Characterizing the robustness of Bayesian adaptive experimental designs to active learning bias

Sabina J. Sloman SSLOMAN@ANDREW.CMU.EDU
Department of Social and Decision Sciences
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Daniel M. Oppenheimer OPPENHEIMER@CMU.EDU
Departments of Social and Decision Sciences and of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Stephen B. Broomell BROOMELL@GMAIL.COM
Department of Social and Decision Sciences
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Cosma Rohilla Shalizi CSHALIZI@CMU.EDU
Departments of Statistics and of Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213, USA
Santa Fe Institute
Santa Fe, NM 87501, USA

Abstract

Bayesian adaptive experimental design is a form of active learning, which chooses samples to maximize the information they give about uncertain parameters. Prior work has shown that other forms of active learning can suffer from **active learning bias**, where unrepresentative sampling leads to inconsistent parameter estimates. We show that active learning bias can also afflict Bayesian adaptive experimental design, depending on model misspecification. We develop an information-theoretic measure of misspecification, and show that worse misspecification implies more severe active learning bias. We also show that model classes incorporating more "noise" — i.e., specifying higher inherent variance in observations — suffer less from active learning bias, because their predictive distributions are likely to overlap more with the true distribution.

Keywords: active learning, optimal experimental design, Bayesian inference, mathematical modeling, model misspecification

Adaptive sampling methods choose data points in sequence to be as informative as possible (Cavagnaro et al., 2010; Ryan et al., 2016). In machine learning, this is called *active learning* (Kanamori, 2002; Settles, 2012).[1] Despite their advantages, adaptive sampling schemes can produce training sets that are highly unrepresentative of the target distribution

1. We will use "adaptive sampling" and "active learning" interchangeably.

(Farquhar et al., 2021), and the estimates made from adaptively-sampled data may not generalize to the target. This phenomenon is **active learning bias** (Farquhar et al., 2021).

We study the presence of active learning bias in a class of active learning methods called information theoretic active learning, where the modeler must pick an objective function capturing the informativeness of observations (Houlsby et al., 2011). These methods have been studied in parallel fields, such as computerized adaptive testing (Owen, 1969), clinical research (Whitehead and Brunier, 1995) and cognitive modeling (Cavagnaro et al., 2010), where they are referred to as **Bayesian** sequential (or **adaptive**) **experimental design** (Drovandi et al., 2013; Ryan et al., 2016) or adaptive design optimization (Cavagnaro et al., 2010).

1. Active learning bias and model misspecification

The advantages of active learning methods are usually expounded assuming that the model class is well-specified, i.e., that the true data-generating distribution is a member of the class (MacKay, 1992; Kanamori, 2002; Dasgupta, 2004; Sugiyama, 2005; Myung et al., 2013). Yet in most modeling enterprises, this assumption is not credible: the exact form of the true data-generating process is difficult if not impossible to know, and models are deliberately simplified tractable approximations.

While sampling bias and vulnerability to model misspecification are often discussed as two separate limitations of active learning methods, it turns out they are deeply related. In particular, sampling bias, or covariate shift, can amplify bias when the model class is misspecified (Sugiyama et al., 2008; Wen et al., 2014; Spencer et al., 2021). The implication of this is that active learning methods can increase bias when the model class is misspecified (Sugiyama, 2005; Bach, 2006).

2. Contributions of the present work

We apply insights on active learning bias from the machine learning literature to Bayesian adaptive experimental designs, and investigate how the extent of active learning bias varies with degree of model misspecification. We also demonstrate how the amount of observational "noise" specified by a model can mitigate active learning bias, since the former affects the model's degree of misspecification.

3. Preliminaries and notation

A modeler wants to predict some variable $y \in \mathbb{R}^m$ using another variable $x \in \mathbb{R}^d$. Given x, y always follows the same distribution, $y|x \sim f(x)$. We call f the **true model**. We assume that x is fully observable and follows some distribution g.

The modeler specifies a **hypothesized model class** that predicts $y|x \sim m(x, \theta)$, with $\theta \in \Theta$; the variable θ contains the **parameters** of the model class, which live in the **parameter space** Θ. Θ is fixed, i.e. determined in advance of the data and unchanging in response to them. m is a probabilistic function, whose form is also presumed to be fixed (e.g., logistic regression). The hypothesized model class $m(x, \Theta)$ is thus comprised of the set of distributions $\{m(x, \theta) \ : \ \theta \in \Theta\}$.

The model estimation problem is to find $\theta^* \in \Theta$ which minimizes the risk, i.e., such that

$$\theta^* \equiv \underset{\theta \in \Theta}{\operatorname{argmin}} \, \mathbf{E}\left[\mathscr{L}(m(x,\theta), y)\right] \tag{1}$$

for some suitable loss function $\mathscr{L}$, which takes as inputs a predictive distribution and a realized value for y. Here, the expectation is under the true data-generating distribution, i.e., the product of g and f. $m(x, \theta^*)$ is an instance of the hypothesized model class, the **best-fitting model**.[2]

To estimate θ^*, the modeler gets a set of samples $\mathbf{x} \in \mathbb{R}^{n \times d}$ and observes outcomes $\mathbf{y} \in \mathbb{R}^{n \times m}$. We consider only distributional estimates of θ^*, essentially assigning a probability to each $\theta \in \Theta$ that it is the risk minimizer:

$$p(\theta) \equiv P(\theta = \theta^* | \mathbf{x}, \mathbf{y}) \tag{2}$$

We refer to the corresponding predictive distribution for $y|x$ as $m(x, \hat{\theta})$.[3] We call $m(x, \hat{\theta})$ the **trained** or **estimated** model.

Random sampling or passive learning techniques draw x values IIDly from the population distribution g. Adaptive sampling or active learning techniques *actively* construct $\mathbf{x}$ to maximize the concentration of p. Both then get $\mathbf{y}|\mathbf{x} \sim f(\mathbf{x})$. We write $m(x, \hat{\theta}_{adaptive})$ for the result of an adaptive procedure, and $m(x, \hat{\theta})$, unmodified, for passive learning.

Note that risk, the modeler's objective, is defined using the distribution g, the **target distribution** of inputs whose consequences the modeler ultimately wants to predict. Active learning bias (ALB) occurs when, averaging over data sets $x \sim g$ and corresponding observations $y|x \sim f$, $\mathbf{E}\left[\mathscr{L}(m(x, \hat{\theta}_{adaptive}), y)\right] > \mathbf{E}\left[\mathscr{L}(m(x, \theta^*), y)\right]$.

Finally, we say that a hypothesized model class is **misspecified** when the true model is not in the class, i.e., $f(x) \neq m(x, \theta)$ for all $\theta \in \Theta$. Consistent with prior literature, we say "model misspecification," even though it is the hypothesized model *class* that is wrong.

4. Bayesian adaptive experimental design

A natural way to construct p is via Bayesian inference. We begin with a prior distribution p_0, and, at the t^{th} step, where we observe x_t and y_t, we use Bayes's rule to update it recursively

$$p_t(\theta) = p_{t-1}(\theta) \frac{m(y_t | x_t, \theta)}{\int_\theta m(y_t | x_t, \theta) \; p_{t-1}(\theta) \; d\theta} \tag{3}$$

Here, $m(y_t|x_t, \theta)$ indicates the likelihood of observation y_t under the distribution $m(x_t, \theta)$. $p_t(\theta)$ thus implicitly involves the whole history of inputs and responses to date, $\mathbf{x}_t$ and $\mathbf{y}_t$ respectively.

2. Related work writes of selecting a *hypothesis* from a *hypothesis class* (Dasgupta, 2004; Golovin et al., 2010). Our "hypothesized model class" matches their "hypothesis class," and selecting θ^* from Θ amounts to selecting a hypothesis from the hypothesis class. Our terminology is similar to Shiffrin and Chandramouli (2016), who also distinguish between model classes and model instances in the context of Bayesian modeling.
3. We use this notation for brevity, but note that there need not be a single $\theta \in \Theta$ giving exactly this distribution.

For Bayesian adaptive designs, at every step t we define *the modeler's* expected utility of an input x over the current posterior distribution of θ:

$$\mathbf{E}\left[u_t(x)\right] = \int_\theta \int_y u(x, y, \theta)\ m(y|x, \theta)\ p_{t-1}(\theta)\ dy\ d\theta \tag{4}$$

Bayesian adaptive experimental designs pick the maximizer of $\mathbb{E}[u_t(x)]$ as the next value of x. We pick a u that encourages precision of parameter estimates:[4]

$$u(x, y, \theta) = \log \frac{p_{t-1}(\theta|y, x)}{p_{t-1}(\theta)} \tag{5}$$

5. Results

We here compare the behavior of Bayesian adaptive experimental designs in the case of a misspecified vs. well-specified class of polynomial regression models. In the extended version of this paper (Sloman et al., 2022), we show that our findings apply to both a toy classification problem and a preference-learning problem.

Figures 1a–1c show the results of several simulations in which f is a degree-two polynomial regression model, i.e., $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 100)$. In the underparameterized case (Figure 1a), m, the functional form corresponding to the hypothesized model class, is linear in x. In the fully parameterized case (Figure 1b), m is quadratic in x. In the overparameterized case (Figure 1c), m is cubic. In all cases, the additive noise is (correctly) specified as $\epsilon \sim \mathcal{N}(0, 100)$. Thus, by our definition, only the underparameterized case is misspecified.

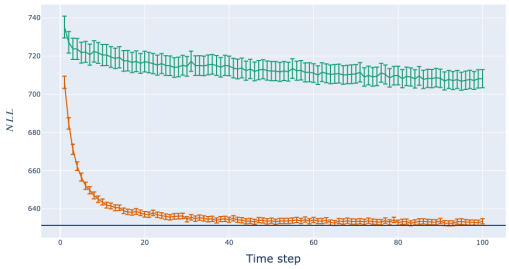
Behavior in the misspecified (underparameterized) case is markedly different from the behavior in the other two cases. Risk was measured as the negative log likelihood (NLL) of 100 observations drawn randomly from the target distribution, $g = \text{Unif}(0, 100)$.[5] For the well-specified model classes, adaptive sampling reduces the risk more quickly than random sampling, though both converge on the same risk as θ^*.[6] However, under misspecification, Bayesian adaptive designs lead to worse generalization than random sampling.

5.1 ALB depends on degree of misspecification

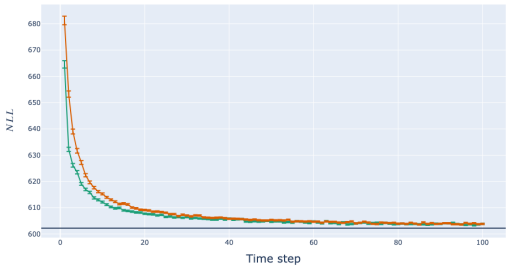
We now investigate how the extent of ALB varies with the degree of misspecification. There is a natural information-theoretic way to measure the latter, using the Kullback-Leibler divergence (KLD). Concretely, we define the degree of misspecification as the expectation, under g, of the KLD between the true f and the best-fitting model:

$$\mathscr{D}_{model} \equiv \int_x KLD(f(x), m(x, \theta^*))\ g(x)\ dx \tag{6}$$

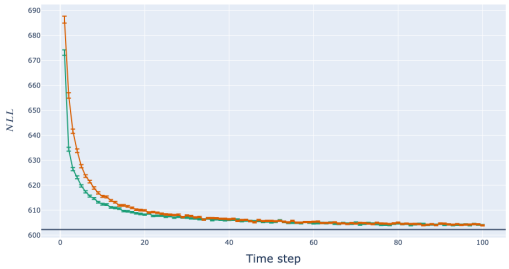
4. With this choice of u, $\mathbf{E}\left[u_t\right]$ is the mutual information between the next observation and the parameter θ, regarded as a random variable distributed according to p (Bernardo, 1979; Myung et al., 2013; Ryan et al., 2016).
5. When using adaptive sampling, we identified the optimal design at each time step t using the `iminuit` package (Dembinski et al., 2020) and numerically integrating the expected utility function shown in equation 4 for 4,000 samples from the parameter distribution at $t - 1$.
6. We found θ^* by taking the OLS solution to a regression of 1,001 evenly spaced points across the domain and their expectation under the true model.



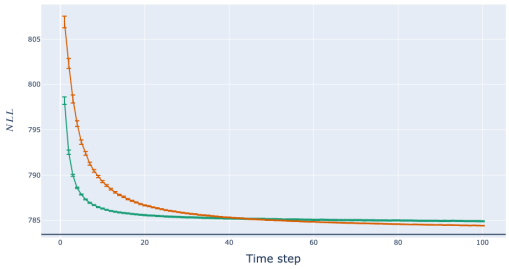
(a) m is linear, with additive noise $\epsilon \sim N(0, 100)$.



(b) m is quadratic, with additive noise $\epsilon \sim N(0, 100)$.



(c) m is cubic, with additive noise $\epsilon \sim N(0, 100)$.



(d) m is linear, with additive noise $\epsilon \sim N(0, 1000)$.

Figure 1: Risk incurred across 100 steps of Bayesian updating. Risk is approximated as the negative log likelihood (NLL) of 100 observations from the target distribution. The generating model is always a degree-two polynomial with parameters drawn from a $\mathcal{N}([0, 0, 0], \text{diag}(100, 10, .1))$ distribution. Lines are means across 1,000 simulated experiments, with error bars showing ± 1 standard error around the mean. Horizontal lines show the risk achievable by $m(x, \theta^*)$. Green lines (▬) show results using Bayesian adaptive design optimization during model estimation. Orange lines (▬) show results using random sampling from the target distribution.

We define ALB as

$$ALB \equiv \frac{\mathbf{E}\left[\mathscr{L}(m(x, \hat{\theta}_{adaptive}), y)\right]}{\mathbf{E}\left[\mathscr{L}(m(x, \theta^*), y)\right]} - 1 \tag{7}$$

ALB is thus proportion of expected risk in excess of what the best parameter value would deliver.

We analyzed the ALB values from the simulated experiments shown in Figure 1a as a function of the corresponding $\mathscr{D}_{model}$ values. (To calculate ALB, we approximated the risk by the NLL of 100 observations from the target distribution. $\mathbf{E}\left[\mathscr{L}(m(x, \hat{\theta}_{adaptive}), y)\right]$ was calculated as the NLL assigned by the estimated model after 100 steps of Bayesian adaptive design optimization — the same values shown as the rightmost point of the green

line in Figure 1a.) We found a very strong correlation between the two of .95 ($p = .00$). In other words, more misspecification predicted more ALB.

One way to adjust $\mathscr{D}_{model}$ is by specifying a "noisier" model class, i.e., to force m to predict more inherent variation in its outputs. As the predictive distribution corresponding to the hypothesized model class becomes more dispersed, it will overlap more with the data-generating distribution, leading to a lower divergence from f to the hypothesized model class, and thus to lower misspecification. This suggests that if the degree of ALB depends on $\mathscr{D}_{model}$, then ALB should be reduced, or even eliminated, by specifying noisier models.

Figure 1d shows the evolution of risk for the misspecified model class of the form shown in Figure 1a. However, instead of additive noise distributed $\mathscr{N}(0, 100)$, the hypothesized model class incorporates additive noise distributed $\mathscr{N}(0, 1000)$. Figure 1d shows that the incorporation of this additional noise nearly eliminates the ALB, which is reminiscent of theoretical results in Sugiyama (2005) that active learning is robust to model classes that are only slightly misspecified.[7]

6. Discussion

We demonstrated that active learning bias can afflict Bayesian adaptive experimental designs, that the extent of active learning bias varies with the degree of model misspecification, and that the amount of observational noise specified by the model class mitigates the extent of active learning bias. Subsequent work should investigate the existence of mathematical bounds on the amount of noise needed for a model class to be tolerant to active learning bias,[8] as well as the applicability of our findings in more realistic modeling scenarios.

Author note

The extended version of this paper is under review at NeurIPS 2022.

Acknowledgments

SJS was supported by a Tata Consultancy Services (TCS) Fellowship while contributing to this work.

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562 (Towns et al., 2014). Specifically, it used the Bridges-2 system, which is supported by NSF award number ACI-1928147, at the Pittsburgh Supercomputing Center.

Finally, this work benefited enormously from discussion with Marina Dubova and Daniel Cavagnaro.

7. We also found the same qualitative relationship between degree of misspecification, ALB and anticipated observational noise for a model misspecified as quadratic in x when the generating model was cubic. These results are reported in the extended version of this paper (Sloman et al., 2022).
8. Such a study could perhaps draw from theoretical results on the tolerance of parameterized bandit learning algorithms to misspecification (Kannan et al., 2018; Bogunovic and Krause, 2021).

References

Francis Bach. Active learning for misspecified generalized linear models. Technical Report N15/06/MM, École des Mines de Paris, 2006.

Jose M. Bernardo. Expected Information as Expected Utility. *The Annals of Statistics*, 7 (3):686–690, 1979. doi: 10.1214/aos/1176344689.

Ilija Bogunovic and Andreas Krause. Misspecified Gaussian Process Bandit Optimization. *arXiv:2111.05008 [cs]*, 2021.

Daniel R. Cavagnaro, Jay I. Myung, Mark A. Pitt, and Janne V. Kujala. Adaptive Design Optimization: A Mutual Information-Based Approach to Model Discrimination in Cognitive Science. *Neural Computation*, 22(4):887–905, 2010. doi: 10.1162/neco.2009.02-09-959.

Sanjoy Dasgupta. Analysis of a greedy active learning strategy. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing 17 [NIPS 2004]*, pages 337–344, Cambridge, Massachusetts, 2004. MIT Press.

Hans Dembinski, Piti Ongmongkolkul, Christoph Deil, Henry Schreiner, Matthew Feickert, Andrew, Chris Burr, Jason Watson, Fabian Rost, Alex Pearce, Lukas Geiger, Bernhard M. Wiedemann, Christoph Gohlke, Gonzalo, Jonas Drotleff, Jonas Eschle, Ludwig Neste, Marco Edward Gorelli, Max Baak, Omar Zapata, and odidev. Scikit-Hep/Iminuit. *Zenodo*, 2020. doi: 10.5281/zenodo.3949207.

Christopher C. Drovandi, James M. McGree, and Anthony N. Pettitt. Sequential Monte Carlo for Bayesian sequentially designed experiments for discrete data. *Computational Statistics & Data Analysis*, 57(1):320–335, 2013. doi: 10.1016/j.csda.2012.05.014.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On Statistical Bias In Active Learning: How and When to Fix It. In *Proceedings of the International Conference on Learning Representations (ICLR) 2021*, 2021.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-Optimal Bayesian Active Learning with Noisy Observations. In John Lafferty, C. K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing 23 [NIPS 2010]*, pages 766–774, Cambridge, Massachusetts, 2010. MIT Press.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. *arXiv:1112.5745 [cs, stat]*, 2011.

Takafumi Kanamori. Statistical Asymptotic Theory of Active Learning. *Annals of the Institute of Statistical Mathematics*, 54(3):459–475, 2002.

Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A Smoothed Analysis of the Greedy Algorithm for the Linear Contextual Bandit Problem. *arXiv:1801.03423 [cs]*, 2018.

David J. C. MacKay. Information-Based Objective Functions for Active Data Selection. *Neural Computation*, 4(4):590–604, 1992. doi: 10.1162/neco.1992.4.4.590.

Jay I. Myung, Daniel R. Cavagnaro, and Mark A. Pitt. A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, 57(3-4):53–67, 2013. doi: 10.1016/j.jmp.2013.05.005.

Roger J. Owen. A Bayesian Approach to Tailored Testing. Research Bulletin RB-69-92, Educational Testing Service, Princeton, NJ, 1969.

Elizabeth G. Ryan, Christopher C. Drovandi, James M. McGree, and Anthony N. Pettitt. A Review of Modern Computational Algorithms for Bayesian Optimal Design. *International Statistical Review*, 84(1):128–154, 2016. doi: 10.1111/insr.12107.

Burr Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.

Richard Shiffrin and Suyog Chandramouli. Model Selection, Data Distributions, and Reproducibility. In Harald Atmanspacher and Sabine Maasen, editors, *Reproducibility: Principles, Problems, Practices, and Prospects*, pages 115–140. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2016. doi: 10.1002/9781118865064.ch6.

Sabina J. Sloman, Daniel M. Oppenheimer, Stephen B. Broomell, and Cosma Rohilla Shalizi. Characterizing the robustness of Bayesian adaptive experimental designs to active learning bias. *arXiv:2205.13698 [stat]*, 2022.

Jonathan Spencer, Sanjiban Choudhury, Arun Venkatraman, Brian Ziebart, and J. Andrew Bagnell. Feedback in Imitation Learning: The Three Regimes of Covariate Shift. *arXiv:2102.02872 [cs, stat]*, 2021.

Masashi Sugiyama. Active Learning for Misspecified Models. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing 18 [NIPS 2005]*, pages 1305–1312, Cambridge, Massachusetts, 2005. MIT Press.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008. doi: 10.1007/s10463-008-0197-x.

John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkins-Diehr. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering*, 16(5):62–74, 2014. doi: 10.1109/MCSE.2014.80.

Junfeng Wen, Chun-Nam Yu, and Russell Greiner. Robust Learning under Uncertain Test Distributions: Relating Covariate Shift to Model Misspecification. In *Proceedings of the 31^{s}t International Conference on Machine Learning*, volume 32. JMLR: W&CP, 2014.

John Whitehead and Hazel Brunier. Bayesian Decision Procedures for Dose Determining Experiments. *Statistics in Medicine*, 14(9):885–893, 1995. doi: 10.1002/sim.4780140904.