M. Steyvers, H. Tejeda, G. Kerrigan, and P. Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119, 2022.

T. Sühr, S. Samadi, and C. Farronato. A dynamic model of performative human-ml collaboration: Theory and empirical evidence. *ArXiv*, abs/2405.13753, 2024. URL `https://api.semanticscholar.org/CorpusID:269982203`.

R. Verma and E. Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, pages 22184–22202. PMLR, 2022.

R. Verma, D. Barrejón, and E. Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 11415–11434. PMLR, 2023.

A Proof of Regret Guarantee

For simplicity, we drop the a suffix and look at a single parameter θ^*. Note that all the following also apply to w.

Define $C_t = \{\theta : \|\theta - \hat{\theta}_t\|_{M_t} \leq \frac{\sigma}{\kappa}\sqrt{2d\log\left(\frac{1+2td}{\delta}\right)} = \beta(t)\}$. Let $\tau = \min_{t\in[T]} : \lambda_{min}(M_t) \geq 1$. It was shown by Li et al. (2017) that with probability $1-\delta$, $\tau = O\left((d + \log 1/\delta)/\sigma_0^2\right)$ (recalling $\sigma_0 = \lambda_{min}\mathbb{E}_{x\sim\mathcal{D}}xx^\top > 0$).

We present two key lemmas from Li et al. (2017) on generalized linear bandits.

Lemma A.1 (Lemma 3 of Li et al. (2017)). *With probability* $1-\delta$*, for all* $t \geq \tau$*,* $\theta^* \in C_t$*.*

Lemma A.2 (Lemma 2 of Li et al. (2017)). *For all* $t > \tau$

$$\sum_{s=\tau+1}^{t} \|x_s\|_{M_t^{-1}} \leq \sqrt{2(t-\tau)d\log\frac{t}{d}}$$

These three results lead to the following two corollaries, corresponding to two corollaries given by Agrawal and Devanur (2016) in the linear bandit case.

Corollary A.3 (Corollary 1 of Agrawal and Devanur (2016)). *Let* $\bar{\theta} \in C_t$*. Then,*

$$\sum_{s=\tau}^{T} |x_t^\top\bar{\theta} - x_t^\top\theta^*| \leq \beta(T)\sqrt{2Td\log\frac{T}{d}}$$

Proof.

$$\begin{aligned}\sum_{s=\tau}^{T} |x_t^\top\bar{\theta} - x_t^\top\theta^*| &\leq \sum_{t=\tau}^{T}\|\bar{\theta} - \theta^*\|_{V_t}\|x_t\|_{V_t^{-1}} \\ &\leq \beta(T)\sqrt{2dT\log\frac{T}{d}}\end{aligned}$$

The first line comes from a known matrix-norm inequality (Lemma 7 of Agrawal and Devanur (2016)).

The second line comes from Lemmas A.1 and A.2. □

Via the definition of the optimistic estimate:

Corollary A.4 (Corollary 2 of Agrawal and Devanur (2016)). *With probability* $1-\delta$*, for all* $t \geq \tau$*,* $\mu(x_t^\top\tilde{\theta}_t) \geq \mu(x_t^\top\theta^*)$*, and*

$$\sum_{t=1}^{T} \mu(x_t^\top\tilde{\theta}_t) - \mu(x_t^\top\theta^*) \leq L_\mu\beta(T)\sqrt{2dT\log\frac{T}{d}}$$

Proof. The first part comes from the assumption that μ is an increasing function. Thus, $\mu(x_t^\top \tilde{\theta}_t) \geq \mu(x_t^\top \theta^*)$ by Lemma A.1 and the definition of $\tilde{\theta}$ as an optimistic estimator.

The second part follows from the assumption that μ is $L_\mu-$Lipschitz. So, $\sum_{t=1}^T \mu(x_t^\top \tilde{\theta}_t) - \mu(x_t^\top \theta^*) \leq \sum_{t=1}^T L_\mu(x_t^\top \tilde{\theta}_t - x_t^\top \theta^*)$, and the result follows from Corollary A.3. □

Now we have the tools we need to prove the regret bound.

Corollary A.5. *Given Z, the algorithm achieves the following with probability $1-\delta$:*

$$regret(T) = O\left(\left(\frac{OPT}{B}+1\right)\frac{L_\mu d\sigma}{\kappa}\sqrt{T\log\frac{T}{d\delta}\log\frac{T}{d}}\right)$$

Proof. We follow the proof steps presented in Agrawal and Devanur (2016), extending the claims to the generalized linear model when necessary.

Let T_{stop} be the stopping time of the algorithm. Let $R'(T) = O\left(\frac{d\sigma}{\kappa}\sqrt{T\log\frac{T}{d\delta}\log\frac{T}{d}}\right)$. Fix a. Also define $T_a = \{\tau < s < T_{stop} : a_t = a\}$. Via the Azuma-Hoeffding inequality,

$$\left|\sum_{s=\tau+1}^{T_{stop}} c_s - \mu(x_s^\top w_{a_t}^*)\right| \leq R'(T)$$

$$\left|\sum_{s=\tau+1}^{T_{stop}} r_s - \mu(x_s^\top \theta_{a_t}^*)\right| \leq R'(T)$$

Additionally, recalling Corollary A.4, with probability $1-\delta$, $\sum_{t=\tau+1}^{T_{stop}} \mu(x_t^\top \tilde{\theta}_{a_t,t}) - \mu(x_t^\top \theta_{a_t}^*) \leq L_\mu R'(T)$ (and similarly for w). Therefore, as in the linear case, a bound on the estimated reward with $\tilde{\theta}$ can serve as a proxy for the bound with θ^*.

Define $\tilde{r}_t = \mu(x_t^\top \tilde{\theta}_{a_t,t})$ and $\tilde{c}_t = \mu(x_t^\top \tilde{w}_{a_t,t})$.

Lemma A.6 (Lemma 8 of Agrawal and Devanur (2016))**.**

$$\sum_{t=\tau}^{T_{stop}} \mathbb{E}[\tilde{r}_t] \geq \frac{T_{stop}}{T} OPT + Z \sum_{t=\tau}^{T_{stop}} \gamma_t \mathbb{E}[\tilde{c}_t - B/T]$$

Proof. Let a^* be the action taken by the optimal static policy at t. By Corollary A.4, for any x_t, $\mu(x_t^\top \tilde{\theta}_{t,a^*}) \geq \mu(x_t^\top \theta_{a^*}^*)$. Therefore, $\mathbb{E}[\mu(x_t^\top \tilde{\theta}_{t,a^*})] \geq \text{OPT}/T$ and $\mathbb{E}[\mu(x_t^\top \tilde{w}_{a^*,T})] \leq B/T$ (taking the expectation over the choice of x_t, conditioned on the history). However, since the algorithm chooses the optimal optimistic action:

$$\begin{aligned}
\tilde{r}_t - Z\gamma_t\tilde{c}_t &\geq \mu(x_t^\top \tilde{\theta}_{t,a^*}) - Z\gamma_t\mu(x_t^\top \tilde{w}_{t,a^*}) \\
\mathbb{E}[\tilde{r}_t - Z\gamma_t\tilde{c}_t] &\geq \mathbb{E}[\mu(x_t^\top \tilde{\theta}_{t,a^*})] - Z\gamma_t\mathbb{E}[\mu(x_t^\top \tilde{w}_{t,a^*})] \\
&\geq \frac{\text{OPT}}{T} - Z\gamma_t\frac{B}{T}
\end{aligned}$$

Sum to T_{stop} to get the Lemma statement. □

The rest of the proof follows identically to Agrawal and Devanur (2016). □

B Training Details for Neural Algorithm

For both the ImageNet16H data and the Knapsack data, the Neural algorithm trained three separate neural networks with the same architecture. They consist of an input layer with the same dimension as the context, a hidden layer of dimension 50, and a single output layer. Note that this means that in both experiments, the dimension of the linear system is 50. There is a ReLU activation between the input layer and the hidden layer, and a Sigmoid activation on the output. The weights of the networks were updated every 10 steps for the Knapsack data. For the ImageNet16H data, the weights were initially updated every 20 steps, decreasing to every 100 steps after time step 4000. Both were trained with mini-batches of size 500 using the Adam optimizer with learning rate 0.0005 for the Knapsack data and 0.0001 for the ImageNet16H data. The experiments were run on Google Colab servers using their T4 GPU.

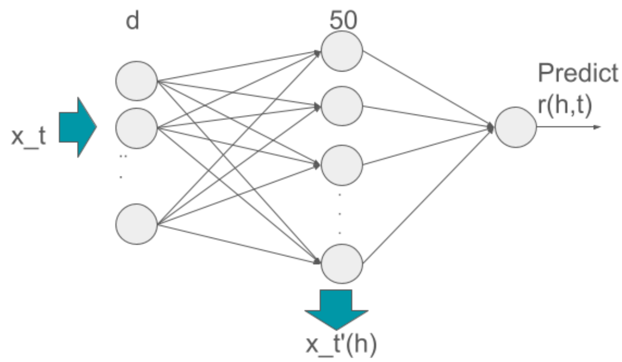

Figure 7: The architecture for computing the embedding for the human arm. An identical architecture exists for the model arm and the cost.

After updating the network weights, the embeddings for all previous contexts are recomputed using the new networks. These new embeddings are used to recompute the estimated parameters per Definition 4.2 and 4.1. As noted in (Riquelme et al., 2018), this may not be practical in applications where all previous contexts cannot be stored, either due to space constraints or legal concerns. In these settings, one can continue to use the previous embeddings and apply weights which decrease the influence of old embeddings on the linear system over time.

C Bandit Feedback Experiments

Overall, we do not observe a significant difference in performance between the bandit feedback setting and the full information setting. With random reward and cost functions, the average performance in the full information setting is slightly better, as seen in Figure 8. Interestingly, in the Knapsack dataset, the linear algorithm seemed to perform slightly better in the pure bandit setting (as shown in Figure 9). This may indicate that the full information setting overexplored the human arm.

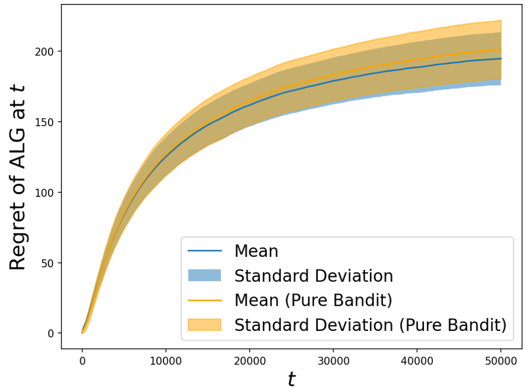

Figure 8: The experiment described in Figure 2 with the Pure Bandit setting included. Mean and standard deviation of the regret over 100 trials. The reward and cost functions are sampled uniformly at random from $[0, 1]^d$ for each trial.The algorithm is run over $T = 50000$ random contexts with $B = 8000$. Then, the reward received by OPT is computed for the same contexts.

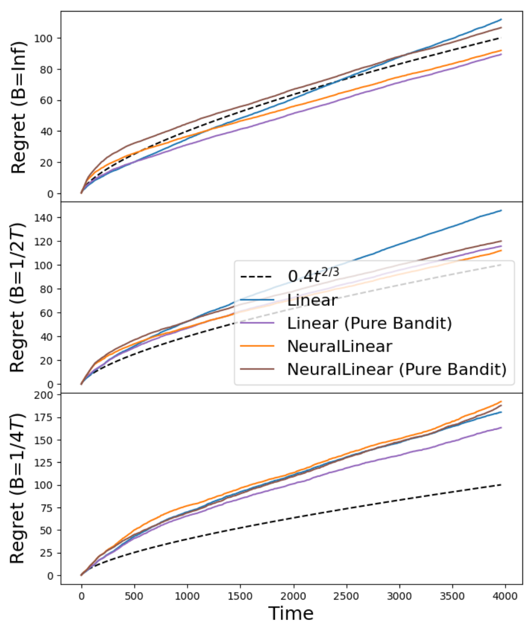

Figure 9: The experiment described in Figure 5 with the Pure Bandit setting included. For clarity, only the means are plotted.