\left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{g}}_t - \bar{\mathbf{g}}_t\right]\right\|_\infty \leq 2\eta G^2$. Together with $\left\|\mathbf{w}_t - \mathbf{w}^*\right\|_1 \leq 2B$, this implies,

$$\mathbb{E}_{D,A}\left[\left(\bar{\mathbf{g}}_t - \widetilde{\mathbf{g}}_t\right)^T \left(\mathbf{w}_t - \mathbf{w}^*\right)\right] \leq 4\eta G^2.$$

Summing over $t = 1, .., m$, and taking the expectations, we obtain the lemma. □

C.11. Proof of Lemma 4.2

We will use two auxiliary lemmas. The first will help us bound the infinity norm of the data point estimator.

Lemma C.17. *For every distribution* $\left(q_1, .., q_d\right)$ *where* $q_i \geq 0$ *and* $i \in \left[d\right]$, *we have* $\left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{x}}_t^2\right]\right\|_\infty \leq \max_i \frac{1}{k}\mathbb{E}_{D,A}\left[\widetilde{\mathbf{x}}_{t,r}^2\left[i\right]\right] + \frac{k-1}{k}\mathbb{E}_D\left[\|\mathbf{x}\|_\infty\right]^2$.

The second will help us bound the square of the estimator of the inner product (minus the label).

Lemma C.18. *Using our sampling method we have* $\mathbb{E}_{D,A}\left[\widetilde{\phi_t}^2\right] \leq 4B^2$.

Now, the lemma follows directly from these lemmas, using the independence of $\widetilde{\mathbf{x}}_t$ and $\widetilde{\phi_t}$ given $\mathbf{x}_t$ and the assumption of $\|\mathbf{x}\|_\infty \leq 1$.

Proof of Lemma C.17. From the definition of $\widetilde{\mathbf{x}}_t$, we have

$$\begin{aligned} &\left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{x}}_t^2\right]\right\|_\infty \\ &\quad= \left\|\mathbb{E}_{D,A}\left[\left(\frac{1}{k}\sum_{r=1}^{k}\widetilde{\mathbf{x}}_{t,r}\right)^2\right]\right\|_\infty \\ &\quad= \frac{1}{k^2}\left\|\sum_{r=1}^{k}\mathbb{E}_{D,A}\left[\widetilde{\mathbf{x}}_{t,r}^2\right] + \sum_{r\neq s}^{k}\mathbb{E}_{D,A}\left[\widetilde{\mathbf{x}}_{t,r}\right]^2\right\|_\infty , \end{aligned}$$

where we used the fact that $\widetilde{\mathbf{x}}_{t,r}$ and $\widetilde{\mathbf{x}}_{t,s}$ are independent of each other. Now using the triangle inequality on the infinity norm and the fact that $\mathbb{E}_{D,A}\left[\widetilde{\mathbf{x}}_{t,r}\right] = \mathbb{E}_D\left[\mathbf{x}\right]$, we have

$$\begin{aligned} &\left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{x}}_t^2\right]\right\|_\infty \leq \\ &\qquad \max_i \frac{1}{k}\mathbb{E}_{D,A}\left[\widetilde{\mathbf{x}}_{t,r}^2\left[i\right]\right] + \frac{k-1}{k}\left\|\mathbb{E}_D\left[\mathbf{x}\right]^2\right\|_\infty . \end{aligned}$$

Using the convexity of the infinity norm and Jensen's inequality, the lemma follows. □

Proof of Lemma C.18. Recalling $|y_t| \leq B$ and using the inequality $(a-b)^2 \leq 2\left(a^2+b^2\right)$, by a straightforward calculation we obtain:

$$\begin{aligned} \mathbb{E}_{D,A}\left[\widetilde{\phi_t}^2\right] &= \mathbb{E}_{D,A}\left[\left(\frac{w_{t,j}}{p_j}\mathbf{x}_t\left[j_t\right] - y_t\right)^2\right] \\ &\leq 2\mathbb{E}_{D,A}\left[\left(\frac{w_{t,j}}{p_j}\mathbf{x}_t\left[j_t\right]\right)^2 + y_t^2\right] \\ &\leq 2\sum_{j=1}^{d}\frac{1}{p_j}w_{t,j}^2\mathbb{E}_D\left[x_j^2\right] + 2B^2 \\ &\leq 2\sum_{j=1}^{d}\frac{\|\mathbf{w}_t\|_1}{|w_{t,j}|}w_{t,j}^2 + 2B^2 \\ &\leq 2\left\|\mathbf{w}_t\right\|_1\sum_{j=1}^{d}|w_{t,j}| + 2B^2 \\ &\leq 4B^2. \end{aligned}$$

□

C.12. Proof of Lemma 4.3

The optimization problem is equivalent to

$$\begin{aligned} \underset{q_i}{\text{minimize}} \quad & \max_i \frac{1}{q_i}\mathbb{E}_D\left[x_i^2\right] \\ \text{subject to} \quad & \sum_{i=1}^{d} q_i = 1,\ \forall i\ q_i \geq 0. \end{aligned}$$

Let $C_i = \frac{\mathbb{E}_D\left[x_i^2\right]}{q_i}$. Note that $q_i = \frac{\mathbb{E}_D\left[x_i^2\right]}{\sum_{j=1}^{d}\mathbb{E}_D\left[x_j^2\right]}$ if, and only if, all C_i are equal. Assume by contradiction that not all C_i are equal, yet they still yield the minimal value for $\max_i \frac{1}{q_i}\mathbb{E}_D\left[x_i^2\right]$. Let $I = \{i|C_i = max_jC_j\}$, and i_0 be an index for which $C_{i_0} < max_jC_j$, which exists, by our assumption. For $\Delta > 0$, consider a new set of q_i'-s, such that $q_{i_0}' = q_{i_0} - \Delta$, and $q_i' = q_i + \frac{\Delta}{|I|}$ for $i \in I$. For a small enough Δ, still $C_{i_0}' < max_jC_j'$. Note that this is still a valid assignment of probabilities because $\sum_{i=1}^{d} q_i' = 1$ and all $q_i' > 0$ for a small enough Δ. However, max_jC_j' is smaller than max_jC_j, in contradiction to the assumption. Therefore, all C_i are equal and the minimal value is attained when $q_i = \frac{\mathbb{E}_D\left[x_i^2\right]}{\sum_{j=1}^{d}\mathbb{E}_D\left[x_j^2\right]}$.

C.13. Proof of Theorem 4.4

If $m \geq \log 2d$, we have $\eta \leq \frac{1}{2G}$ and the theorem follows directly from Theorem 4.1, Lemma 4.2, equation (3) and the calculated q_i-s in Lemma 4.3.

C.14. Proof of Theorem 4.5

The main goal of the proof is to bound the expected squared infinity-norm of the gradient estimator from above. By using Lemma 4.2, all that remains is to upper bound $\left\|\mathbb{E}_{D,A}\left[\widetilde{\mathbf{x}}_{t,r}^2\right]\right\|_\infty$ as we do in the next lemma.

Lemma C.19. *For all $t > m_1$, the following bound holds with probability 1 if $\epsilon = 1$ and with probability $\geq 1-\delta$, if $\epsilon \leq 1$*

$$\left\|\mathbb{E}_{D,A_2}\left[\widetilde{\mathbf{x}}_{t,r}^2\right]\right\|_\infty \leq 4\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_1 + \frac{20}{3}d\epsilon.$$

The proof can be found in Appendix C.15.

In the Lasso scenario it is sufficient to use one bound (compare to Lemma C.5 in the Ridge scenario) as we are able to join the two regimes of ϵ by ensuring $\epsilon \leq 1$ (Algorithm 4, line 4). Using this bound, the proof of the theorem is straightforward. First, using Theorem 4.1 on the second phase of the algorithm, we have

$$\mathbb{E}_{D,A_2}\left[L_{\mathcal{D}}\left(\bar{\mathbf{w}}\right)\right] - L_{\mathcal{D}}\left(\mathbf{w}^*\right) \leq B\left(\frac{\log 2d}{\eta m_2} + 5\eta G^2\right). \tag{8}$$

Now we use Lemma C.19, plug it into Lemma 4.2 and have $G^2 \leq 4B^2\left(\frac{4}{k}\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_1 + \frac{20}{3k}d\epsilon + 1\right)$ with probability 1 if $\epsilon = 1$ and with probability $\geq 1-\delta$, if $\epsilon \leq 1$. We continue by denoting $\widehat{G^2} = 4B^2\left(\frac{4}{k}\left\|2\mathbf{A} + \frac{10}{3}\epsilon\right\|_1 + \frac{20}{3k}d\epsilon + 1\right)$ and by using equation (5) we obtain $G^2 \leq \widehat{G^2}$. Plugging $\eta = \sqrt{\frac{\log 2d}{\widehat{G^2}5m_2}} =$

$\sqrt{\frac{k\log 2d}{20B^2 m_2\left(8\|\mathbf{A}\|_1+20d\epsilon+k\right)}}$ into equation (8), we have

$$
\begin{aligned}
&\mathbb{E}_{D,A_2}\left[L_{\mathcal{D}}\left(\bar{\mathbf{w}}\right)\right]-L_{\mathcal{D}}\left(\mathbf{w}^*\right)\\
&\quad\leq B\left(\frac{\log 2d}{m_2\eta}+5\eta G^2\right)\\
&\quad\leq B\left(\frac{\log 2d}{m_2\eta}+5\eta\widehat{G^2}\right)\\
&\quad\leq 2B\sqrt{\frac{5\widehat{G^2}\log 2d}{m_2}}\\
&\quad\leq 4B^2\sqrt{\frac{5\left(4\left\|2\mathbf{A}+\frac{10}{3}\epsilon\right\|_1+\frac{20}{3}d\epsilon+k\right)\log 2d}{km_2}}.
\end{aligned}
$$

Using

$$
\begin{aligned}
\left\|2\mathbf{A}+\frac{10}{3}\epsilon\right\|_1 &\leq \left\|4\mathbb{E}_D\left[\mathbf{x}^2\right]+\frac{14}{6}\epsilon+\frac{10}{3}\epsilon\right\|_1\\
&\leq 4\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_1+\frac{17}{3}d\epsilon,
\end{aligned}
\tag{9}
$$

we have

$$
\begin{aligned}
&\mathbb{E}_{D,A_2}\left[L_{\mathcal{D}}\left(\bar{\mathbf{w}}\right)\right]-L_{\mathcal{D}}\left(\mathbf{w}^*\right)\\
&\leq 4B^2\sqrt{\frac{5\left(16\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_1+\frac{68}{3}d\epsilon+\frac{20}{3}d\epsilon+k\right)\log 2d}{km_2}}\\
&\leq 4B^2\sqrt{\frac{5\left(16\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_1+\frac{88}{3}d\epsilon+k\right)\log 2d}{km_2}}.
\end{aligned}
$$

If $\epsilon=1$, we have

$$
\begin{aligned}
&4B^2\sqrt{\frac{5\left(16\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_1+\frac{88}{3}d\epsilon+k\right)\log 2d}{km_2}}\\
&\qquad\qquad\qquad\leq 61B^2\sqrt{\frac{d\log 2d}{km_2}}
\end{aligned}
$$

with probability 1. Otherwise plugging in $\epsilon=\min\left(\frac{d\log\frac{2d}{\delta}}{(k+1)m_1},1\right)$ finishes the proof.

C.15. Proof of Lemma C.19

Using the definition of $\widetilde{\mathbf{x}}_{t,r}$,

$$
\begin{aligned}
\left\|\mathbb{E}_{D,A_2}\left[\widetilde{\mathbf{x}}_{t,r}^2\right]\right\|_\infty &= \max_i \mathbb{E}_{D,A_2}\left[\widetilde{\mathbf{x}}_{t,r}^2\left[i\right]\right]\\
&= \max_i \frac{1}{q_i}\mathbb{E}_D\left[x_i^2\right]\\
&= \sum_{j=1}^d\left(A\left[j\right]+\frac{13}{6}\epsilon\right)\max_i\frac{\mathbb{E}_D\left[x_i^2\right]}{A\left[i\right]+\frac{13}{6}\epsilon}.
\end{aligned}
$$

Using equations (5), we have

$$
\begin{aligned}
&\left\|\mathbb{E}_{D,A_2}\left[\widetilde{\mathbf{x}}_{t,r}^2\right]\right\|_\infty\\
&\leq \sum_{j=1}^d\left(2\mathbb{E}_D\left[x_j^2\right]+\frac{7}{6}\epsilon+\frac{13}{6}\epsilon\right)\max_i\frac{\mathbb{E}_D\left[x_i^2\right]}{\frac{1}{2}\mathbb{E}_D\left[x_i^2\right]-\frac{5}{3}\epsilon+\frac{13}{6}\epsilon}\\
&\leq 4\sum_{j=1}^d\left(\mathbb{E}_D\left[x_j^2\right]+\frac{5}{3}\epsilon\right)\max_i\frac{\mathbb{E}_D\left[x_i^2\right]}{\mathbb{E}_D\left[x_i^2\right]+\epsilon}\\
&\leq 4\sum_{j=1}^d\left(\mathbb{E}_D\left[x_j^2\right]+\frac{5}{3}\epsilon\right)\max_i\frac{\mathbb{E}_D\left[x_i^2\right]}{\mathbb{E}_D\left[x_i^2\right]}\\
&\leq 4\left\|\mathbb{E}_D\left[\mathbf{x}^2\right]\right\|_1+\frac{20}{3}d\epsilon.
\end{aligned}
$$

If $\epsilon=1$, as equations (5) hold with probability 1, this bound also holds with probability 1. If $\epsilon\leq 1$, this bound holds with probability $\geq 1-\delta$.