Supplementary Appendix to Optimal sampling in unbiased active learning

Henrik Imberg **Johan Jonasson** **Marina Axelson-Fisk**

Department of Mathematical Sciences,
Chalmers University of Technology and University of Gothenburg,
SE-412 96 Gothenburg, Sweden.

A Lemmas

We present in this section two lemmas that are needed for our main results. Proofs are provided in Appendix B.

Lemma 1
Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ *and consider the function*

$$f(\boldsymbol{\pi}) = \sum_{i=1}^{N} \frac{\omega_i}{\pi_i}\,, \quad \omega_i > 0\,,$$

subject to the constraints

$$\sum_{i=1}^{N} \pi_i = 1\,,$$
$$\pi_i \in (0,1)\,, \quad i = 1, \ldots, N\,.$$

Then, the function $f(\boldsymbol{\pi})$ *is minimised by choosing* π_i *according to*

$$\pi_i = \frac{\sqrt{\omega_i}}{\sum_{j=1}^{N} \sqrt{\omega_j}}\,, \quad i = 1, \ldots, N\,.$$

Lemma 2
Under assumptions A1 - A3 in Section 3, it holds that

$$\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) = \boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0))\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1} + o(t^{-1})\,,$$

where for vector-valued random variables $o(\cdot)$ *is interpreted elementwise as its scalar analogue.*

B Proofs

B.1 Proof of Lemma 1

Using the method of Lagrange multipliers, we introduce the auxiliary function

$$\Lambda(\boldsymbol{\pi}, \lambda) = f(\boldsymbol{\pi}) + \lambda h(\boldsymbol{\pi})\,, \quad h(\boldsymbol{\pi}) = \sum_{i=1}^{N} \pi_i - 1\,.$$

Critical points of the Lagrangian are found by solving the equation system

$$\nabla\Lambda(\boldsymbol{\pi}, \lambda) = \mathbf{0} \quad \Leftrightarrow \quad \begin{cases} h(\boldsymbol{\pi}) = 0 \\ -\nabla_{\boldsymbol{\pi}} f(\boldsymbol{\pi}) = \lambda \nabla_{\boldsymbol{\pi}} h(\boldsymbol{\pi}) \end{cases}\,.$$

Since $\frac{\partial f(\boldsymbol{\pi})}{\partial \pi_i} = -\omega_i/\pi_i^2$ and $\frac{\partial h(\boldsymbol{\pi})}{\partial \pi_i} = 1$, this implies that $\lambda = \omega_1/\pi_1^2 = \ldots = \omega_N/\pi_N^2$, and further that

$$\pi_i \propto \sqrt{\omega_i}\,.$$

By the constraints $\pi_i > 0$ and $\sum_{i=1}^N \pi_i = 1$, we obtain

$$\pi_i = \frac{\sqrt{\omega_i}}{\sum_{j=1}^N \sqrt{\omega_j}}\,. \tag{S.1}$$

Thus, the point $(\boldsymbol{\pi}^*, \lambda^*)$ with entries π_i^* defined according to (S.1) and $\lambda^* = \omega_1/\pi_1^{*2}$ is a stationary point of $\Lambda(\boldsymbol{\pi}, \lambda)$. Hence, $\boldsymbol{\pi}^*$ is a stationary point of $f(\boldsymbol{\pi})$ under the specified constraints. Furthermore, the Hessian of $f(\boldsymbol{\pi})$ is positive definite on the domain specified by $\pi_i \in (0, 1)$, so $\boldsymbol{\pi}^*$ is a local minimum. By convexity, this implies that $\boldsymbol{\pi}^*$ is the global minimum of $f(\boldsymbol{\pi})$ under the specified constraints. □

B.2 Proof of Lemma 2

The result follows from Binder (1983). A sketch of the proof is provided here for completeness.

We will need the following error bounds, deduced from assumptions A2 and A3:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0 &= \mathcal{O}_p(t^{-1/2})\,, \\ \frac{1}{N}\left(\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0) - \boldsymbol{S}(\boldsymbol{\theta}_0)\right) &= \mathcal{O}_p(t^{-1/2})\,, \\ \frac{1}{N}\left(\hat{\boldsymbol{H}}(\boldsymbol{\theta}_0) - \boldsymbol{H}(\boldsymbol{\theta}_0)\right) &= o_p(1)\,, \end{aligned} \tag{S.2}$$

where for vector and matrix-valued random variables $\mathcal{O}_p(\cdot)$ and $o_p(\cdot)$ are interpreted elementwise as their scalar analogues. These error bounds in turn imply that

$$o_p(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) = o_p(\mathcal{O}_p(t^{-1/2})) = o_p(t^{-1/2})\,, \tag{S.3}$$

$$\frac{1}{N}\hat{\boldsymbol{H}}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) = \frac{1}{N}\boldsymbol{H}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) + o_p(t^{-1/2})\,. \tag{S.4}$$

We point out that the limiting procedure is such that $N \to \infty$, $t \to \infty$ and $N - t \to \infty$, and that we implicitly assume the existence of a sequence of (hypothetical) populations $\mathcal{P}_1, \mathcal{P}_2, \ldots$ of increasing sizes $N_1, N_2, \ldots$ such that A2 and A3 holds. We do, however, omit the dependence on this sequence from the notation.

Under the assumed regularity conditions, we first note that $\hat{\boldsymbol{\theta}}_t$ may be defined as the solution to the estimating equation $\hat{\boldsymbol{S}}(\boldsymbol{\theta}) = \boldsymbol{0}$, where $\hat{\boldsymbol{S}}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}}\hat{\ell}_t(\boldsymbol{\theta})$. A Taylor expansion of $\hat{\boldsymbol{S}}(\boldsymbol{\theta})$ in a neighbourhood of the optimal parameter $\boldsymbol{\theta}_0$ gives

$$\frac{1}{N}\hat{\boldsymbol{S}}(\boldsymbol{\theta}) = \frac{1}{N}\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0) + \frac{1}{N}\hat{\boldsymbol{H}}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + o(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\,.$$

Taking $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_t$ and using that $\hat{\boldsymbol{S}}(\hat{\boldsymbol{\theta}}_t) = \boldsymbol{0}$, we obtain by (S.3) and (S.4) that

$$\begin{aligned} \boldsymbol{0} = \frac{1}{N}\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0) + \frac{1}{N}\hat{\boldsymbol{H}}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) + o_p(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) &\Leftrightarrow -\frac{1}{N}\hat{\boldsymbol{H}}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) = \frac{1}{N}\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0) + o_p(t^{-1/2}) \\ &\Leftrightarrow -\frac{1}{N}\boldsymbol{H}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) = \frac{1}{N}\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0) + o_p(t^{-1/2})\,. \end{aligned}$$

Taking variances and using (S.2), we get

$$\frac{\boldsymbol{H}(\boldsymbol{\theta}_0)}{N}\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0)\frac{\boldsymbol{H}(\boldsymbol{\theta}_0)}{N} = \frac{\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0))}{N^2} + o(t^{-1})\,,$$

where we have used that $\boldsymbol{H}(\boldsymbol{\theta}_0) = \boldsymbol{H}(\boldsymbol{\theta}_0)^T$ since the Hessian is a symmetric matrix. The desired result is now obtained by a final rearrangement of the terms:

$$\begin{aligned} \mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) &= \left(N\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\right)\frac{\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0))}{N^2}\left(N\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\right) + o(t^{-1}) \\ &= \boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0))\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1} + o(t^{-1})\,. \end{aligned}$$

□

B.3 Proof of Proposition 1

Using multinomial sampling, i.e. for $\boldsymbol{Q}_t := (Q_{t,1}, \ldots, Q_{t,N}) \sim \text{Multinomial}(1, \boldsymbol{\pi}_t)$, we have that

$$\begin{aligned}\text{Var}_{\boldsymbol{\pi}}(Q_{t,i}) &= \pi_{t,i}(1-\pi_{t,i})\,,\\ \text{Cov}_{\boldsymbol{\pi}}(Q_{t,i}, Q_{t,j}) &= -\pi_{t,i}\pi_{t,j}\,.\end{aligned}$$

Taking $\boldsymbol{y} := (y_1, \ldots, y_N)$ as fixed, we therefore have that

$$\begin{aligned}\text{Var}_{\boldsymbol{\pi}}\left(\frac{1}{t}\sum_{i\in\mathcal{P}}\frac{Q_{t,i}}{\pi_{t,i}}\ell(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})\right) &= \frac{1}{t^2}\left[\sum_{i\in\mathcal{P}}\frac{\pi_{t,i}(1-\pi_{t,i})}{\pi_{t,i}^2}\ell_i^2 - \sum_{\substack{i,j\in\mathcal{P}\\ i\neq j}}\frac{\pi_{t,i}\pi_{t,j}}{\pi_{t,i}\pi_{t,j}}\ell_i\ell_j\right]\\ &= \frac{1}{t^2}\left[\sum_{i\in\mathcal{P}}\frac{\ell(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})^2}{\pi_{t,i}} - \sum_{i,j\in\mathcal{P}}\ell(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})\ell(y_j, \boldsymbol{x}_j, \boldsymbol{\theta})\right]\\ &= \frac{1}{t^2}\sum_{i\in\mathcal{P}}\frac{\ell(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})^2}{\pi_{t,i}} + k\,,\end{aligned}$$

where k is a constant not depending on $\boldsymbol{\pi}_t$. The anticipated variance (5) is thus given by

$$\text{E}_{\boldsymbol{\theta}}\left[\text{Var}_{\boldsymbol{\pi}}\left(\frac{1}{t}\sum_{i\in\mathcal{P}}\frac{Q_{t,i}}{\pi_{t,i}}\ell(Y_i^*, \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}})\big|\boldsymbol{Y}^*\right)\right] = \frac{1}{t^2}\sum_{i\in\mathcal{P}}\frac{\text{E}_{\boldsymbol{\theta}}[\ell(Y_i^*, \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}})^2]}{\pi_{t,i}} + k_2\,,$$

where the outer expectation is taken respect to $\{Y_i^*\}_{i\in\mathcal{P}}$ under the model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ and k_2 is a constant not depending on $\boldsymbol{\pi}_t$. The desired result is now obtained by application of Lemma 1. □

B.4 Proof of Proposition 2

First, a second order Taylor expansion of the total loss in a neighbourhood of the optimal parameter $\boldsymbol{\theta}_0$ gives

$$\ell_0(\boldsymbol{\theta}) = \ell_0(\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}}\ell_0(\boldsymbol{\theta})^T\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathbf{0}}) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathbf{0}})^T\boldsymbol{H}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathbf{0}}) + o(||\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathbf{0}}||^2)\,. \tag{S.5}$$

We next note that $\boldsymbol{\theta}_0$, under the assumed regularity conditions, may be defined as the solution to the estimating equation $\nabla_{\boldsymbol{\theta}}\ell_0(\boldsymbol{\theta}) = \mathbf{0}$, implying that the second term in (S.5) vanishes. Taking $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_t$, we thus get

$$\text{E}_{\boldsymbol{\pi}}\left[\frac{1}{N}\ell_0(\hat{\boldsymbol{\theta}}_t)\right] = \frac{1}{N}\ell_0(\boldsymbol{\theta}_0) + \frac{1}{2N}\text{E}_{\boldsymbol{\pi}}\left[(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_{\mathbf{0}})^T\boldsymbol{H}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_{\mathbf{0}})\right] + o(t^{-1})\,, \tag{S.6}$$

using the fact that $o_p(||\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_{\mathbf{0}}||) = o_p(t^{-1/2})$, as implied by assumption A2. By properties on quadratic forms in random variables (Mathai, 1992), we may next write

$$\begin{aligned}\frac{1}{N}\text{E}_{\boldsymbol{\pi}}\left[(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_{\mathbf{0}})^T\boldsymbol{H}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_{\mathbf{0}})\right] &= \frac{1}{N}\,\text{tr}\left(\boldsymbol{H}(\boldsymbol{\theta}_0)\text{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0)\right) + \frac{1}{N}\text{E}_{\boldsymbol{\pi}}\left[(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0)^T\right]H(\boldsymbol{\theta}_0)\text{E}_{\boldsymbol{\pi}}\left[(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0)\right]\\ &= \frac{1}{N}\,\text{tr}\left(\boldsymbol{H}(\boldsymbol{\theta}_0)\text{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0)\right) + o(t^{-1})\,,\end{aligned} \tag{S.7}$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix, and the second equality follows from assumption A2. By Lemma 2, we also have that

$$\text{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0) = \boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\text{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0))\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1} + o(t^{-1})\,,$$

where, in analogy with the proof of Proposition 1 (Appendix B.3), the variance-covariance matrix of $\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0)$ is given by

$$\text{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0)) = \frac{1}{t^2}\sum_{s=1}^{t}\sum_{i=1}^{N}\frac{\boldsymbol{s}_i(\boldsymbol{\theta}_0)\boldsymbol{s}_i(\boldsymbol{\theta}_0)^T}{\pi_{s,i}} + K\,,$$

and K is a constant matrix not depending on $\boldsymbol{\pi}_{1:t}$. Using the linearity and cyclic properties of the trace, we may now write

$$\begin{aligned}\frac{1}{N}\operatorname{tr}\left(\boldsymbol{H}(\boldsymbol{\theta}_0)\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0)\right) &= \operatorname{tr}\left(\frac{1}{N}\boldsymbol{H}(\boldsymbol{\theta}_0)\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\left(\frac{1}{t^2}\sum_{s=1}^{t}\sum_{i=1}^{N}\frac{\boldsymbol{s}_i(\boldsymbol{\theta}_0)\boldsymbol{s}_i(\boldsymbol{\theta}_0)^T}{\pi_{s,i}}+K\right)\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}+o(t^{-1})\right)\\ &= \operatorname{tr}\left(\frac{1}{Nt^2}\sum_{s=1}^{t}\sum_{i=1}^{N}\frac{\boldsymbol{s}_i(\boldsymbol{\theta}_0)\boldsymbol{s}_i(\boldsymbol{\theta}_0)^T}{\pi_{s,i}}\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}+K_2+o(t^{-1})\right)\\ &= \operatorname{tr}\left(\frac{1}{Nt^2}\sum_{s=1}^{t}\sum_{i=1}^{N}\frac{\boldsymbol{s}_i(\boldsymbol{\theta}_0)^T\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{s}_i(\boldsymbol{\theta}_0)}{\pi_{s,i}}\right)+k_3+o(t^{-1})\\ &= \frac{1}{Nt^2}\sum_{s=1}^{t}\sum_{i=1}^{N}\frac{\boldsymbol{s}_i(\boldsymbol{\theta}_0)^T\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{s}_i(\boldsymbol{\theta}_0)}{\pi_{s,i}}+k_3+o(t^{-1})\,, \end{aligned}\tag{S.8}$$

where k_3 is a constant not depending on $\boldsymbol{\pi}_{1:t}$. Combining (S.6) - (S.8), we obtain the desired result:

$$\mathrm{E}_{\boldsymbol{\pi}}\left[\frac{1}{N}\ell_0(\hat{\boldsymbol{\theta}}_t)\right]=\frac{1}{N}\ell_0(\boldsymbol{\theta}_0)+\frac{1}{2Nt^2}\sum_{s=1}^{t}\sum_{i=1}^{N}\frac{\boldsymbol{s}_i(\boldsymbol{\theta}_0)^T\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{s}_i(\boldsymbol{\theta}_0)}{\pi_{s,i}}+k_4+o(t^{-1})\,,$$

where k_4 is a constant not depending on $\boldsymbol{\pi}_{1:t}$. This proves the first part of the theorem. The optimality of the sampling scheme (8) now follows by application of Lemma 1. □

B.5 Proof of Proposition 3

Under assumption A2 and by application of the Delta theorem (DasGupta, 2008), we have, for a differentiable function $g:\mathbb{R}^p\to\mathbb{R}$, that

$$\mathrm{Var}_{\boldsymbol{\pi}}(g(\hat{\boldsymbol{\theta}}_t)-g(\boldsymbol{\theta}_0))=\nabla_{\boldsymbol{\theta}}g(\boldsymbol{\theta})^T\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}}g(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}+o(t^{-1})\,,\tag{S.9}$$

provided that the first term of the right hand side is greater than zero. By Lemma 2, we also have that

$$\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0)=\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0))\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}+o(t^{-1})\,,\tag{S.10}$$

where, in analogy with the proof of Proposition 1 (Appendix B.3), the variance-covariance matrix of $\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0)$ is given by

$$\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{S}}(\boldsymbol{\theta}_0))=\frac{1}{t^2}\sum_{s=1}^{t}\sum_{i=1}^{N}\frac{\boldsymbol{s}_i(\boldsymbol{\theta}_0)\boldsymbol{s}_i(\boldsymbol{\theta}_0)^T}{\pi_{s,i}}+K\,,\tag{S.11}$$

and K is a constant matrix not depending on $\boldsymbol{\pi}_{1:t}$.

Next, a Taylor expansion of $\mu(\boldsymbol{x},\boldsymbol{\theta})$ in a neighbourhood of $\boldsymbol{\theta}_0$ gives

$$\mu(\boldsymbol{x},\boldsymbol{\theta})=\mu(\boldsymbol{x},\boldsymbol{\theta}_0)+\nabla_{\boldsymbol{\theta}}\mu(\boldsymbol{x},\boldsymbol{\theta})^T\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+o(||\boldsymbol{\theta}-\boldsymbol{\theta}_0||)\,,$$

Thus, for a single term $\mu(\boldsymbol{x},\hat{\boldsymbol{\theta}}_t)-\mu(\boldsymbol{x},\boldsymbol{\theta}_0)$, we have that

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\pi}}\left[\mu(\boldsymbol{x},\hat{\boldsymbol{\theta}}_t)-\mu(\boldsymbol{x},\boldsymbol{\theta}_0)\right] &= \mathrm{E}_{\boldsymbol{\pi}}\left[\mu(\boldsymbol{x},\boldsymbol{\theta}_0)+\nabla_{\boldsymbol{\theta}}\mu(\boldsymbol{x},\boldsymbol{\theta})^T\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0)+o_p(||\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0||)-\mu(\boldsymbol{x},\boldsymbol{\theta}_0)\right]\\ &= \mathrm{E}_{\boldsymbol{\pi}}\left[\nabla_{\boldsymbol{\theta}}\mu(\boldsymbol{x},\boldsymbol{\theta})^T\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0)+o_p(||\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0||)\right]\\ &= \nabla_{\boldsymbol{\theta}}\mu(\boldsymbol{x},\boldsymbol{\theta})^T\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\mathrm{E}_{\boldsymbol{\pi}}\left[(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0)\right]+\mathrm{E}_{\boldsymbol{\pi}}\left[o_p(||\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_0||)\right]\\ &= o(t^{-1/2})\,, \end{aligned}\tag{S.12}$$

where the last step follows from assumption A2.

Now, combining (S.12) with (S.9) - (S.11) and taking $g(\boldsymbol{\theta}) = \mu(\boldsymbol{x}, \boldsymbol{\theta})$, we get

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\pi}}\left[\Big(\mu(\boldsymbol{x},\hat{\boldsymbol{\theta}}_t) - \mu(\boldsymbol{x},\boldsymbol{\theta}_0)\Big)^2\right] &= \mathrm{Var}_{\boldsymbol{\pi}}\Big(\mu(\boldsymbol{x},\hat{\boldsymbol{\theta}}_t) - \mu(\boldsymbol{x},\boldsymbol{\theta}_0)\Big) + \Big(\mathrm{E}_{\boldsymbol{\pi}}\Big[\mu(\boldsymbol{x},\hat{\boldsymbol{\theta}}_t) - \mu(\boldsymbol{x},\boldsymbol{\theta}_0)\Big]\Big)^2 \\
&= \mathrm{Var}_{\boldsymbol{\pi}}\Big(\mu(\boldsymbol{x},\hat{\boldsymbol{\theta}}_t) - \mu(\boldsymbol{x},\boldsymbol{\theta}_0)\Big) + o(t^{-1}) \\
&= \nabla_{\boldsymbol{\theta}}\mu(\boldsymbol{x},\boldsymbol{\theta})^T\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\mathrm{Var}_{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}}\mu(\boldsymbol{x},\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o(t^{-1}) \\
&= ... \\
&= \frac{1}{t^2}\sum_{s=1}^{t}\sum_{i=1}^{N}\frac{d_i(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta}_0)}{\pi_s(i)} + k_2 + o(t^{-1})\,,
\end{aligned} \tag{S.13}$$

where

$$d_i(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta}_0) = \Big(\nabla_{\boldsymbol{\theta}}\mu(\boldsymbol{x},\boldsymbol{\theta})^T\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{s}_i(\boldsymbol{\theta}_0)\Big)^2$$

and k_2 is a constant not depending on $\boldsymbol{\pi}_{1:t}$. Considering all the predictions $\{\mu(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_t)\}_{i\in\mathcal{P}}$, the result in (S.13) generalises to

$$\mathrm{E}_{\boldsymbol{\pi}}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(\mu(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_t) - \mu(\boldsymbol{x}_i,\boldsymbol{\theta}_0)\Big)^2\right] = \frac{1}{Nt^2}\sum_{s=1}^{t}\sum_{i=1}^{N}\frac{d_i(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta}_0)}{\pi_s(i)} + k_3 + o(t^{-1})\,,$$

where $d_i(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta})$ now is given by

$$d_i(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta}) = ||\boldsymbol{M}(\boldsymbol{\theta}_0)\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{s}_i(\boldsymbol{\theta}_0)||^2\,,$$

and k_3 is a constant not depending on $\boldsymbol{\pi}_{1:t}$. This proves the first part of the theorem. The optimality of the sampling scheme (10) now follows by application of Lemma 1. □

B.6 Proof of Corollary 1

Before giving the proof of Corollary 1, we need the following preliminaries concerning the exponential family and generalised linear models. See McCullagh and Nelder (1989) for additional details.

Preliminaries

For a random variable Y_i in an exponential family with parameters (ζ_i, ϕ), we can write the density function on the form

$$f(y_i;\zeta_i,\phi) = \exp\left\{\frac{y_i\zeta_i - b(\zeta_i)}{\phi} + \gamma(y_i,\phi)\right\},$$

where ζ_i is the called the canonical parameter, ϕ is a dispersion parameter, and $b(\cdot)$ is a function satisfying

$$\begin{aligned}
b'(\zeta_i) &= \mathrm{E}[Y_i]\,, \\
b''(\zeta_i) &= \frac{1}{\phi}\mathrm{Var}(Y_i)\,.
\end{aligned} \tag{S.14}$$

In generalised linear models, the mean of the response variable Y_i given the predictors $\boldsymbol{x}_i$ is related to the parameter $\boldsymbol{\theta}$ through the relation $\mathrm{E}_{\boldsymbol{\theta}}[Y_i|\boldsymbol{x}_i] =: \mu(\boldsymbol{x}_i,\boldsymbol{\theta}) = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\theta})$ for some function $g(\cdot)$, referred to as the link function. Similarly, the canonical parameter ζ_i is related to the parameter $\boldsymbol{\theta}$ through $h(\zeta_i) = \boldsymbol{x}_i^T\boldsymbol{\theta}$ for some function $h(\cdot)$. Using the canonical link function, $h(\cdot)$ is simply the identity function and $\zeta_i = \boldsymbol{x}_i^T\boldsymbol{\theta}$.

Given independent observations $(\boldsymbol{x}_i, y_i)$, $i = 1,\ldots,N$, and using the canonical link function, we can write the log-likelihood of $\boldsymbol{\theta}$ as

$$\sum_{i=1}^{N}\frac{y_i\boldsymbol{x}_i^T\boldsymbol{\theta} - b(\boldsymbol{x}_i^T\boldsymbol{\theta})}{\phi} + \gamma(y_i,\phi)\,,$$

and the maximum likelihood estimator is defined as the maximiser of this function. However, for the purpose of estimating the parameter $\boldsymbol{\theta}$, we note that the dispersion parameter ϕ and the terms $\gamma(y_i, \phi)$ can be ignored, and we may equivalently define the maximum likelihood estimator of $\boldsymbol{\theta}$ as the minimiser of the loss

$$\ell_0(\boldsymbol{\theta}) = -\sum_{i=1}^{N} y_i \boldsymbol{x}_i^T \boldsymbol{\theta} - b(\boldsymbol{x}_i^T \boldsymbol{\theta}) \,.$$

The corresponding score vector and Hessian matrix are given by

$$\boldsymbol{S}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} \ell_0(\boldsymbol{\theta}) = -\sum_{i=1}^{N} y_i \boldsymbol{x}_i - b'(\boldsymbol{x}_i^T \boldsymbol{\theta}) \boldsymbol{x}_i = -\sum_{i=1}^{N} (y_i - \mu(\boldsymbol{x}_i, \boldsymbol{\theta})) \boldsymbol{x}_i \,, \tag{S.15}$$

$$\boldsymbol{H}(\boldsymbol{\theta}) = \sum_{i=1}^{N} b''(\boldsymbol{x}_i^T \boldsymbol{\theta}) \boldsymbol{x}_i \boldsymbol{x}_i^T = \frac{1}{\phi} \boldsymbol{X}^T \boldsymbol{V}(\boldsymbol{\theta}) \boldsymbol{X} \,,$$

where $\boldsymbol{V}(\boldsymbol{\theta})$ is the diagonal matrix with entries $\mathrm{Var}_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_i)$, $i = 1, \ldots, N$. Within this class of models, the statistical leverage score pertaining to instance i is given by the i-th diagonal element of the matrix

$$\boldsymbol{V}^{1/2} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{1/2} \,,$$

which we also may write as

$$\begin{aligned} h_{ii}(\boldsymbol{\theta}) &:= \left(\boldsymbol{V}^{1/2} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{1/2} \right)_{ii} \\ &= \mathrm{SD}_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_i) \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{x}_i \mathrm{SD}_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_i) \\ &\propto \mathrm{Var}_{\boldsymbol{\theta}}(Y_i|\boldsymbol{x}_i) \boldsymbol{x}_i^T \boldsymbol{H}(\boldsymbol{\theta})^{-1} \boldsymbol{x}_i \,, \end{aligned}$$

which for a linear regression model with constant error variance simplifies further to $h_{ii} = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i$.

We are now ready to present the proof of Corollary 1.

Proof of Corollary 1a

We first note that assumption A1 is immediately fulfilled for the class of the models under consideration. Assuming that A2 and A3 also hold we may apply the result of Proposition 2, stating that the expectation of the total loss $\ell_0(\hat{\boldsymbol{\theta}}_t)$ of the active learning algorithm with respect to the subsampling mechanism can be expressed as

$$\mathrm{E}_{\boldsymbol{\pi}} \left[\frac{1}{N} \ell_0(\hat{\boldsymbol{\theta}}_t) \right] = \frac{1}{N} \ell_0(\boldsymbol{\theta}_0) + \frac{1}{2Nt^2} \sum_{s=1}^{t} \sum_{i=1}^{N} \frac{c_i(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}_0)}{\pi_{s,i}} + k + o(t^{-1}) \,,$$

where $c_i(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}) = s_i(\boldsymbol{\theta})^T \boldsymbol{H}(\boldsymbol{\theta})^{-1} s_i(\boldsymbol{\theta})$, $\boldsymbol{H}(\boldsymbol{\theta})$ is the Hessian matrix of the total loss, $\boldsymbol{s}_i(\boldsymbol{\theta})$ the gradient of the individual loss $\ell_i(\boldsymbol{\theta})$, and k is a constant not depending on $\boldsymbol{\pi}_{1:t}$.

Taking $\boldsymbol{\theta}_0 = \boldsymbol{\theta}$, the anticipated generalisation error can thus be expressed as

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta}} \left[\mathrm{E}_{\boldsymbol{\pi}} \left[\frac{1}{N} \ell_0(\hat{\boldsymbol{\theta}}_t) \big| \boldsymbol{Y}^* \right] \right] &= \frac{1}{N} E_{\boldsymbol{\theta}} \left[\ell_0(\boldsymbol{\theta}_0) \right] + \mathrm{E}_{\boldsymbol{\theta}} \left[\frac{1}{2Nt^2} \sum_{s=1}^{t} \sum_{i=1}^{N} \frac{c_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})}{\pi_{s,i}} + k \right] + o(t^{-1}) \\ &= \frac{1}{N} E_{\boldsymbol{\theta}} \left[\ell_0(\boldsymbol{\theta}_0) \right] + \frac{1}{2Nt^2} \sum_{s=1}^{t} \sum_{i=1}^{N} \frac{\mathrm{E}_{\boldsymbol{\theta}} \left[c_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta}) \right]}{\pi_{s,i}} + k_2 + o(t^{-1}) \,, \end{aligned}$$

where the outer expectation is taken respect to $\{Y_i^*\}_{i \in \mathcal{P}}$ under the model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, and k_2 is a constant not depending on $\boldsymbol{\pi}_{1:t}$. By application of Lemma 1, this expectation is minimised by choosing

$$\pi_{s,i} \propto \sqrt{\mathrm{E}_{\boldsymbol{\theta}} \left[c_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta}) \right]}$$

for all $i \in \mathcal{P}$ and $s = 1, \ldots, t$, normalised so that $\sum_{i \in \mathcal{P}} \pi_{s,i} = 1$. What remains is to show that this is equivalent to choosing

$$\pi_{s,i} \propto \sqrt{\mathrm{Var}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i) \boldsymbol{x}_i^T \boldsymbol{H}^{-1} \boldsymbol{x}_i}$$

within the specified class of models, i.e. that

$$\mathrm{E}_{\boldsymbol{\theta}}\left[c_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})\right] \propto \mathrm{Var}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{x}_i^T\boldsymbol{H}^{-1}\boldsymbol{x}_i\,,$$

where $\boldsymbol{H} \propto \boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}$.

According to the preliminaries, we may, for a generalised linear model with canonical link function, write the total loss as $\ell_0(\boldsymbol{\theta}) = -\sum_{i=1}^N y_i\boldsymbol{x}_i^T\boldsymbol{\theta} - b(\boldsymbol{x}_i^T\boldsymbol{\theta})$, ignoring terms that do not involve the parameter $\boldsymbol{\theta}$, and the individual losses as $\ell(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = -(y_i\boldsymbol{x}_i^T\boldsymbol{\theta} - b(\boldsymbol{x}_i^T\boldsymbol{\theta}))$. We further have that the Hessian matrix of the total loss is given by $\boldsymbol{H}(\boldsymbol{\theta}) = \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{X}$, and, in analogy with (S.15), that the gradient of loss pertaining to instance i is given by

$$\boldsymbol{s}_i(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}}\ell(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = -(y_i - \mu_i)\boldsymbol{x}_i\,,$$

where $\mu_i := \mu(\boldsymbol{x}_i, \boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[Y_i^*|\boldsymbol{x}_i]$. Taking expectation with respect to Y_i^*, we thus have that

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}}[c_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})] : &= \mathrm{E}_{\boldsymbol{\theta}}\left[s_i(\boldsymbol{\theta})^T\boldsymbol{H}(\boldsymbol{\theta})^{-1}s_i(\boldsymbol{\theta})\right] \\ &= \mathrm{E}_{\boldsymbol{\theta}}[(Y_i^* - \mu_i)^2]\boldsymbol{x}_i^T\boldsymbol{H}^{-1}\boldsymbol{x}_i \\ &= \mathrm{Var}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{x}_i^T\boldsymbol{H}^{-1}\boldsymbol{x}_i \\ &\propto \mathrm{Var}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}_i\,. \end{aligned} \tag{S.16}$$

This gives the desired result and completes the first part of the proof. For a linear regression model with constant error variance, we further have that $\mathrm{Var}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i) =: \sigma^2$ does not depend on i. From this, it follows that $\boldsymbol{V} = \sigma^2\boldsymbol{I}_{N\times N}$, where $\boldsymbol{I}_{N\times N}$ is an $(N \times N)$ identity matrix, and that $\boldsymbol{H} \propto \boldsymbol{X}^T\boldsymbol{X}$. This now implies that

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{x}_i^T\boldsymbol{H}^{-1}\boldsymbol{x}_i &\propto \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i \\ &=: h_{ii}\,,\end{aligned}$$

which is known as the statistical leverage score for linear regression (Rawlings et al., 1998). □

Proof of Corollary 1b

We first note that assumptions A1 and A4 are immediately fulfilled for the class of the models under consideration. Assuming that A2 and A3 also hold we may apply the result of Proposition 3, stating that the mean squared error of the predictions $\{\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_t)\}$ can be expressed as

$$\mathrm{E}_{\boldsymbol{\pi}}\left[\frac{1}{N}\sum_{i=1}^N\left(\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_t) - \mu(\boldsymbol{x}_i, \boldsymbol{\theta}_0)\right)^2\right] = \frac{1}{Nt^2}\sum_{s=1}^t\sum_{i=1}^N\frac{d_i(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}_0)}{\pi_{s,i}} + k + o(t^{-1})\,,$$

where $d_i(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}) = ||\boldsymbol{M}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_i(\boldsymbol{\theta})||^2$, $\boldsymbol{H}(\boldsymbol{\theta})$ is the Hessian matrix of the total loss, $\boldsymbol{s}_i(\boldsymbol{\theta})$ the gradient of the individual loss $\ell_i(\boldsymbol{\theta})$, $\boldsymbol{M}(\boldsymbol{\theta})$ the matrix with rows $\nabla_{\boldsymbol{\theta}}\mu(\boldsymbol{x}_i, \boldsymbol{\theta})^T$, and k is a constant not depending on $\boldsymbol{\pi}_{1:t}$.

Taking $\boldsymbol{\theta}_0 = \boldsymbol{\theta}$, the anticipated mean squared error of the predictions $\{\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_t)\}$ can thus be expressed as

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}}\left[\mathrm{E}_{\boldsymbol{\pi}}\left[\frac{1}{N}\sum_{i=1}^N\left(\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_t) - \mu(\boldsymbol{x}_i, \boldsymbol{\theta})\right)^2\Big|\boldsymbol{Y}^*\right]\right] &= \mathrm{E}_{\boldsymbol{\theta}}\left[\frac{1}{Nt^2}\sum_{s=1}^t\sum_{i=1}^N\frac{d_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})}{\pi_{s,i}} + k\right] + o(t^{-1}) \\ &= \frac{1}{Nt^2}\sum_{s=1}^t\sum_{i=1}^N\frac{\mathrm{E}_{\boldsymbol{\theta}}\left[d_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})\right]}{\pi_{s,i}} + k_2 + o(t^{-1})\,,\end{aligned}$$

where the outer expectation is taken respect to $\{Y_i^*\}_{i\in\mathcal{P}}$ under the model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, and k_2 is a constant not depending on $\boldsymbol{\pi}_{1:t}$. By application of Lemma 1, the anticipated asymptotic mean squared error of the predictions $\{\mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_t)\}$ is thus minimised by choosing

$$\pi_{s,i} \propto \sqrt{\mathrm{E}_{\boldsymbol{\theta}}\left[d_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})\right]}$$

for all $i \in \mathcal{P}$ and $s = 1, \ldots, t$, normalised so that $\sum_{i\in\mathcal{P}}\pi_{s,i} = 1$. What remains is to show that this is equivalent to choosing

$$\pi_{s,i} \propto ||\mathrm{SD}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{V}\boldsymbol{X}\boldsymbol{H}^{-1}\boldsymbol{x}_i||$$

within the specified class of models, i.e. that

$$\mathrm{E}_{\boldsymbol{\theta}}\left[d_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})\right] \propto ||\mathrm{SD}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{V}\boldsymbol{X}\boldsymbol{H}^{-1}\boldsymbol{x}_i||^2 \, .$$

According to the preliminaries, we may, for a generalised linear model with canonical link function, write the total loss as $\ell_0(\boldsymbol{\theta}) = -\sum_{i=1}^N y_i\boldsymbol{x}_i^T\boldsymbol{\theta} - b(\boldsymbol{x}_i^T\boldsymbol{\theta})$, ignoring terms that do not involve the parameter $\boldsymbol{\theta}$, and the individual losses as $\ell(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = -(y_i\boldsymbol{x}_i^T\boldsymbol{\theta} - b(\boldsymbol{x}_i^T\boldsymbol{\theta}))$. In analogy (S.15), the gradient of the loss pertaining to instance i is given by

$$\boldsymbol{s}_i(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}}\ell(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = -(y_i - \mu_i)\boldsymbol{x}_i \, ,$$

where $\mu_i = \mu(\boldsymbol{x}_i, \boldsymbol{\theta}) := \mathrm{E}_{\boldsymbol{\theta}}[Y_i^*|\boldsymbol{x}_i]$. By (S.14), we also have that the gradient of the mean function $\mu(\boldsymbol{x}_i, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is given by

$$\nabla_{\boldsymbol{\theta}}\mu(\boldsymbol{x}_i, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}b'(\boldsymbol{x}_i^T\boldsymbol{\theta}) = b''(\boldsymbol{x}_i^T\boldsymbol{\theta})\boldsymbol{x}_i = \frac{1}{\phi}\mathrm{Var}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{x}_i \, ,$$

and consequently that $\boldsymbol{M}(\boldsymbol{\theta}) = \frac{1}{\phi}\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{X}$. The coefficients $d_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})$ can thus be written as

$$d_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta}) = ||\boldsymbol{M}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}\ell(Y_i^*, \boldsymbol{x}_i, \boldsymbol{\theta})||^2 = ||\frac{1}{\phi}\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{X}\boldsymbol{H}(\boldsymbol{\theta})^{-1}(Y_i^* - \mu_i)\boldsymbol{x}_i||^2 \, .$$

Taking expectation with respect to Y_i^*, we now get

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\theta}}[d_i(\boldsymbol{Y}^*, \boldsymbol{X}, \boldsymbol{\theta})] &= \mathrm{E}_{\boldsymbol{\theta}}\left[||\frac{1}{\phi}\boldsymbol{V}\boldsymbol{X}\boldsymbol{H}^{-1}(Y_i^* - \mu_i)\boldsymbol{x}_i||^2\right] \\
&= \frac{1}{\phi^2}\mathrm{E}_{\boldsymbol{\theta}}[(Y_i^* - \mu_i)^2]\boldsymbol{x}_i^T\boldsymbol{H}^{-1}\boldsymbol{X}^T\boldsymbol{V}^2\boldsymbol{X}\boldsymbol{H}^{-1}\boldsymbol{x}_i \\
&= \frac{1}{\phi^2}\mathrm{Var}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{x}_i^T\boldsymbol{H}^{-1}\boldsymbol{X}^T\boldsymbol{V}^2\boldsymbol{X}\boldsymbol{H}^{-1}\boldsymbol{x}_i \\
&= \frac{1}{\phi^2}||\mathrm{SD}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{V}\boldsymbol{X}\boldsymbol{H}^{-1}\boldsymbol{x}_i||^2 \\
&\propto ||\mathrm{SD}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{V}\boldsymbol{X}\boldsymbol{H}^{-1}\boldsymbol{x}_i||^2 \, .
\end{aligned}$$

This gives the desired result and completes the first part of the proof.

For a linear regression model with constant error variance, we further have that $\mathrm{SD}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i) =: \sigma$ does not depend on i. From this, it follows that $\boldsymbol{V} = \sigma^2\boldsymbol{I}_{N\times N}$, where $\boldsymbol{I}_{N\times N}$ is an $(N \times N)$ identity matrix, and that $\boldsymbol{H} \propto \boldsymbol{X}^T\boldsymbol{X}$. This now implies that

$$\begin{aligned}
||\mathrm{SD}_{\boldsymbol{\theta}}(Y_i^*|\boldsymbol{x}_i)\boldsymbol{V}\boldsymbol{X}\boldsymbol{H}^{-1}\boldsymbol{x}_i||^2 &\propto ||\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i||^2 \\
&= \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i \\
&= \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i \\
&=: h_{ii} \, ,
\end{aligned}$$

which is known as the statistical leverage score for linear regression (Rawlings et al., 1998). □

C Additional experiment results

In all tables and figures that follow, 'Prop. 1' refers to unbiased active learning with a sampling scheme that is optimised to minimise the anticipated variance of the estimated loss, 'Cor. 1a' to leverage sampling, optimised to minimise the anticipated generalisation error in terms of total loss of the active learning algorithm, 'Cor. 1b' to sampling optimised to minimise the anticipated mean squared error of the predictions, 'Prob. un.' to probabilistic uncertainty sampling (Chu 2011, Ganti and Gray 2012), 'Det. un.' to deterministic uncertainty sampling (Lewis and Gale, 1994), and 'Uniform' to uniform or simple random sampling, i.e. passive learning.

Table S1: Descriptive statistics of benchmark datasets.

Feature	Abalone	Australian	E. coli	German	Red Wine	White Wine
Number of records	4177	690	9090	1000	1599	4898
n (%) in minority class	1447 (34.6%)	307 (44.5%)	3861 (42.5%)	300 (30.0%)	217 (13.6%)	1060 (21.6%)
Number of predictors[a]	10	35	15	24	11	11
Performance under optimal model[b]						
Accuracy[c]	0.75	0.88	0.81	0.78	0.88	0.80
AUC	0.83	0.94	0.85	0.82	0.88	0.79

[a] After re-coding of categorical predictors and removal of redundant variables.

[b] Using L_2-penalised logistic regression on the entire data set.

[c] Using 50% probability cut-off.

AUC, area under the receiver operating characteristic curve.

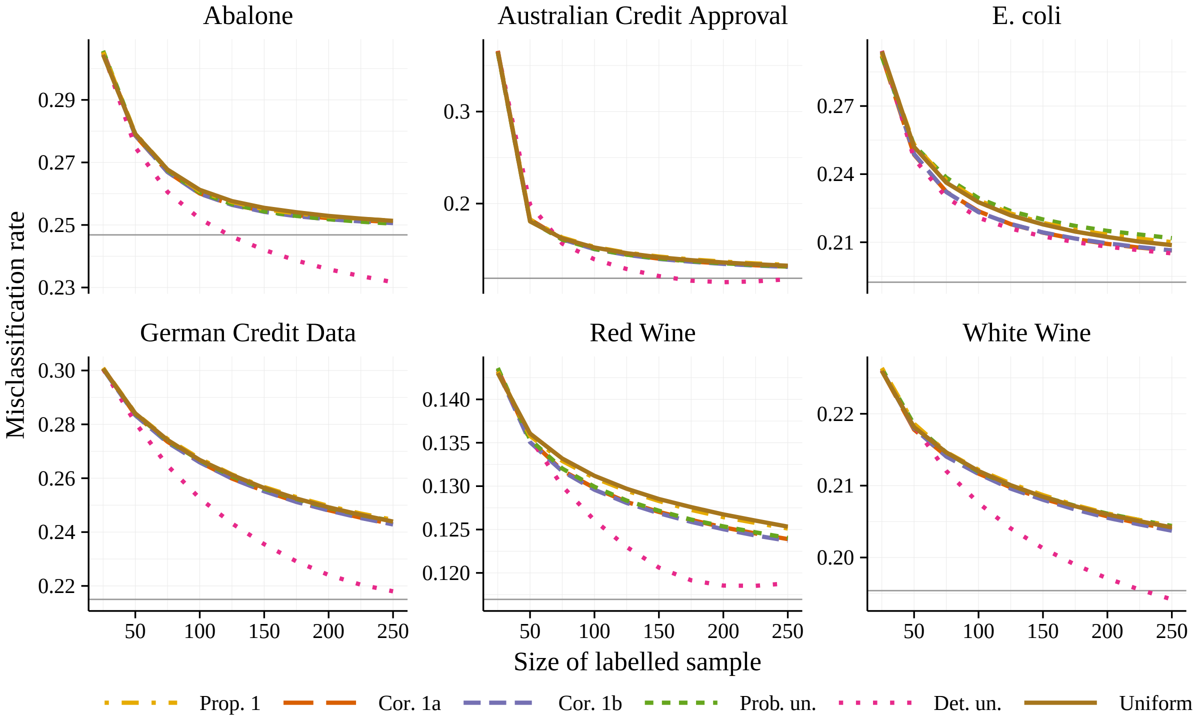

Figure S1: Average misclassification rate in 10 000 active learning experiments. The grey solid line shows the performance when using the entire dataset for training.

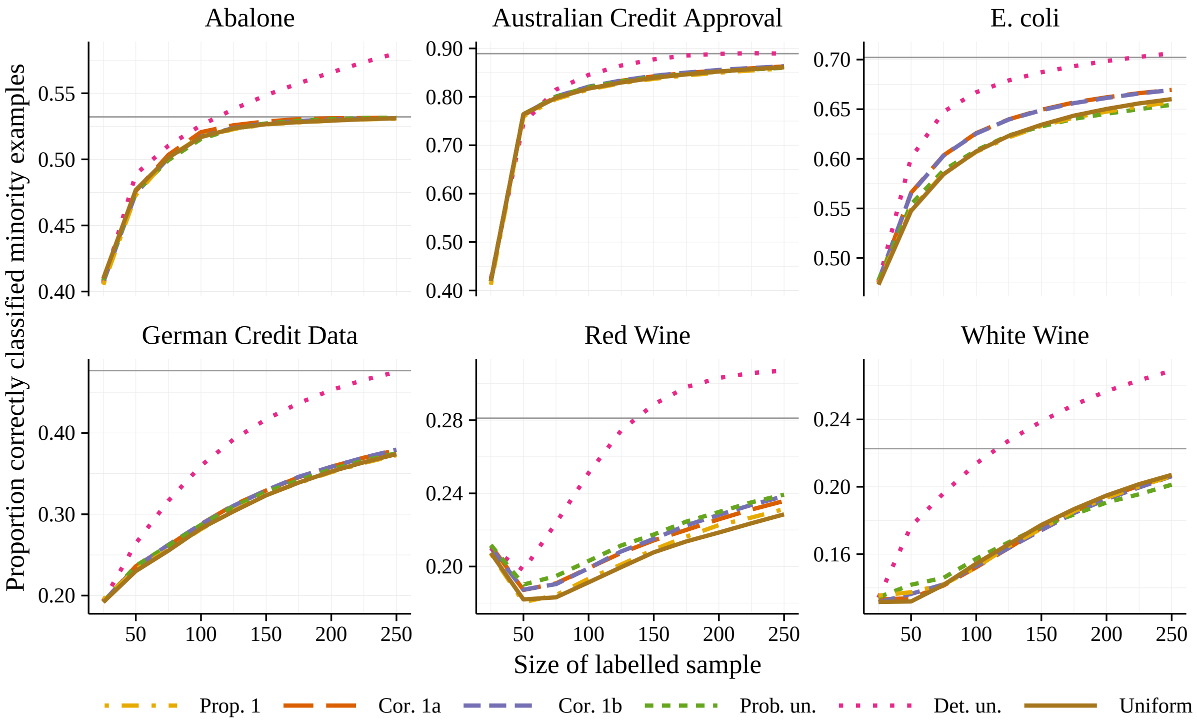

Figure S2: Average of the proportion correctly classified minority examples in 10 000 active learning experiments. The grey solid line shows the performance when using the entire dataset for training.

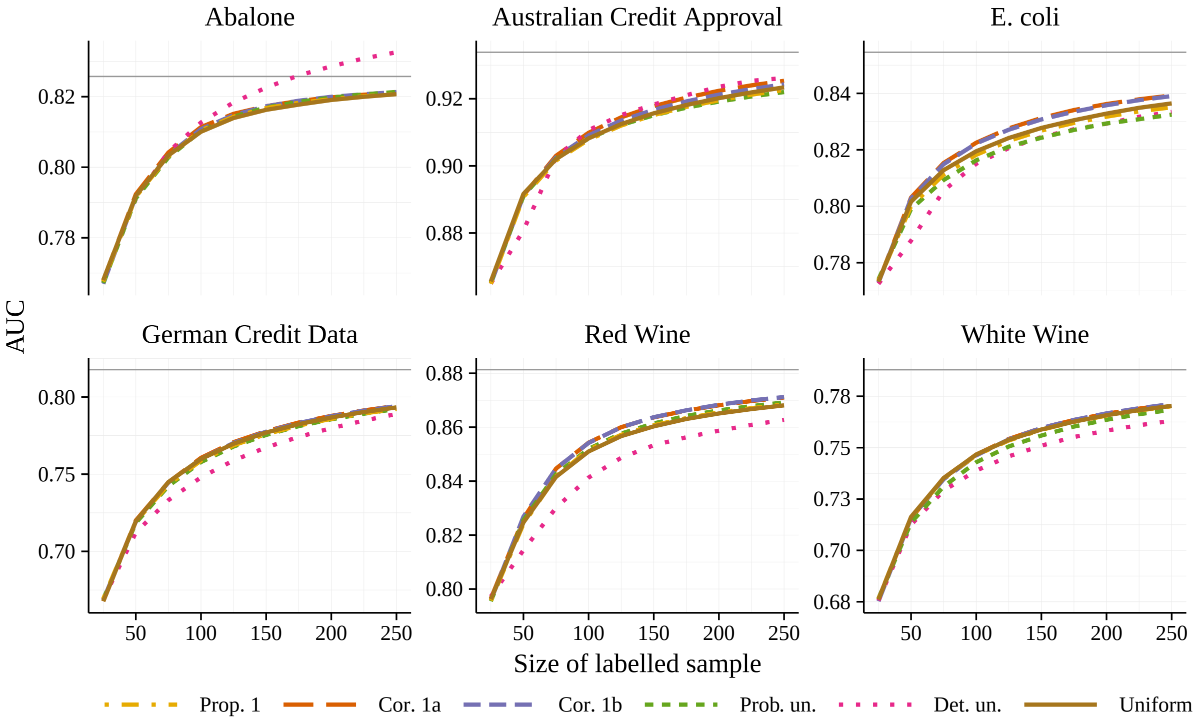

Figure S3: Mean AUC in 10 000 active learning experiments. The grey solid line shows the performance when using the entire dataset for training. AUC, area under the receiver operating characteristic curve.

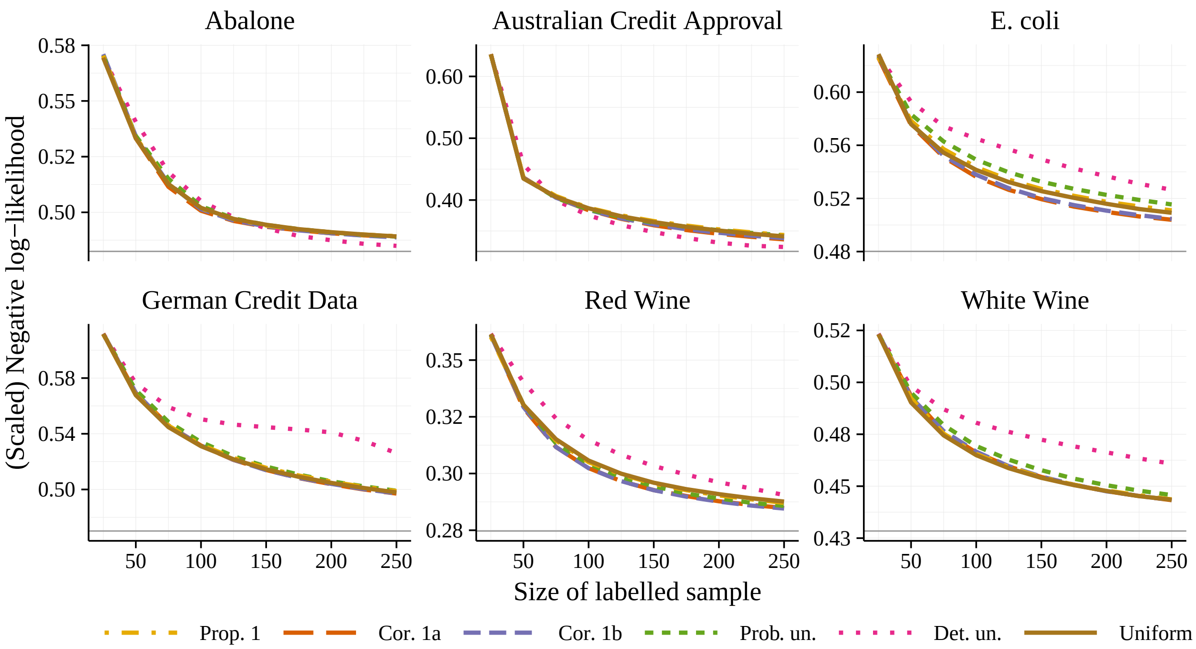

Figure S4: Median of the negative log-likelihood (scaled by a factor $1/N$) of the predicted class probabilities in 10 000 active learning experiments. The grey solid line shows the performance when using the entire dataset for training.

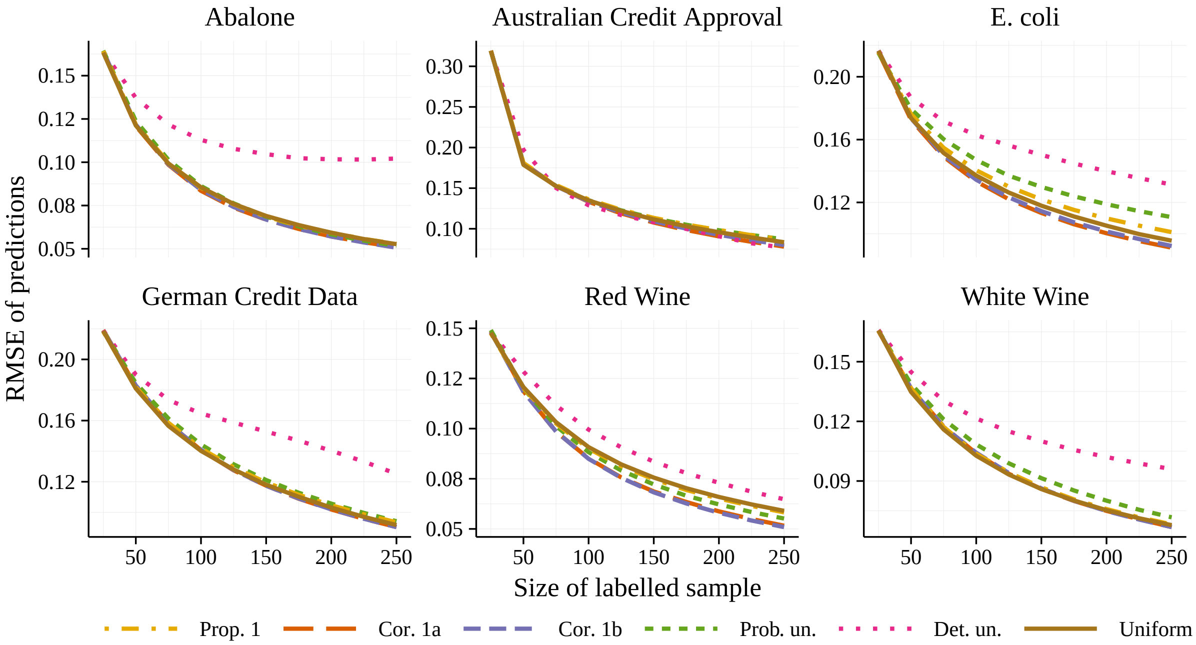

Figure S5: Average root mean squared error (RMSE) of the predictions in 10 000 active learning experiments, as compared to the predictions obtained when using the entire datasets for training. The RMSE was computed as $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{p}_i - p_i\right)^2}$, where $\hat{p}_i = P_{\hat{\boldsymbol{\theta}}_t}(Y_i = 1|\boldsymbol{x}_i)$ and $p_i = P_{\boldsymbol{\theta}_0}(Y_i = 1|\boldsymbol{x}_i)$.

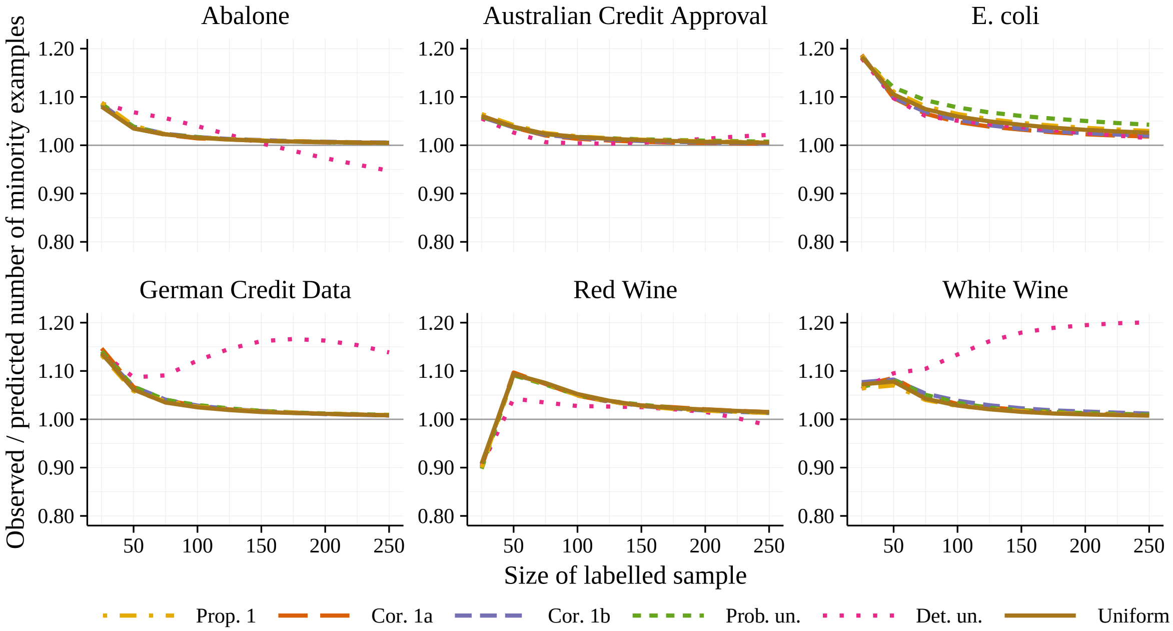

Figure S6: Mean of the ratio between the observed and predicted (expected) number of minority examples in 10 000 active learning experiments, computed as $\sum_{i=1}^{N} y_i / \sum_{i=1}^{N} \hat{p}_i$, where $\hat{p}_i = P_{\hat{\boldsymbol{\theta}}_t}(Y_i = 1|\boldsymbol{x}_i)$ and $Y = 1$ is coded as the minority class. The grey solid line represents the ideal performance, i.e. perfect agreement between observed and expected number of minority examples. A value greater than 1 corresponds to a bias in the predicted probabilities towards the majority class, and a value smaller than 1 to a bias towards the minority class.

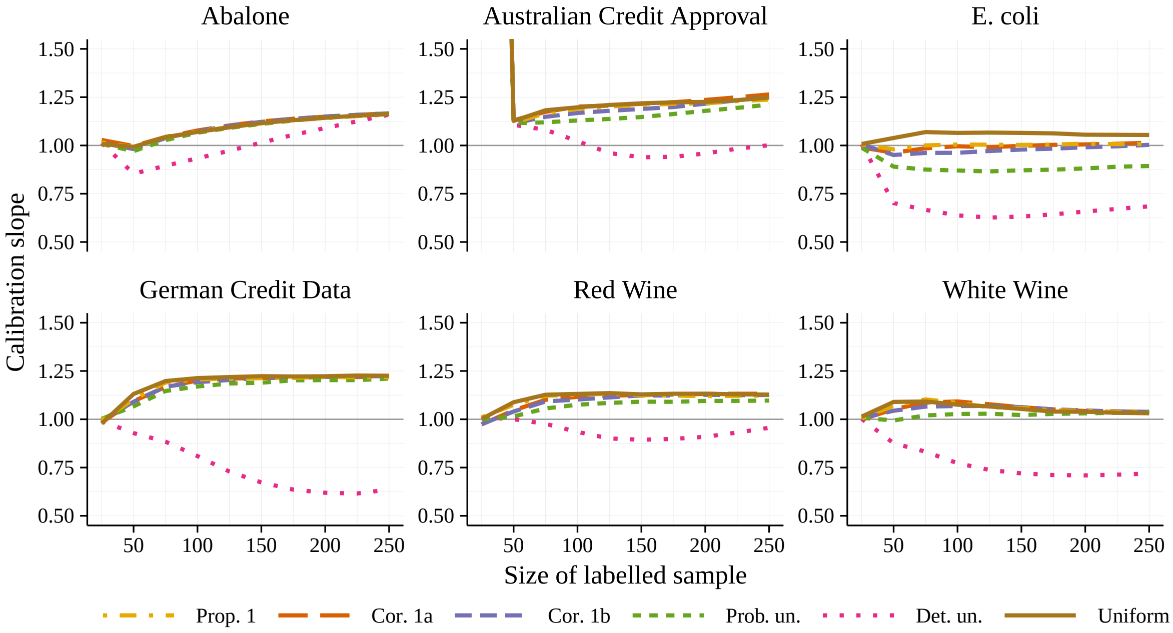

Figure S7: Median of the calibration slope between the observed outcomes and predicted probabilities in 10 000 active learning experiments, computed as described in Steyerberg and Vergouwe (2014). The grey solid line represents the ideal performance. A value > 1 correspond to conservative predictions that are shrunk towards the overall mean, and a value < 1 to overfitting in the sense that the predicted class probabilities are too extreme: low predictions too low and high predictions too high.

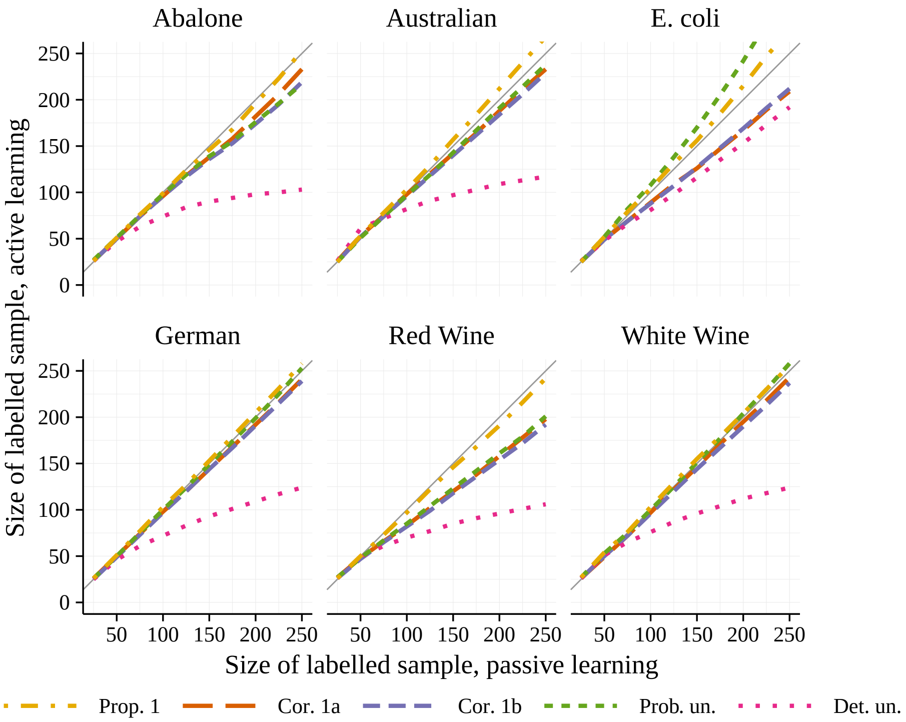

Figure S8: Number of training examples needed for active learning to achieve equal performance in terms of misclassification rate as passive learning with a given sample size. The diagonal line represents no improvement compared to passive learning.

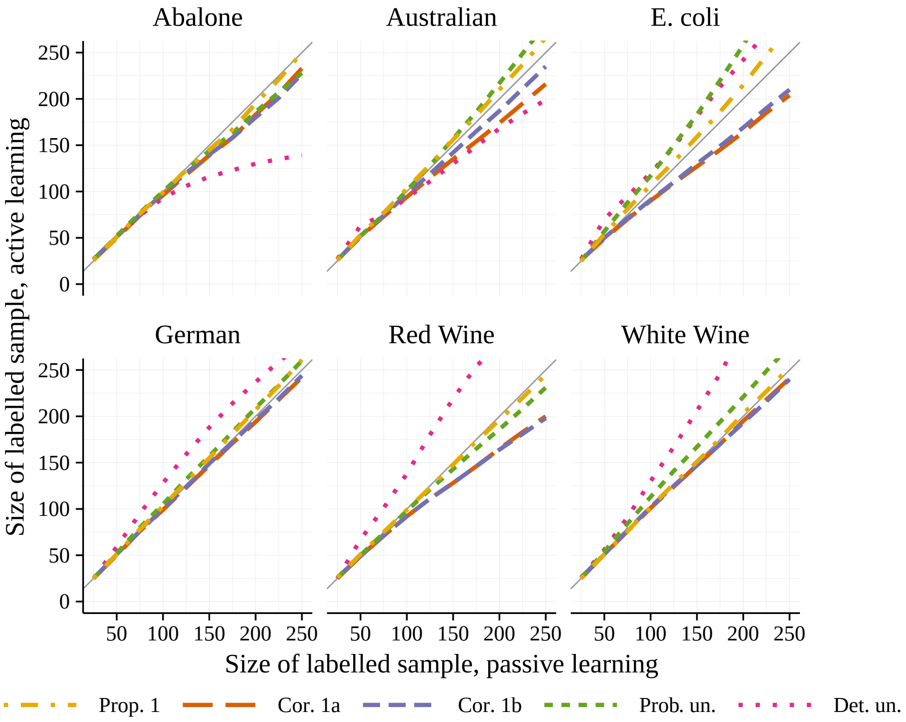

Figure S9: Number of training examples needed for active learning to achieve equal performance in terms of AUC as passive learning with a given sample size. The diagonal line represents no improvement compared to passive learning. AUC, area under the receiver operating characteristic curve.

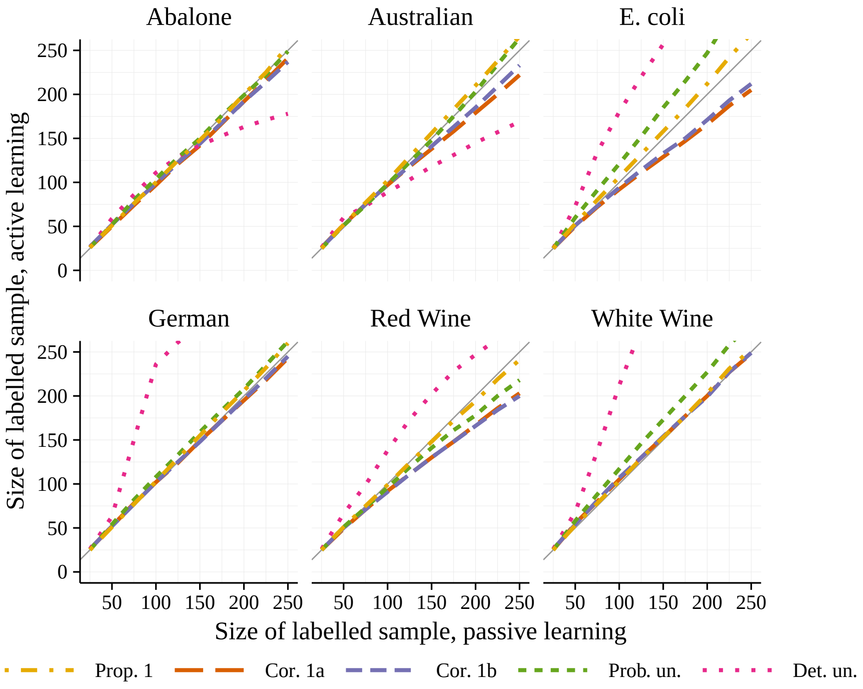

Figure S10: Number of training examples needed for active learning to achieve equal performance in terms of the negative log-likelihood of the predictions as passive learning with a given sample size. The diagonal line represents no improvement compared to passive learning.

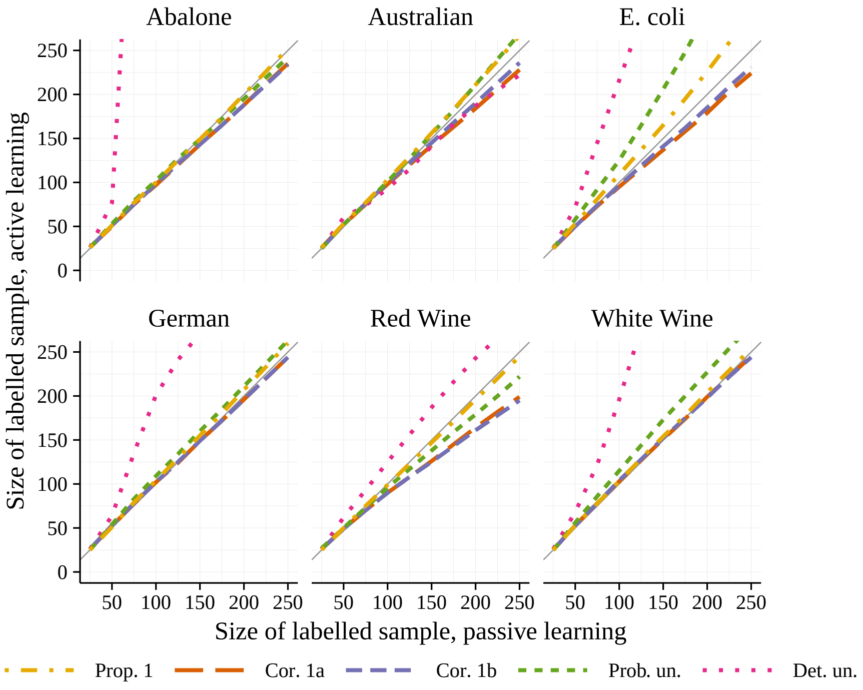

Figure S11: Number of training examples needed for active learning to achieve equal performance in terms of the RMSE of the predictions as passive learning with a given sample size. The diagonal line represents no improvement compared to passive learning. The RMSE was computed as $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{p}_i - p_i\right)^2}$, where $\hat{p}_i = P_{\hat{\boldsymbol{\theta}}_t}(Y_i = 1|\boldsymbol{x}_i)$ and $p_i = P_{\boldsymbol{\theta}_0}(Y_i = 1|\boldsymbol{x}_i)$.

Table S2: Label complexity of active vs. passive learning, presented as the relative increase (ratio > 1) or decrease (ratio < 1) in the sample size needed for active learning to achieve equal performance as passive learning with $n = 250$ training examples.

		Ratio of sample sizes using active vs. passive learning				
Dataset	Performance metric	Prop. 1	Cor. 1a	Cor. 1b	Prob. un.	Det. un.
Abalone	AUC	1.00	0.93	0.91	0.91	0.56
	Misclassification rate	1.01	0.93	0.88	0.87	0.41
	Negative log-likelihood	1.02	0.96	0.95	1.00	0.71
	RMSE of predictions	1.01	0.94	0.94	0.97	>2.00
Australian Credit Approval	AUC	1.06	0.86	0.94	1.12	0.80
	Misclassification rate	1.07	0.93	0.92	0.95	0.47
	Negative log-likelihood	1.06	0.89	0.93	1.06	0.68
	RMSE of predictions	1.07	0.91	0.94	1.08	0.89
E. coli	AUC	1.11	0.82	0.84	1.39	1.18
	Misclassification rate	1.10	0.84	0.85	1.32	0.77
	Negative log-likelihood	1.08	0.82	0.85	1.28	1.56
	RMSE of predictions	1.16	0.90	0.92	1.58	>2.00
German Credit Data	AUC	1.04	0.97	0.98	1.04	1.11
	Misclassification rate	1.03	0.96	0.96	1.01	0.50
	Negative log-likelihood	1.04	0.97	0.98	1.05	1.29
	RMSE of predictions	1.04	0.98	0.98	1.05	1.36
Red Wine	AUC	0.98	0.80	0.79	0.92	1.27
	Misclassification rate	0.97	0.79	0.77	0.80	0.42
	Negative log-likelihood	0.96	0.81	0.80	0.87	1.12
	RMSE of predictions	0.98	0.80	0.78	0.89	1.14
White Wine	AUC	1.01	0.96	0.96	1.10	1.68
	Misclassification rate	1.02	0.97	0.95	1.03	0.50
	Negative log-likelihood	1.01	0.99	1.00	1.13	>2.00
	RMSE of predictions	1.02	0.98	0.98	1.12	>2.00

AUC, area under the receiver operating characteristic curve; RMSE, root mean squared error.