Supplementary material to the paper
Selective sampling algorithms for cost-sensitive multiclass prediction

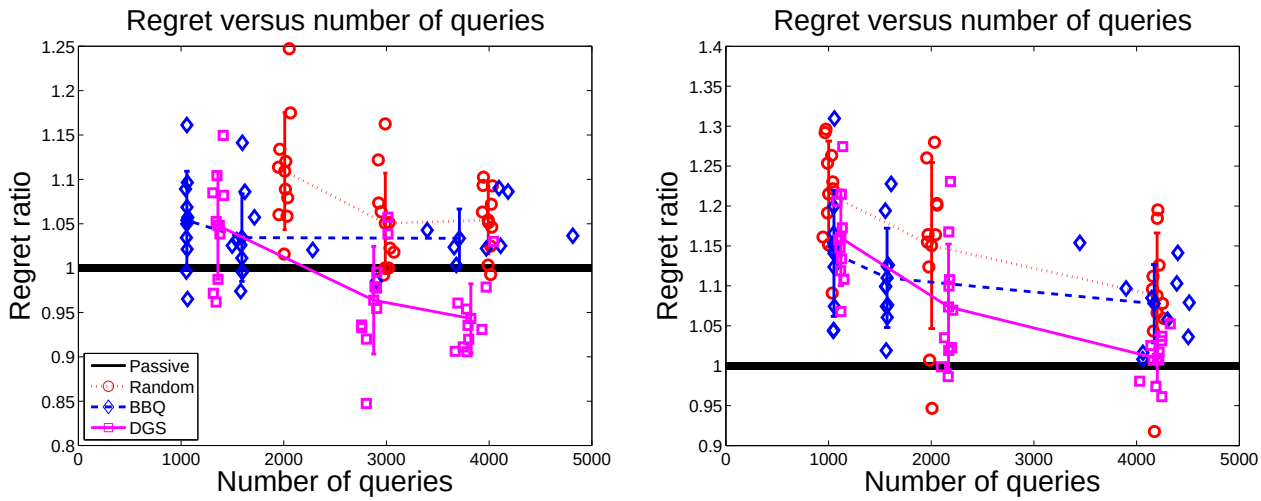
A. Numerical simulation details

We have displayed enlarged versions of our earlier results in Figures 2 and 3 for an easier visualization of the results. See Figures 4 and 5 for these plots.

We will now describe the details of our data generating procedure. As mentioned earlier, we used synthetic data generated according to a mixture of Gaussians. Our intuition was to have each cluster roughly correspond to one group, but with enough overlap so that there is adequate noise in the problem. We started by picking Gaussian random vectors in $\mathbb{R}^{1000}$ as our cluster means. However, due to concentration of measure, this gives rise to means that are far apart, and nearly orthogonal. The resulting classification problems from such means tended to be relatively noiseless and easy to solve with extremely few queries. To avoid this, we started by generating Gaussian random vectors in 10 dimensions, with mean 0 and standard deviation $I_{10\times 10}/\sqrt{10}$, so that the means have roughly unit norm. We then apply a random rotation to these weights in order to embed them into 1000 dimensions. For each sample, we first picked a mean vector uniformly at random from $1, 2, \ldots, K$. We then picked a random Gaussian vector with the mean as the selected cluster mean and standard deviation $(10/\sqrt{1000})I_{1000\times 1000}$. We tried other multipliers on the variance as well, but the results were stable within a reasonable range. As another robustness test, we added a certain fraction of random x vectors centered around the origin with the same variance. Again, the results were found to be fairly stable to such changes. For each x, we picked the label based on our generative model (1). As mentioned before, Figure 2 uses the exponential link function for the multinomial logit model while Figure 3 uses $\mathbb{P}(Y = i \mid x) \propto (x^T W_i^*)^2$. It might appear curious that the regret ratio has not gone up by much despite the model mismatch. While the ratio does seem fairly stable, the actual regret was substantially higher in this case, both for active and passive learning.

B. Proofs of main results

We start by giving a high-level outline of our proof. As remarked earlier, our proofs rely on conditional probability estimation. We start by formalizing this claim. Specifically, we provide two results in Proposition 1 and Lemma 2, which capture the rate at which our weights and our predicted probabilities correspond to their true versions under W^*. At a high level, our Assumption 1 regarding the strong convexity of the link function is crucial for this part, because otherwise we do not get good estimates of the weight matrix W^*. Qualitatively, our estimation rates are $\widetilde{O}(1/N_t)$ after we have made N_t queries. The next step is to relate the error in conditional probability estimation with the regret under our cost-sensitive loss (14). While this cannot be done in general, we use our generative model (1) to make this link. Specifically, following similar intuition in earlier works (Cesa-Bianchi et al., 2009; Dekel et al., 2010), we discard all the T_ϵ points which are too hard to resolve. On the remaining points, it is rather easy to control the regret of the points where we query the labels by using properties of our update rule. This intuition is formalized in Lemma 4. Everything up until this point is a property of the update rule (9) and applies for all query criteria. The remaining step is to control the regret on the points where we issue no queries, and this is where the query rule comes in. By design, it will turn out that we pay no regret on the points where we do not query, and this part heavily exploits the small error in our conditional probability estimates. We also provide bounds on the number of queries we make for our rules. The important intuition here is that all our rules involve the quantity $\|x_t\|_{M_t^{-1}}$, which decays suitably over time. By understanding how the decay of this quantity relates with the tolerance ϵ below which we do not account for regret, we obtain bounds on our query complexity.



$K = 5$ $K = 10$

Figure 4. Plots showing the ratio of active to passive regret, as a function of the number of queries. Left panel shows $K = 5$ and right panel shows $K = 10$.

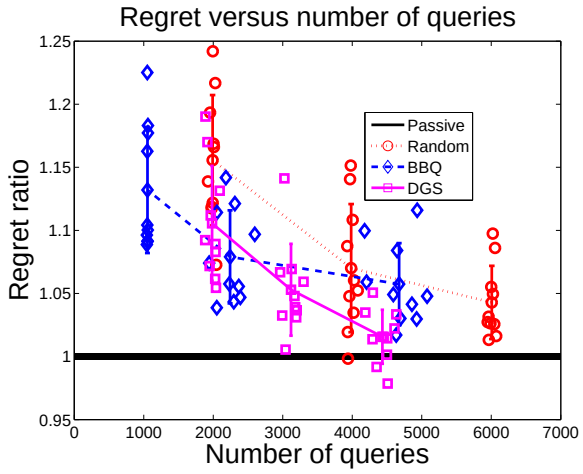


Figure 5. Plot showing the ratio of active to passive regret, as a function of the number of queries in a model mismatch scenario. $K = 10$ in this experiment.

We start with a proposition regarding the convergence of the weight matrix learned by Algorithm 1 to the optimal weight matrix. Then we will state some important lemmas that will be used to establish our main results, followed by the proofs of our theorems.

B.1. Convergence of weight matrices

In order to describe the results succinctly, we introduce the following notation for a positive definite matrix M

$$\|W\|_M^2 = \sum_{i=1}^K \|W_i\|_M^2,$$

as an extension of the Mahalanobis norm to matrices. It is clear that with this definition, for any vector $x \in \mathbb{R}^d$ we have

$$\|Wx\|_2^2 = \sum_{i=1}^K \langle W_i, x\rangle^2 \le \sum_{i=1}^K \|W_i\|_M^2 \|x\|_{M^{-1}}^2 = \|W\|_M^2 \|x\|_{M^{-1}}^2. \tag{19}$$

Proposition 1. *Under Assumptions 1 and 2, the iterates of Algorithm 1 satisfy with probability* $1-\delta$

$$\|W_t - W^*\|_{M_t} \le \frac{2}{\gamma_\ell}\sqrt{3 + 2\log\left(1 + \frac{2R^2\gamma_\ell}{\gamma}\right)}\sqrt{dK\log t}\sqrt{\log(dKt/\delta)} + \sqrt{\frac{2\gamma\omega^2}{\gamma_\ell}},$$

uniformly for all $t = 1, 2, \ldots, T$.

Proposition 1 is a property of the update rule (9) and does not rely on the query conditions. The proof uses standard techniques for the analysis of online convex optimization algorithms along with martingale concentration.

Proof of Proposition 1 By the definition of W_t, first-order optimality conditions for convex optimization guarantee that

$$\left\langle \gamma W_t + \sum_{s=1}^{t-1} Z_s \nabla \ell(W_t x_s, y_s), W - W_t \right\rangle \ge 0, \quad \text{for all } W \in \mathcal{W}.$$

Recalling the definition (4) and using the optimality condition with $W = W^*$, we obtain the condition

$$\gamma \langle W_t, W^* - W_t\rangle + \sum_{s=1}^{t-1} Z_s \left\langle \nabla\Phi(W_t x_s)x_s^T - y_s x_s^T, W^* - W_t\right\rangle \ge 0.$$

Let us define the shorthand $\xi_s = y_s - \nabla\Phi(W^* x_s)$. Recall the definition of the sigma-field $\mathcal{F}_s$ which is generated by x_1 through x_s, along with the observed y values up to round $s-1$. Then it is clear that ξ_s is measurable with respect to $\mathcal{F}_{s+1}$. Furthermore, the definition (1) of our probabilistic model guarantees that $\mathbb{E}[\xi_s \mid \mathcal{F}_s] = 0$, meaning that ξ_s is a martingale difference sequence adapted to the filtration $\{\mathcal{F}_{s+1}\}$. In terms of this shorthand, we can now rewrite the optimality condition as

$$\gamma \langle W_t, W^* - W_t\rangle + \sum_{s=1}^{t-1} Z_s \left\langle \nabla\Phi(W_t x_s)x_s^T - \nabla\Phi(W^* x_s)x_s^T + \xi_s x_s^T, W^* - W_t\right\rangle \ge 0.$$

Rearranging terms, we obtain

$$\begin{aligned}\sum_{s=1}^{t-1} Z_s \left\langle \xi_s x_s^T, W^* - W_t \right\rangle &\geq \sum_{s=1}^{t-1} Z_s \left\langle \nabla\Phi(W_t x_s) x_s^T - \nabla\Phi(W^* x_s) x_s^T, W_t - W^* \right\rangle + \gamma \left\langle W_t, W_t - W^* \right\rangle \\ &= \sum_{s=1}^{t-1} Z_s \left\langle \nabla\Phi(W_t x_s) - \nabla\Phi(W^* x_s), W_t x_s - W^* x_s \right\rangle + \gamma \left\langle W_t, W_t - W^* \right\rangle \\ &\geq \gamma_\ell \sum_{s=1}^{t-1} Z_s \|W_t x_s - W^* x_s\|_2^2 + \gamma \|W_t - W^*\|_2^2 + \gamma \left\langle W^*, W_t - W^* \right\rangle,\end{aligned}$$

where the last inequality follows from the strong convexity Assumption 1. Recalling the definition of the matrix M_t (7) as well as our boundedness Assumption 3, we can further simplify the above inequality to

$$\sum_{s=1}^{t-1} Z_s \left\langle \xi_s, W^* x_s - W_t x_s \right\rangle \geq \gamma_\ell \|W_t - W^*\|_{M_t}^2 - 2\gamma\omega^2. \tag{20}$$

We now focus on the right hand side of the inequality. Observe that

$$\begin{aligned}\sum_{s=1}^{t-1} Z_s \left\langle \xi_s, W^* x_s - W_t x_s \right\rangle &= \sum_{i=1}^{K} \sum_{s=1}^{t-1} Z_s \xi_{s,i} \left\langle W^*_{t,i} - W^i_t, x_s \right\rangle \\ &\leq \sum_{i=1}^{K} \| \sum_{s=1}^{t-1} Z_s \xi_{s,i} x_s \|_{M_t^{-1}} \|W^*_{t,i} - W^i_t\|_{M_t}.\end{aligned}$$

To control each term in the sum, we use a tail inequality for vector-valued martingales from Fillipi et al. (Filippi et al., 2010). In particular, invoking Lemma 1 in the Appendix A.1 of the paper with constants $c_m = R$, $\lambda_0 = \gamma/\gamma_\ell$ and $R = 2$ yields for any $0 < \delta < 1/e$ and $t \geq 2$ the following bound with probability at least $1 - \delta/K$

$$\| \sum_{s=1}^{t-1} Z_s \xi_{s,i} x_s \|_{M_t^{-1}} \leq 2\sqrt{3 + 2\log(1 + 2R^2\gamma_\ell/\gamma)} \sqrt{d \log t} \sqrt{\log(dK/\delta)},$$

for all $i = 1, 2, \ldots, K$. Taking a union bound over all the classes yields with probability at least $1 - \delta$

$$\begin{aligned}\sum_{s=1}^{t-1} Z_s \left\langle \xi_s, W^* x_s - W_t x_s \right\rangle &\leq 2\sqrt{3 + 2\log(1 + 2R^2\gamma_\ell/\gamma)} \sqrt{d \log t} \sqrt{\log(dK/\delta)} \sum_{i=1}^{K} \|W^*_{t,i} - W^i_t\|_{M_t} \\ &\leq 2\sqrt{3 + 2\log(1 + 2R^2\gamma_\ell/\gamma)} \sqrt{dK \log t} \sqrt{\log(dK/\delta)} \|W^* - W_t\|_{M_t},\end{aligned}$$

where the final inequality follows from the definition of $\|W^* - W_t\|_{M_t}$ and the fact $\sum_{i=1}^K a_i \leq \sqrt{K}\sqrt{\sum_{i=1}^K a_i^2}$ for $a_i \geq 0$. Plugging the above inequality in our earlier bound (20), we have shown that with probability at least $1 - \delta$

$$\gamma_\ell \|W_t - W^*\|_{M_t}^2 \leq 2\sqrt{3 + 2\log(1 + 2R^2\gamma_\ell/\gamma)} \sqrt{dK \log t} \sqrt{\log(dK/\delta)} \|W^* - W_t\|_{M_t} + 2\gamma\omega^2.$$

We can now solve the quadratic inequality to obtain a high probability upper bound on $\|W_t - W^*\|$. Rearranging terms, and taking another union bound over the rounds $t = 1, 2, \ldots, T$ completes the proof. □

We conclude the section with a technical lemma which is in a similar vein as Proposition 1, and will be needed for some of our following proofs.

Lemma 2. *Under conditions of Theorem 1, with probability at least* $1 - 4\delta \log T$ *for some* $0 < \delta < 1/e$ *and for* $T \geq 3$*, we have*

$$\sum_{t=1}^{T} Z_t \|W_t x_t - W^* x_t\|_2^2 \leq \frac{8d(\gamma_\ell + \gamma)}{\gamma_\ell^2 \gamma} \log\left(\frac{R^2 \gamma_\ell T}{\gamma} + 1\right) + \frac{112 R\omega}{\gamma_\ell^2} \log\frac{1}{\delta}.$$

The key difference between the lemma and Proposition 1 is that the proposition gives a bound on the error in weight matrices, which immediately allows us to bound the error in predictions on *any* future data point. In contrast, Lemma 2 only concerns with bounding the sums of errors in predictions over the data points the algorithm actually queries. However, doing so allows us to get bounds that are sharper in factors of d and K in some applications of the result. The proof of this lemma is somewhat involved, and is deferred until the end. For now, we proceed with proving our main theorems, which requires a better understanding of the regret (14).

B.2. A useful regret decomposition

In the following results, we assume that both the above high-probability upper bounds hold deterministically, and bound the probability of error at the very end. We will now present a series of lemmas that provide a decomposition for the multiclass classification loss. The results can be seen as analogues of previous such decompositions in the binary case (Cesa-Bianchi et al., 2009; Dekel et al., 2010), but the techniques involved are somewhat different in the multiclass setting. Before stating the results, we recall our earlier definitions (10). We also recall the definition of the σ-field $\mathcal{F}_t = \sigma\{x_1, \ldots, x_t, y_s \ : \ 1 \leq s < t, \ Z_s = 1\}$. Our results will involve the previously definition notation T_ϵ (15).

Lemma 3. *For any* $\epsilon \in [0, 1]$*, we have the following*

$$\sum_{t=1}^{T} (\mathbb{E}[C(y, \hat{y}_t) \mid \mathcal{F}_t] - \mathbb{E}[C(y, y_t^*) \mid \mathcal{F}_t]) = \epsilon T_\epsilon + \mathcal{T}_{T,\epsilon}^1 + \mathcal{T}_{T,\epsilon}^2,$$

where

$$\mathcal{T}_{T,\epsilon}^1 = \sum_{t=1}^{T} (1 - Z_t) \mathbb{1}\left\{S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(\hat{y}_t) \geq \epsilon\right\} (S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_r}(\hat{y}_t)), \quad \textit{and}$$

$$\mathcal{T}_{T,\epsilon}^2 = \sum_{t=1}^{T} Z_t \mathbb{1}\left\{S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(\hat{y}_t) \geq \epsilon\right\} (S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(\hat{y}_t)) \tag{21}$$

Proof. From the definition (6), we see that the regret in the expected costs is directly linked with the score function since

$$\sum_{t=1}^{T} (\mathbb{E}[C(y, \hat{y}_t) \mid \mathcal{F}_t] - \mathbb{E}[C(y, y_t^*) \mid \mathcal{F}_t]) = \sum_{t=1}^{T} S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(\hat{y}_t).$$

Here we used the fact that $\sum_{i=1}^{K} (\nabla \Phi(W^* x))_i = 1$, so that the additive term $C_{\max}$ in the definition of score function cancels in the definition of the regret. We now break up our analysis over the rounds where $0 < S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(\hat{y}_t) \leq \epsilon$ and where it is greater than ϵ. On the first case, the expected regret is clearly at most ϵ. Furthermore, the number of such rounds is at most T_ϵ. This is because either we have $S_{W^*}^{x_t}(\hat{y}_t) = S_{W^*}^{x_t}(y_t^*)$,

in which case we incur no regret or we should have $S^{x_t}_{W^*}(\hat{y}_t) \geq S^{x_t}_{W^*}(y_t^{'})$. Hence, we are guaranteed to have the scores of $y_t^{'}$ and y_t^* within ϵ if the scores of $\hat{y}_t$ and y_t^* are unequal but within ϵ. Recalling the definition (15), this yields the first term in our decomposition.

The second and third terms result simply from further breaking our analysis over rounds where we do not query and query respectively. This completes the proof of the lemma. □

We next tackle $\mathcal{T}^2_{T,\epsilon}$ in our decomposition above. This term is incurred on the rounds where we make queries, and will be identical for all of our query rules. The impact of the specific query rules is only on $\mathcal{T}^1_{T,\epsilon}$, that is on guaranteeing small regret on rounds where we do not query. Recall that we are still assuming that the bound of Lemma 2 holds deterministically in this lemma.

Lemma 4.

$$\mathcal{T}^2_{T,\epsilon} \leq \frac{32\sigma^2(C)\gamma_u^2(\gamma_\ell+\gamma)}{\gamma_\ell^2\gamma\epsilon}\, d\log\left(\frac{R^2\gamma_\ell T}{\gamma}+1\right) + \frac{448\gamma_u^2\sigma^2(C)}{\gamma_\ell^2\epsilon}\log\frac{1}{\delta}.$$

Proof. We begin by observing that under the conditions of the decomposition, we have that

$$\begin{aligned}\mathcal{T}^2_{T,\epsilon} &= \sum_{t=1}^{T} Z_t \mathbb{1}\left\{S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W^*}(\hat{y}_t) \geq \epsilon\right\}\left(S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W^*}(\hat{y}_t)\right)\\ &\leq \sum_{t=1}^{T} Z_t \frac{(S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W^*}(\hat{y}_t))^2}{\epsilon}.\end{aligned}$$

Furthermore, by the definitions (10), we have that

$$S^{x_t}_{W_t}(\hat{y}_t) - S^{x_t}_{W_t}(y_t^*) \geq 0.$$

Hence, we can conclude that

$$0 \leq S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W^*}(\hat{y}_t) \leq S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W^*}(\hat{y}_t) + S^{x_t}_{W_t}(\hat{y}_t) - S^{x_t}_{W_t}(y_t^*).$$

Since both sides are non-negative, we can square to futher obtain

$$\begin{aligned}&\frac{(S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W^*}(\hat{y}_t))^2}{\epsilon}\\ &\leq \frac{\left(S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W^*}(\hat{y}_t) + S^{x_t}_{W_t}(\hat{y}_t) - S^{x_t}_{W_t}(y_t^*)\right)^2}{\epsilon}\\ &\leq 2\frac{(S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W_t}(y_t^*))^2}{\epsilon} + 2\frac{(S^{x_t}_{W^*}(\hat{y}_t) - S^{x_t}_{W_t}(\hat{y}_t))^2}{\epsilon}. \qquad (22)\end{aligned}$$

We focus on the first term above, since the treatment for the second is identical. To do so, we now unwrap the definition of the score function and observe that

$$\begin{aligned}\frac{(S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W_t}(y_t^*))^2}{\epsilon} &= \frac{\left(\sum_{j=1}^K((\nabla\Phi(W^*x_t))_j - (\nabla\Phi(W_tx_t))_j)(C_{\max} - C(j,y_t^*))\right)^2}{\epsilon}\\ &= \frac{\left(\sum_{j=1}^K((\nabla\Phi(W^*x_t))_j - (\nabla\Phi(W_tx_t))_j)(-C(j,y_t^*))\right)^2}{\epsilon},\end{aligned}$$

where the second equality follows since $\sum_{j=1}^{K}((\nabla\Phi(W^*x_t))_j - (\nabla\Phi(W_tx_t))_j) = 0$. To proceed further, we recall our earlier notation $\bar{C}_i = \sum_{j=1}^{K} C(j,i)/K$. Since the above inequality is invariant to any translation of the costs involving class y_t^* by a constant independent, of j, we further obtain

$$\begin{aligned}\frac{(S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W_t}(y_t^*))^2}{\epsilon} &= \frac{\left(\sum_{j=1}^{K}((\nabla\Phi(W^*x_t))_j - (\nabla\Phi(W_tx_t))_j)(\bar{C}_{y_t^*} - C(j,y_t^*))\right)^2}{\epsilon} \\ &\leq \frac{\left(\sum_{j=1}^{K}((\nabla\Phi(W^*x_t))_j - (\nabla\Phi(W_tx_t))_j)^2\right)\left(\sum_{j=1}^{K}(\bar{C}_{y_t^*} - C(j,y_t^*))^2\right)}{\epsilon},\end{aligned}$$

where the inequality is a consequence of Cauchy-Shwartz inequality. We can further use Lipschitz continuity of $\nabla\Phi$ to obtain

$$\begin{aligned}\frac{(S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W_t}(y_t^*))^2}{\epsilon} &\leq \frac{\gamma_u^2}{\epsilon}\|W^*x_t - W_tx_t\|_2^2\|\bar{C}_{y_t^*} - C_{y_t^*}\|_2^2 \\ &\leq \frac{\gamma_u^2}{\epsilon}\|W^*x_t - W_tx_t\|_2^2\sigma^2(C),\end{aligned}$$

where we obtain the last step by recalling the definition (16) of $\sigma^2(C)$. Since the same bound also holds for the differences in scores on $\hat{y}_t$, we can plug the above bound into our earlier inequality (22) and obtain

$$\frac{(S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W^*}(\hat{y}_t))^2}{\epsilon} \leq 4\frac{\gamma_u^2\sigma^2(C)}{\epsilon}\|W^*x_t - W_tx_t\|_2^2.$$

Summing the bound over all the queried rounds and invoking Lemma 2 completes the proof. □

B.3. Proofs of Theorems 1 and 2

We are now in a position to prove our main results. In both the theorems, it only remains to control the term $\mathcal{T}^1_{T,\epsilon}$ given our work so far. As we will see, both the query criteria BBQ_ϵ and DGS are designed so that this term will actually be zero. The second part of the proof consists of bounding the number of queries. This turns out to be rather straightforward for the BBQ_ϵ rule, but significantly more involved for the DGS rule.

Proof of Theorem 1 We focus on the regret, which requires us to understand $\mathcal{T}^1_{T,\epsilon}$. To this end, we note that from the proof of Lemma 4, we have

$$\begin{aligned}S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W^*}(\hat{y}_t) &\leq S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W^*}(\hat{y}_t) - S^{x_t}_{W_t}(y_t^*) + S^{x_t}_{W_t}(\hat{y}_t) \\ &= \sum_{j=1}^{K}(\bar{C}_{y_t^*} - C(j,y_t^*))((\nabla\Phi(W^*x_t))_j - (\nabla\Phi(W_tx_t))_j) \\ &\qquad - \sum_{j=1}^{K}(\bar{C}_{\hat{y}_t} - C(j,\hat{y}_t))((\nabla\Phi(W^*x_t))_j - (\nabla\Phi(W_tx_t))_j) \\ &\leq 2\sigma(C)\,\gamma_u\|W_tx_t - W^*x_t\|_2.\end{aligned}$$

For the BBQ_ϵ query criterion, the above term is at most ϵ when we do not query the label y_t. Consequently, we incur regret only if $S^{x_t}_{W^*}(y_t^*) - S^{x_t}_{W^*}(\hat{y}_t) \leq \epsilon$. Since this quantity is guaranteed to be at least ϵ on the summands in $\mathcal{T}^1_{T,\epsilon}$, we see that either $Z_t = 0$ or the indicator of the event in $\mathcal{T}^1_{T,\epsilon}$ is zero. As a result, $\mathcal{T}^1_{T,\epsilon} = 0$, which completes the proof of the regret bound.

As for the bound on the number of queries, proceed similarly as the earlier analysis of Cesa-Bianchi et al. (Cesa-Bianchi et al., 2009). We observe that by the query condition, we have

$$N_T = \sum_{t\,:\,4\sigma^2(C)\gamma_u^2\theta_t^2\|x_t\|^2_{M_t^{-1}}\geq\epsilon^2} 1 \leq \sum_{t\,:\,4\sigma^2(C)\gamma_u^2\theta_t^2\|x_t\|^2_{M_t^{-1}}\geq\epsilon^2} \frac{4\sigma^2(C)\gamma_u^2\theta_t^2\|x_t\|^2_{M_t^{-1}}}{\epsilon^2}$$
$$\leq \frac{4\sigma^2(C)\gamma_u^2\theta_T^2}{\epsilon^2}\sum_{t=1}^{T} Z_t\|x_t\|^2_{M_t^{-1}}.$$

Further applying Lemma 5 from the appendix completes the proof of the theorem. □

We now establish the result for the DGS selection rule

Proof of Theorem 2 The proof relies on the following observation which is a consequence of the definition (6) and the Lipschitz continuity of the mapping $\nabla\Phi$ from Assumption 2

$$|S^{x_t}_{W^*}(i) - S^{x_t}_{W_t}(i)| \leq \sigma(C)\gamma_u \|W_t x_t - W^* x_t\|_2 \leq \sigma(C)\gamma_u \|W_t - W^*\|_{M_t}\|x_t\|_{M_t^{-1}}, \tag{23}$$

for all $i = 1, 2, \ldots, K$. Now let us suppose that on a round t, we have that $S^{x_t}_{W^*}(\hat{y}_t) < S^{x_t}_{W^*}(y^*_t)$. Then using the above bound, we see that

$$0 > S^{x_t}_{W^*}(\hat{y}_t) - S^{x_t}_{W^*}(y^*_t) \geq S^{x_t}_{W_t}(\hat{y}_t) - S^{x_t}_{W_t}(y^*_t) - 2\sigma(C)\gamma_u \|W_t - W^*\|_{M_t}\|x_t\|_{M_t^{-1}}$$
$$\geq S^{x_t}_{W_t}(\hat{y}_t) - S^{x_t}_{W_t}(y^{''}_t) - 2\sigma(C)\gamma_u \|W_t - W^*\|_{M_t}\|x_t\|_{M_t^{-1}} \geq 0,$$

on the rounds where we do not query. Hence, we have a contradiction unless $S^{x_t}_{W^*}(\hat{y}_t) - S^{x_t}_{W^*}(y^*_t) \leq 0$ on the rounds where we do not query, meaning that $\mathcal{T}^1_{T,\epsilon}$ is zero once again. This completes the proof of the regret bound.

The proof of the query bound is a little more involved in this case. We break up our analysis over the cases where $\hat{y}_t = y^*_t$ and the ones where they disagree. Starting with the latter, we see that for any $\epsilon > 0$ we have

$$\sum_{t=1}^{T} Z_t \mathbb{1}\{\hat{y}_t \neq y^*_t\} = \sum_{t=1}^{T} Z_t \left(\mathbb{1}\{\hat{y}_t \neq y^*_t, (S^{x_t}_{W^*}(y^*_t) - S^{x_t}_{W^*}(\hat{y}_t) \leq \epsilon\}\right.$$
$$\left. + \mathbb{1}\{\hat{y}_t \neq y^*_t, S^{x_t}_{W^*}(y^*_t) - S^{x_t}_{W^*}(\hat{y}_t) \geq \epsilon\}\right)$$
$$\leq \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t \neq y^*_t, S^{x_t}_{W^*}(y^*_t) - S^{x_t}_{W^*}(y^{'}_t) \leq \epsilon\right\}$$
$$+ \sum_{t=1}^{T} Z_t \mathbb{1}\{\hat{y}_t \neq y^*_t, S^{x_t}_{W^*}(y^*_t) - S^{x_t}_{W^*}(\hat{y}_t) \geq \epsilon\}. \tag{24}$$

We focus on controlling the second sum, which can be done by invoking Equation 23 twice, once with $i = y^*_t$ and once with $i = \hat{y}_t$. Since $S^{x_t}_{W_t}(\hat{y}_t) \geq S^{x_t}_{W_t}(y^*_t)$, we obtain the upper bound

$$S^{x_t}_{W^*}(\hat{y}_t) \leq S^{x_t}_{W^*}(y^*_t) \leq S^{x_t}_{W^*}(\hat{y}_t) + 2\sigma(C)\gamma_u\theta_t\|x_t\|_{M_t^{-1}}. \tag{25}$$

Combining this with our earlier upper bound (24), we further obtain

$$\sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t \neq y_t^*\right\} \leq \sum_{t=1}^{T} \mathbb{1}\left\{\hat{y}_t \neq y_t^*, S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(y_t^{'}) \leq \epsilon\right\} + \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t \neq y_t^*, 2\sigma(C)\gamma_u \theta_t \|x_t\|_{M_t^{-1}} \geq \epsilon\right\}$$
$$\leq \sum_{t=1}^{T} \mathbb{1}\left\{\hat{y}_t \neq y_t^*, S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(y_t^{'}) \leq \epsilon\right\} + \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t \neq y_t^*\right\} \frac{4\sigma^2(C)\gamma_u^2\theta_t^2 \|x_t\|^2_{M_t^{-1}}}{\epsilon^2}. \quad (26)$$

We now analyze the other case where $\hat{y}_t = y_t^*$. In this case, our query condition guarantees that

$$\begin{aligned}
\sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t = y_t^*\right\} &= \sum_{t=1}^{T} \mathbb{1}\left\{\hat{y}_t = y_t^*, S_{W_t}^{x_t}(\hat{y}_t) - S_{W_t}^{x_t}(y_t^{''}) \leq 2\sigma(C)\gamma_u\, \theta_t \|x_t\|_{M_t^{-1}}\right\} \\
&= \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t = y_t^*, S_{W_t}^{x_t}(y_t^*) - S_{W_t}^{x_t}(y_t^{''}) \leq 2\sigma(C)\gamma_u\, \theta_t \|x_t\|_{M_t^{-1}}\right\} \\
&\overset{(*)}{\leq} \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t = y_t^*, S_{W_t}^{x_t}(y_t^*) - S_{W_t}^{x_t}(y_t^{''}) \leq 4\sigma(C)\gamma_u\, \theta_t \|x_t\|_{M_t^{-1}}\right\} \\
&\leq \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t = y_t^*, S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(y_t^{'}) \leq 4\sigma(C)\gamma_u\, \theta_t \|x_t\|_{M_t^{-1}}\right\} \\
&\leq \sum_{t=1}^{T} \mathbb{1}\left\{\hat{y}_t = y_t^*, S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(y_t^{'}) \leq \epsilon\right\} \\
&+ \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t = y_t^*, \epsilon \leq S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(y_t^{'}) \leq 4\sigma(C)\gamma_u\, \theta_t \|x_t\|_{M_t^{-1}}\right\} \\
&\leq \sum_{t=1}^{T} \mathbb{1}\left\{\hat{y}_t = y_t^*, S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{t}(y_t^{'}) \leq \epsilon\right\} + \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t = y_t^*\right\} \frac{16\sigma^2(C)\gamma_u^2\theta_t^2 \|x_t\|^2_{M_t^{-1}}}{\epsilon^2}.
\end{aligned}$$

In the above display, the inequality $(*)$ follows from using Proposition 1 to establish the closeness of $S_{W_t}^{x_t}(i)$ and $S_{W^*}^{x_t}(i)$ for $i = y_t^*$ and $i = y_t^{'}$. Adding this to our earlier bound (26), we obtain the bound on the number of queries as

$$\begin{aligned}
N_T &= \sum_{t=1}^{T} Z_t = \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t \neq y_t^*\right\} + \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t = y_t^*\right\} \\
&\leq \sum_{t=1}^{T} \mathbb{1}\left\{\hat{y}_t \neq y_t^*, S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(y_t^{'}) \leq \epsilon\right\} + \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t \neq y_t^*\right\} \frac{4\sigma^2(C)\gamma_u^2\theta_t^2 \|x_t\|^2_{M_t^{-1}}}{\epsilon^2} \\
&+ \sum_{t=1}^{T} \mathbb{1}\left\{\hat{y}_t = y_t^*, S_{W^*}^{x_t}(y_t^*) - S_{W^*}^{x_t}(y_t^{'}) \leq \epsilon\right\} + \sum_{t=1}^{T} Z_t \mathbb{1}\left\{\hat{y}_t = y_t^*\right\} \frac{16\sigma^2(C)\gamma_u^2\theta_t^2 \|x_t\|^2_{M_t^{-1}}}{\epsilon^2} \\
&\leq T_\epsilon + \sum_{t=1}^{T} Z_t \frac{16\gamma_u^2\theta_t^2 \|x_t\|^2_{M_t^{-1}}}{\epsilon^2}.
\end{aligned}$$

Finally, invoking Lemma 5 completes the proof. □

C. Proof of Lemma 1

Proof of Lemma 1

The proof follows almost directly from the definitions. Suppose we were to predict a class i for a given data point i. Recalling our notation $C_{\max} = \max_{a,b} C(a,b)$, the expected loss incurred is

$$\sum_{j=1}^{K} \mathbb{P}(Y = j \mid x) C(j,i) = \sum_{j=1}^{K} (\nabla\Phi(W^* x))_j C(j,i) = \sum_{j=1}^{K} (\nabla\Phi(W^* x))_j (C(j,i) - C_{\max}) + C_{\max},$$

where the last equality follows since $\sum_j (\nabla\Phi(W^* x))_j = 1$. The above quantity is easily seen to be $C_{\max} - S^x_{W^*}(i)$. Hence, picking the class maximizing $S^x_{W^*}$ minimizes the expected loss pointwise, meaning that it is the Bayes optimal prediction.

□

D. Auxiliary results for Theorems 1 and 2

In this appendix, we collect many auxiliary technical results and proofs that are used throughout the paper in our proofs.

D.1. Sums of quadratic forms

We start with a simple lemma. The lemma is an adaptation of Lemma 11 in Hazan et al. (2007). Our statement is slightly different since our matrices are off by one time index, as opposed to theirs.

Lemma 5.

$$\sum_{t=1}^{T} Z_t \|x_t\|^2_{M_t^{-1}} \le \frac{\gamma_\ell + \gamma}{\gamma}\, d \log\left(\frac{R^2 \gamma_\ell T}{\gamma} + 1\right)$$

Proof. The proof is a slight adaptation of Lemma 11 from Hazan et al. (2007). Note that invoking Lemma 11 from that paper, we can conclude that

$$\sum_{t=1}^{T} Z_t \|x_t\|^2_{M_{t+1}^{-1}} \le d \log\left(\frac{R^2 \gamma_\ell T}{\gamma} + 1\right).$$

Also observe that using the Sherman-Morrison-Woodbury matrix identity, we have that

$$\begin{aligned}
Z_t \|x_t\|^2_{M_{t+1}^{-1}} &= Z_t x_t^T M_{t+1}^{-1} x_t \\
&= (Z_t x_t)^T (M_t + Z_t x_t x_t^T)^{-1} (Z_t x_t) \\
&= Z_t \|x_t\|^2_{M_t^{-1}} - Z_t x_t^T \left(\frac{M_t^{-1} x_t x_t^T M_t^{-1}}{1 + x_t^T M_t^{-1} x_t}\right) x_t \\
&= Z_t \left(\|x_t\|^2_{M_t^{-1}} - \frac{\|x_t\|^4_{M_t^{-1}}}{1 + \|x_t\|^2_{M_t^{-1}}}\right) \\
&= Z_t \frac{\|x_t\|^2_{M_t^{-1}}}{1 + \|x_t\|^2_{M_t^{-1}}}.
\end{aligned}$$

Rearranging terms, we obtain

$$Z_t\|x_t\|^2_{M_t^{-1}} = Z_t \frac{\|x_t\|^2_{M_{t+1}^{-1}}}{1 - \|x_t\|^2_{M_{t+1}^{-1}}} \leq Z_t \frac{\gamma_\ell + \gamma}{\gamma} \|x_t\|^2_{M_{t+1}^{-1}} .$$

Here the last inequality follows since $Z_t\|x_t\|^2_{M_{t+1}^{-1}} \leq \frac{1}{1+\gamma/\gamma_\ell}$. Combining these facts yields the statement of the lemma. □

D.2. Proof of Lemma 2

In order to prove the lemma, we will need a couple of additional technical results that we state next. The first is a martingale convergence result, which will allow us to relate the LHS of Lemma 2 with the surrogate loss 4 incurred by our algorithm. The next result bounds precisely this surrogate loss. We begin with the martingale result.

Lemma 6. *Suppose that the labels are generated according to the probabilistic model* (1) *and Assumption 3 holds. Then for any* $0 < \delta < 1/e$ *and* $T \geq 3$*, with probability at least* $1 - 4\delta \log(T)$ *we have the following bound*

$$\sum_{t=1}^{T} Z_t \mathrm{D}_\Phi(W_t x_t, W^* x_t) \leq 2 \sum_{t=1}^{T} Z_t (\ell(W_t; (x_t, y_t)) - \ell(W^*; (x_t, y_t)) + \frac{56 R \omega}{\gamma_\ell} \log \frac{1}{\delta}.$$

The next lemma concerns the surrogate loss regret of the update rule (9). In terms of the online learning literature, the update rule is what is often called the *follow the leader* strategy. While the proof technique for bounding the regret of this strategy under our assumptions is quite standard (Kalai & Vempala, 2005), we include a proof for completeness.

Lemma 7.

$$\sum_{t=1}^{T} Z_t (\ell(W_t; (x_t, y_t)) - \ell(W^*; (x_t, y_t))) \leq \frac{4(1+\gamma)\, d}{\gamma_\ell \gamma} \log \left(\frac{R^2 \gamma_\ell T}{\gamma} + 1 \right)$$

We now prove Lemma 2 using the above results. We provide the proofs of Lemma 6 and 7 following that.

Proof of Lemma 2 The proof proceeds by relating the squared deviation $\|W^* x_t - W_t x_t\|_2^2$ to the Bregman divergence of the function Φ under Assumptions 1 and 2. For a convex function f, the Bregman divergence, denoted by $\mathrm{D}_f(u, v)$ is the difference between the function f and its first-order Taylor approximation. More formally,

$$\mathrm{D}_f(u, v) = f(u) - f(v) - \langle \nabla f(v), u - v \rangle .$$

It is easily seen that Assumptions 1 and 2 correspond to quadratic lower and upper bounds respectively on the Bregman divergence of Φ. That is,

$$\frac{\gamma_\ell}{2}\|u - v\|_2^2 \leq \mathrm{D}_\Phi(u, v) \leq \frac{\gamma_u}{2}\|u - v\|_2^2, \quad \text{for all } u, v \in S. \tag{27}$$

In our current context, we use Assumption 1 to conclude

$$\sum_{t=1}^{T} Z_t \|W^* x_t - W_t x_t\|_2^2 \leq \frac{2}{\gamma_\ell} \sum_{t=1}^{T} Z_t \mathrm{D}_\Phi(W_t x_t, W^* x_t)$$

The above inequality allows us to invoke Lemmas 6 and 7 in turn which completes the proof. □

Proof of Lemma 6 Consider the random variable

$$\nu_t = Z_t \left[\mathrm{D}_\Phi(W_t x_t, W^* x_t) - (\ell(W_t; (x_t, y_t)) - \ell(W^*; (x_t, y_t))) \right].$$

In order for our proof, it will be convenient to work with the simplified form of the random variable obtained by using the definition (4) of the loss function.

$$\begin{aligned}
\nu_t &= Z_t \left[\mathrm{D}_\Phi(W_t x_t, W^* x_t) - (\ell(W_t; (x_t, y_t)) - \ell(W^*; (x_t, y_t))) \right] \\
&= Z_t \left[\mathrm{D}_\Phi(W_t x_t, W^* x_t) - (\Phi(W_t x_t) - y_t^T W_t x_t - \Phi(W^* x_t) - y_t^T W^* x_t) \right] \\
&= Z_t \left[\Phi(W_t x_t) - \Phi(W^* x_t) - \langle \nabla\Phi(W^* x_t), W_t x_t - W^* x_t \rangle - (\Phi(W_t x_t) - y_t^T W_t x_t - \Phi(W^* x_t) - y_t^T W^* x_t) \right] \\
&= Z_t \left\langle y_t - \nabla\Phi(W^* x_t), W_t x_t - W^* x_t \right\rangle . \qquad (28)
\end{aligned}$$

Here the second equality uses the definition of the Bregman divergence. Now recalling our earlier definition of the σ-fields $\mathcal{F}_t$, it is clear that ν_t is measurable with respect to $\mathcal{F}_{t+1}$. Furthermore, its conditional expectation conditioned on $\mathcal{F}_t$ is zero, since W_t, Z_t and x_t are measurable with respect to $\mathcal{F}_t$ and $\mathbb{E}[y_t \mid \mathcal{F}_t] = \nabla\Phi(W^* x_t)$. Hence the sequence ν_t is a martingale difference sequence with respect to the filtration $\mathcal{F}_t$. In order to prove the lemma, we just need to show that this sequence concentrates around its expectation. We do so by appealing to a form of Freedman's inequality (Freedman, 1975) presented in Kakade & Tewari (2009). In order to use the result, we need bounds on the value and the conditional variance of the random variable ν_t. We start with the bound on the value. Based on Equation 28, we have

$$\begin{aligned}
|\nu_t| &\leq |\left\langle y_t - \nabla\Phi(W^* x_t), W_t x_t - W^* x_t \right\rangle| \\
&\leq \|y_t - \nabla\Phi(W^* x_t)\|_1 \|W_t x_t - W^* x_t\|_\infty \\
&\leq 2(2R\omega).
\end{aligned}$$

Here the last inequality follows since y_t is a canonical basis vector, $\nabla\Phi(W^* x_t)$ is a probability distribution over $\mathbb{R}^K$ and $x_t^T W_t^i$ as well as $x_t^T W_i^*$ are bounded by $R\omega$ for $i = 1, 2, \ldots, K$ by Assumption 3. Hence we have obtained the upper bound

$$|\nu_t| \leq 4R\omega. \qquad (29)$$

Reasoning similarly for the conditional variance, we observe that

$$\begin{aligned}
\mathbb{E}[\nu_t^2 \mid \mathcal{F}_t] &\leq Z_t \mathbb{E}\left[\left\langle y_t - \nabla\Phi(W^* x_t), W_t x_t - W^* x_t \right\rangle^2 \mid \mathcal{F}_t \right] \\
&\leq 4 Z_t \|W_t x_t - W^* x_t\|_\infty^2 \\
&\leq 4 Z_t \|W_t x_t - W^* x_t\|_2^2 \\
&\leq \frac{8}{\gamma_\ell} Z_t \mathrm{D}_\Phi(W_t x_t, W^* x_t).
\end{aligned}$$

Now we appeal to Lemma 3 of Kakade & Tewari (2009), which yields for any $\delta < 1/e$ and $T \geq 3$, with probability at least $1 - 4\delta \log(T)$

$$\begin{aligned}\sum_{t=1}^{T}\nu_t &\le \max\left\{2\sqrt{\sum_{t=1}^{T}\frac{8}{\gamma_\ell}Z_t\mathrm{D}_\Phi(W_tx_t,W^*x_t)},12R\omega\sqrt{\log(1/\delta)}\right\}\sqrt{\log(1/\delta)}\\ &\le 4\sqrt{\frac{2}{\gamma_\ell}\sum_{t=1}^{T}Z_t\mathrm{D}_\Phi(W_tx_t,W^*x_t)\log\frac{1}{\delta}}+12R\omega\log\frac{1}{\delta}\\ &\le \frac{1}{2}\sum_{t=1}^{T}Z_t\mathrm{D}_\Phi(W_tx_t,W^*x_t)+\left(12R\omega\ +\frac{16}{\gamma_\ell}\right)\log\frac{1}{\delta},\end{aligned}$$

where the last inequality follows by Cauchy-Shwartz inequality. Recalling the definition of ν_t and our assumptions that $R\omega \ge 1$ as well as $\gamma_\ell \le 1$ completes the proof. □

Proof of Lemma 7 We follow the proof technique, which is an inductive argument introduced by Kalai & Vempala (2005). The proof reasons via an auxiliary sequence of fictitious iterates:

$$\tilde{W}_{t+1} = \arg\min_{W\in\mathcal{W}}\left\{\sum_{s=1}^{t+1}Z_s\ell(Wx_s,y_s)+\gamma\|W\|_F^2\right\}. \tag{30}$$

The main idea is that $\tilde{W}_t$ is an iterate sequence which cannot be played by the algorithm, since it relies on the unknown data point (x_t, y_t). However, it turns out that our iterates W_t are not too different from $\tilde{W}_t$, and the sequence $\tilde{W}_t$ has a low regret since it can see the data point (x_t, y_t) at which the regret is measured. The second claim can be found, for example, in Lemma 2.1 of Shalev-Shwartz (2012). That is, we are guaranteed that

$$\sum_{t=1}^{T}Z_t(\ell(\tilde{W}_t;(x_t,y_t))-\ell(W^*;(x_t,y_t)))\le 0.$$

Hence we focus on showing the closeness of the two sequences. Taking the optimality conditions for W_t and $\tilde{W}_t$, we see that

$$\begin{aligned}\left\langle\sum_{s=1}^{t-1}Z_s(\nabla\Phi(W_tx_s)^Tx_s-y_s^Tx_s)+\gamma W_t,\tilde{W}_t-W_t\right\rangle\ge 0\\ \left\langle\sum_{s=1}^{t}Z_s(\nabla\Phi(\tilde{W}_tx_s)^Tx_s-y_s^Tx_s)+\gamma\tilde{W}_t,W_t-\tilde{W}_t\right\rangle\ge 0.\end{aligned}$$

Adding the two inequalities, and rearranging we obtain

$$\sum_{s=1}^{t-1}Z_s\left\langle\nabla\Phi(W_tx_s)-\nabla\Phi(\tilde{W}_tx_s),\tilde{W}_tx_s-W_tx_s\right\rangle+Z_t\left\langle\nabla\Phi(\tilde{W}_tx_t)-y_t,W_tx_t-\tilde{W}_tx_t\right\rangle-\gamma\|W_t-\tilde{W}_t\|_F^2\ge 0.$$

By Assumption 2, the above inequality further yields

$$\begin{aligned}Z_t\left\langle\nabla\Phi(\tilde{W}_tx_t)-y_t,W_tx_t-\tilde{W}_tx_t\right\rangle &\ge \gamma_\ell\,Z_s\sum_{s=1}^{t-1}\|\tilde{W}_tx_s-W_tx_s\|_2^2+\gamma\|W_t-\tilde{W}_t\|_F^2\\ &= \gamma_\ell\,\|W_t-\tilde{W}_t\|_{M_t^2},\end{aligned}$$

where the last line uses the definition (7) of M_t. On the other hand, since $\nabla\Phi(\tilde{W}_t x_t)$ is a probability distribution over $\mathbb{R}^K$ and y_t is a canonical basis vector, we can also conclude

$$\begin{aligned}\left\langle \nabla\Phi(\tilde{W}_t x_t) - y_t, W_t x_t - \tilde{W}_t x_t \right\rangle &\le \|\nabla\Phi(\tilde{W}_t x_t) - y_t\|_1 \|W_t x_t - \tilde{W}_t x_t\|_\infty \\ &\le 2\|W_t x_t - \tilde{W}_t x_t\|_2 \\ &\le 2\|W_t - \tilde{W}_t\|_{M_t} \|x_t\|_{M_t^{-1}}.\end{aligned}$$

Combining the above two displays finally yields the desired inequality

$$\|W_t - \tilde{W}_t\|_{M_t} \le \frac{2Z_t}{\gamma_\ell} \|x_t\|_{M_t^{-1}}.$$

We are almost done now. All we need is to bound the difference between the regret of W_t and $\tilde{W}_t$ by using the above inequality. This will be done by exploiting the Lipschitz property of our loss function. We observe that we have

$$\begin{aligned}\sum_{t=1}^T Z_t(\ell(W_t;(x_t,y_t)) - \ell(\tilde{W}_t;(x_t,y_t))) &= \sum_{t=1}^T Z_t(\Phi(W_t x_t) - y_t^T W_t x_t - \Phi(\tilde{W}_t x_t) - y_t^T \tilde{W}_t x_t) \\ &\le \sum_{t=1}^T Z_t \left\langle \nabla\Phi(W_t x_t)^T x_t - y_t^T x_t, W_t - \tilde{W}_t \right\rangle \\ &= \sum_{t=1}^T Z_t \left\langle \nabla\Phi(W_t x_t) - y_t, W_t x_t - \tilde{W}_t x_t \right\rangle \\ &\le \sum_{t=1}^T Z_t \, 2\|W_t x_t - \tilde{W}_t x_t\|_2 \\ &\le \sum_{t=1}^T \frac{4Z_t}{\gamma_\ell} \|x_t\|^2_{M_t^{-1}}.\end{aligned}$$

Appealing to Lemma 5 completes the proof. □