Not All Data Are Created Equal - Lessons From Sampling Theory For Adaptive Machine Learning

Julian Rodemann, Sebastian Fischer, Lennart Schneider, Malte Nalenz, Thomas Augustin

Department of Statistics, LMU Munich

`julian@stat.uni-muenchen.de`

Example 1: (Causal) Bayesian Optimization

Pseudo-Code: Bayesian Optimization (BO)

1: create an initial design $D = \{(\boldsymbol{x}_{(i)}, \Psi_{(i)})\}_{i=1,\dots,n_{init}}$ of size n_{init}
2: **while** termination criterion is not met **do**
3: **train** a surrogate model (SM) on data D
4: **propose** $\boldsymbol{x}_{new} = \arg\max_{\boldsymbol{x}} AF(SM(\boldsymbol{x}))$, $AF(\cdot)$ an acquisition function
5: **evaluate** Ψ on $\boldsymbol{x}_{new}$
6: **update** $D \leftarrow D \cup (\boldsymbol{x}_{new}, \Psi(\boldsymbol{x}_{new}))$
7: **end while**
8: **return** $\arg\min_{\boldsymbol{x}\in D} \Psi(\boldsymbol{x})$ and respective Ψ_{min}

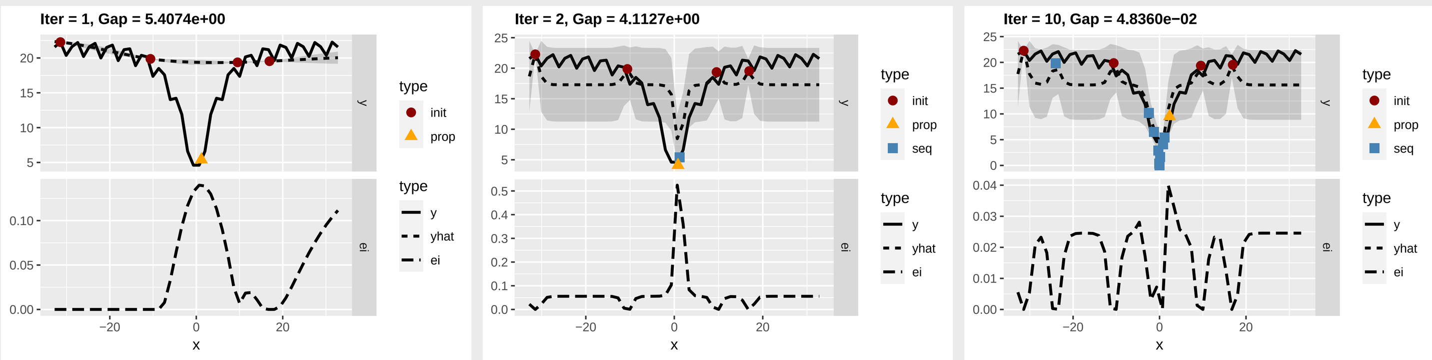

Figure 1: Visualization of univariate Bayesian Optimization with Gaussian Process as Surrogate Modle.

Causal Bayesian Optimization [1]: Prior to $\boldsymbol{x}_{new} = \arg\max_{\boldsymbol{x}} AF(SM(\boldsymbol{x}))$, an optimal set of covariates $\boldsymbol{ES}$ based on a DAG to intervene on is returned and $\boldsymbol{x}_{new} \in \boldsymbol{ES}$.

Sampling Bias: Global generalization suffers from the BO-induced covariate shift, see distribution of proposed points (blue) in figure 1. Moreover, some covariates vary more than others in the sample obtained by causal BO, resulting in a *de facto* stratified sample. [2], [4] and more specifically [9] discuss implications of such a feedback covariate shift for conformal prediction.

Example 2: Self-Training

Pseudo-Code: Self-Training (ST)

1: **require:** labeled data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \in (\mathcal{X} \times \mathcal{Y})^n$, unlabeled data $\mathcal{U} = \{(\boldsymbol{x}_i, \mathcal{Y})\}_{i=n+1}^{m} \in \left(\mathcal{X} \times 2^{\mathcal{Y}}\right)^{m-n}$ from same distribution with $\mathcal{Y}$ categorical and a classifier.
2: **while** stopping criterion is not met **do**
3: **train** classifier on D to obtain $\widehat{y(\boldsymbol{x})}$
4: **predict** on $\mathcal{U}$: $\hat{y}_i = \widehat{y(\boldsymbol{x}_i)}$ $\forall i \in \{n+1, \dots, m\}$
5: **select** subset $C \subseteq \{(\boldsymbol{x}_i, \hat{y}_i)\}_{i=n+1}^{m}$ according to a confidence measure
6: **update** $D \leftarrow D \cup C$
7: **end while**
8: **return** final fit

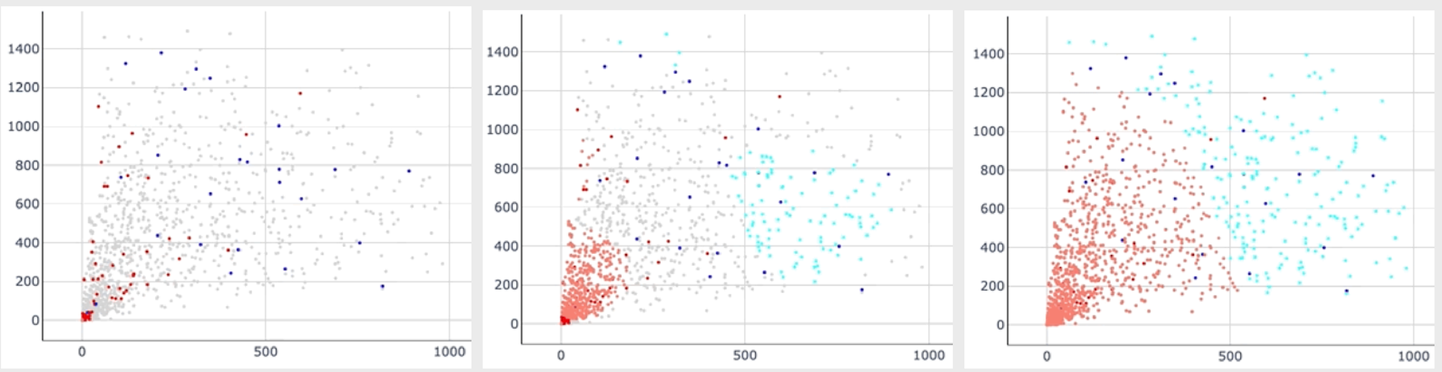

Figure 2: Covariate scatter plot from self-training of a support vector machine. Dark blue/red: labels y_i in $\mathcal{D}$. Light blue/red: predicted labels $\hat{y}_i$ in $\mathcal{U}$. Credits: `gist.github.com/SolClover`

Sampling Bias: Covariate distribution may differ from initial (and final) distribution, see figure 2. Depending on the stopping criterion, a covariate shift may harm interpretability of the model. E.g., regions in the covariate space where data is scarce are detrimental to reliable estimates of partial dependencies [5].

Definition 1 (Partial Dependencies [5]) *Let $\mathcal{X} = \mathcal{X}_I \cup X_R$, where $\mathcal{X}_I$ are the covariates of interest. $\hat{f}_I(x_I) = \int \hat{f}(x_I, X_R) d\mathbb{P}(X_R)$ is said to be a partial dependence function.*

Example 3: Active Learning

Idea: Similar to self-training, except that an oracle is available, thus $\hat{y}$ in C is replaced by ground-truth y when used to update D in line 6 of pseudo-code above. While self-trainig usually selects instances with high confidence, active learning queries points of high uncertainty.

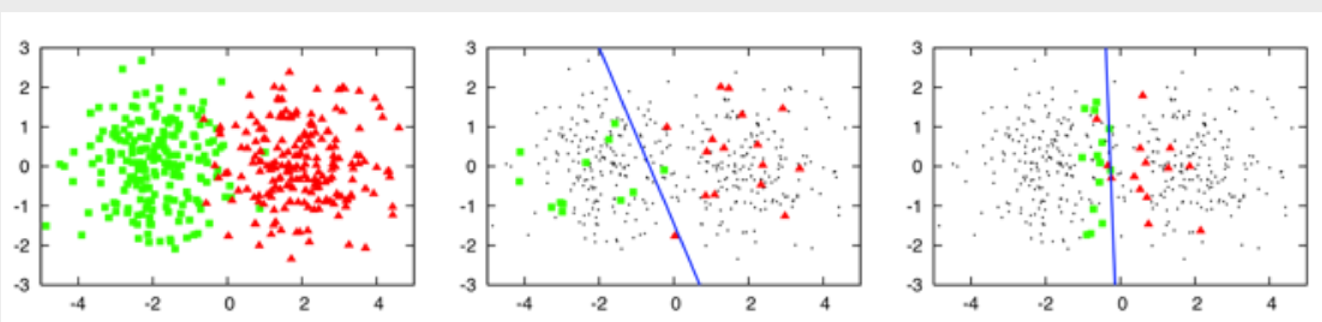

Figure 3: Active learning of a binary classification problem (test distribution: left). Random sampling (middle) is less efficient than uncertainty sampling (right). Credits: [8]

References

[1] Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3155–3164. PMLR, 2020.

[2] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

[3] Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2017. `http://archive.ics.uci.edu/ml`.

[4] Clara Fannjiang, Stephen Bates, Anastasios N Angelopoulos, Jennifer Listgarten, and Michael I Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

[5] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[6] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[7] Michael Langberg and Leonard J Schulman. Universal ε-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 598–607. SIAM, 2010.

[8] David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *ACM SIGIR Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.

[9] Samuel Stanton, Wesley Maddox, and Andrew Gordon Wilson. Bayesian optimization with conformal coverage guarantees. *arXiv preprint arXiv:2210.12496*, 2022.

Lesson 1: Explore-Exploit-Weights

Idea

- Inclusion probability of proposed points is analytically not available due to myopic optimization of AF
- Weight by potential gain of information at time of proposal as expressed by the surrogate model's epistemic uncertainty (= standard errors) and compare to standard errors at n randomly sampled points by empirical distribution function → surrogate-based weights (definition 2)
- Use random forest as surrogate and include weights as drawing probability in bootstrapping
- **possible extension to causal BO:** *a posteriori* stratification (weights proportional to strata size) to estimate effects on population level.

Definition 2 (Surrogate-based Weights) *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be an iid. post hoc sample and the surrogate model's standard errors $\hat{s}(\boldsymbol{x}) = \sqrt{\mathrm{Var}(\widehat{\mu}(\boldsymbol{x})_t)}$, with $\mu(\cdot)_t$ the surrogate model's prediction function with date of birth (dob) t. Let $\boldsymbol{x}^* \in \{\boldsymbol{x}_{n_{init}+1}, \dots, \boldsymbol{x}_{n_{init}+t}\}$ be a proposed point with dob t. Furthermore, let $F_{\hat{s}(\boldsymbol{x})}(\bullet)$ be the empirical distribution function of $\hat{s}(\boldsymbol{x}_1), \dots, \hat{s}(\boldsymbol{x}_n)$. Name its value at $\hat{s}(\boldsymbol{x}^*)$ standard error distribution value (SED): $SED(\boldsymbol{x}^*) = F_{\hat{s}(\boldsymbol{x})}(\hat{s}(\boldsymbol{x}^*))$. Set $SED(\boldsymbol{x}_i) = 1\ \forall \boldsymbol{x}_i \in \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n_{init}}\}$. The weights*

$$w_j = \frac{SED(\boldsymbol{x}_j)}{\sum_{i=1}^{n_{init}+t} SED(\boldsymbol{x}_i)}$$

with $i, j \in \{1, \dots, n_{init}, n_{init}+1, \dots, n_{init}+t\}$ shall be called surrogate-based weights.

Preliminary Results

Hypothesis: Weighted Surrogates (WS) are better global (i.e., on whole parameter space) approximates of 2D synthetic functions than unweighted surrogates (US).

Experiments: We run 40 BO replications with random forest on six well-established synthetic benchmark functions and compare MSE on an *iid.* random sample ($N = 10000$) of weighted and unweighted surrogates with dob $t \in \{30, 60, 90, 120\}$.

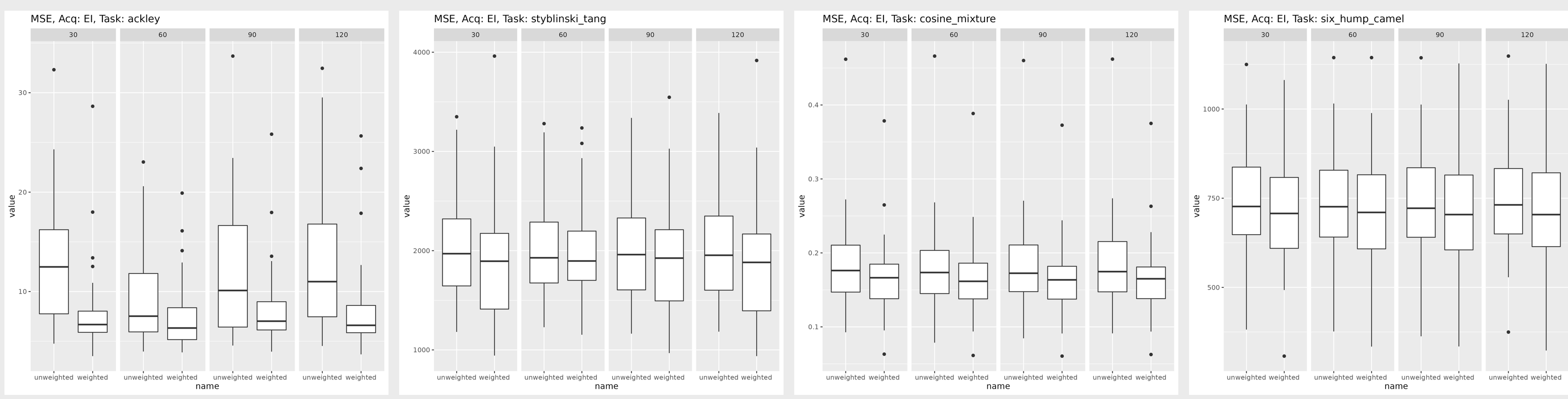

Figure 4: MSEs of unweighted and weighted surrogates (random forest with case weights) with dob $t \in \{30, 60, 90, 120\}$ for Ackley, Styblinski-Tang, Cosine Mixture and Six-Hump-Camel benchmark function. AF: Expected Improvement.

Lesson 2: Sampling-Sensitive Stopping

Compare selected data $\boldsymbol{S} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_t\} \subseteq \mathcal{X}$ and a hypothetical *i.i.d.* sample $\boldsymbol{U} = \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_t\} \subseteq \mathcal{X}$.

Kernel Two-Sample Test: We use a non-parametric tests proposed by [6] to determine if two samples are drawn from different distributions. Its test statistic is the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space (RKHS), and is called the maximum mean discrepancy (MMD). An unbiased estimate of the MMD for $\boldsymbol{S}$ and $\boldsymbol{U}$ is

$$\widehat{MMD} = \left(\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k_\sigma(\boldsymbol{u}_i, \boldsymbol{u}_j) + \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k_\sigma(\boldsymbol{s}_i, \boldsymbol{s}_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j \neq i}^{m} k_\sigma(\boldsymbol{s}_j, \boldsymbol{u}_i) \right)^{\frac{1}{2}}$$

with the radial basis function kernel $k_\sigma(\cdot, \cdot)$ with σ set to the median euclidean distance between sample points.

Stopping Criterion: Stop as soon as we reject the H_0 of $\boldsymbol{S}$ and $\boldsymbol{U}$ following the same distributions.

Experiments: We self-train a support vector machine on subsamples of the `wine` dataset [3] with varying size. We observe SVM's test accuracy and the samples' MMD, see figure 4, and identify two prototypical cases: 1. High accuracy and low covariate shift go hand in hand, see first plot. 2. High accuracy and low covariate shift are competing goals, see second plot. Plots 3 and 4 highlight situations, where the MMD helps deciding among similarly well-performing learners (e.g., from iteration 3 and 19 in plot 3).

Figure 5: Test Accuracy of self-trained SVM and Maximum Mean Discrepancy of respective sample.

Lesson 3: *Post Hoc* Coresets

Motivation: Find a subsample of selected training data that balances covariate shift and predictive performance → sacrifice ϵ performance in order to mitigate covariate shift → Coresets, see e.g. [7].

Definition 3 ($(1+\epsilon)$-Coreset) *Let $X \in \mathbb{R}^{n \times p}$ be a set of data points, $\mathcal{R}(\cdot)$ an empirical risk function and $\beta \in \mathbb{R}^p$ the parameters of a (parametric) learner. Then a set $S \in \mathbb{R}^{k \times p}$ is called a $(1+\epsilon)$-Coreset of X for $\mathcal{R}(\cdot)$ if $k < n$ and*

$$\forall \beta \in \mathbb{R}^p : |\mathcal{R}(X, \beta) - \mathcal{R}(S, \beta)| \leq \epsilon \cdot \mathcal{R}(X, \beta)$$

Naive Approach: Find lowest ϵ such that the above inequality holds for a subsample with inverse inclusion probabilities (estimated through selection criterion, see lesson 1) → *expensive*

Sampling Based Coreset Constructions: Importance sampling with inverse sensitivity scores (worst-case importance for approximating the objective function on X) as inclusion probabilities [7].

Open Issue: How can we combine sensitivity scores and inverse inclusion probabilites for subsampling?