Feedback Coding for Active Learning: Supplementary Materials

A PROOFS OF ANALYTICAL RESULTS

A.1 Proof of Proposition 2.1

Proof. Our proof follows closely to that of Singh et al. (2009) for the capacity of the one-bit quantized Gaussian channel. We start by writing $I(L;Y) = H(Y) - H(Y \mid L)$, where H denotes the entropy of a discrete random variable (Cover and Thomas, 2006). $H(Y)$ is maximized at 1 bit, when $p(Y=1) = p(Y=-1) = 0.5$. Expanding $H(Y \mid L)$, we have $H(Y \mid L) = \mathbb{E}_{p_L}[h_b(p(Y=1 \mid L))] = \mathbb{E}_{p_L}[h_b(f(L))]$.

For distribution p_L, consider its symmetrized distribution $\widetilde{p}_L(\ell) = \frac{1}{2}p_L(\ell) + \frac{1}{2}p_L(-\ell)$ and the expectation of any even function $e(\cdot)$ over $\widetilde{p}_L(\ell)$:

$$\begin{aligned}
E_{\widetilde{p}_L}[e(L)] &= \int_{-\infty}^{\infty}\left(\frac{1}{2}p_L(\ell) + \frac{1}{2}p_L(-\ell)\right)e(\ell)d\ell \\
&= \frac{1}{2}\int_{-\infty}^{\infty} p_L(\ell)e(\ell)d\ell + \frac{1}{2}\int_{-\infty}^{\infty} p_L(-\ell)e(\ell)d\ell \\
&= \frac{1}{2}\int_{-\infty}^{\infty} p_L(\ell)e(\ell)d\ell + \frac{1}{2}\int_{-\infty}^{\infty} p_L(\ell)e(-\ell)d\ell && \text{change of variables} \\
&= \frac{1}{2}\int_{-\infty}^{\infty} p_L(\ell)e(\ell)d\ell + \frac{1}{2}\int_{-\infty}^{\infty} p_L(\ell)e(\ell)d\ell && e(\ell) \text{ is even} \\
&= \frac{1}{2}E_{p_L}[e(L)] + \frac{1}{2}E_{p_L}[e(L)] \\
&= E_{p_L}[e(L)]
\end{aligned}$$

Observe that h_b is symmetric about 0.5, i.e. for $x \in [-0.5, 0.5]$, $h_b(0.5+x) = h_b(0.5-x)$. Combining this with the fact that $f(\ell) - 0.5$ is an odd function (i.e. $f(-\ell) - 0.5 = -(f(\ell) - 0.5)$), we have

$$h_b(f(-\ell)) = h_b(f(-\ell) - 0.5 + 0.5) = h_b(-(f(\ell) - 0.5) + 0.5) = h_b((f(\ell) - 0.5) + 0.5) = h_b(f(\ell))$$

and so $h_b(f(\ell))$ is an even function. Therefore, the conditional entropy $H(Y \mid L)$ is equivalent when L is distributed as p_L or $\widetilde{p}_L$, i.e. $E_{\widetilde{p}_L}[h_b(f(L))] = E_{p_L}[h_b(f(L))]$.

We also have

$$\begin{aligned}
E_{\widetilde{p}_L}[f(L)] &= E_{\widetilde{p}_L}[f(L) - 0.5] + 0.5 \\
&= \int_{-\infty}^{\infty}\left(\frac{1}{2}p_L(\ell) + \frac{1}{2}p_L(-\ell)\right)(f(\ell) - 0.5)d\ell + 0.5 \\
&= \frac{1}{2}\int_{-\infty}^{\infty} p_L(\ell)(f(\ell) - 0.5)d\ell + \frac{1}{2}\int_{-\infty}^{\infty} p_L(-\ell)(f(\ell) - 0.5)d\ell + 0.5 \\
&= \frac{1}{2}\int_{-\infty}^{\infty} p_L(\ell)(f(\ell) - 0.5)d\ell + \frac{1}{2}\int_{-\infty}^{\infty} p_L(\ell)(f(-\ell) - 0.5)d\ell + 0.5 && \text{change of variables} \\
&= \frac{1}{2}\int_{-\infty}^{\infty} p_L(\ell)(f(\ell) - 0.5)d\ell - \frac{1}{2}\int_{-\infty}^{\infty} p_L(\ell)(f(\ell) - 0.5)d\ell + 0.5 && (f(\ell) - 0.5) \text{ is odd} \\
&= 0.5
\end{aligned}$$

and so under $\widetilde{p}_L$, $p(Y=1) = \mathbb{E}_{\widetilde{p}_L}[f(L)] = 0.5$ and $H(Y)$ is maximized at 1 bit.

Combining these facts, we have

$$I(\widetilde{p}_L, f) = 1 - E_{\widetilde{p}_L}[h_b(f(L))] = 1 - E_{p_L}[h_b(f(L))] \geq h_b(E_{p_L}[f(L)]) - E_{p_L}[h_b(f(L))] = I(p_L, f)$$

and so symmetrizing a distribution can only increase $I(L;Y)$. Furthermore, since ℓ^2 is even we have $\mathbb{E}_{\widetilde{p}_L}[L^2] = \mathbb{E}_{p_L}[L^2]$. Therefore, when evaluating the capacity of channel with transition probability f under power constraint P, we only consider symmetric distributions since for every $p_L \in \mathcal{C}_P$ there exists a symmetric distribution $\widetilde{p}_L \in \mathcal{C}_P$ satisfying $I(\widetilde{p}_L, f) \geq I(p_L, f)$. We solve for the capacity-achieving distribution over the set of symmetric distributions in $\mathcal{C}_P$:

$$\begin{aligned} p_L^* &= \underset{\substack{\mathbb{E}_{p_L}[L^2] \leq P \\ p_L(\ell) = p_L(-\ell)}}{\arg\max} \quad I(p_L, f) && (1) \\ &= \underset{\substack{\mathbb{E}_{p_L}[L^2] \leq P \\ p_L(\ell) = p_L(-\ell)}}{\arg\max} \quad 1 - E_{p_L}[h_b(f(L))] \\ &= \underset{\substack{\mathbb{E}_{p_L}[L^2] \leq P \\ p_L(\ell) = p_L(-\ell)}}{\arg\min} \quad E_{p_L}[h_b(f(L))] && (2) \end{aligned}$$

Since $h_b(f(\ell))$ is even, $h_b(f(\ell)) = h_b(f(|\ell|)) = h_b(f(\sqrt{\ell^2}))$. Omitting calculations, we have

$$\frac{d^2}{du^2} h_b(f(\sqrt{u})) = (\log_2 e) \frac{\tanh(\frac{\sqrt{u}}{2}) \operatorname{sech}^2(\frac{\sqrt{u}}{2})}{16\sqrt{u}}$$

which is non-negative for $u > 0$ and therefore $h_b(f(\sqrt{u}))$ (which is continuous on $u \geq 0$) is convex on $u \geq 0$. We then have

$$E_{p_L}[h_b(f(L))] = E_{p_L}\left[h_b\big(f\big(\sqrt{L^2}\big)\big)\right] \overset{(a)}{\geq} h_b\Big(f\Big(\sqrt{E_{p_L}[L^2]}\Big)\Big) \overset{(b)}{\geq} h_b\big(f\big(\sqrt{P}\big)\big)$$

where Jensen's inequality is used in (a) (Cover and Thomas, 2006), with equality if and only if L^2 is constant, and (b) results from the power constraint $E_{p_L}[L^2] \leq P$ and the fact that $h_b(f(\sqrt{u}))$ is monotonically decreasing for $u \geq 0$. For symmetric p_L, equality in (a) is achieved if $p_L = B_t$ for some $t > 0$. By setting $t = \sqrt{P}$, equality in (b) is also achieved, and so $B_{\sqrt{P}}$ minimizes (2) (and therefore maximizes (1)). The maximum value in (1), which is equal to capacity C, is then

$$I(B_{\sqrt{P}}, f) = 1 - \mathbb{E}_{B_{\sqrt{P}}}[h_b(f(L))] = 1 - \frac{1}{2} h_b(f(\sqrt{P})) - \frac{1}{2} h_b(f(-\sqrt{P})) = 1 - h_b(f(\sqrt{P})). \qquad \square$$

A.2 Proof of Proposition 2.2

Proof. Since p_θ is log-concave, then $p_{\theta|\mathcal{L}_{n-1}}(\theta) \propto p_\theta(\theta) \prod_{i=1}^{n-1} p(Y = y_i \mid x_i, \theta)$ is also log-concave since it is the product of log-concave functions (Saumard and Wellner, 2014). Since marginals of log-concave distributions are log-concave (Lovász and Vempala, 2007), $L_n = x_n^T \theta$ is log-concave for any x_n under the distribution $p_{\theta|\mathcal{L}n-1}$. However, we know from Proposition 2.1 that p_L^* for logistic regression is a sum of mass points, which is not log-concave. Therefore no x_n exists which can induce p_L^* from h. $\square$

A.3 Proof of Theorem 3.1

Proof. In the following, suppose that $p_L \in \mathcal{C}_P$, and let $H_{p_L}(Y) = h_b(\mathbb{E}_{p_L}[f(L)])$ and $H_{p_L}(Y \mid L) = \mathbb{E}_{p_L}[h_b(f(L))]$. $f(\ell)$ is K_1-Lipschitz, where $K_1 = 0.25$, and $h_b(f(\ell))$ is K_2-Lipschitz, where $K_2 \approx 0.32$.

$$\begin{aligned} |I(p_L, f) - I(B_t, f)| &= |H_{p_L}(Y) - H_{p_L}(Y \mid L) - (H_{B_t}(Y) - H_{B_t}(Y \mid L))| \\ &\leq |H_{p_L}(Y) - H_{B_t}(Y))| + |H_{p_L}(Y \mid L) - H_{B_t}(Y \mid L)| \\ &= |h_b(\mathbb{E}_{p_L}[f(L)]) - h_b(\mathbb{E}_{B_t}[f(L)])| + \left| \int_\ell h_b(f(\ell)) p_L(\ell) d\ell - \int_\ell h_b(f(\ell)) B_t(\ell) d\ell \right| \end{aligned}$$

Assume that there exists $\varepsilon \in (0, 0.5)$ such that $\varepsilon \leq \mathbb{E}_{p_L}[f(L)] \leq 1-\varepsilon$. For $\ell \in (\varepsilon, 1-\varepsilon)$, h_b is $\log_2 \frac{1-\varepsilon}{\varepsilon}$-Lipschitz. Since $\varepsilon < \mathbb{E}_{p_L}[f(L)] < 1-\varepsilon$ by assumption and B_t satisfies $\varepsilon < \mathbb{E}_{B_t}[f(L)] < 1-\varepsilon$ since $\mathbb{E}_{B_t}[f(L)] = 0.5$, we have

$$\begin{aligned}|h_b(\mathbb{E}_{p_L}[f(L)]) - h_b(\mathbb{E}_{B_t}[f(L)])| &\leq \log_2\Big(\frac{1-\varepsilon}{\varepsilon}\Big)|\mathbb{E}_{p_L}[f(L)] - \mathbb{E}_{B_t}[f(L)]| \\ &= \log_2\Big(\frac{1-\varepsilon}{\varepsilon}\Big)\Big|\int_\ell f(\ell)p_L(\ell)d\ell - \int_\ell f(\ell)B_t(\ell)d\ell\Big|\end{aligned}$$

which implies

$$|I(p_L, f) - I(B_t, f)| \leq \log_2\Big(\frac{1-\varepsilon}{\varepsilon}\Big)\Big|\int_\ell f(\ell)p_L(\ell)d\ell - \int_\ell f(\ell)B_t(\ell)d\ell\Big| + \Big|\int_\ell h_b(f(\ell))p_L(\ell)d\ell - \int_\ell h_b(f(\ell))B_t(\ell)d\ell\Big| \quad (3)$$

To continue, we use the following result from Villani (2008): defining $P_1(\mathbb{R}) \coloneqq \{\mu' : \mathbb{E}_{\mu'}[|L|] < \infty\}$, for any $\mu, \nu \in P_1(\mathbb{R})$ we have

$$\sup_{\|f\|_{\text{Lip}} \leq 1} \int_\ell f(\ell)\mu(\ell)d\ell - \int_\ell f(\ell)\nu(\ell)d\ell = W_1(\mu, \nu).$$

Therefore, for any K-Lipschitz function g we have that $\frac{g}{K}$ is 1-Lipschitz and so

$$\begin{aligned}\Big|\int_\ell g(\ell)\mu(\ell)d\ell - \int_\ell g(\ell)\nu(\ell)d\ell\Big| &= K\Big|\int_\ell \frac{g(\ell)}{K}\mu(\ell)d\ell - \int_\ell \frac{g(\ell)}{K}\nu(\ell)d\ell\Big| \\ &= K\max\Big\{\int_\ell \frac{g(\ell)}{K}\mu(\ell)d\ell - \int_\ell \frac{g(\ell)}{K}\nu(\ell)d\ell, \int_\ell \frac{-g(\ell)}{K}\mu(\ell)d\ell - \int_\ell \frac{-g(\ell)}{K}\nu(\ell)d\ell\Big\} \\ &\leq K \sup_{\|f\|_{\text{Lip}} \leq 1} \int_\ell f(\ell)\mu(\ell)d\ell - \int_\ell f(\ell)\nu(\ell)d\ell \\ &\leq KW_1(\mu, \nu) \\ &\leq KW_2(\mu, \nu) \qquad (4)\end{aligned}$$

where the last inequality is from $W_1(\mu, \nu) \leq W_2(\mu, \nu)$ (Villani, 2008).

To apply this inequality to both expressions in (3), we first verify that $p_L, B_t \in P_1(\mathbb{R})$. $\mathbb{E}_{B_t}[|L|] = t < \infty$, and

$$\mathbb{E}_{p_L}[|L|] = \mathbb{E}_{p_L}\Big[\sqrt{L^2}\Big] \overset{(a)}{\leq} \sqrt{\mathbb{E}_{p_L}[L^2]} \overset{(b)}{\leq} \sqrt{P} < \infty$$

where (a) results from Jensen's inequality with the concavity of $\sqrt{\cdot}$, and (b) is since $\mathbb{E}_{p_L}[L^2] \leq P$ by assumption and $\sqrt{\cdot}$ is monotonically increasing. Applying (4) separately to both terms in (3), we have

$$|I(p_L, f) - I(B_t, f)| \leq \Big(K_1 \log_2\Big(\frac{1-\varepsilon}{\varepsilon}\Big) + K_2\Big)W_2(p_L, B_t) \quad (5)$$

Finally, we compute a valid value of ε for all $p_L \in \mathcal{C}_P$. First note that $f(\ell) < 0.5$ for $\ell < 0$ and $f(\ell) \geq 0.5$ for $\ell \geq 0$, implying that $f(\ell) \leq f(|\ell|)\ \forall \ell$. Next note that $f(\sqrt{u})$ is concave on $u \geq 0$, since for any $u, v \in [0, \infty)$ and any $0 < \phi < 1$

$$f(\sqrt{\phi u + (1-\phi)v}) \geq f(\phi\sqrt{u} + (1-\phi)\sqrt{v})$$

since f is monotonically increasing and $\sqrt{\cdot}$ is concave.

$$\geq \phi f(\sqrt{u}) + (1-\phi)f(\sqrt{v})$$

since f is concave on $\mathbb{R}_{\geq 0}$. This can be shown by considering

$$\frac{d^2}{du^2}f(u) = \frac{e^u}{(1+e^u)^3}(1-e^u) \leq 0\ \forall u \geq 0$$

Combining these facts, we have

$$\begin{aligned}\mathbb{E}_{p_L}[f(L)] &\le \mathbb{E}_{p_L}[f(|L|)] && \text{since } f(\ell) \le f(|\ell|)\\ &= \mathbb{E}_{p_L}\left[f\big(\sqrt{L^2}\big)\right]\\ &\le f\left(\sqrt{\mathbb{E}_{p_L}[L^2]}\right) && \text{from Jensen's inequality with the concavity of } f(\sqrt{\cdot})\\ &\le f(\sqrt{P})\end{aligned}$$

since $f(\sqrt{\cdot})$ is monotonically increasing, and by assumption $E_{p_L}[L^2] \le P$. Similarly, $\mathbb{E}_{p_L}[1 - f(L)] \le f(\sqrt{P})$, and therefore we can set $\varepsilon = 1 - f(\sqrt{P})$. Applying this choice of ε to (5) we have

$$|I(p_L, f) - I(B_t, f)| \le \left(K_1 \log_2\left(\frac{f(\sqrt{P})}{1 - f(\sqrt{P})}\right) + K_2\right) W_2(p_L, B_t)$$

and can set $K_P = K_1 \log_2\left(\frac{f(\sqrt{P})}{1-f(\sqrt{P})}\right) + K_2$ to obtain $|I(p_L, f) - I(B_t, f)| \le K_P W_2(p_L, B_t)$.

Recall that $C = \max_{p_L \in \mathcal{C}_P} I(p_L, f) = I(B_{\sqrt{P}}, f)$ and $\widetilde{C}_n = \max_{x \in \mathcal{U}_n} I(p_{L_n|\mathcal{L}_{n-1}}, f)$. By assumption, P is selected such that $p_{L_n|\mathcal{L}_{n-1}} \in \mathcal{C}_P$ for any $x \in \mathcal{U}_n$, which implies $I(p_{L_n|\mathcal{L}_{n-1}}, f) \le C$ for any $x \in \mathcal{U}_n$ and hence $\widetilde{C}_n \le C$. Combining these facts, we have

$$\widetilde{C}_n - I(p_{L_n|\mathcal{L}_{n-1}}, f) \le C - I(p_{L_n|\mathcal{L}_{n-1}}, f) = |I(B_{\sqrt{P}}, f) - I(p_{L_n|\mathcal{L}_{n-1}}, f)| \le K_P W_2(p_{L_n|\mathcal{L}_{n-1}}, B_{\sqrt{P}}). \qquad \square$$

A.4 Proof of Proposition 3.2

Proof. Adopting notation from Mérigot (2011), let S denote a finite set of points in $\mathbb{R}$, and $w\colon S \to \mathbb{R}$ a weight vector. Define $\mathrm{Vor}_S^w(p) = \{\ell : \|\ell - p\|_2^2 - w(p) \le \|\ell - q\|_2^2 - w(q) \;\; \forall q \in S\}$.

Let μ be a given probability measure with density p_L. Consider $S = \{-t, t\}$, with the corresponding measure $B_t = \sum_{p \in S} \frac{1}{2}\delta_p = \frac{1}{2}\delta_{-t} + \frac{1}{2}\delta_t$. Let $w^*(-t) = 2t\,\mathrm{med}_{p_L}(L)$, and $w^*(t) = -2t\,\mathrm{med}_{p_L}(t)$. We have

$$\begin{aligned}\mathrm{Vor}_S^{w^*}(-t) &= \{\ell : \|\ell + t\|_2^2 - w^*(-t) \le \|\ell - q\|_2^2 - w^*(q) \;\; \forall q \in \{-t, t\}\}\\ &= \{\ell : \|\ell + t\|_2^2 - w^*(-t) \le \|\ell - t\|_2^2 - w^*(t)\}\\ &= \{\ell : \|\ell + t\|_2^2 - 2t\,\mathrm{med}_{p_L}(L) \le \|\ell - t\|_2^2 + 2t\,\mathrm{med}_{p_L}(L)\}\\ &= \{\ell : \ell \le \mathrm{med}_{p_L}(L)\}\end{aligned}$$

and similarly $\mathrm{Vor}_S^{w^*}(t) = \{\ell : \ell \ge \mathrm{med}_{p_L}(L)\}$. We have

$$\int_{\mathrm{Vor}_S^{w^*}(-t)} p_L(\ell)d\ell = \int_{\ell \le \mathrm{med}_{p_L}(L)} p_L(\ell)d\ell = \frac{1}{2}$$

and similarly $\int_{\mathrm{Vor}_S^{w^*}(t)} p_L(\ell)d\ell = \frac{1}{2}$. Therefore, w^* is *adapted* to (μ, B_t). By Theorem 2 of Mérigot (2011), a map $T_S^{w^*} : \mathbb{R} \to \mathbb{R}$ exists which realizes an optimal transport between μ and B_t. By Mérigot (2011) Theorem 1, we have

$$\begin{aligned}W_2^2(\mu, B_t) &= \int_{\mathrm{Vor}_S^{w^*}(-t)} \|\ell + t\|_2^2\, p_L(\ell)d\ell + \int_{\mathrm{Vor}_S^{w^*}(t)} \|\ell - t\|_2^2\, p_L(\ell)d\ell\\ &= \int_{\ell \le \mathrm{med}_{p_L}(L)} \|\ell + t\|_2^2\, p_L(\ell)d\ell + \int_{\ell \ge \mathrm{med}_{p_L}(L)} \|\ell - t\|_2^2\, p_L(\ell)d\ell\\ &= \mathbb{E}_{p_L}[L^2] + t^2 - 2t\left(\int_{\ell \ge \mathrm{med}_{p_L}(L)} \ell\, p_L(\ell)d\ell - \int_{\ell \le \mathrm{med}_{p_L}(L)} \ell\, p_L(\ell)d\ell\right)\\ &= \mathbb{E}_{p_L}[L^2] + t^2 - 2t\left(\int_{\ell \ge \mathrm{med}_{p_L}(L)} (\ell - \mathrm{med}_{p_L}(L))\, p_L(\ell)d\ell + \int_{\ell \le \mathrm{med}_{p_L}(L)} (\mathrm{med}_{p_L}(L) - \ell)\, p_L(\ell)d\ell\right)\\ &= \mathbb{E}_{p_L}[L^2] + t^2 - 2t\left(\int_{\ell \ge \mathrm{med}_{p_L}(L)} |\ell - \mathrm{med}_{p_L}(L)|\, p_L(\ell)d\ell + \int_{\ell \le \mathrm{med}_{p_L}(L)} |\ell - \mathrm{med}_{p_L}(L)|\, p_L(\ell)d\ell\right)\\ &= \mathbb{E}_{p_L}[L^2] + t^2 - 2t\,\mathbb{E}_{p_L}[|L - \mathrm{med}_{p_L}(L)|]\end{aligned} \qquad \square$$

A.5 Proof of Corollary 3.2.1

Proof. Let $p_L \sim \mathcal{N}(\mu, \sigma^2)$. We have $\mathbb{E}_{p_L}[L^2] = \mathbb{E}_{p_L}[L]^2 + \mathrm{Var}_{p_L}(L) = \mu^2 + \sigma^2$, and $\mathbb{E}_{p_L}[|L - \mathrm{med}_{p_L}(L)|] = \mathbb{E}_{p_L}[|L - \mu|] = \sigma\sqrt{\frac{2}{\pi}}$ (Winkelbauer, 2014). Hence $W_2^2(p_L, B_t) = \mathbb{E}_{p_L}[L^2] + t^2 - 2t\,\mathbb{E}_{p_L}[|L - \mathrm{med}_{p_L}(L)|] = \mu^2 + \sigma^2 + t^2 - 2\sqrt{\frac{2}{\pi}}t\sigma$. Completing the square, we have the desired result. □

B EXPERIMENT DETAILS

B.1 Selection of Power Constraint

Recall that APM-LR minimizes an objective function consisting of a mixture of two terms, reprinted below:

$$\pi_n(\mathcal{L}_{n-1}) = \underset{x \in \mathcal{U}_n}{\arg\min}\ (\mu_n^T x)^2 + \left(\sqrt{x^T \Sigma_n x} - \sqrt{\frac{2}{\pi} P_n}\right)^2. \tag{6}$$

The first term in (6), which is independent of P_n, encourages x to lie orthogonal to the hyperplane posterior mean, μ_n. For all such x satisfying $\mu_n^T x = 0$, we have $\mathbb{E}[L_n] = \mu_n^T x = 0$ and

$$\mathbb{E}[L_n^2] = (\mu_n^T x)^2 + x^T \Sigma_n x = x^T \Sigma_n x \leq B^2 \lambda_1(\Sigma_n)$$

where expectations are taken with respect to $p_{L_n|\mathcal{L}_{n-1}}$. Therefore $P_n = B^2\lambda_1(\Sigma_n)$ is a valid power constraint for the set of examples that induce zero-mean input distributions. This set arguably contains the "best" candidate examples, since if $(\mu_n^T x)^2 \gg 0$ then the objective in (6) will be large. For this reason we set $P_n = B^2\lambda_1(\Sigma_n)$ in our experiments, as opposed to the power constraint of $B^2\lambda_1(\mu_n\mu_n^T + \Sigma_n)$ which is valid for all examples but is loose for examples encouraged by the first term in (6).

B.2 Dataset Information

In Table 1 we describe the datasets used in our experiments. Several datasets have multiple classes: in this case, we select a two-class dataset partition by either grouping individual classes together into super-classes, or simply training on a subset of the classes. In our experiments we treat each class partition as its own dataset, and refer to each partition by a nickname. All datasets except for *clouds*, *cross*, and *horseshoe* come from the UCI Machine Learning Repository (Dua and Graff, 2017); several UCI datasets have additional citations, which are listed next to their names.

Nickname	Dataset	Class partition	# of features	# of examples
vehicle-full	Vehicle Silhouettes (Siebert, 1987)	$Y = -1$: 'saab' or 'opel' $Y = 1$: 'bus' or 'van'	18	846
vehicle-cars	Vehicle Silhouettes (Siebert, 1987)	$Y = -1$: 'saab' $Y = 1$: 'opel'	18	429
vehicle-transport	Vehicle Silhouettes (Siebert, 1987)	$Y = -1$: 'bus' $Y = 1$: 'van'	18	417
letterDP	Letter Recognition	$Y = -1$: 'D' $Y = 1$: 'P'	16	1608
letterEF	Letter Recognition	$Y = -1$: 'E' $Y = 1$: 'F'	16	1543
letterIJ	Letter Recognition	$Y = -1$: 'I' $Y = 1$: 'J'	16	1502
letterMN	Letter Recognition	$Y = -1$: 'M' $Y = 1$: 'N'	16	1575
letterUV	Letter Recognition	$Y = -1$: 'U' $Y = 1$: 'V'	16	1577
letterVY	Letter Recognition	$Y = -1$: 'V' $Y = 1$: 'Y'	16	1550
austra	Australian Credit Approval	$Y = -1$: '0' $Y = 1$: '1'	14	690
wdbc	Breast Cancer Wisconsin (Diagnostic)	$Y = -1$: 'M' $Y = 1$: 'B'	30	569
clouds	Synth1 (Yang and Loog, 2018)	$Y = -1$: '-1' $Y = 1$: '1'	2	600
cross	Synth2 (Yang and Loog, 2018)	$Y = -1$: '-1' $Y = 1$: '1'	2	600
horseshoe	Synth3 (Yang and Loog, 2018)	$Y = -1$: '-1' $Y = 1$: '1'	2	600

Table 1: Full dataset information

B.3 Baseline Methods Details

Below we elaborate on the BALD and InfoGain baseline selection methods:

InfoGain We can directly approximate information gain $I(\theta; Y \mid \mathcal{L}_{n-1})$ with a Monte Carlo approximation over s samples from $p_{\theta|\mathcal{L}_{n-1}} \sim \mathcal{N}(\mu_n, \Sigma_n)$:

$$\begin{aligned} I(\theta; Y \mid \mathcal{L}_{n-1}) &= h_b(\mathbb{E}_{p_{\theta|\mathcal{L}_{n-1}}}[f(\theta^T x_n)]) - \mathbb{E}_{p_{\theta|\mathcal{L}_{n-1}}}[h_b(f(\theta^T x_n))] \\ &\approx h_b\left(\frac{1}{s}\sum_{i=1}^{s} f(\theta_i^T x_n)\right) - \frac{1}{s}\sum_{i=1}^{s} h_b\Big(f(\theta_i^T x_n)\Big) \qquad \theta_i \sim p_{\theta|\mathcal{L}_{n-1}} \\ &\approx h_b\left(\frac{1}{s}\sum_{i=1}^{s} f(\theta_i^T x_n)\right) - \frac{1}{s}\sum_{i=1}^{s} h_b\Big(f(\theta_i^T x_n)\Big) \qquad \theta_i \sim \mathcal{N}(\mu_n, \Sigma_n) \end{aligned} \tag{7}$$

Our "InfoGain" baseline selects the example $x_n \in \mathcal{U}_n$ that maximizes the expression in (7), computed in $O(sd)$ time per candidate example.

BALD Consider a probit regression label distribution $p(Y = 1 \mid L) = \Phi(L)$, where Φ is the standard normal cumulative distribution function. For $p_L \sim \mathcal{N}(\mu, \sigma^2)$, Houlsby et al. (2011) use a Taylor expansion in the BALD algorithm to approximate $I(p_L, \Phi(L))$ as

$$I(p_L, \Phi(L)) \approx h_b\left(\Phi\left(\frac{\mu}{\sqrt{\sigma^2+1}}\right)\right) - \frac{D\exp\left(-\frac{\mu^2}{2(\sigma^2+D^2)}\right)}{\sqrt{\sigma^2+D^2}} \tag{8}$$

where $D = \sqrt{\frac{\pi \ln 2}{2}}$. By equalizing derivatives at $L = 0$, we can approximate $f(L) \approx \Phi(kL)$ where $k = \sqrt{\frac{\pi}{8}}$ (Bishop, 2006). Define $\widetilde{L} = kL$ and note that $\widetilde{L} \sim \mathcal{N}(\widetilde{\mu}, \widetilde{\sigma}^2)$ for $\widetilde{\mu} = k\mu$ and $\widetilde{\sigma}^2 = k^2\sigma^2$. We can then use the BALD approximation in (8) for logistic regression:

$$\begin{aligned}
I(p_L, f(L)) &\approx I(p_L, \Phi(kL)) \\
&= h_b(\mathbb{E}_{p_L}(\Phi(kL))) - \mathbb{E}_{p_L}(h_b(\Phi(kL))) \\
&= h_b(\mathbb{E}_{p_{\widetilde{L}}}(\Phi(\widetilde{L}))) - \mathbb{E}_{p_{\widetilde{L}}}(h_b(\Phi(\widetilde{L}))) \\
&= I(p_{\widetilde{L}}, \Phi(\widetilde{L})) \\
&\approx h_b\left(\Phi\left(\frac{k\mu}{\sqrt{k^2\sigma^2+1}}\right)\right) - \frac{D\exp\left(-\frac{k^2\mu^2}{2(k^2\sigma^2+D^2)}\right)}{\sqrt{k^2\sigma^2+D^2}}
\end{aligned}$$

Approximating $p_{\theta|\mathcal{L}_{n-1}} \sim \mathcal{N}(\mu_n, \Sigma_n)$, we have $p_{L_n|\mathcal{L}_{n-1}} \sim \mathcal{N}(\mu_n^T x_n, x_n^T \Sigma_n x_n)$ and so we can approximate

$$I(p_{L_n|\mathcal{L}_{n-1}}, f(L)) \approx h_b\left(\Phi\left(\frac{k\mu_n^T x_n}{\sqrt{k^2 x_n^T \Sigma_n x_n + 1}}\right)\right) - \frac{D\exp\left(-\frac{k^2(\mu_n^T x_n)^2}{2(k^2 x_n^T \Sigma_n x_n + D^2)}\right)}{\sqrt{k^2 x_n^T \Sigma_n x_n + D^2}} \tag{9}$$

where $D = \sqrt{\frac{\pi \ln 2}{2}}$ and $k = \sqrt{\frac{\pi}{8}}$. Our "BALD" baseline method selects the example $x_n \in \mathcal{U}_n$ that maximizes the expression in (9), computed in $O(d^2)$ time per candidate example.

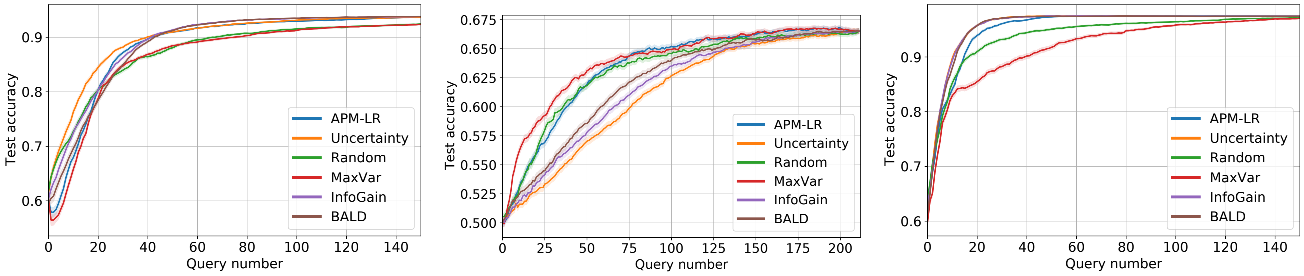
Summary For completeness, below we summarize all selection methods used in our experiments. For any method utilizing a normal approximation to the hyperplane posterior, let $p_{\theta|\mathcal{L}_{n-1}} \sim \mathcal{N}(\mu_n, \Sigma_n)$. Let $\widehat{\theta}_{n-1} = \mathrm{A}(\mathcal{L}_{n-1})$, $D = \sqrt{\frac{\pi \ln 2}{2}}$, and $k = \sqrt{\frac{\pi}{8}}$.

$$\begin{aligned}
\textit{APM-LR}:&\quad x_n = \operatorname*{arg\,min}_{x \in \mathcal{U}_n}\ (\mu_n^T x)^2 + \left(\sqrt{x^T \Sigma_n x} - \sqrt{\frac{2}{\pi} P_n}\right)^2 \qquad (10) \\
\textit{Uncertainty}:&\quad x_n = \operatorname*{arg\,min}_{x \in \mathcal{U}_n}\ x^T \widehat{\theta}_{n-1} \\
\textit{Random}:&\quad \text{Select } x_n \text{ uniformly at random from } \mathcal{U}_n \\
\textit{MaxVar}:&\quad x_n = \operatorname*{arg\,max}_{x \in \mathcal{U}_n}\ x^T \Sigma_n x \\
\textit{InfoGain}:&\quad x_n = \operatorname*{arg\,max}_{x \in \mathcal{U}_n}\ h_b\left(\frac{1}{s}\sum_{i=1}^{s} f(\theta_i^T x_n)\right) - \frac{1}{s}\sum_{i=1}^{s} h_b\Big(f(\theta_i^T x_n)\Big) \qquad \theta_i \sim \mathcal{N}(\mu_n, \Sigma_n) \\
\textit{BALD}:&\quad x_n = \operatorname*{arg\,max}_{x \in \mathcal{U}_n}\ h_b\left(\Phi\left(\frac{k\mu_n^T x_n}{\sqrt{k^2 x_n^T \Sigma_n x_n + 1}}\right)\right) - \frac{D\exp\left(-\frac{k^2(\mu_n^T x_n)^2}{2(k^2 x_n^T \Sigma_n x_n + D^2)}\right)}{\sqrt{k^2 x_n^T \Sigma_n x_n + D^2}}
\end{aligned}$$

B.4 Extended Test Accuracy Results

Below we plot average holdout test accuracy against number of queried examples, excluding one initial seed point selected uniformly at random per class. Error bars show ±1 standard error over 150 trials per method. For visual clarity, we display different numbers of queried examples for each dataset.

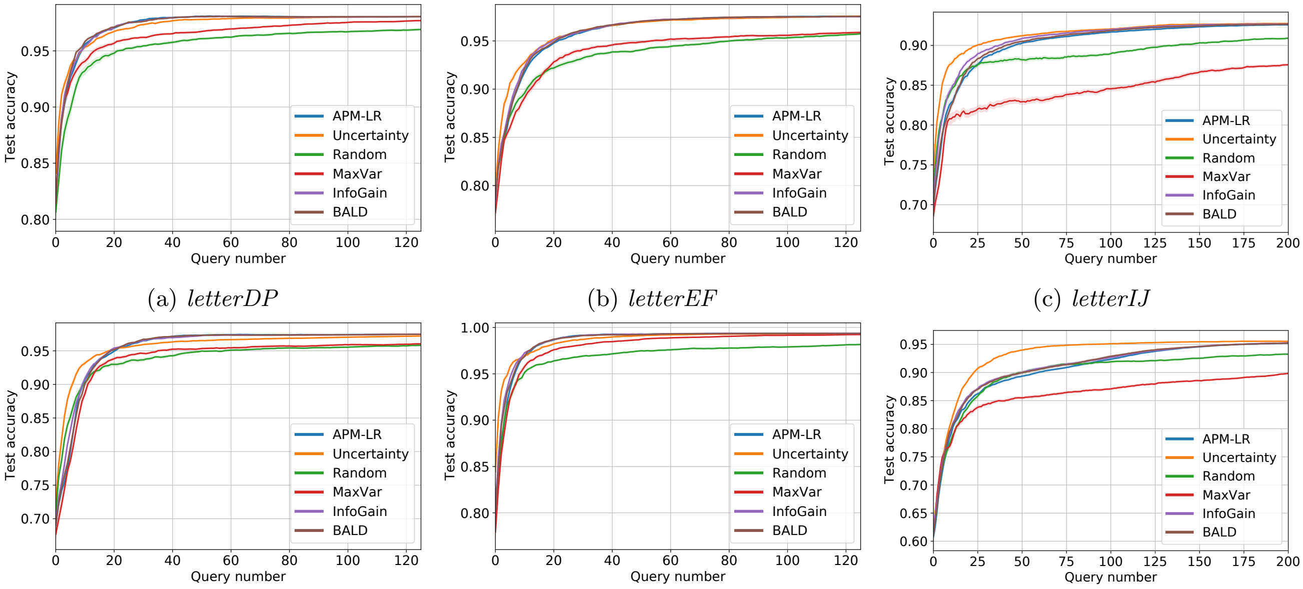
Figure 1 shows test accuracy across several two-class partitions of the Vehicle Silhouettes dataset (see Table 1). In *vehicle-cars*, Uncertainty, InfoGain, and BALD fail to perform as well as MaxVar, Random, and APM-LR. As noted in Yang and Loog (2018), there are cases where Random sampling — or more generally, selection methods that encourage dataset exploration — can outperform methods that maximize information. In *vehicle-cars*, it's possible that the "exploration" component in APM-LR encourages the selection of satisfactory examples, which we investigate further in Section B.6.



(a) *vehicle-full* (b) *vehicle-cars* (c) *vehicle-transport*

Figure 1: Test accuracy on "Vehicle Silhouettes"

Figure 2 shows test accuracy across several two-class partitions of the Letter Recognition dataset. All partitions show similar trends to *letterDP*, which was included in the paper body.

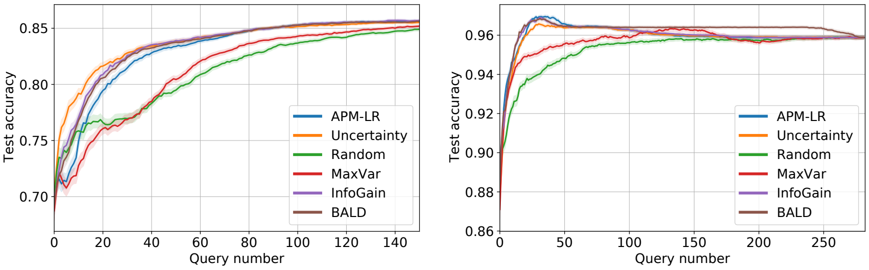


(a) *letterDP* (b) *letterEF* (c) *letterIJ*

(d) *letterMN* (e) *letterUV* (f) *letterVY*

Figure 2: Test accuracy on "Letter Recognition"

Figure 3 shows test accuracy across the remaining UCI datasets in Table 1. On *wdbc*, the active methods appear to have an average test accuracy that peaks early and then gradually decreases. While this behavior merits further investigation, we note that it is possible in some cases for a selected subset of the full data pool to generalize better than when training on the entire pool (Ma et al., 2018).

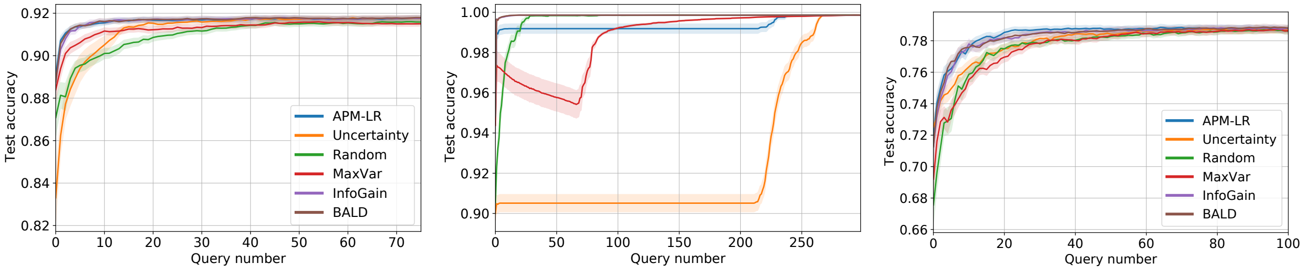


(a) *austra*

(b) *wdbc*

Figure 3: Miscellaneous UCI datasets

Figure 4 shows test accuracy across several synthetic datasets. On *clouds* and *cross*, Uncertainty sampling is outperformed by the other baseline active learning methods, except MaxVar.



(a) *clouds*

(b) *cross*

(c) *horseshoe*

Figure 4: Synthetic datasets

B.5 Extended Computational Cost Results

All experiments were run on Intel Xeon Gold 6226 CPUs at 2.7 GHz. In Table 2 we present for all datasets the cumulative compute time (in seconds) needed for each method to select the first 40 examples (excluding seed points). In this first table, we exclude the compute time needed to retrain the logistic regression model and perform the VariationalEM posterior update after each example is selected, since these steps are common to all selection methods. While some methods do not directly utilize the variational posterior in selecting examples, we perform variational posterior updates for all data selection methods since we consider the variational posterior to be part of the Bayesian model produced by the training routine.

	APM-LR	Uncertainty	BALD	InfoGain	Random	MaxVar
vehicle-full	0.173	0.077	2.212	7.166	0.003	0.061
vehicle-cars	0.089	0.039	1.078	3.462	0.002	0.030
vehicle-transport	0.087	0.037	1.036	3.306	0.002	0.029
letterDP	0.336	0.149	4.230	12.755	0.005	0.118
letterEF	0.318	0.143	4.044	12.188	0.005	0.113
letterIJ	0.314	0.139	3.941	11.879	0.004	0.110
letterMN	0.331	0.147	4.170	12.531	0.005	0.117
letterUV	0.330	0.145	4.129	12.429	0.005	0.115
letterVY	0.318	0.143	4.063	12.284	0.004	0.114
austra	0.150	0.063	1.770	5.089	0.003	0.050
wdbc	0.125	0.052	1.480	6.415	0.003	0.042
clouds	0.119	0.053	1.522	2.735	0.002	0.041
cross	0.125	0.053	1.521	2.722	0.002	0.040
horseshoe	0.116	0.053	1.517	2.731	0.002	0.040

Table 2: **Cumulative selection time:** comparison of median cumulative time (s) for each method to select the first 40 examples (excluding seed points).

Table 3 isolates the compute time needed for performing VariationalEM at each input, summed over the first 40 examples. Interestingly, methods which are primarily focused on data space exploration (MaxVar, Random) require more time for variational posterior updating than exploitation methods (Uncertainty). Since VariationalEM is an iterative procedure that we run with an adaptive stopping rule (with convergence defined as the relative variational parameter difference falling below 1e$-$6 between iterations), it presumably requires more iterations to adjust to significant changes in the posterior distribution due to variability in examples. Although less accurate of an approximation than VariationalEM, using a Laplace posterior approximation instead would have a constant update time per method (Jaakkola and Jordan, 2000).

	APM-LR	Uncertainty	BALD	InfoGain	Random	MaxVar
vehicle-full	10.088	4.540	10.118	9.729	7.469	18.064
vehicle-cars	5.420	4.412	5.605	5.475	3.280	4.558
vehicle-transport	9.609	5.814	9.289	9.083	11.216	21.058
letterDP	7.618	6.412	6.904	6.758	10.694	11.851
letterEF	6.866	5.701	6.320	6.160	11.302	10.755
letterIJ	7.367	5.724	6.924	6.708	10.019	9.846
letterMN	8.190	6.281	7.615	7.375	10.082	13.236
letterUV	8.029	6.556	7.137	7.075	10.746	12.585
letterVY	7.463	5.760	7.142	6.910	8.234	9.975
austra	12.513	6.451	12.009	11.645	8.541	13.580
wdbc	17.966	10.880	14.183	13.874	20.763	29.778
clouds	1.221	1.172	1.156	1.322	3.201	5.318
cross	1.386	2.417	1.474	1.537	3.138	4.453
horseshoe	0.996	0.908	0.863	0.931	0.802	1.208

Table 3: **Cumulative VariationalEM time:** comparison of median cumulative time (s) for each method to perform VariationalEM over the first 40 examples (excluding seed points).

Table 4 depicts the total compute time needed for selecting each example, performing VariationalEM, and retraining the logistic regression classifier at each iteration, summed over the first 40 examples. The median time needed for retraining the logistic regression classifier lies within 0.01 to 0.03 seconds across all methods and datasets, and therefore contributes only marginally to the total. While the spread of running times is more narrow than it would be when only evaluating selection time, the same general trend holds that InfoGain is more expensive than BALD and APM-LR.

	APM-LR	Uncertainty	BALD	InfoGain	Random	MaxVar
vehicle-full	10.288	4.637	12.365	16.943	7.493	18.148
vehicle-cars	5.532	4.474	6.727	8.980	3.306	4.616
vehicle-transport	9.721	5.876	10.341	12.419	11.238	21.116
letterDP	7.992	6.583	11.139	19.534	10.730	11.995
letterEF	7.215	5.868	10.396	18.414	11.330	10.887
letterIJ	7.716	5.892	10.896	18.619	10.048	9.981
letterMN	8.561	6.455	11.813	19.991	10.124	13.374
letterUV	8.399	6.724	11.294	19.552	10.781	12.724
letterVY	7.802	5.931	11.233	19.233	8.260	10.118
austra	12.690	6.538	13.801	16.804	8.574	13.655
wdbc	18.130	10.968	15.711	20.323	20.787	29.842
clouds	1.358	1.241	2.706	4.122	3.224	5.385
cross	1.534	2.490	3.028	4.291	3.159	4.515
horseshoe	1.134	0.978	2.405	3.741	0.819	1.264

Table 4: **Cumulative running time:** comparison of median cumulative run time (s) for each method to select each example, perform VariationalEM, and retrain the logistic regression classifier over the first 40 examples (excluding seed points).

B.6 Failure Mode Analysis

While in many cases APM-LR performs comparably to InfoGain, BALD, and Uncertainty while outperforming Random and MaxVar, the main exception in our experiments is on *vehicle-cars* (Figure 1b), where APM-LR, Random, and MaxVar outperform InfoGain, BALD, and Uncertainty. Conceptually, what differentiates these two classes of methods is that APM-LR, Random, and MaxVar have explicit exploration components to their selection policies, while InfoGain, BALD, and Uncertainty only seek to directly maximize information or uncertainty. As we will demonstrate below, on *vehicle-cars* this difference in exploration correlates with significant differences in generalization performance.

To isolate the effect of each term in APM-LR (eq. (10)) — corresponding to exploitation and exploration — we simulated two pseudo-APM policies where only one of the terms is active at once. In *APM-LR-U*, examples are selected that minimize the first term, which has an action similar to uncertainty sampling:

$$\textit{APM-LR-U}: \quad x_n = \operatorname*{arg\,min}_{x \in \mathcal{U}_n} \; (\mu_n^T x)^2.$$

In *APM-LR-V*, examples are selected that minimize the second term, which prefers examples that probe in directions of high posterior variance:

$$\textit{APM-LR-V}: \quad x_n = \operatorname*{arg\,min}_{x \in \mathcal{U}_n} \; \left(\sqrt{x^T \Sigma_n x} - \sqrt{\frac{2}{\pi} P_n}\right)^2.$$

We start in Figure 5 by plotting generalization performance as in Figure 1b, with the addition of APM-LR-U and APM-LR-V. In all plots below, error bars are removed for visual clarity, and the query horizon spans the entire training sequence (until the training pool is exhausted). As expected, APM-LR-V performs comparably to MaxVar, since both methods prefer examples that probe in directions of large posterior variance. Similarly, APM-LR-U performs comparably to Uncertainty, since both methods minimize distance to a hyperplane estimate (the former using the posterior mean hyperplane, the latter using a MAP estimate). These results support the hypothesis that it is the exploration component of APM-LR which leads to improved performance on *vehicle-cars* over non-exploration methods, including its own exploitation variant APM-LR-U.

We can explore this hypothesis further by directly evaluating metrics for exploitation and exploration of each method. To measure exploitation, in Figure 6, we plot the average distance from each selected example to the

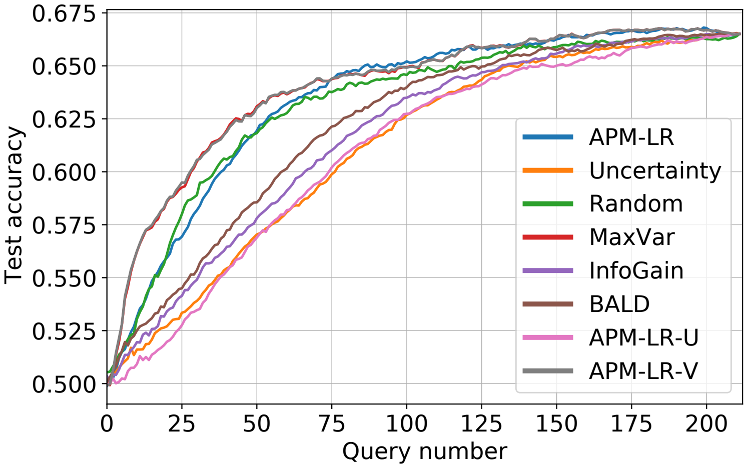


Figure 5: Test accuracy on *vehicle-cars*, over expanded method set.

MAP hyperplane estimate. Since distance from the classifier hyperplane directly corresponds to label uncertainty in logistic regression, this distance is a direct measure of how often a policy selects uncertain examples. By definition, Uncertainty begins by querying examples that are closest to the hyperplane estimate, maximally exploiting the estimate to query examples with the highest model uncertainty. The remaining methods vary in their levels of initial distance from the hyperplane estimate, but all eventually query close to their respective estimates, either by design or due to exhausting the full training pool. Notably, the level of initial distance from the hyperplane corresponds almost exactly to test accuracy performance: high-performing MaxVar and APM-LR-V initially query far from their hyperplane estimates, while the poorly performing Uncertainty queries examples close by.

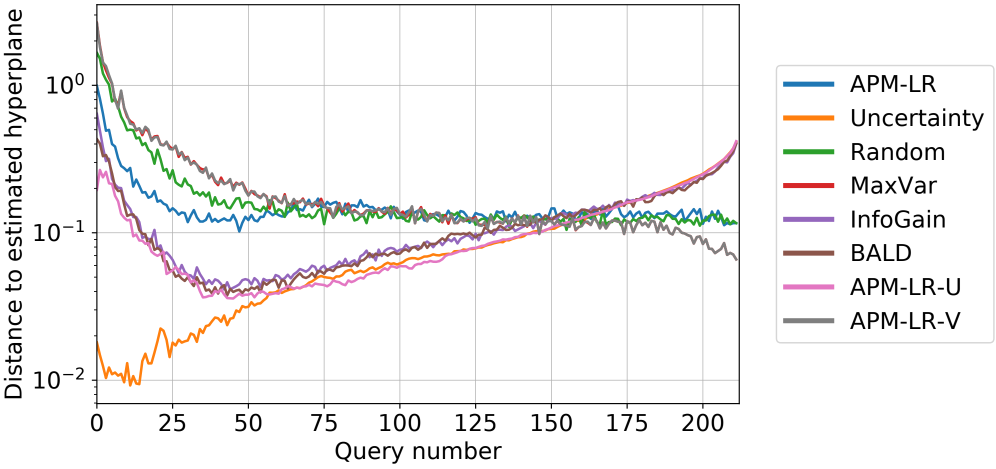
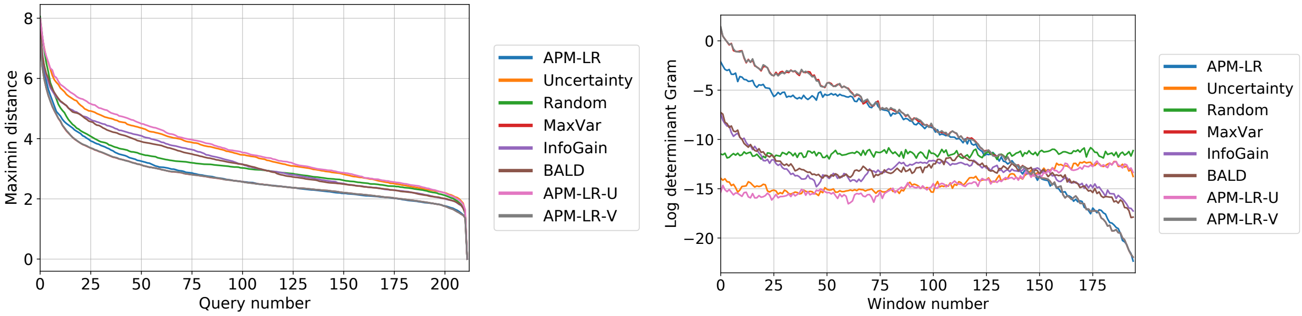


Figure 6: Exploitation metric for *vehicle-cars*: average distance of selected example to estimated hyperplane. Small distances reflect high levels of policy exploitation since this reflects examples being queried that are uncertain with respect to the current hyperplane estimate.

To measure policy exploration, we use two metrics and plot their average values in Figure 7. In the first metric, we measure the Euclidean distance from each unlabeled example to its nearest labeled neighbor, and take the maximum such distance over all unlabeled examples. This quantity measures the worst-case level of isolation of an unlabeled point to its nearest labeled neighbor, with lower values corresponding to higher degrees of policy exploration. A similar quantity is involved in the construction of coresets for active learning to promote diversity among selected examples (Sener and Savarese, 2018). As our second metric, we consider windows of d examples (recall that d denotes the data space dimension) and plot the log determinant of the Gram matrix of the examples selected in each window, which can be used as a measure of example diversity (higher values correspond to higher levels of example diversity) (Ash et al., 2020). In Figure 7a, MaxVar, APM-LR-V, APM-LR, and Random have the lowest average maximin distances, corresponding to lower levels of isolated unlabeled examples. Similarly, these methods generally have large initial Gram matrix log determinants, as depicted in Figure 7b.



(a)

(b)

Figure 7: Exploration metrics for *vehicle-cars*: (a) maximum distance from an unlabeled example to its closest labeled example. Smaller values indicate lower levels of unlabeled data isolation, and correspond to higher levels of exploration. (b) Log determinant of Gram matrix, where larger values correspond to higher levels of exploration.

The ablation of individual terms in APM-LR and direct measurement of exploitation and exploration of each active learning method suggests that when tested on *vehicle-cars*, exploration-based methods outperform methods that do not explicitly optimize for diverse selection. While this extended analysis is limited to a single dataset, it provides evidence that the exploration term in APM-LR can lead to higher levels of performance on a real-world dataset, where methods that do not directly account for exploration might fail.

References

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. (2020). Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, NY. Softcover published in 2016.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian active learning for classification and preference learning.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Lovász, L. and Vempala, S. (2007). The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358.

Ma, Y., Nowak, R., Rigollet, P., Zhang, X., and Zhu, X. (2018). Teacher improves learning by selecting a training subset.

Mérigot, Q. (2011). A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592.

Saumard, A. and Wellner, J. A. (2014). Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45–114. 27134693[pmid].

Sener, O. and Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Siebert, J. (1987). Vehicle recognition using rule based methods. Project report, Turing Institute, Glasgow.

Singh, J., Dabeer, O., and Madhow, U. (2009). On the limits of communication with low-precision analog-to-digital conversion at the receiver. *IEEE Transactions on Communications*, 57(12):3629–3639.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Winkelbauer, A. (2014). Moments and absolute moments of the normal distribution.

Yang, Y. and Loog, M. (2018). A benchmark and comparison of active learning for logistic regression. *Pattern Recognition*, 83:401 – 415.