Unifying Approaches in Data Subset Selection via Fisher Information and Information-Theoretic Quantities

Anonymous authors
Paper under double-blind review

Abstract

The mutual information between predictions and model parameters—also referred to as *expected information gain* or BALD in machine learning—measures informativeness. It is a popular acquisition function in Bayesian active learning and Bayesian optimal experiment design. In data subset selection, i.e. active learning and active sampling, several recent works use Fisher information, Hessians, similarity matrices based on the gradients, or simply the gradient lengths to compute the acquisition scores that guide sample selection. Are these different approaches connected, and if so how? In this paper, we revisit the Fisher information and use it to show how several otherwise disparate methods are connected as approximations of information-theoretic quantities.

1 Introduction

Label and training efficiency are the key to a wider deployment of deep learning. Deep learning generally requires a lot of data, most of which has to be annotated. This is expensive and time-consuming. Together with semi-supervised and unsupervised approaches, active learning helps increase label efficiency. Given access to unlabeled data, *active learning* selects the most informative samples to label for a given model, thus decreasing the number of required annotations. Apart from label efficiency, training deep learning models is expensive and time-consuming, too, and *active sampling* improves training efficiency by filtering the training set to focus on the samples that will be the most informative for the model.

The acquisition functions used to select informative samples can often be traced back to information-theoretic objectives that are known from Bayesian optimal experiment design (Lindley, 1956). The connection between contemporary acquisition functions and basic information-theoretic quantities (short: *information quantities*) is the topic of this work. Amongst them we will examine: for active learning, the *expected information gain (EIG)* (Lindley, 1956; Houlsby et al., 2011; Kirsch et al., 2019)), the *(joint) expected predictive information gain (JEPIG or EPIG, respectively)* (Kirsch et al., 2021b; MacKay, 1992), and, for active sampling, the *information gain (IG)* (Sun et al., 2022) and *(joint) predictive information gain (JPIG or PIG, respectively)* (Mindermann et al., 2022). We will show that many contemporary methods can be seen as approximating information quantities using different approximations, which have different trade-offs.

We develop a unifying perspective on various existing approaches: we look at second-order posterior approximations (Laplace approximations) in §3, which we use to revisit the Fisher information and its approximations in §4. This leads to approximations of the information quantities mentioned above in §5. We pay special attention to the limitations: for example, we will see that approximations that use the trace of the Fisher information do not take redundancies between samples into account. They exhibit the same pathologies as other methods that in essence score points individually, also known as top-K batch acquisition (Kirsch et al., 2021a). In §6, we connect these approximations to approximations that use similarity matrices of the log loss gradients. In §7, we show that (Batch-)BALD and EPIG on the one hand; and BADGE, BAIT, SIMILAR[1], and PRISM[1] on the other hand can be seen as optimizing the same objectives. The difference is that (Batch-)BALD (Houlsby et al., 2011; Kirsch et al., 2019) and EPIG (Kirsch et al., 2021b) operate in prediction space, while Fisher information-based methods operate in weight space: we show that

---

[1] using $\log\det$ objectives

an approximation of EPIG, a transductive active learning objective, using Fisher information, matches the BAIT objective (Ash et al., 2021). We show how BADGE (Ash et al., 2019) can be seen as approximating the EIG, using the connection between similarity matrices and EIG approximations; similarly, we find that submodularity-based approaches (Iyer et al., 2021) such as SIMILAR (Kothawade et al., 2021) and PRISM (Kothawade et al., 2022), which report their best results using the $\log\det$ of similarity matrices, approximate information quantities. We also show that gradient-length-based methods like *expected gradient length* (Settles et al., 2007) and *gradient norm score* (Paul et al., 2021) can be connected to information quantities.

Although our results employ a hierarchy of approximations, we do not examine the error terms in detail. This is in line with how these approximations are used in deep learning, where they often only provide motivation for useful mechanisms. However, we try to identify where these approximations might break, enumerate the limitations they introduce, and raise research questions that can be verified empirically as future work.

In general, we show that various disparate non-Bayesian approaches in data subset selection approximate information quantities from the Bayesian literature, going back to Lindley (1956) and MacKay (1992).

2 Background & Setting

This section introduces the relevant notation and the probabilistic model we use throughout this paper.

Information Theory. We follow the practical notation from Kirsch and Gal (2021). In particular, for entropy, we use an implicit or explicit notation, $\mathrm{H}[X]$ or $\mathrm{H}(\mathrm{p}(X))$, while $\mathrm{H}(\mathrm{p} \parallel \mathrm{q})$ denotes the cross-entropy similar to the Kullback-Leibler divergence $\mathrm{D_{KL}}(\mathrm{p} \parallel \mathrm{q})$, and $\mathrm{H}(p)$ denotes Shannon's information content:

$$\mathrm{H}(p) := -\log p, \tag{1}$$
$$\mathrm{H}[x] := \mathrm{H}(\mathrm{p}(x)), \tag{2}$$
$$\mathrm{H}(\mathrm{p}(X) \parallel \mathrm{q}(X)) := \mathbb{E}_{\mathrm{p}(x)}[\mathrm{H}(\mathrm{q}(x))], \tag{3}$$
$$\mathrm{H}[X] := \mathrm{H}(\mathrm{p}(X)) := \mathrm{H}(\mathrm{p}(X) \parallel \mathrm{p}(X)), \tag{4}$$

where p and q are probabilities distributions, X is a random variable, and x is an outcome. Conditional and joint entropies are defined as usual (note that we take an expectation over y as well):

$$\mathrm{H}[X \mid Y] = \mathbb{E}_{\mathrm{p}(x,y)}[-\log \mathrm{p}(x \mid y)]. \tag{5}$$

The mutual information $\mathrm{I}[X;Y]$ for random variables X and Y is defined as:

$$\mathrm{I}[X;Y] := \mathrm{H}[X] - \mathrm{H}[X \mid Y], \tag{6}$$

and is also referred to as expected uncertainty reduction or expected information gain (Lindley, 1956) because the entropy $\mathrm{H}[X]$ quantifies the uncertainty about the random variable X, and $\mathrm{H}[X \mid Y]$ about X after observing Y (in expectation).

Probabilistic Model. We assume a supervised setting: for inputs X, we have a Bayesian model with parameters Ω that makes predictions Y. What makes the model Bayesian is that the parameters follow a distribution $\mathrm{p}(\omega)$, and we have a probabilistic model:

$$\mathrm{p}(y, \omega \mid x) = \mathrm{p}(y \mid x, \omega)\ \mathrm{p}(\omega). \tag{7}$$

We extend this model to additional data $\mathcal{D} := \{(x_i, y_i)\}_i$ as follows: $\mathrm{p}(\{y_i\}, \omega \mid \{x_i\}) = \mathrm{p}(\{y_i\} \mid \{x_i\}, \omega)\ \mathrm{p}(\omega)$. That is, we examine the common discriminative case where, unlike in the generative case, we do not model $\mathrm{p}(x)$. The corresponding marginal prediction of the model is $\mathrm{p}(y \mid x) = \mathbb{E}_{\mathrm{p}(\omega)}[\mathrm{p}(y \mid x, \omega)]$.

Transductive Objectives. When an objective uses additional data (unlabeled or labeled) to guide acquisitions, we refer to this as a *transductive* objective (Yu et al., 2006; Wang et al., 2020).

Active Learning. To increase label efficiency, instead of labeling data indiscriminately, active learning iteratively selects and acquires labels for the *most informative* unlabeled data from an unlabeled *pool set* according to some underlying *acquisition function*. An acquisition function scores the informativeness of an unlabeled candidate batch $\{x^{\mathrm{acq}}{}_i\}_{i=1}^N$, and the batch that maximizes the score is selected for labeling. After each such acquisition step, the model is re-trained to take the newly labeled data into account. Labels can be acquired individually or in batches (*batch active learning*, Gal et al. (2017)). The *expected information gain (EIG)*

$$\mathrm{I}[\Omega; Y^{\mathrm{acq}} \mid x^{\mathrm{acq}}] \tag{EIG/BALD}$$

and *(joint) expected predictive information gain (EPIG and JEPIG, respectively)*

$$\mathrm{I}[Y^{\mathrm{eval}}; Y^{\mathrm{acq}} \mid X^{\mathrm{eval}}, x^{\mathrm{acq}}], \tag{EPIG}$$
$$\mathrm{I}[\{Y_i^{\mathrm{eval}}\}; Y^{\mathrm{acq}} \mid \{x_i^{\mathrm{eval}}\}, x^{\mathrm{acq}}] \tag{JEPIG}$$

are examples of such acquisition functions.

Active Sampling. To increase training efficiency, instead of training with all samples, active sampling selects the informative samples $\mathcal{D}^{\mathrm{acq}} = \{(x^{\mathrm{acq}}{}_i, y^{\mathrm{acq}}{}_i)\}_{i=1}^N$ from the training set to train on. This can be done statically before training the model, in which case this is also referred to as core-set selection, or more dynamically, in which case it is also referred to as curriculum learning. The *information gain (IG)*

$$\mathrm{I}[\Omega; y^{\mathrm{acq}} \mid x^{\mathrm{acq}}], \tag{IG}$$

and *(joint) predictive information gain (PIG or JPIG, respectively)*

$$\mathrm{I}[Y^{\mathrm{eval}}; y^{\mathrm{acq}} \mid X^{\mathrm{eval}}, x^{\mathrm{acq}}] \quad (= \mathbb{E}_{\hat{\mathrm{p}}(x^{\mathrm{eval}}, y^{\mathrm{eval}})} \mathrm{I}[y^{\mathrm{eval}}; y^{\mathrm{acq}} \mid x^{\mathrm{eval}}, x^{\mathrm{acq}}]), \tag{PIG}$$
$$\mathrm{I}[\{y_i^{\mathrm{eval}}\}; y^{\mathrm{acq}} \mid \{x_i^{\mathrm{eval}}\}, x^{\mathrm{acq}}] \tag{JPIG}$$

are examples of such acquisition functions.

Log Loss. While many active learning and active sampling methods are motivated independently of the underlying loss, we will focus on log losses, such as the common cross-entropy loss or squared error loss, to be principled. These log-losses can be viewed through an information-theoretic lens.

3 Second-Order Posterior Approximation

To compute approximations of information quantities using Fisher information, we need to approximate the posterior $\mathrm{p}(\omega \mid \mathcal{D}, \mathcal{D}^{\mathrm{train}})$, where $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ are additional (new) samples, and we start with $\mathrm{p}(\omega \mid \mathcal{D}^{\mathrm{train}})$ as the "prior" distribution. We will drop $\mathcal{D}^{\mathrm{train}}$ and use $\mathrm{p}(\omega)$ when possible.

To begin, we can complete the square of a second-order Taylor approximation around the log parameter likelihood around a fixed ω^*:

$$\log \mathrm{p}(\omega) \approx \log \mathrm{p}(\omega^*) + \nabla_\omega \log \mathrm{p}(\omega^*)(\omega - \omega^*) + \frac{1}{2}(\omega - \omega^*)^T \nabla_\omega^2 \log \mathrm{p}(\omega^*)(\omega - \omega^*) \tag{8}$$
$$= \frac{1}{2}(\omega - (\omega^* - \nabla_\omega^2 \log \mathrm{p}(\omega^*)^{-1} \nabla_\omega \log \mathrm{p}(\omega^*))^T \nabla_\omega^2 \log \mathrm{p}(\omega^*)(\omega - (\omega^* - \nabla_\omega^2 \log \mathrm{p}(\omega^*)^{-1} \nabla_\omega \log \mathrm{p}(\omega^*))) + \ldots. \tag{9}$$

We can express this in more concise terms by extending the notation of $\mathrm{H}[\cdot]$ to its derivatives:

Notation 3.1. *We write* $\mathrm{H}'[\cdot]$ *for the Jacobian and* $\mathrm{H}''[\cdot]$ *for the Hessian of* $\mathrm{H}[\cdot]$*:*

$$\mathrm{H}'[\cdot] \coloneqq -\nabla_\omega \log \mathrm{p}(\cdot), \tag{10}$$
$$\mathrm{H}''[\cdot] \coloneqq -\nabla_\omega^2 \log \mathrm{p}(\cdot). \tag{11}$$

Then, we have:

$$\mathrm{H}[\omega] \approx \mathrm{H}[\omega^*] + \mathrm{H}'[\omega^*](\omega - \omega^*) + \frac{1}{2}(\omega - \omega^*)^T \; \mathrm{H}''[\omega^*] \, (\omega - \omega^*) \tag{12}$$

$$= \frac{1}{2}(\omega - (\omega^* - \mathrm{H}''[\omega^*]^{-1}\,\mathrm{H}'[\omega^*])^T\;\mathrm{H}''[\omega^*]\,(\omega - (\omega^* - \mathrm{H}''[\omega^*]^{-1}\,\mathrm{H}'[\omega^*])) + \dots. \tag{13}$$

Comparing this to the information content of a multivariate Gaussian distribution:

$$\mathrm{H}[\mathcal{N}(w;\,\mu,\,\Sigma)] = \frac{1}{2}(\omega - \mu)^T\,\Sigma^{-1}\,(\omega - \mu) + \dots, \tag{14}$$

we obtain:

Proposition 3.2. *An approximation of the distribution* $\mathrm{p}(\omega)$ *of* Ω *around some* ω^* *is given by:*

$$\Omega \overset{\approx}{\sim} \mathcal{N}(\omega^* - \mathrm{H}''[\omega^*]^{-1}\,\mathrm{H}'[\omega^*],\ \mathrm{H}''[\omega^*]^{-1}), \tag{15}$$

where $\mathrm{H}''[\omega^*]$ *must be positive-definite. If* ω^* *is also a (global) minimizer of* $\mathrm{H}[\omega]$ *(i.e.* $\mathrm{H}'[\omega^*] = 0$*), we obtain the* Laplace approximation*:*

$$\Omega \overset{\approx}{\sim} \mathcal{N}(\omega^*,\ \mathrm{H}''[\omega^*]^{-1}). \tag{16}$$

Approximation Quality. Obviously, this approximation can be arbitrarily bad depending on $\mathrm{p}(\omega)$ and ω^*—however, given enough data, we expect that $\mathrm{p}(w)$ will concentrate around a maximum a posteriori (MAP) estimate, giving rise to the Laplace approximation. Obviously, insufficient data to reach concentration of the parameters and multimodality in over-parameterized models (Long, 2021) can be issues for active learning and active sampling.

Flat Minimum Intuition. The information-geometric interpretation of the Hessian being positive definite is that the information content (pointwise entropy) is convex around ω^* and, equivalently, that the (log) posterior is concave around ω^*. The latter provides an intuition for the Gaussian approximation: the Hessian measures curvature, and the "flatter" the Hessian, e.g., the smaller the largest eigenvalue or the smaller the determinant, the less the loss changes when ω^* is perturbed. This leads to the search for flat minima as a way to improve generalization (Hinton and Van Camp, 1993; Hochreiter and Schmidhuber, 1994).

Notation 3.3. *To further shorten the notation, we will write* $\mathrm{H}''[\mathcal{D} \mid \omega^*]$ *instead of* $\mathrm{H}''[\{y_i\} \mid \{x_i\}, \omega^*]$*.*

Posterior Approximation of $\Omega \mid \mathcal{D}$. While the Laplace approximation is centered on a (global) minimizer, the approach above can be used for a (potentially low-quality) posterior approximation in general. We can expand $\mathrm{H}[\omega^* \mid \mathcal{D}]$ using Bayes' law and the additivity of the logarithm. That is, we have:

$$\mathrm{H}[\omega^* \mid \mathcal{D}] = \mathrm{H}[\mathcal{D} \mid , \omega^*] + \mathrm{H}[\omega^*] - \mathrm{H}[\mathcal{D}], \tag{17}$$

and then, as $\mathrm{H}[\mathcal{D}]$ is independent of ω:

$$\mathrm{H}'[\omega^* \mid \mathcal{D}] = \mathrm{H}'[\mathcal{D} \mid \omega^*] + \mathrm{H}'[\omega^*] + 0 = \mathrm{H}'[\mathcal{D} \mid \omega^*] + \mathrm{H}'[\omega^*], \tag{18}$$

$$\mathrm{H}''[\omega^* \mid \mathcal{D}] = \mathrm{H}''[\mathcal{D} \mid \omega^*] + \mathrm{H}''[\omega^*]. \tag{19}$$

Proposition 3.4. *The* observed information $\mathrm{H}''[\{y_i\} \mid \{x_i\}, \omega^*]$ *is additive:*

$$\mathrm{H}''[\{y_i\} \mid \{x_i\}, \omega^*] = \sum_i \mathrm{H}''[y_i \mid x_i, \omega^*] = \sum_i -\nabla^2_\omega \log \mathrm{p}(y_i \mid x_i, \omega^*). \tag{20}$$

The observed information is defined with opposite sign compared to other works as it simplifies the exposition.

Uninformative Prior. For a typical Gaussian prior $\mathrm{p}(\omega) \sim \mathcal{N}(\mu, \Sigma)$, we have $\mathrm{H}''[\omega^*] = \Sigma^{-1}$ and $\mathrm{H}''[\omega^* \mid \mathcal{D}] = \mathrm{H}''[\mathcal{D} \mid \omega^*] + \Sigma^{-1}$. For an uninformative prior with "infinite prior variance" $\Sigma^{-1} \to 0$, we have $\mathrm{H}''[\omega^*] = 0$ and $\mathrm{H}''[\omega^* \mid \mathcal{D}] = \mathrm{H}''[\mathcal{D} \mid \omega^*]$.

Proposition 3.5. *The entropy of the second-order approximation of* $\mathrm{p}(\omega)$ *around* ω^* *is*

$$\mathrm{H}[\Omega] \approx -\tfrac{1}{2}\log\det \mathrm{H}''[\omega^*] + C_k, \tag{21}$$

where $C_k = \frac{k}{2}\log 2\pi e$ *is a constant (independent of* $\mathcal{D}$ *and* ω^**) and* k *is the number of dimensions of* ω*.*

4 Fisher Information

The following section revisits the Fisher information and its properties. All proofs can be found in §A. In particular, we look at two special cases with more favorable properties: following Kunstner et al. (2019), when we can write our model as $\mathrm{p}(y \mid \hat{z} = \hat{f}(x;\omega))$, where $\hat{f}(x;\omega)$ are the logits, and $\mathrm{p}(y \mid \hat{z})$ is a distribution from an exponential family, the Fisher information is independent of y, which has useful consequences as we will see; and following Chaudhuri et al. (2015), when we have a Generalized Linear Model (GLM), the Hessian is independent of y. The results for the GLM are also used as an approximation known as *Generalized Gauss-Newton approximation.* Together with other numerical approximations, such as using a diagonal matrix or low-rank factorizations, the Hessian and Fisher information can be efficiently approximated for large deep neural networks (Daxberger et al., 2021).

Definition 4.1. *The* Fisher information $\mathrm{F}(\{x_i\}, \omega^*)$ *is the expectation over the observed information using the model's own predictions for given* $\{x_i\}$ *at* ω^**:*

$$\mathrm{F}(\{x_i\}, \omega^*) = \mathbb{E}_{\mathrm{p}(\{y_i\} \mid \{x_i\}, \omega^*)}[\mathrm{H}''[\{y_i\} \mid \{x_i\}, \omega^*]]. \tag{22}$$

Note that this definition of the Fisher information is also referred to as "empirical" Fisher information in statistics because we do not take an expectation over x but use empirical samples for x (Kunstner et al., 2019).

Proposition 4.2. *Like the observed information, the Fisher information is additive:*

$$\mathrm{F}(\{x_i\}, \omega^*) = \sum_i \mathrm{F}(x_i, \omega^*). \tag{23}$$

There are two other equivalent definitions of the Fisher information:

Proposition 4.3. *The Fisher information is equivalent to:*

$$\mathrm{F}(x, \omega^*) = \mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[\mathrm{H}'[y \mid x, \omega^*]^T \ \mathrm{H}'[y \mid x, \omega^*]] = \mathrm{Cov}[\mathrm{H}'[Y \mid x, \omega^*]]. \tag{24}$$

Note that the expectation is over $\mathrm{p}(y \mid x, \omega^*)$ and not $\mathrm{p}(y \mid x)$. The proof in §A applies two generally useful lemmas:

Lemma 4.4. *For the Jacobian* $\mathrm{H}'[y \mid x, \omega^*]$*, we have:*

$$\mathrm{H}'[y \mid x, \omega^*] = -\frac{\nabla_\omega \mathrm{p}(y \mid x, \omega^*)}{\mathrm{p}(y \mid x, \omega^*)}, \tag{25}$$

and for the Hessian $\mathrm{H}''[y \mid x, \omega^*]$*, we have:*

$$\mathrm{H}''[y \mid x, \omega^*] = \mathrm{H}'[y \mid x, \omega^*]^T \ \mathrm{H}'[y \mid x, \omega^*] - \frac{\nabla^2_\omega \mathrm{p}(y \mid x, \omega^*)}{\mathrm{p}(y \mid x, \omega^*)}. \tag{26}$$

Lemma 4.5. *The following expectations over the model's own predictions vanish:*

$$\mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[\mathrm{H}'[y \mid x, \omega^*]] = 0, \tag{27}$$

$$\mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}\left[\frac{\nabla^2_\omega \mathrm{p}(y \mid x, \omega^*)}{\mathrm{p}(y \mid x, \omega^*)}\right] = 0. \tag{28}$$

Special Case: Exponential Family. Kunstner et al. (2019) show in their appendix that if we split a discriminative model into prelogits $\hat{f}(x;\omega)$ and a predictor $\mathrm{p}(y \mid \hat{z} = \hat{f}(x;\omega))$, the Hessian does not depend on y when $\mathrm{p}(y \mid \hat{z})$ is a distribution from an exponential family (independent of ω). Examples include using a normal distribution for regression parameterized by mean and variance predictions or a categorical distribution via the softmax function. The following statements and proofs follow Kunstner et al. (2019):

Proposition 4.6. *The Fisher information* $\mathrm{F}(\{x_i\}, \omega^*)$ *for a model* $\mathrm{p}(y \mid \hat{z} = \hat{f}(x; \omega^*))$ *is equivalent to:*

$$\mathrm{F}(x, \omega^*) = \nabla_\omega \hat{f}(x; \omega^*)^T \, \mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[\nabla_{\hat{z}}^2 \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)]] \nabla_\omega \hat{f}(x; \omega^*), \tag{29}$$

where $\nabla_{\hat{z}}^2 \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)]$ *is short for* $\nabla_{\hat{z}}^2 \, \mathrm{H}[y \mid \hat{z}]\big|_{\hat{z} = \hat{f}(x; \omega^*)}$.

Proposition 4.7. *The Fisher information* $\mathrm{F}(x, \omega^*)$ *of a model of the form* $\mathrm{p}(y \mid \hat{z} = \hat{f}(x; \omega^*))$ *is independent of* y*, where* $\mathrm{p}(y \mid \hat{z})$ *is an exponential distribution, i.e.,* $\log \mathrm{p}(y \mid \hat{z}) = \hat{z}^T T(y) - A(\hat{z}) + \log h(y)$*:*

$$\mathrm{F}(x, \omega^*) = \nabla_\omega \hat{f}(x; \omega^*)^T \, \nabla_{\hat{z}}^2 A(\hat{z} = \hat{f}(x; \omega^*)) \, \nabla_\omega \hat{f}(x; \omega^*). \tag{30}$$

It is crucial that the exponential distribution not depend on ω*.* This simplifies computing the Fisher information: no expectation over ys is needed anymore. The full outer product may not be needed explicitly either.

To make this more concrete, we will consider two common exponential distributions:

Gaussian Distribution. When $\mathrm{p}(y \mid \hat{z}) = \mathcal{N}(y; \, \hat{z}, \, 1)$, we have $\mathrm{H}''[y \mid \hat{z}] = 1$, and thus

$$\mathrm{F}(x, \omega^*) = \nabla_\omega \hat{f}(x; \omega^*)^T \, \nabla_\omega \hat{f}(x; \omega^*). \tag{31}$$

Categorical Distribution. When $\mathrm{p}(y \mid \hat{z}) = \mathrm{softmax}(\hat{z})_y$, we have $\mathrm{H}''[y \mid \hat{z}] = \mathrm{diag}(\pi) - \pi \, \pi^T$, with $\pi_y = \mathrm{p}(y \mid \hat{z})$, and thus:

$$\mathrm{F}(x, \omega^*) = \nabla_\omega \hat{f}(x; \omega^*)^T \, (\mathrm{diag}(\pi) - \pi \, \pi^T) \, \nabla_\omega \hat{f}(x; \omega^*). \tag{32}$$

Special Case: Generalized Linear Models. Chaudhuri et al. (2015) require that the Hessian is independent of y, which we will make use of later as well. This holds for Generalized Linear Models:

Definition 4.8. *A* generalized linear model (GLM) *is a model* $\mathrm{p}(y \mid \hat{z} = \hat{f}(x; \omega))$ *such that* $\log \mathrm{p}(y \mid \hat{z}) = \hat{z}^T T(y) - A(\hat{z}) + \log h(y)$ *is a distribution of the exponential family, independent of* ω*, and* $\hat{f}(x; \omega) = \omega^T x$ *is linear in the parameters* ω*.*

Proposition 4.9. *The Hessian* $\mathrm{H}''[y \mid x, \omega^*]$ *of a GLM is independent of* y*.*

$$\begin{aligned} \mathrm{H}''[y \mid x, \omega^*] &= \nabla_\omega \hat{f}(x; \omega^*)^T \, \nabla_{\hat{z}}^2 \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)] \, \nabla_\omega \hat{f}(x; \omega^*) & (33) \\ &= \nabla_\omega \hat{f}(x; \omega^*)^T \, \nabla_{\hat{z}}^2 A(w^T x) \, \nabla_\omega \hat{f}(x; \omega^*). & (34) \end{aligned}$$

Proposition 4.10. *For a model such that the Hessian* $\mathrm{H}''[y \mid x, \omega^*]$ *is independent of* y*, we have:*

$$\mathrm{F}(x, \omega^*) = \mathrm{H}''[y^* \mid x, \omega^*] \tag{35}$$

for any y^**, and:*

$$\mathbb{E}_{\mathrm{p}(y \mid x)}[\mathrm{H}''[y \mid x, \omega^*]] = \mathrm{F}(x, \omega^*). \tag{36}$$

Note that the expectation is over $\mathrm{p}(y \mid x)$ and not $\mathrm{p}(y \mid x, \omega^*)$, and $\mathbb{E}_{\mathrm{p}(\{y_i\} \mid \{x_i\})}[\mathrm{H}''[\{y_i\} \mid \{x_i\}, \omega^*]] = \mathrm{F}(\{x_i\}, \omega^*)$ is additive.

Proposition 4.11. *When* $\hat{f}(x; \omega) : \mathbb{R}^D \to \mathbb{R}^C$*, where* C *is the number of classes (outputs) and* D *is the number of input dimensions, and* $\omega \in \mathbb{R}^{D \times C}$*, and assuming the parameters are flattened into a single vector for the Jacobian,* $\nabla_\omega \hat{f}(x; \omega^*) = \mathrm{Id}_C \otimes x^T \in \mathbb{R}^{C \times (C \cdot D)}$*, where* $\otimes$ *denotes the Kronecker product, we have:*

$$\nabla_\omega \hat{f}(x; \omega^*)^T \, \nabla_{\hat{z}}^2 A(\omega^T x) \, \nabla_\omega \hat{f}(x; \omega^*) = \nabla_{\hat{z}}^2 A(w^T x) \otimes x \, x^T. \tag{37}$$

$\mathrm{p}(y \mid x, \omega^*)$ vs $\mathrm{p}(y \mid x)$. Having a GLM solves an important issue we will encounter in §5: approximating the EIG requires taking an expectation over $\mathrm{p}(y \mid x)$ and not $\mathrm{p}(y \mid x, \omega^*)$. One can approximate $\mathrm{p}(y \mid x) \approx \mathrm{p}(y \mid x, \omega^*)$, which can be justified in the limit, but this is likely not a good approximation in the cases interesting for active learning and active sampling. With a GLM, this is not a problem.

Generalized Gauss-Newton Approximation. When we have an exponential family but not a GLM, the equality in Proposition 4.9 is often used as an approximation for the Hessian, i.e., we simply use the Fisher information as an approximation of the Hessian (via Proposition 4.7):

$$\mathrm{H}''[y \mid x, \omega^*] \approx \nabla_\omega \hat{f}(x;\omega^*)^T \, \nabla^2_{\hat{z}} A(\omega^T x) \, \nabla_\omega \hat{f}(x;\omega^*) = \mathrm{F}(x, \omega^*). \tag{38}$$

This is known as *Generalized Gauss-Newton (GGN) approximation* (Kunstner et al., 2019; Immer et al., 2020). This approximation has the advantage that it is always positive semidefinite unlike the true Hessian.

Last-Layer Approaches. GLMs are often used in deep active learning (Gal et al., 2017; Ash et al., 2019; 2021). If we split the model into $\mathrm{p}(y \mid x, \omega) = \mathrm{p}(y \mid z = \omega^T f(x))$, where $z = f(x)$ are the embeddings, and treat the encoder $f(x)$ as fixed, we obtain a GLM that only uses the embeddings and last-layer weights.

5 Approximating Information Quantities

We use the results so far to derive various approximations for information quantities using the Hessian and Fisher information. This will help us connect information quantities to existing literature in §7.

5.1 Approximate Expected Information Gain

The expected information gain is a popular acquisition function in active learning, where it is also known as BALD, and in Bayesian optimal experimental design (Gal et al., 2017; Houlsby et al., 2011; Lindley, 1956).

We can approximate the EIG $\mathrm{I}[\Omega; \{Y_i^{\text{acq}}\} | \{x_i^{\text{acq}}\}]$ of acquisition candidates $\{x_i^{\text{acq}}\}$ using the Fisher information:

$$\begin{aligned}
\mathrm{I}[\Omega; \{Y_i^{\text{acq}}\} \mid \{x_i^{\text{acq}}\}] &= \mathrm{H}[\Omega] - \mathrm{H}[\Omega \mid \{Y_i^{\text{acq}}\}, \{x_i^{\text{acq}}\}] && (39)\\
&= \mathrm{H}[\Omega] - \mathbb{E}_{\mathrm{p}(\{y_i^{\text{acq}}\} | \{x_i^{\text{acq}}\})}[\mathrm{H}[\Omega \mid \{y_i^{\text{acq}}\}, \{x_i^{\text{acq}}\}]] && (40)\\
&\approx -\tfrac{1}{2} \log\det \mathrm{H}''[\omega^*] - \mathbb{E}_{\mathrm{p}(\{y_i^{\text{acq}}\} | \{x_i^{\text{acq}}\})}[-\tfrac{1}{2} \log\det \mathrm{H}''[\omega \mid \{y_i^{\text{acq}}\}, \{x_i^{\text{acq}}\}]] && (41)\\
&= \tfrac{1}{2}\, \mathbb{E}_{\mathrm{p}(\{y_i^{\text{acq}}\} | \{x_i^{\text{acq}}\})}[\log\det\big((\mathrm{H}''[\{y_i^{\text{acq}}\} \mid \{x_i^{\text{acq}}\}, \omega^*] + \mathrm{H}''[\omega^*])\, \mathrm{H}''[\omega^*]^{-1}\big)] && (42)\\
&= \tfrac{1}{2}\, \mathbb{E}_{\mathrm{p}(\{y_i^{\text{acq}}\} | \{x_i^{\text{acq}}\})}[\log\det\big(\mathrm{H}''[\{y_i^{\text{acq}}\} \mid \{x_i^{\text{acq}}\}, \omega^*]\; \mathrm{H}''[\omega^*]^{-1} + Id\big)]. && (43)
\end{aligned}$$

The constant C_k from Proposition 3.5 cancels out as we subtract two entropy terms.

Generalized Linear Model. When we have a GLM, we can use Proposition 4.10 to obtain:

$$\begin{aligned}
\mathrm{I}[\Omega; \{Y_i^{\text{acq}}\} \mid \{x_i^{\text{acq}}\}] &\approx \ldots = \tfrac{1}{2}\, \mathbb{E}_{\mathrm{p}(\{y_i^{\text{acq}}\} | \{x_i^{\text{acq}}\})}[\log\det\big(\mathrm{H}''[\{y_i^{\text{acq}}\} \mid \{x_i^{\text{acq}}\}, \omega^*]\; \mathrm{H}''[\omega^*]^{-1} + Id\big)] && (44)\\
&= \tfrac{1}{2} \log\det\big(\mathrm{F}(\{x_i^{\text{acq}}\}, \omega^*)\; \mathrm{H}''[\omega^*]^{-1} + Id\big). && (45)
\end{aligned}$$

We can upper-bound the log determinant and obtain:

$$\begin{aligned}
\mathrm{I}[\Omega; \{Y_i^{\text{acq}}\} \mid \{x_i^{\text{acq}}\}] &\approx \tfrac{1}{2} \log\det\big(\mathrm{F}(\{x_i^{\text{acq}}\}, \omega^*)\; \mathrm{H}''[\omega^*]^{-1} + Id\big) && (46)\\
&\leq \tfrac{1}{2}\, \mathrm{tr}\big(\mathrm{F}(\{x_i^{\text{acq}}\}, \omega^*)\; \mathrm{H}''[\omega^*]^{-1}\big) && (47)\\
&= \tfrac{1}{2} \sum_i \mathrm{tr}\big(\mathrm{F}({x^{\text{acq}}}_i, \omega^*)\; \mathrm{H}''[\omega^*]^{-1}\big)\,. && (48)
\end{aligned}$$

where we have used the following inequality (proof in §B.1):

Lemma 5.1. *For symmetric, positive-semidefinite matrices A, we have (with equality iff $A = 0$):*

$$\log\det(A + Id) \leq \mathrm{tr}(A). \tag{49}$$

General Case & Exponential Family. For the general case, we need to make a strong approximation:

$$\mathrm{p}(\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}) \approx \mathrm{p}(\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}, \omega^*), \tag{50}$$

which likely holds for a mostly converged posterior, but probably not for cases with little data. Alternatively, we could use the GGN approximation when we have an exponential family for the same result (but not an upper bound). See §B.1 for the derivation.

Proposition 5.2 (EIG)**.** *The expected information gain can be approximately upper bounded via:*

$$\mathrm{I}[\Omega; \{Y_i^{acq}\} \mid \{x_i^{acq}\}, \mathcal{D}^{train}] \overset{\approx}{\leq} \tfrac{1}{2}\log\det\left(\sum_i \mathrm{F}(x^{acq}, \omega^*)\ \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}]^{-1} + Id\right) \tag{51}$$

$$\leq \tfrac{1}{2}\sum_i \operatorname{tr}\left(\mathrm{F}(x^{acq}{}_i, \omega^*)\, \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}]^{-1}\right). \tag{52}$$

Furthermore, we have the following:

$$\arg\max_{\{x_i^{acq}\}}\{\mathrm{I}[\Omega; \{Y_i^{acq}\} \mid \{x_i^{acq}\}, \mathcal{D}^{train}]\} = \arg\max_{\{x_i^{acq}\}}\{-\mathrm{H}[\Omega \mid \{Y_i^{acq}\}, \{x_i^{acq}\}, \mathcal{D}^{train}]\}\ ,\ \textit{and} \tag{53}$$

$$-\mathrm{H}[\Omega \mid \{Y_i^{acq}\}, \{x_i^{acq}\}, \mathcal{D}^{train}] \overset{\approx}{\leq} \tfrac{1}{2}\log\det\left(\sum_i \mathrm{F}(x^{acq}, \omega^*) + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}]\right) - C_k. \tag{54}$$

The second statement follows from Equation (39), since $\mathrm{H}[\Omega \mid \mathcal{D}^{\mathrm{train}}]$ is constant, and provides an alternative objective when we are only interested in optimization of the EIG. We use it together to show a connection to the expected gradient length approach in active learning in §7.

Batch Acquisition Pathologies. Importantly, this approximation of the EIG using the trace is additive, whereas the one using the log determinant is not. This means that the trace approximation ignores dependencies between samples and only leads to naive top-k batch acquisition; see Kirsch et al. (2019; 2021a) for details. On the other hand, the log determinant of the Fisher information version might well capture these dependencies.

5.2 Approximate Information Gain

Following the same steps, we can also approximate the information gain, which is useful for active sampling:

Proposition 5.3 (IG)**.** *The* information gain $\mathrm{I}[\Omega; \{y_i^{acq}\} \mid \{x_i^{acq}\}, \mathcal{D}^{train}] = \mathrm{H}[\Omega \mid \mathcal{D}^{train}] - \mathrm{H}[\Omega \mid \{y_i^{acq}\}, \{x_i^{acq}\}, \mathcal{D}^{train}]$ *can be approximately upper bounded via:*

$$\mathrm{I}[\Omega; \{y_i^{acq}\} \mid \{x_i^{acq}\}, \mathcal{D}^{train}] \overset{\approx}{\leq} \tfrac{1}{2}\log\det\left(\mathrm{H}''[\{y_i^{acq}\} \mid \{x_i^{acq}\}, \omega^*]\ \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}]^{-1} + Id\right) \tag{55}$$

$$\leq \tfrac{1}{2}\sum_i \operatorname{tr}\left(\mathrm{H}''[\{y_i^{acq}\} \mid \{x_i^{acq}\}, \omega^*]\ \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}]^{-1}\right). \tag{56}$$

Furthermore, we have the following:

$$\arg\max_{\{x_i^{acq}\}}\{\mathrm{I}[\Omega; \{y_i^{acq}\} \mid \{x_i^{acq}\}, \mathcal{D}^{train}]\} = \arg\max_{\{x_i^{acq}\}}\{-\mathrm{H}[\Omega \mid \{y_i^{acq}\}, \{x_i^{acq}\}, \mathcal{D}^{train}]\},\ \textit{and} \tag{57}$$

$$-\mathrm{H}[\Omega \mid \{y_i^{acq}\}, \{x_i^{acq}\}, \mathcal{D}^{train}] \overset{\approx}{\leq} \tfrac{1}{2}\log\det\left(\mathrm{H}''[\{y_i^{acq}\} \mid \{x_i^{acq}\}, \omega^*] + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}]\right) - C_k. \tag{58}$$

Comparison to EIG. Importantly, when we have a GLM or use the GGN approximation, this approximation of the IG is equal to the one of the EIG. This tells us that active learning on a GLM with the EIG approximation will work as well as if we had access to the labels. On the other hand, active sampling via IG with the GGN approximation will not work better than the equivalent active learning approach without labels.

5.3 Approximate Expected Predictive Information Gain

In transductive active learning, we have access to an (empirical) distribution $\hat{\mathrm{p}}(x^{\text{eval}})$ (e.g., the pool set) and want to find x^{acq} that maximizes the *expected predictive information gain (EPIG)* (Kirsch et al., 2021b):

$$\arg\max_{x^{\text{acq}}} \mathrm{I}[Y^{\text{eval}}; Y^{\text{acq}} \mid X^{\text{eval}}, x^{\text{acq}}, \mathcal{D}^{\text{train}}] = \arg\max_{x^{\text{acq}}} \mathbb{E}_{\hat{\mathrm{p}}(x^{\text{eval}})} \mathrm{I}[Y^{\text{eval}}; Y^{\text{acq}} \mid x^{\text{eval}}, x^{\text{acq}}, \mathcal{D}^{\text{train}}], \tag{59}$$

We expand the objective as follows:

$$\mathrm{I}[Y^{\text{eval}}; Y^{\text{acq}} \mid X^{\text{eval}}, x^{\text{acq}}] = \mathrm{I}[\Omega; Y^{\text{eval}} \mid X^{\text{eval}}] - \mathrm{I}[\Omega; Y^{\text{eval}} \mid X^{\text{eval}}, Y^{\text{acq}}, x^{\text{acq}}], \tag{60}$$

where $\mathrm{I}[\Omega; Y^{\text{eval}} \mid X^{\text{eval}}]$ can be removed from the objective because it is independent of x^{acq}, thus, optimizing EPIG is equivalent to *minimizing* $\mathrm{I}[\Omega; Y^{\text{eval}} \mid X^{\text{eval}}, Y^{\text{acq}}, x^{\text{acq}}]$:

$$\arg\max_{x^{\text{acq}}} \mathrm{I}[Y^{\text{eval}}; Y^{\text{acq}} \mid X^{\text{eval}}, x^{\text{acq}}] = \arg\min_{x^{\text{acq}}} \mathrm{I}[\Omega; Y^{\text{eval}} \mid X^{\text{eval}}, Y^{\text{acq}}, x^{\text{acq}}]. \tag{61}$$

Following Proposition 5.2, this can be approximated by:

$$\begin{aligned} &\mathrm{I}[\Omega; Y^{\text{eval}} \mid X^{\text{eval}}, Y^{\text{acq}}, x^{\text{acq}}] \\ &\quad \approx \tfrac{1}{2}\, \mathbb{E}_{\mathrm{p}(y^{\text{eval}}, y^{\text{acq}} \mid x^{\text{eval}}, x^{\text{acq}})\, \hat{\mathrm{p}}(x^{\text{eval}})} [\log\det \left(\mathrm{H}''[y^{\text{eval}} \mid x^{\text{eval}}, \omega^*] \,(\mathrm{H}''[y^{\text{acq}} \mid x^{\text{acq}}, \omega^*] + \mathrm{H}''[\omega^*])^{-1} + Id\right)]. \end{aligned} \tag{62}$$

Generalized Linear Model. For a generalized linear model, we can drop the expectation and obtain:

$$\begin{aligned} &\mathrm{I}[\Omega; Y^{\text{eval}} \mid X^{\text{eval}}, Y^{\text{acq}}, x^{\text{acq}}] \\ &\quad \approx \tfrac{1}{2}\, \mathbb{E}_{\hat{\mathrm{p}}(x^{\text{eval}})} [\log\det \left(\mathrm{F}(x^{\text{eval}}, \omega^*)\, (\mathrm{F}(x^{\text{acq}}, \omega^*) + \mathrm{H}''[\omega^*])^{-1} + Id\right)] && (63) \\ &\quad \leq \tfrac{1}{2} \log\det \left(\mathbb{E}_{\hat{\mathrm{p}}(x^{\text{eval}})}[\mathrm{F}(x^{\text{eval}}, \omega^*)]\, (\mathrm{F}(x^{\text{acq}}, \omega^*) + \mathrm{H}''[\omega^*])^{-1} + Id\right) && (64) \\ &\quad \leq \tfrac{1}{2} \operatorname{tr} \left(\mathbb{E}_{\hat{\mathrm{p}}(x^{\text{eval}})} \left[\mathrm{F}(x^{\text{eval}}, \omega^*)\right] (\mathrm{F}(x^{\text{acq}}, \omega^*) + \mathrm{H}''[\omega^*])^{-1}\right), && (65) \end{aligned}$$

where we have again used the concavity of the log determinant.

General Case & Exponential Family. To our knowledge, there is no rigorous way to obtain a similar result in the general case as the Fisher information for an acquisition candidate now lies within an inverted term. Obviously, the GGN approximation can be applied when we have an exponential family, which leads back to the above GLM result as an approximation. See §B.2 for more details.

Proposition 5.4 (EPIG)**.** *For a generalized linear model (or with the GGN approximation), we have:*

$$\arg\max_{\{x_i^{acq}\}} \mathrm{I}[Y^{eval}; \{Y_i^{acq}\} \mid X^{eval}, \{x_i^{acq}\}, \mathcal{D}^{train}] = \arg\min_{\{x_i^{acq}\}} \mathrm{I}[\Omega; Y^{eval} \mid X^{eval}, \{Y_i^{acq}\}, \{x_i^{acq}\}, \mathcal{D}^{train}], \tag{66}$$

with

$$\begin{aligned} &\mathrm{I}[\Omega; Y^{eval} \mid X^{eval}, \{Y_i^{acq}\}, \{x_i^{acq}\}, \mathcal{D}^{train}] \\ &\quad \overset{\approx}{\leq} \tfrac{1}{2} \log\det \left(\mathbb{E}_{\mathrm{p}(x^{eval})}[\mathrm{F}(x^{eval}, \omega^*)]\, (\mathrm{F}(\{x_i^{acq}\}, \omega^*) + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}])^{-1} + Id\right) && (67) \\ &\quad \leq \tfrac{1}{2} \operatorname{tr} \left(\mathbb{E}_{\mathrm{p}(x^{eval})} \left[\mathrm{F}(x^{eval}, \omega^*)\right] (\mathrm{F}(\{x_i^{acq}\}, \omega^*) + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}])^{-1}\right). && (68) \end{aligned}$$

Approximations for JEPIG, PIG and JPIG. The results can be found in §B.

6 Similarity Matrices and One-Sample Approximations of the Fisher Information

Many active learning methods do not use Fisher information, but use a kernel-based approach using similarity matrices of the loss gradients $\mathrm{H}'[y^* \mid x]$, where y^* is either the true label or a hypothesized label $y*$ when no label is available (usually using the $\arg\max$ prediction of the model).

Connection to Fisher Information. Crucially, given $\mathcal{D} = \{(y_i, x_i)_i\}$, if we let

$$\hat{\mathrm{H}}'[\mathcal{D} \mid \omega^*] := \begin{pmatrix} \vdots \\ \mathrm{H}'[y_i \mid x_i, \omega^*] \\ \vdots \end{pmatrix} \tag{69}$$

be a "data matrix" of the Jacobians, then $\hat{\mathrm{H}}'[\mathcal{D} \mid \omega^*] \hat{\mathrm{H}}'[\mathcal{D} \mid \omega^*]^T$ yields the similarity matrix $S[\mathcal{D} \mid \omega^*]$ using the Euclidean inner product:

$$S[\mathcal{D} \mid \omega]_{ij} := \langle \mathrm{H}'[y_i \mid x_i, \omega^*], \mathrm{H}'[y_j \mid x_j, \omega^*] \rangle = \hat{\mathrm{H}}'[\mathcal{D} \mid \omega^*] \hat{\mathrm{H}}'[\mathcal{D} \mid \omega^*]^T. \tag{70}$$

If we sample the $\{y_i\} \sim \mathrm{p}(\{y_i\} \mid \{x_i\}, \omega^*)$, the "flipped" product $\hat{\mathrm{H}}'[\mathcal{D} \mid \omega^*]^T \hat{\mathrm{H}}'[\mathcal{D} \mid \omega^*]$ yields a one-sample estimate of the Fisher information $\mathrm{F}(\{x_i\}, \omega^*)$:

$$\mathrm{F}(\{x_i\}, \omega^*) = \sum_i \mathrm{F}(x_i, \omega^*) = \mathbb{E}_{\mathrm{p}(\{y_i\} \mid \{x_i\}, \omega^*)} \sum_i \mathrm{H}'[y_i \mid x_i \omega^*]^T \, \mathrm{H}'[y_i \mid x_i \omega^*] \tag{71}$$

$$= \mathbb{E}_{\mathrm{p}(\{y_i\} \mid \{x_i\}, \omega^*)} \hat{\mathrm{H}}'[\mathcal{D} \mid \omega^*]^T \hat{\mathrm{H}}'[\mathcal{D} \mid \omega^*]. \tag{72}$$

If we do not sample but instead use the arg max class, we only obtain a biased estimate (Kunstner et al., 2019, §B).

Connection to the Expected Information Gain. More importantly, when we define an inner product $\langle \cdot, \cdot \rangle_{\mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]}$ using the Hessian $\mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]$, we can connect the similarity matrix, which uses this inner product, to our information gain approximations:

$$S_{\mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]}[\mathcal{D} \mid \omega^*] := \hat{\mathrm{H}}'[\mathcal{D} \mid \omega^*] \, \mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]^{-1} \hat{\mathrm{H}}'[\mathcal{D} \mid \omega^*]^T \tag{73}$$

Specifically, we apply the matrix determinant lemma $\det(AB + M) = \det M \det(Id + BM^{-1}B)$ to obtain:

Proposition 6.1. *Given $\mathcal{D}^{train}$, $\{x_i^{acq}\}$ and (sampled) $\{y_i^{acq}\}$, we have for the EIG:*

$$\mathrm{I}[\Omega; \{Y_i^{acq}\} \mid \{x_i^{acq}\}, \mathcal{D}^{train}] \overset{\approx}{\leq} \tfrac{1}{2} \log\det \left(S_{\mathrm{H}''[\omega^* \mid \mathcal{D}^{train}]}[\mathcal{D}^{acq} \mid \omega^*] + Id \right) \tag{74}$$

Proposition 6.2. *Assuming an uninformative posterior $\mathrm{H}''[\omega^* \mid \mathcal{D}^{train}] = \lambda Id$ for $\lambda \to 0$, and given $\mathcal{D}^{train}$, $\{x_i^{acq}\}$, and (sampled) $\{y_i^{acq}\}$, we have for the EIG (before taking $\lambda \to 0$):*

$$\mathrm{I}[\Omega; \{Y_i^{acq}\} \mid \{x_i^{acq}\}, \mathcal{D}^{train}] \overset{\approx}{\leq} \tfrac{1}{2} \log\det \left(S[\mathcal{D}^{acq} \mid \omega^*] + \lambda Id \right) - \tfrac{|\mathcal{D}^{acq}|}{2} \log \lambda. \tag{75}$$

The second term is constant independently of $\{x_i^{acq}\}$, and we can maximize $\{x_i^{acq}\}$ in $\log\det \left(S[\mathcal{D}^{acq} \mid \omega^*] + \lambda Id \right)$ and then take the limit $\lambda \to 0$. Thus, we can use the following proxy objective:

$$\log\det \left(S[\mathcal{D}^{acq} \mid \omega^*] \right). \tag{76}$$

Connection to Other Approximate Information Quantities. Interestingly, we can use the above approach to obtain valid approximations for the predictive information gains (EPIG and JEPIG) because the terms that would go towards $-\infty$ cancel out; see §C for details.

7 Connection to Other Acquisition Functions in the Literature

Here, we use the results so far to connect approaches in the literature to information quantities explicitly.

7.1 BAIT in "Gone Fishing" (Ash et al., 2021) and ActiveSetSelect in "Convergence Rates of Active Learning for Maximum Likelihood Estimation" (Chaudhuri et al., 2015)

While Chaudhuri et al. (2015) and Ash et al. (2021) use a similar objective, Chaudhuri et al. (2015) apply it only to GLMs, whereas Ash et al. (2021) use it for deep learning and introduce it as the *BAIT* objective:

$$\arg\min_{\{x_i^{\text{acq}}\}} \operatorname{tr}\left(\left(\mathrm{F}(\{x_i^{\text{acq}}\},\omega^*) + \mathrm{F}(x^{\text{train}},\omega^*) + \lambda I\right)^{-1} \mathrm{F}(\{x_i^{\text{eval}}\},\omega^*)\right), \tag{BAIT}$$

where λ is a hyperparameter. While Ash et al. (2021) use DNNs, they only use the last layer to compute the Fisher information. The last layer of the DNNs together with appropriate activation functions and losses constitute a generalized linear model. Following Proposition 5.4, Ash et al. (2021) thus perform transductive active learning (using the pool set as an evaluation set) and approximate EPIG as information quantity:

Proposition 7.1. *Both Chaudhuri et al. (2015) and Ash et al. (2021) perform transductive active learning, approximating EPIG:*

$$\arg\max_{x^{acq}} \mathrm{I}[Y^{eval}; Y^{acq} \mid X^{eval}, x^{acq}] \approx \arg\min_{x^{acq}} \mathrm{tr}(\mathrm{F}(x^{eval},\omega^*)\,(\sum_i \mathrm{F}(x^{acq}{}_i,\omega^*) + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}])^{-1}), \tag{77}$$

with $\mathrm{H}''[\omega^* \mid \mathcal{D}^{train}] = \mathrm{H}''[\mathcal{D}^{train} \mid \omega^*] + \mathrm{H}''[\omega^*]$ *and* $\mathrm{H}''[\omega^*] = \lambda Id$*.*

Proof. This follows immediately for GLM (last-layer approaches) when we expand $\mathrm{H}''[\omega^* \mid \mathcal{D}^{\text{train}}]$. ☐

However, following §4, it does not seem that this approach will translate beyond a last-layer approach for DNNs. It thus remains an open question to find a principled approach for the general case and to go beyond last-layer active learning when using Fisher information without using the GGN approximation.

7.2 BADGE (Ash et al., 2019)

BADGE maximizes an approximation of the EIG.

Following §6, BADGE uses one-sample estimates with hard pseudo-labels $y^{\text{acq}} = \arg\max_y \mathrm{p}(y \mid x^{\text{acq}}, \omega^*)$, and its sample selection is motivated by a k-DPP (Kulesza and Taskar, 2011) on the similarity matrix $S[\mathcal{D}^{\text{acq}} \mid \omega^*]$ using the pseudo-labels. This can be seen as approximating $\mathrm{I}[\Omega; \{Y_i^{\text{acq}}\} \mid \{x_i^{\text{acq}}\}, \mathcal{D}^{\text{train}}]$ with an uninformative posterior distribution following Proposition 6.1. However, BADGE actually uses k-MEANS++ (Arthur and Vassilvitskii, 2006; Ostrovsky et al., 2013) instead of a k-DPP to select samples to further speed up acquisition: it uses the Jacobians from the data matrix directly and samples a diverse batch based on the Euclidean distance between the Jacobians. We leave a comparison between k-DDP and k-means++ to future work and refer to the ablation in Ash et al. (2019) to see that the performance matches.

7.3 SIMILAR (Kothawade et al., 2021) and PRISM (Kothawade et al., 2022)

The LogDet objective maximizes an approximation of the EIG, and the LogDetMI objective of EPIG.

The results based on the log determinant of similarity matrices are reported as among the best in the experiments of Kothawade et al. (2021) and Kothawade et al. (2022). The LogDet objective $\log\det S[\mathcal{D}^{\text{acq}} \mid \omega^*]$ exactly matches the EIG approximation in §6.2. Furthermore, in §D.1, we show that the LogDetMI objective matches an approximation of JEPIG and simillary, re-derive the LogDetCMI objective.

7.4 Expected Gradient Length

Expected Gradient Length (EGL) (Settles et al., 2007; Settles, 2009) is usually defined for non-Bayesian models. Originally, it was an expectation over the gradient norm. In more recent literature (Huang et al., 2016), it is introduced using the squared gradient norm:

$$\mathbb{E}_{\mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*)} \left\| \mathrm{H}'[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*] \right\|^2. \tag{EGL}$$

Using a diagonal approximation of the Fisher information, we obtain:

Proposition 7.2. *The EIG for a candidate sample x^{acq} approximately lower-bounds the EGL:*

$$2\, \mathrm{I}[\Omega; Y^{acq} \mid x^{acq}] \overset{\approx}{\leq} \mathbb{E}_{\mathrm{p}(y^{acq} \mid x^{acq}, \omega^*)} \left\| \mathrm{H}'[y^{acq} \mid x^{acq}, \omega^*] \right\|^2 + const.. \tag{78}$$

7.5 Deep Learning on a Data Diet

Paul et al. (2021) use the gradient length of given labeled samples x, y (averaged over multiple training runs) to select a subset of informative samples for training:

$$\mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D}^{\mathrm{train}})} \left\| \mathrm{H}'[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega] \right\|^2, \tag{GraNd}$$

which they call the *gradient norm score (GraNd)*.

Proposition 7.3. *The IG for a candidate sample x^{acq} approximately lower-bounds the gradient norm (GraNd) score at ω^* up to a second-order term:*

$$2\, \mathrm{I}[\Omega; y^{acq} \mid x^{acq}] \overset{\approx}{\leq} \mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D}^{train})}[\left\| \mathrm{H}'[y^{acq} \mid x^{acq}, \omega] \right\|^2] - \mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D}^{train})}[\mathrm{tr}\left(\frac{\nabla^2_\omega\, \mathrm{p}(y \mid x, \omega)}{\mathrm{p}(y \mid x, \omega)} \right)] + const. \tag{79}$$

The second term might not be negligible. Hence, GraNd (the first term on the left) might deviate from the information gain. This raises the question how the information gain compares to GraNd in practice.

8 Conclusion

We have looked at Fisher information and Laplace approximations and have derived weight-space approximations for the expected information gain and expected predictive information gain. This has allowed us to connect various information quantities to objectives already used in the literature.

Last-Layer Active Learning. Methods that only use last-layer Fisher information or similar perform active learning on the embeddings only, despite that feature learning is arguably the most important feature of deep neural networks. However, these approaches can find great use with large pre-trained models, which are only fine-tuned on new data domains anyway (Tran et al., 2022).

Batch Acquisition Pathologies. Approaches that use the matrix trace instead of the log determinant are additive in the batch candidates $\{x_i^{\mathrm{acq}}\}$ and can thus by definition not take redundancies between batch candidates into account, leading to failures detailed in Kirsch et al. (2019; 2021a).

Weight vs. Prediction Space. Ash et al. (2019; 2021); Kothawade et al. (2021; 2022) approximate the relevant information quantity in weight space, while Kirsch et al. (2021b); Houlsby et al. (2011); Kirsch et al. (2019); Mindermann et al. (2022) approximate the relevant information quantities in prediction space.

Informativeness. Taking a step back, we have seen that a Bayesian perspective using information quantities connects seemingly disparate literature. Although Bayesian methods are often seen as separate from regular active learning and active sampling, the sometimes fuzzy notion of "informativeness" expressed through various different objectives in non-Bayesian settings collapses to the same couple of information quantities, which were, in principle, already well known by Lindley (1956) and MacKay (1992).

References

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, 2006.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds, 2019.

Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with fisher embeddings, 2021.

Kamalika Chaudhuri, Sham Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation, 2015.

Thomas M Cover and A Thomas. Determinant inequalities via information theory. *SIAM journal on Matrix Analysis and Applications*, 9(3):384–392, 1988.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, 7, 1994.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. Active learning for speech recognition: the power of gradients, 2016.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization, 2020.

Rishabh Iyer, Ninad Khargoankar, Jeff Bilmes, and Himanshu Asanani. Submodular combinatorial information measures with applications in machine learning. In *Algorithmic Learning Theory*, pages 722–754. PMLR, 2021.

Andreas Kirsch and Yarin Gal. A practical & unified notation for information-theoretic quantities in ml. *arXiv preprint arXiv:2106.12062*, 2021.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, pages 7024–7035, 2019.

Andreas Kirsch, Sebastian Farquhar, Parmida Atighehchian, Andrew Jesson, Frederic Branchaud-Charron, and Yarin Gal. Stochastic batch acquisition for deep active learning, 2021a.

Andreas Kirsch, Tom Rainforth, and Yarin Gal. Test distribution-aware active learning: A principled approach against distribution shift and outliers, 2021b.

Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. Similar: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems*, 34:18685–18697, 2021.

Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. Prism: A rich class of parameterized submodular information measures for guided data subset selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10238–10246, 2022.

Alex Kulesza and Ben Taskar. k-dpps: Fixed-size determinantal point processes. In *ICML*, 2011.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent, 2019.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

Quan Long. Multimodal information gain in bayesian design of experiments, 2021.

David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4): 590–604, 1992.

Sören Mindermann, Jan Brauner, Muhammed Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt, 2022.

Rafail Ostrovsky, Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59(6):1–22, 2013.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training, 2021.

Burr Settles. Active learning literature survey. 2009.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Advances in neural information processing systems*, 20, 2007.

Shengyang Sun, Daniele Calandriello, Huiyi Hu, Ang Li, and Michalis Titsias. Information-theoretic online memory selection for continual learning, 2022.

Dustin Tran, Jeremiah Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, et al. Plex: Towards reliability using pretrained large model extensions. *arXiv preprint arXiv:2207.07411*, 2022.

Chaoqi Wang, Shengyang Sun, and Roger Grosse. Beyond marginal uncertainty: How accurately can bayesian regression models estimate posterior predictive correlations?, 2020.

Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pages 1081–1088, 2006.

A Fisher Information: Additional Derivations & Proofs

Proposition 4.2. *Like the observed information, the Fisher information is additive:*

$$\mathrm{F}(\{x_i\}, \omega^*) = \sum_i \mathrm{F}(x_i, \omega^*). \tag{23}$$

Proof. This follows immediately from $Y_i \perp\!\!\!\perp Y_j \mid x_i, x_j, \omega^*$ for $i \neq j$ and the additivity of the observed information:

$$\mathrm{F}(\{x_i\}, \omega^*) = \mathbb{E}_{\mathrm{p}(\{y_i\} \mid \{x_i\}, \omega^*)}[\mathrm{H}''[\{y_i\} \mid \{x_i\}, \omega^*]] = \mathbb{E}_{\mathrm{p}(\{y_i\} \mid \{x_i\}, \omega^*)}[\sum_i \mathrm{H}''[y_i \mid x_i, \omega^*]] \tag{80}$$

$$= \sum_i \mathbb{E}_{\mathrm{p}(y_i \mid x_i, \omega^*)}[\mathrm{H}''[y_i \mid x_i, \omega^*]] = \sum_i \mathrm{F}(x_i, \omega^*). \tag{81}$$

□

Proposition 4.3. *The Fisher information is equivalent to:*

$$\mathrm{F}(x, \omega^*) = \mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[\mathrm{H}'[y \mid x, \omega^*]^T \ \mathrm{H}'[y \mid x, \omega^*]] = \mathrm{Cov}[\mathrm{H}'[Y \mid x, \omega^*]]. \tag{24}$$

To prove Proposition 4.3, we use the two lemmas below:

Lemma 4.4. *For the Jacobian* $\mathrm{H}'[y \mid x, \omega^*]$*, we have:*

$$\mathrm{H}'[y \mid x, \omega^*] = -\frac{\nabla_\omega \, \mathrm{p}(y \mid x, \omega^*)}{\mathrm{p}(y \mid x, \omega^*)}, \tag{25}$$

and for the Hessian $\mathrm{H}''[y \mid x, \omega^*]$*, we have:*

$$\mathrm{H}''[y \mid x, \omega^*] = \mathrm{H}'[y \mid x, \omega^*]^T \ \mathrm{H}'[y \mid x, \omega^*] - \frac{\nabla^2_\omega \, \mathrm{p}(y \mid x, \omega^*)}{\mathrm{p}(y \mid x, \omega^*)}. \tag{26}$$

Proof. The result follows immediately from the application of the rules of multivariate calculus. □

Lemma 4.5. *The following expectations over the model's own predictions vanish:*

$$\mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[\mathrm{H}'[y \mid x, \omega^*]] = 0, \tag{27}$$

$$\mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}\left[\frac{\nabla^2_\omega \, \mathrm{p}(y \mid x, \omega^*)}{\mathrm{p}(y \mid x, \omega^*)}\right] = 0. \tag{28}$$

Proof. We use the previous equivalences and rewrite the expectations as integral; the results follows:

$$\mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[\mathrm{H}'[y \mid x, \omega^*]] = \mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[-\nabla_\omega \log \mathrm{p}(y \mid x, \omega^*)] = -\,\mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}\left[\frac{\nabla_\omega \, \mathrm{p}(y \mid x, \omega^*)}{\mathrm{p}(y \mid x, \omega^*)}\right] \tag{82}$$

$$= -\int \nabla_\omega \, \mathrm{p}(y \mid x, \omega^*) \, dy = -\nabla_\omega \int \mathrm{p}(y \mid x, \omega^*) \, dy = -\nabla_\omega 1 = 0, \tag{83}$$

$$\mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}\left[\frac{\nabla^2_\omega \, \mathrm{p}(y \mid x, \omega^*)}{\mathrm{p}(y \mid x, \omega^*)}\right] = \int \nabla^2_\omega \, \mathrm{p}(y \mid x, \omega^*) \, dy = \nabla^2_\omega \int \mathrm{p}(y \mid x, \omega^*) \, dy = \nabla^2_\omega 1 = 0. \tag{84}$$

□

Proof of Proposition 4.3. With the previous lemma, we have the following.

$$\mathrm{Cov}[\mathrm{H}'[Y \mid x, \omega^*]] = \mathbb{E}[\mathrm{H}'[Y \mid x, \omega^*]^T \ \mathrm{H}'[Y \mid x, \omega^*]] - \underbrace{\mathbb{E}[\mathrm{H}'[Y \mid x, \omega^*]^T]}_{=0} \ \underbrace{\mathbb{E}[\mathrm{H}'[Y \mid x, \omega^*]]}_{=0} \tag{85}$$

$$= \mathbb{E}[\mathrm{H}'[Y \mid x, \omega^*]^T \; \mathrm{H}'[Y \mid x, \omega^*]]. \tag{86}$$

For the expectation over the Hessian, we plug Lemma 4.4 into Lemma 4.5 and obtain:

$$\mathrm{F}(x, \omega^*) = \mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[\mathrm{H}''[y \mid x, \omega^*]] = \mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}\left[\mathrm{H}'[y \mid x, \omega^*]^T \; \mathrm{H}'[y \mid x, \omega^*] - \frac{\nabla^2_\omega \, \mathrm{p}(y \mid x, \omega^*)}{\mathrm{p}(y \mid x, \omega^*)}\right] \tag{87}$$

$$= \mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[\mathrm{H}'[y \mid x, \omega^*]^T \; \mathrm{H}'[y \mid x, \omega^*]] - 0 = \mathrm{Cov}[\mathrm{H}'[Y \mid x, \omega^*]]. \tag{88}$$

□

A.1 Special Case: Exponential Family

Proposition 4.6. *The Fisher information* $\mathrm{F}(\{x_i\}, \omega^*)$ *for a model* $\mathrm{p}(y \mid \hat{z} = \hat{f}(x; \omega^*))$ *is equivalent to:*

$$\mathrm{F}(x, \omega^*) = \nabla_\omega \hat{f}(x; \omega^*)^T \, \mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[\nabla^2_{\hat{z}} \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)]] \nabla_\omega \hat{f}(x; \omega^*), \tag{29}$$

where $\nabla^2_{\hat{z}} \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)]$ *is short for* $\nabla^2_{\hat{z}} \, \mathrm{H}[y \mid \hat{z}]\big|_{\hat{z} = \hat{f}(x; \omega^*)}$.

Proof. We apply the second equivalence in Proposition 4.3 twice:

$$\mathrm{F}(x, \omega^*) = \mathrm{Cov}[\mathrm{H}'[Y \mid x, \omega^*]] = \mathrm{Cov}[\nabla_\omega \hat{f}(x; \omega^*)^T \; \nabla_{\hat{z}} \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)] \; \nabla_\omega \hat{f}(x; \omega^*)] \tag{89}$$

$$= \nabla_\omega \hat{f}(x; \omega^*)^T \; \mathrm{Cov}[\nabla_{\hat{z}} \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)]] \; \nabla_\omega \hat{f}(x; \omega^*) \tag{90}$$

$$= \nabla_\omega \hat{f}(x; \omega^*)^T \; \mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[\nabla^2_{\hat{z}} \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)]] \; \nabla_\omega \hat{f}(x; \omega^*) \tag{91}$$

□

A.2 Special Case: Generalized Linear Models

Proposition 4.9. *The Hessian* $\mathrm{H}''[y \mid x, \omega^*]$ *of a GLM is independent of* y.

$$\mathrm{H}''[y \mid x, \omega^*] = \nabla_\omega \hat{f}(x; \omega^*)^T \; \nabla^2_{\hat{z}} \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)] \; \nabla_\omega \hat{f}(x; \omega^*) \tag{33}$$

$$= \nabla_\omega \hat{f}(x; \omega^*)^T \; \nabla^2_{\hat{z}} A(w^T x) \; \nabla_\omega \hat{f}(x; \omega^*). \tag{34}$$

Proof.

$$\mathrm{H}''[y \mid x, \omega^*] \tag{92}$$

$$= \nabla_\omega [\mathrm{H}'[y \mid x, \omega^*]] \tag{93}$$

$$= \nabla_\omega [\nabla_{\hat{z}} \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)] \; \nabla_\omega \hat{f}(x; \omega^*)] \tag{94}$$

$$= \nabla_{\hat{z}} \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)] \; \underbrace{\nabla^2_\omega \hat{f}(x; \omega^*)}_{= \nabla^2_\omega [w^T x] = 0} + \nabla_\omega \hat{f}(x; \omega^*)^T \; \nabla^2_{\hat{z}} \, \mathrm{H}[y \mid \hat{z} = \hat{f}(x; \omega^*)] \; \nabla_\omega \hat{f}(x; \omega^*) \tag{95}$$

$$= \nabla_\omega \hat{f}(x; \omega^*)^T \; \nabla^2_{\hat{z}} A(w^T x) \; \nabla_\omega \hat{f}(x; \omega^*). \tag{96}$$

□

Proposition 4.11. *When* $\hat{f}(x; \omega) : \mathbb{R}^D \to \mathbb{R}^C$*, where* C *is the number of classes (outputs) and* D *is the number of input dimensions, and* $\omega \in \mathbb{R}^{D \times C}$*, and assuming the parameters are flattened into a single vector for the Jacobian,* $\nabla_\omega \hat{f}(x; \omega^*) = \mathrm{Id}_C \otimes x^T \in \mathbb{R}^{C \times (C \cdot D)}$*, where* $\otimes$ *denotes the Kronecker product, we have:*

$$\nabla_\omega \hat{f}(x; \omega^*)^T \; \nabla^2_{\hat{z}} A(\omega^T x) \; \nabla_\omega \hat{f}(x; \omega^*) = \nabla^2_{\hat{z}} A(w^T x) \otimes x \, x^T. \tag{37}$$

Proof. We begin with a few statements that lead to the conclusion step by step, where $x \in \mathbb{R}^D$, $A \in \mathbb{R}^{C \times C}$, $G \in \mathbb{R}^{C \times (C \cdot D)}$:

$$
\begin{aligned}
(x\,x^T)_{ij} &= x_i\,x_j && (97)\\
(Id_C \otimes x^T)_{c,d\,D+i} &= x_i \cdot \mathbb{1}\{c = d\} && (98)\\
(G^T\,A\,G)_{ij} &= \sum_{k,l} G_{ki}\,A_{kl}\,G_{lj} && (99)\\
(A \otimes x\,x^T)_{c\,D+i,d\,D+j} &= A_{cd}\,x_i\,x_j, && (100)\\
((Id_C \otimes x^T)^T\,A\,(Id_C \otimes x^T))_{c\,D+i,d\,D+j} &= \sum_{k,l} (Id_C \otimes x^T)_{k,c\,D+i}\,A_{kl}\,(Id_C \otimes x^T)_{l,d\,D+j} && (101)\\
&= \sum_{k,l} x_i \cdot \mathbb{1}\{k = c\}\,A_{kl}\,x_j \cdot \mathbb{1}\{l = d\} && (102)\\
&= x_i\,A_{cd}\,x_j && (103)\\
&= (A \otimes x\,x^T)_{c\,D+i,d\,D+j}. && (104)\\
\implies \nabla_\omega \hat{f}(x;\omega^*)^T\,\nabla^2_{\hat{z}} A(\omega^T x)\,\nabla_\omega \hat{f}(x;\omega^*) &= \nabla^2_{\hat{z}} A(w^T x) \otimes x\,x^T. && (105)
\end{aligned}
$$

☐

Proposition 4.10. *For a model such that the Hessian $\mathrm{H}''[y \mid x, \omega^*]$ is independent of y, we have:*

$$\mathrm{F}(x, \omega^*) = \mathrm{H}''[y^* \mid x, \omega^*] \tag{35}$$

for any y^*, and:

$$\mathbb{E}_{\mathrm{p}(y \mid x)}[\mathrm{H}''[y \mid x, \omega^*]] = \mathrm{F}(x, \omega^*). \tag{36}$$

Proof. This follows directly from Proposition 4.7. In particular, we have:

$$\mathrm{F}(x, \omega^*) = \mathbb{E}_{\mathrm{p}(y \mid x, \omega^*)}[\mathrm{H}''[y \mid x, \omega^*]] = \mathrm{H}''[y^* \mid x, \omega^*], \tag{106}$$

where we have fixed y^* to an arbitrary value. ☐

B Approximating Information Quantities

B.1 Approximate Expected Information Gain

Lemma 5.1. *For symmetric, positive-semidefinite matrices A, we have (with equality iff $A = 0$):*

$$\log \det(A + Id) \leq \operatorname{tr}(A). \tag{49}$$

Proof. When A is positive semidefinite and symmetric, its eigenvalues $(\lambda_i)_i$ are real and non-negative. Moreover, $A + Id$ has eigenvalues $(\lambda_i + 1)_i$; $\det(A + Id) = \prod_i (\lambda_i + 1)$; and $\operatorname{tr} A = \sum_i \lambda_i$. These properties easily follow from the respective eigenvalue decomposition. Thus, we have:

$$\log \det(A + Id) \leq \log \prod_i (\lambda_i + 1) = \sum_i \log(\lambda_i + 1) \leq \sum_i \lambda_i = \operatorname{tr}(A), \tag{107}$$

where we have used $\log(x + 1) \leq x$ iff equality for $x = 0$. ☐

General Case. In the main text, we only skimmed the general case and mentioned the main assumption. Here, we look at the general case in detail.

For the general case, we need to make strong approximations to be able to pursue a similar derivation. First, we cannot drop the expectation; instead, we note that the log determinant is a concave function on the positive-semidefinite symmetric cone (Cover and Thomas, 1988), and we can use Jensen's inequality on the log determinant term from Equation (43) as follows:

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{p}(\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\})}[\log\det\left(\mathrm{H}''[\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}, \omega^*]\ \mathrm{H}''[\omega^*]^{-1} + Id\right)] && (108)\\
&\quad\le \log\det\left(\mathbb{E}_{\mathrm{p}(\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\})}[\mathrm{H}''[\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}, \omega^*]]\ \mathrm{H}''[\omega^*]^{-1} + Id\right). && (109)
\end{aligned}
$$

Second, we need use the following approximation:

$$
\mathrm{p}(\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}) \approx \mathrm{p}(\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}, \omega^*) \tag{110}
$$

to obtain a Fisher information and use its additivity. That is, we obtain:

$$
\mathbb{E}_{\mathrm{p}(\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}, \omega^*)}[\mathrm{H}''[\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}, \omega^*]] = \mathrm{F}(\{x_i^{\mathrm{acq}}\}, \omega^*) = \sum_i \mathrm{F}({x^{\mathrm{acq}}}_i, \omega^*). \tag{111}
$$

Plugging all of this together and applying Lemma 5.1, we obtain the same final approximation:

$$
\begin{aligned}
\mathrm{I}[\Omega; \{Y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}] = \ldots &\approx \tfrac{1}{2}\, \mathbb{E}_{\mathrm{p}(\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\})}[\log\det\left(\mathrm{H}''[\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}, \omega^*]\ \mathrm{H}''[\omega^*]^{-1} + Id\right)] && (112)\\
&\le \tfrac{1}{2} \log\det\left(\mathbb{E}_{\mathrm{p}(\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\})}[\mathrm{H}''[\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}, \omega^*]]\ \mathrm{H}''[\omega^*]^{-1} + Id\right) && (113)\\
&\approx \tfrac{1}{2} \log\det\left(\mathbb{E}_{\mathrm{p}(\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}, \omega^*)}[\mathrm{H}''[\{y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}, \omega^*]]\ \mathrm{H}''[\omega^*]^{-1} + Id\right) && (114)\\
&= \tfrac{1}{2} \log\det\left(\sum_i \mathrm{F}({x^{\mathrm{acq}}}_i, \omega^*)\ \mathrm{H}''[\omega^*]^{-1} + Id\right) && (115)\\
&\le \tfrac{1}{2} \sum_i \operatorname{tr}\left(\mathrm{F}({x^{\mathrm{acq}}}_i, \omega^*)\ \mathrm{H}''[\omega^*]^{-1}\right). && (116)
\end{aligned}
$$

Unlike in the case of generalized linear models, a stronger assumption was necessary to reach the same result. Alternatively, we could use the GGN approximation, which leads to the same result. It does not lead to an upper bound, however.

B.2 Approximate Expected Predicted Information Gain

In the main text, we only briefly referred to not knowing a principled way to arrive at the same result of Proposition 5.4 for the general case. This is because unlike for the expected information gain, the Fisher information for an acquisition candidate now lies within an matrix inversion. Even if we use the fact that $\log\det(Id + X\,Y^{-1})$ is concave in X and convex in Y, we end up with:

$$
\begin{aligned}
&\ldots && (117)\\
&\approx \tfrac{1}{2}\, \mathbb{E}_{\mathrm{p}(y^{\mathrm{eval}}, y^{\mathrm{acq}} \mid x^{\mathrm{eval}}, x^{\mathrm{acq}})\ \mathrm{p}(x^{\mathrm{eval}})}[\log\det\left(\mathrm{H}''[y^{\mathrm{eval}} \mid x^{\mathrm{eval}}, \omega^*]\, (\mathrm{H}''[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*] + \mathrm{H}''[\omega^*])^{-1} + Id\right)] && (118)\\
&\le \tfrac{1}{2}\, \mathbb{E}_{\mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}})}[\log\det\left(\mathbb{E}_{\mathrm{p}(y^{\mathrm{eval}}, x^{\mathrm{eval}})}[\mathrm{H}''[y^{\mathrm{eval}} \mid x^{\mathrm{eval}}, \omega^*]]\, (\mathrm{H}''[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*] + \mathrm{H}''[\omega^*])^{-1} + Id\right)] && (119)\\
&\ge \tfrac{1}{2} \log\det\left(\mathbb{E}_{\mathrm{p}(y^{\mathrm{eval}}, x^{\mathrm{eval}})}[\mathrm{H}''[y^{\mathrm{eval}} \mid x^{\mathrm{eval}}, \omega^*]]\, (\mathbb{E}_{\mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}})}[\mathrm{H}''[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*]] + \mathrm{H}''[\omega^*])^{-1} + Id\right) && (120)\\
&= \tfrac{1}{2} \log\det\left(\mathbb{E}_{\mathrm{p}(x^{\mathrm{eval}})}[\mathrm{F}(x^{\mathrm{eval}}, \omega^*)]\, (\mathrm{F}(x^{\mathrm{acq}}, \omega^*) + \mathrm{H}''[\omega^*])^{-1} + Id\right) && (121)\\
&\ge \tfrac{1}{2} \operatorname{tr}\left(\mathbb{E}_{\mathrm{p}(x^{\mathrm{eval}})}[\mathrm{F}(x^{\mathrm{eval}}, \omega^*)]\, (\mathrm{F}(x^{\mathrm{acq}}, \omega^*) + \mathrm{H}''[\omega^*])^{-1}\right). && (122)
\end{aligned}
$$

Note the $\le \ldots \ge$, which invalidates the chain. Obviously, the errors could cancel out, but a principled statement seems hardly possible using this deduction.

B.3 Approximate Predictive Information Gain

Similar to Proposition 5.3, we can approximate the predictive information gain. We assume that we have access to an (empirical) distribution $\hat{\mathrm{p}}(x^{\mathrm{eval}}, y^{\mathrm{eval}})$:

Proposition B.1. *We have (where we take the expectation over* $\hat{\mathrm{p}}(x^{eval}, y^{eval})$*):*

$$\arg\max_{x^{acq}} \mathrm{I}[Y^{eval}; y^{acq} \mid X^{eval}, x^{acq}, \mathcal{D}^{train}] = \arg\min_{x^{acq}} \mathrm{I}[\Omega; Y^{eval} \mid X^{eval}, y^{acq}, x^{acq}, \mathcal{D}^{train}] \tag{123}$$

with

$$\begin{aligned}
&\mathrm{I}[\Omega; Y^{eval} \mid X^{eval}, y^{acq}, x^{acq}, \mathcal{D}^{train}] && (124)\\
&\quad = \mathbb{E}_{\hat{\mathrm{p}}(x^{eval}, y^{eval})}\, \mathrm{I}[\Omega; y^{eval} \mid x^{eval}, y^{acq}, x^{acq}, \mathcal{D}^{train}]\\
&\quad \overset{\approx}{\leq} \mathbb{E}_{\hat{\mathrm{p}}(x^{eval}, y^{eval})}\big[\tfrac{1}{2} \log\det\big(\mathrm{H}''[y^{eval} \mid x^{eval}, \omega^*]\,(\mathrm{H}''[y^{acq} \mid x^{acq}, \omega^*] + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}])^{-1} + Id\big)\big] && (125)\\
&\quad \leq \tfrac{1}{2} \log\det\big(\mathbb{E}_{\hat{\mathrm{p}}(x^{eval}, y^{eval})}[\mathrm{H}''[y^{eval} \mid x^{eval}, \omega^*]]\,(\mathrm{H}''[y^{acq} \mid x^{acq}, \omega^*] + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}])^{-1} + Id\big) && (126)\\
&\quad \leq \tfrac{1}{2} \mathrm{tr}\big(\mathbb{E}_{\hat{\mathrm{p}}(x^{eval}, y^{eval})}[\mathrm{H}''[y^{eval} \mid x^{eval}, \omega^*]]\,(\mathrm{H}''[y^{acq} \mid x^{acq}, \omega^*] + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}])^{-1}\big). && (127)
\end{aligned}$$

All of this follows immediately. Only for the second inequality, we need to use Jensen's inequality and that the log determinant is on the positive semidefinite symmetric cone (Cover and Thomas, 1988). Like for the information gain, there is no difference between having access to labels or not when we have a GLM or use the GGN approximation.

B.4 Approximate Joint (Expected) Predictive Information Gain

A comparison of EPIG and JEPIG shows that JEPIG does not require an expectation over $\hat{\mathrm{p}}(x^{\mathrm{eval}})$ but uses a set of *evaluation samples* $\{x_i^{\mathrm{eval}}\}$. As such, we can easily adapt Proposition 5.4 to JEPIG and obtain:

Proposition B.2 (JEPIG)**.** *For a generalized linear model (or with the GGN approximation), we have:*

$$\arg\max_{\{x_i^{acq}\}} \mathrm{I}[\{Y_i^{eval}\}; \{Y_i^{acq}\} \mid \{x_i^{eval}\}, \{x_i^{acq}\}, \mathcal{D}^{train}] = \arg\min_{\{x_i^{acq}\}} \mathrm{I}[\Omega; \{Y_i^{eval}\} \mid \{x_i^{eval}\}, \{Y_i^{acq}\}, \{x_i^{acq}\}, \mathcal{D}^{train}] \tag{128}$$

with

$$\begin{aligned}
&\mathrm{I}[\Omega; \{Y_i^{eval}\} \mid \{x_i^{eval}\}, \{Y_i^{acq}\}, \{x_i^{acq}\}, \mathcal{D}^{train}]\\
&\quad \overset{\approx}{\leq} \tfrac{1}{2}\log\det\big(\mathrm{F}(\{x_i^{eval}\}, \omega^*)\,(\mathrm{F}(\{x_i^{acq}\}, \omega^*) + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}])^{-1} + Id\big) && (129)\\
&\quad \leq \tfrac{1}{2}\mathrm{tr}\big(\mathrm{F}(\{x_i^{eval}\}, \omega^*)\,(\mathrm{F}(\{x_i^{acq}\}, \omega^*) + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}])^{-1}\big). && (130)
\end{aligned}$$

And similarly for JPIG, we obtain:

Proposition B.3 (JPIG)**.** *For a generalized linear model (or with the GGN approximation), we have:*

$$\arg\max_{\{x_i^{acq}\}} \mathrm{I}[\{y_i^{eval}\}; \{y_i^{acq}\} \mid \{x_i^{eval}\}, \{x_i^{acq}\}, \mathcal{D}^{train}] = \arg\min_{\{x_i^{acq}\}} \mathrm{I}[\Omega; \{y_i^{eval}\} \mid \{x_i^{eval}\}, \{y_i^{acq}\}, \{x_i^{acq}\}, \mathcal{D}^{train}] \tag{131}$$

with

$$\begin{aligned}
&\mathrm{I}[\Omega; \{y_i^{eval}\} \mid \{x_i^{eval}\}, \{Y_i^{acq}\}, \{x_i^{acq}\}, \mathcal{D}^{train}]\\
&\quad \overset{\approx}{\leq} \tfrac{1}{2}\log\det\big(\mathrm{H}''[\{y_i^{eval}\} \mid \{x_i^{eval}\}, \omega^*]\,(\mathrm{H}''[\{y_i^{acq}\} \mid \{x_i^{acq}\}, \omega^*] + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}])^{-1} + Id\big) && (132)\\
&\quad \leq \tfrac{1}{2}\mathrm{tr}\big(\mathrm{H}''[\{y_i^{eval}\} \mid \{x_i^{eval}\}, \omega^*]\,(\mathrm{H}''[\{y_i^{acq}\} \mid \{x_i^{acq}\}, \omega^*] + \mathrm{H}''[\omega^* \mid \mathcal{D}^{train}])^{-1}\big). && (133)
\end{aligned}$$

C Similarity Matrices and One-Sample Approximations of the Fisher Information

Proposition 6.1. *Given $\mathcal{D}^{train}$, $\{x_i^{acq}\}$ and (sampled) $\{y_i^{acq}\}$, we have for the EIG:*

$$\mathrm{I}[\Omega; \{Y_i^{acq}\} \mid \{x_i^{acq}\}, \mathcal{D}^{train}] \lessapprox \tfrac{1}{2}\log\det\left(S_{\mathrm{H}''[\omega^* \mid \mathcal{D}^{train}]}[\mathcal{D}^{acq} \mid \omega^*] + Id\right) \tag{74}$$

Proof.

$$\begin{aligned}
\mathrm{I}[\Omega; \{Y_i^{\mathrm{acq}}\} \mid \{x_i^{\mathrm{acq}}\}, \mathcal{D}^{\mathrm{train}}] &\lessapprox \tfrac{1}{2}\log\det\left(\mathrm{F}(\{x_i^{\mathrm{acq}}\}, \omega^*)\ \mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]^{-1} + Id\right) && (134)\\
&= \tfrac{1}{2}\log\det\left((\mathrm{F}(\{x_i^{\mathrm{acq}}\}, \omega^*) + \mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}])\ \mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]^{-1}\right) && (135)\\
&\approx \tfrac{1}{2}\log\det\left((\hat{\mathrm{H}}'[\mathcal{D}^{\mathrm{acq}} \mid \omega^*]^T\, \hat{\mathrm{H}}'[\mathcal{D}^{\mathrm{acq}} \mid \omega^*] + \mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}])\ \mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]^{-1}\right) && (136)\\
&= \tfrac{1}{2}\log\det\left(\hat{\mathrm{H}}'[\mathcal{D}^{\mathrm{acq}} \mid \omega^*]\, \mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]^{-1} \hat{\mathrm{H}}'[\mathcal{D}^{\mathrm{acq}} \mid \omega^*]^T + Id\right) && (137)\\
&= \tfrac{1}{2}\log\det\left(S_{\mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]}[\mathcal{D}^{\mathrm{acq}} \mid \omega^*] + Id\right), && (138)
\end{aligned}$$

where we have used the matrix determinant lemma:

$$\det(AB + M) = \det(BM^{-1}A + Id)\ \det M. \tag{139}$$

□

Connection to the Joint (Expected) Predictive Information Gain. Following eq. (60), JEPIG can be decomposed as the difference between two EIG terms, which we can further split into three terms that are all only conditioned on $\mathcal{D}^{\mathrm{train}}$:

$$\begin{aligned}
&\mathrm{I}[\{Y_i^{\mathrm{eval}}\}; Y^{\mathrm{acq}} \mid \{x_i^{\mathrm{eval}}\}, x^{\mathrm{acq}}, \mathcal{D}^{\mathrm{train}}] && (140)\\
&\quad = \mathrm{I}[\Omega; \{Y_i^{\mathrm{eval}}\} \mid \{x_i^{\mathrm{eval}}\}, \mathcal{D}^{\mathrm{train}}] - \mathrm{I}[\Omega; \{Y_i^{\mathrm{eval}}\} \mid \{x_i^{\mathrm{eval}}\}, Y^{\mathrm{acq}}, x^{\mathrm{acq}}, \mathcal{D}^{\mathrm{train}}]\\
&\quad = \mathrm{I}[\Omega; \{Y_i^{\mathrm{eval}}\} \mid \{x_i^{\mathrm{eval}}\}, \mathcal{D}^{\mathrm{train}}] - \mathrm{I}[\Omega; \{Y_i^{\mathrm{eval}}\}, Y^{\mathrm{acq}} \mid \{x_i^{\mathrm{eval}}\}, x^{\mathrm{acq}}, \mathcal{D}^{\mathrm{train}}] + \mathrm{I}[\Omega; Y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \mathcal{D}^{\mathrm{train}}] && (141)
\end{aligned}$$

Using Proposition 6.1, we can approximate this as:

$$\begin{aligned}
&\mathrm{I}[\{Y_i^{\mathrm{eval}}\}; Y^{\mathrm{acq}} \mid \{x_i^{\mathrm{eval}}\}, x^{\mathrm{acq}}, \mathcal{D}^{\mathrm{train}}] && (142)\\
&= \tfrac{1}{2}\log\det\left(S_{\mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]}[\mathcal{D}^{\mathrm{eval}} \mid \omega^*] + Id\right) - \tfrac{1}{2}\log\det\left(S_{\mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]}[\mathcal{D}^{\mathrm{acq}}, \mathcal{D}^{\mathrm{eval}} \mid \omega^*] + Id\right) && (143)\\
&\quad + \tfrac{1}{2}\log\det\left(S_{\mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]}[\mathcal{D}^{\mathrm{acq}} \mid \omega^*] + Id\right) && (144)
\end{aligned}$$

Further, we can use the approximation in Proposition 6.2 and find that the $\log\lambda$ terms cancel because we have $|\mathcal{D}^{\mathrm{acq}}| + |\mathcal{D}^{\mathrm{train}}| = |\mathcal{D}^{\mathrm{acq}} \cup \mathcal{D}^{\mathrm{train}}|$ obviously. Taking the limit, we obtain:

$$\begin{aligned}
&\mathrm{I}[\{Y_i^{\mathrm{eval}}\}; Y^{\mathrm{acq}} \mid \{x_i^{\mathrm{eval}}\}, x^{\mathrm{acq}}, \mathcal{D}^{\mathrm{train}}] && (145)\\
&\quad \approx \tfrac{1}{2}\log\det\left(S[\mathcal{D}^{\mathrm{eval}} \mid \omega^*] + \lambda Id\right) - \tfrac{1}{2}\log\det\left(S[\mathcal{D}^{\mathrm{acq}}, \mathcal{D}^{\mathrm{eval}} \mid \omega^*] + \lambda Id\right) + \tfrac{1}{2}\log\det\left(S[\mathcal{D}^{\mathrm{acq}} \mid \omega^*] + \lambda Id\right) && (146)\\
&\quad \to \tfrac{1}{2}\log\det\left(S[\mathcal{D}^{\mathrm{eval}} \mid \omega^*]\right) - \tfrac{1}{2}\log\det\left(S[\mathcal{D}^{\mathrm{acq}}, \mathcal{D}^{\mathrm{eval}} \mid \omega^*]\right) + \tfrac{1}{2}\log\det\left(S[\mathcal{D}^{\mathrm{acq}} \mid \omega^*]\right). && (147)
\end{aligned}$$

Finally, the first term is independent of $\mathcal{D}^{\mathrm{acq}}$, and if we are interested in approximately maximizing JEPIG, we can maximize as proxy:

$$\log\det\left(S_{\mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]}[\mathcal{D}^{\mathrm{acq}} \mid \omega^*] + Id\right) - \log\det\left(S_{\mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]}[\mathcal{D}^{\mathrm{acq}}, \mathcal{D}^{\mathrm{eval}} \mid \omega^*] + Id\right), \tag{148}$$

or

$$\log\det\left(S[\mathcal{D}^{\mathrm{acq}} \mid \omega^*]\right) - \log\det\left(S[\mathcal{D}^{\mathrm{acq}}, \mathcal{D}^{\mathrm{eval}} \mid \omega^*]\right). \tag{149}$$

D Connection to Other Acquisition Functions in the Literature

D.1 SIMILAR (Kothawade et al., 2021) and PRISM (Kothawade et al., 2022)

Connection to LogDetMI. If we apply the Schur decomposition to $\log\det S[\mathcal{D}^{\text{acq}}, \mathcal{D}^{\text{eval}}, \mid \omega^*]$ from eq. (149), we obtain the following:

$$\begin{aligned}&\log\det S[\mathcal{D}^{\text{acq}}, \mathcal{D}^{\text{eval}} \mid \omega^*] &(150)\\ &\quad= \log\det S[\mathcal{D}^{\text{eval}} \mid \omega^*] + \log\det(S[\mathcal{D}^{\text{acq}} \mid \omega^*] - S[\mathcal{D}^{\text{acq}}; \mathcal{D}^{\text{eval}} \mid \omega^*] S[\mathcal{D}^{\text{eval}} \mid \omega^*]^{-1} S[\mathcal{D}^{\text{eval}}; \mathcal{D}^{\text{acq}} \mid \omega^*]),\end{aligned}$$

where $S[\mathcal{D}^{\text{acq}}; \mathcal{D}^{\text{eval}} \mid \omega^*]$ is the non-symmetric similarity matrix between $\mathcal{D}^{\text{acq}}$ and $\mathcal{D}^{\text{eval}}$ etc.

Dropping $\log\det S[\mathcal{D}^{\text{eval}} \mid \omega^*]$ which is independent of $\mathcal{D}^{\text{acq}}$, we can also maximize:

$$\log\det S[\mathcal{D}^{\text{acq}} \mid \omega^*] - \log\det(S[\mathcal{D}^{\text{acq}} \mid \omega^*] - S[\mathcal{D}^{\text{acq}}; \mathcal{D}^{\text{eval}} \mid \omega^*] S[\mathcal{D}^{\text{eval}} \mid \omega^*]^{-1} S[\mathcal{D}^{\text{eval}}; \mathcal{D}^{\text{acq}} \mid \omega^*],) \tag{151}$$

which is exactly the LogDetMI objective from SIMILAR (Kothawade et al., 2021) and PRISM (Kothawade et al., 2022).

We can further rewrite this objective by extracting $S[\mathcal{D}^{\text{acq}} \mid \omega^*]$ from the second term, obtaining:

$$\begin{aligned}&\log\det S[\mathcal{D}^{\text{acq}} \mid \omega^*] - \log\det(S[\mathcal{D}^{\text{acq}} \mid \omega^*] - S[\mathcal{D}^{\text{acq}}; \mathcal{D}^{\text{eval}} \mid \omega^*] S[\mathcal{D}^{\text{eval}} \mid \omega^*]^{-1} S[\mathcal{D}^{\text{eval}}; \mathcal{D}^{\text{acq}} \mid \omega^*]) &(152)\\ &\quad= -\log\det(Id - S[\mathcal{D}^{\text{acq}} \mid \omega^*]^{-1} S[\mathcal{D}^{\text{acq}}; \mathcal{D}^{\text{eval}} \mid \omega^*] S[\mathcal{D}^{\text{eval}} \mid \omega^*]^{-1} S[\mathcal{D}^{\text{eval}}; \mathcal{D}^{\text{acq}} \mid \omega^*]). &(153)\end{aligned}$$

Connection to LogDetCMI. Using information-theoretic decompositions, it is easy to show that:

$$\begin{aligned}&\mathrm{I}[\{Y_i^{\text{eval}}\}; Y^{\text{acq}} \mid \{x_i^{\text{eval}}\}, x^{\text{acq}}, \{Y_i\}, \{x_i\}, \mathcal{D}^{\text{train}}] &(154)\\ &\quad= \mathrm{I}[\{Y_i^{\text{eval}}\}; Y^{\text{acq}}, \{Y_i\} \mid \{x_i^{\text{eval}}\}, x^{\text{acq}}, \{x_i\}, \mathcal{D}^{\text{train}}] - \mathrm{I}[\{Y_i^{\text{eval}}\}; \{Y_i\} \mid \{x_i\}, \mathcal{D}^{\text{train}}]. &(155)\end{aligned}$$

These are two JEPIG terms, and using above approximations, including (153), leads to the LogDetCMI objective from Kothawade et al. (2021) and Kothawade et al. (2022):

$$\log \frac{\det(Id - S[\mathcal{D}^{\text{acq}} \mid \omega^*]^{-1} S[\mathcal{D}^{\text{acq}}; \mathcal{D}^{\text{eval}} \mid \omega^*] S[\mathcal{D}^{\text{eval}} \mid \omega^*]^{-1} S[\mathcal{D}^{\text{eval}}; \mathcal{D}^{\text{acq}} \mid \omega^*])}{\det(Id - S[\mathcal{D}^{\text{acq}}, \mathcal{D} \mid \omega^*]^{-1} S[\mathcal{D}^{\text{acq}}, \mathcal{D}; \mathcal{D}^{\text{eval}} \mid \omega^*] S[\mathcal{D}^{\text{eval}} \mid \omega^*]^{-1} S[\mathcal{D}^{\text{eval}}; \mathcal{D}^{\text{acq}}, \mathcal{D} \mid \omega^*])}. \tag{156}$$

D.2 Expected Gradient Length

Proposition 7.2. *The EIG for a candidate sample x^{acq} approximately lower-bounds the EGL:*

$$2\,\mathrm{I}[\Omega; Y^{acq} \mid x^{acq}] \overset{\approx}{\leq} \mathbb{E}_{\mathrm{p}(y^{acq} \mid x^{acq}, \omega^*)} \left\| \mathrm{H}'[y^{acq} \mid x^{acq}, \omega^*] \right\|^2 + const.. \tag{78}$$

Proof. The EIG is equal to the conditional entropy up to a constant term, via eq. (54) in Proposition 5.2:

$$\mathrm{I}[\Omega; Y^{\text{acq}} \mid x^{\text{acq}}] \overset{\approx}{\leq} \tfrac{1}{2} \log\det\left(\mathrm{F}(x^{\text{acq}}, \omega^*) + \mathrm{H}''[\omega^* \mid \mathcal{D}^{\text{train}}]\right) + \text{const}. \tag{157}$$

We apply a diagonal approximation for the Fisher information and Hessian, noting that the determinant of the diagonal matrix upper-bounds the determinant of the full matrix:

$$\begin{aligned}&\leq \tfrac{1}{2} \log\det\left(\mathrm{F}_{diag}(x^{\text{acq}}, \omega^*) + \mathrm{H}''_{diag}[\omega^* \mid \mathcal{D}^{\text{train}}]\right) + \text{const}. &(158)\\ &= \tfrac{1}{2} \sum_k \log\left(\mathrm{F}_{diag,kk}(x^{\text{acq}}, \omega^*) + \mathrm{H}''_{diag,kk}[\omega^* \mid \mathcal{D}^{\text{train}}]\right) + \text{const}. &(159)\end{aligned}$$

We use $\log x \leq x - 1$ and that $\mathrm{H}''[\omega^* \mid \mathcal{D}^{\text{train}}]$ is constant:

$$\leq \tfrac{1}{2} \sum_k \left(\mathrm{F}_{diag,kk}(x^{\text{acq}}, \omega^*) + \mathrm{H}''_{diag,kk}[\omega^* \mid \mathcal{D}^{\text{train}}]\right) + \text{const}. \tag{160}$$

$$\leq \tfrac{1}{2} \sum_k \mathrm{F}_{diag,kk}(x^{\mathrm{acq}}, \omega^*) + \text{const.} \tag{161}$$

From Proposition 4.3, we know that the Fisher information is equivalent to the outer product of the Jacobians: $\mathrm{F}(x^{\mathrm{acq}}, \omega^*) = \mathbb{E}_{\mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*)}[\mathrm{H}'[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*]\ \mathrm{H}'[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*]^T]$, and we finally obtain for the diagonal elements:

$$= \tfrac{1}{2} \sum_k \mathbb{E}_{\mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*)} \left[\mathrm{H}'_k[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*]^2\right] + \text{const.} \tag{162}$$

$$= \tfrac{1}{2}\, \mathbb{E}_{\mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*)} \left[\left\|\mathrm{H}'[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*]\right\|^2\right] + \text{const.} \tag{163}$$

□

D.3 Deep Learning on a Data Diet

Proposition 7.3. *The IG for a candidate sample x^{acq} approximately lower-bounds the gradient norm (GraNd) score at ω^* up to a second-order term:*

$$2\,\mathrm{I}[\Omega; y^{acq} \mid x^{acq}] \overset{\approx}{\leq} \mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D}^{train})}[\left\|\mathrm{H}'[y^{acq} \mid x^{acq}, \omega]\right\|^2] - \mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D}^{train})}[\mathrm{tr}\left(\frac{\nabla^2_\omega\, \mathrm{p}(y \mid x, \omega)}{\mathrm{p}(y \mid x, \omega)}\right)] + const. \tag{79}$$

Proof. For any fixed ω^*, the IG is equal to the conditional entropy up to a constant term, via Proposition 5.3:

$$\mathrm{I}[\Omega; Y^{\mathrm{acq}} \mid x^{\mathrm{acq}}] \overset{\approx}{\leq} \tfrac{1}{2} \log\det\left(\mathrm{H}''[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*] + \mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]\right) + \text{const.} \tag{164}$$

As in the previous proof, we apply a diagonal approximation for the Hessian, noting that the determinant of the diagonal matrix upper-bounds the determinant of the full matrix:

$$\leq \tfrac{1}{2} \log\det\left(\mathrm{H}''_{diag}[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*] + \mathrm{H}''_{diag}[\omega^* \mid \mathcal{D}^{\mathrm{train}}]\right) + \text{const.} \tag{165}$$

$$= \tfrac{1}{2} \sum_k \log\left(\mathrm{H}''_{diag,kk}[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*] + \mathrm{H}''_{diag,kk}[\omega^* \mid \mathcal{D}^{\mathrm{train}}]\right) + \text{const.} \tag{166}$$

Again, we use $\log x \leq x - 1$ and that $\mathrm{H}''[\omega^* \mid \mathcal{D}^{\mathrm{train}}]$ is constant:

$$\leq \tfrac{1}{2} \sum_k \left(\mathrm{H}''_{diag,kk}[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*] + \mathrm{H}''_{diag,kk}[\omega^* \mid \mathcal{D}^{\mathrm{train}}]\right) + \text{const.} \tag{167}$$

$$\leq \tfrac{1}{2} \sum_k \mathrm{H}''_{diag,kk}[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*] + \text{const.} \tag{168}$$

From Lemma 4.4, we know that the Hessian is equivalent to the outer product of the Jacobians plus a second-order term: $\mathrm{H}''[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*] = \mathrm{H}'[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*]\ \mathrm{H}'[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*]^T - \frac{\nabla^2_\omega\, \mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*)}{\mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*)}$, and we finally obtain for the diagonal elements:

$$= \tfrac{1}{2} \sum_k \mathrm{H}'_k[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*]^2 - \tfrac{1}{2}\,\mathrm{tr}\left(\frac{\nabla^2_\omega\, \mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*)}{\mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*)}\right) + \text{const.} \tag{169}$$

$$= \tfrac{1}{2}\left\|\mathrm{H}'[y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*]\right\|^2 - \tfrac{1}{2}\,\mathrm{tr}\left(\frac{\nabla^2_\omega\, \mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*)}{\mathrm{p}(y^{\mathrm{acq}} \mid x^{\mathrm{acq}}, \omega^*)}\right) + \text{const.} \tag{170}$$

Taking an expectation over $\omega^* \sim \mathrm{p}(\omega^* \mid \mathcal{D}^{\mathrm{train}})$ yields the statement. □