Bayesian Regression from Multiple Sources of Weak Supervision

Putra Manggala [1] **Holger H. Hoos** [2 3] **Eric Nalisnick** [1]

Abstract

We describe a Bayesian approach to weakly supervised regression. Our proposed framework propagates uncertainty from the weak supervision to an aggregated predictive distribution. We use a generalized Bayes procedure to account for the supervision being weak and therefore likely misspecified.

1. Introduction

The success of supervised learning crucially depends on the availability of high-quality labels. However, obtaining these labels can be expensive, time consuming, and privacy-intrusive. *Weak supervision* (Zhou, 2018) seeks to solve this problem by learning from an abundance of cheap but low-quality labels. For instance, *data programming* (Ratner et al., 2016) is one popular method for generating such labels: Domain experts write a series of programmatic functions—usually based on heuristics, crude assumptions, or inductive biases—that can quickly generate approximate supervision.

A problem with existing weak supervision methodologies is that they account only for the uncertainty and noise in the labels, with the goal of being model-agnostic. However, there is also uncertainty in the model parameters, and downstream performance likely depends on the interaction between label and model uncertainties. There is no principled feedback loop that allows the resulting predictive distribution to inform the modeling of the upstream label and model uncertainties.

We propose such a framework that models the uncertainties in weak supervision. We leverage the Bayesian paradigm due to its ability to incorporate prior knowledge and to quantify model uncertainty. We show how existing techniques to cope with model misspecification can naturally manage the misspecification in the labeling process. Moreover, we propose a series of unsupervised objectives to tune the final predictive distribution. Each objective instantiates distinct beliefs about the labeling process and its fidelity, allowing the user to select the most appropriate assumptions for the application.

Our core contributions are:

1. We define a framework that re-casts weak supervision within the Bayesian paradigm. We show how existing techniques for handling model misspecification can account for the likely misspecification in the labeling process.

2. We formulate a scalable methodology, using optimal transport to combine weak sources. Our aggregation method is modular, so that additional weak sources can be incorporated efficiently without rerunning the entire inference process.

3. We propose a set of interpretable tuning objectives that enable modellers to encode their meta-beliefs about the weak supervision. The Pareto front of these objectives allows for a trade-off between sharpness, diversity, and calibration of the weak data generating processes. We empirically validate our framework on simulated data.

2. Setting of Interest: Weak Supervision

Notation. Matrices are denoted by upper-case, bold-face letters (e.g. $\boldsymbol{Y}$), vectors using lower-case bold-face (e.g. $\boldsymbol{y}$), and scalars by regular letters (e.g. y or Y). We use italics to differentiate observations and constants (e.g. $\boldsymbol{y}$, $\boldsymbol{\theta}$) from random variables (e.g. $\mathbf{y}$, $\boldsymbol{\theta}$).

We consider the task of regression: predicting labels $\mathrm{y} \in \mathcal{Y} \subseteq \mathbb{R}$ from their paired K-dimensional covariates $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^K$. Throughout this paper, we assume access to a large (potentially unlimited) number of samples from $\mathbf{x} \sim P^*(\mathbf{x})$, the true generative process of the features. The true labels are generated conditionally via $\mathrm{y} \sim P^*(\mathrm{y}|\mathbf{x})$, but we assume we do not have access to this distribution *nor do we have access to samples from it*. The latter condition is the key assumption that motivates weak supervision.

To overcome the absence of labels, we specify weak supervision using L sets of weak examples: $\{\widetilde{\boldsymbol{D}}_1, \ldots, \widetilde{\boldsymbol{D}}_L\}$ where $\widetilde{\boldsymbol{D}}_l = (\widetilde{\boldsymbol{y}}_l, \boldsymbol{X}_l)$. We denote the collection of all weak exam-

[1]University of Amsterdam, Amsterdam, The Netherlands [2]Leiden University, Leiden, The Netherlands [3]University of British Columbia, Vancouver, Canada. Correspondence to: Putra Manggala <p.manggala@uva.nl>.

In Machine Learning for Data: Automated Creation, Privacy, Bias workshop at ICML 2021.

ples as $\widetilde{\boldsymbol{D}} = \bigcup_{l=1}^{L} \widetilde{\boldsymbol{D}}_l$. Each set is sampled from a unique weak generative process: $\widetilde{\boldsymbol{y}}_l \sim \psi_l(\widetilde{\mathrm{y}}|\boldsymbol{x})$, where ψ_l is usually constructed by a domain expert. These labeling mechanisms are all but always *misspecified*: $\psi_l(\widetilde{\mathrm{y}}|\boldsymbol{x}) \neq P^*(\mathrm{y}|\mathbf{x})$ for all l. While all feature sets are sampled from $P^*(\mathbf{x})$, note that each $\boldsymbol{X}_l$ is likely unique, possibly covering disjoint subsets of $\mathcal{X}$. This situation arises, because experts often have specialized knowledge, based on which they can provide supervision in distinct regions of $\mathcal{X}$ and thus combine their expertise effectively.

Our goal is to obtain a predictive model that, despite learning from weak data $\widetilde{\boldsymbol{D}}$, can usefully approximate the true generative process $P^*(\mathrm{y}|\mathbf{x})$. The naïve approach is to simply train a model using $\widetilde{\boldsymbol{D}}$. Many existing approaches (Gupta and Manning, 2014; Bunescu and Mooney, 2007; Ratner et al., 2016) improve upon the naïve solution by checking for consensus across $\psi_1, \ldots, \psi_L$. For example, Bach et al. (2019) use a generative model that produces a specific weight per weak instance. A weight near zero indicates that the labeling functions were not in agreement, and therefore the associated instance should not be influential in training.

3. Weak Bayesian Learning

We now describe our novel formulation of using weak supervision within the Bayesian workflow. Starting with the naïve formulation, one could simply fit a Bayesian model to $\widetilde{\boldsymbol{D}}$,

$$p(\theta \mid \widetilde{\boldsymbol{D}}) = \frac{p(\theta) \cdot \prod_{l=1}^{L} p(\widetilde{\boldsymbol{D}}_l \mid \theta)}{p(\widetilde{\boldsymbol{D}}_1, \ldots, \widetilde{\boldsymbol{D}}_L)} \tag{1}$$

and use the posterior predictive distribution at test time: $p(\mathrm{y} \mid \mathbf{x}, \widetilde{\boldsymbol{D}}) = \int p(\mathrm{y} \mid \mathbf{x}, \theta) \cdot p(\theta \mid \widetilde{\boldsymbol{D}})\, d\theta$. The prior $p(\theta)$ could be set to be diffuse or to incorporate another source of information or constraints.

3.1. Generalized Bayes for Misspecified Supervision

We aim to mitigate the influence of misspecified labels by automatically determining how much information should be extracted from each weak source and propose a generalized Bayes formulation akin to the *safe Bayes* framework (Grünwald and Van Ommen, 2017). In safe Bayes, the likelihood is equipped with an additional parameter η: $p(\mathrm{y}|\mathbf{x}, \theta)^{\eta}$. When $\eta = 0$, the data's influence is removed, and the posterior becomes the prior. When $\eta = 1$, the traditional Bayes rule is recovered.

The formulation in Equation 1 is well-suited to the generalized Bayes framework, since each labeling process is represented by a distinct term in the product likelihood. We can apply the safe Bayes framework analogously:

$$p(\theta \mid \widetilde{\boldsymbol{D}}; \boldsymbol{\eta}) \propto p(\theta) \cdot \prod_{l=1}^{L} p(\widetilde{\boldsymbol{D}}_l \mid \theta)^{\eta_l}, \tag{2}$$

where $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_L\}$ are termed the *learning rates* (Grünwald and Van Ommen, 2017), as they control the degree to which the model learns from the given data. With the (generalized) posterior in hand, we can then use the predictive distribution as our model of interest:

$$p(\mathbf{y} \mid \mathbf{X}, \widetilde{\boldsymbol{D}}; \boldsymbol{\eta}) = \int_{\theta} p(\mathbf{y} \mid \mathbf{X}, \theta) \cdot p(\theta \mid \widetilde{\boldsymbol{D}}; \boldsymbol{\eta})\, d\theta \tag{3}$$

The learning rates serve a similar role as described above but act on each weak source independently. When $\eta_l = 0$, the lth weak source does not contribute to the posterior and in turn the predictive distribution. On the other hand, when $\boldsymbol{\eta} = \mathbf{1}$, the naïve formulation in Equation 1 is recovered.

3.2. Tuning the Learning Rates

Having defined the learning rates η_l, we now need some objective by which to set them. Of course, if we had access to samples from $P^*(\mathrm{y}|\mathbf{x})$, we would use these to set $\boldsymbol{\eta}$, following Grünwald and Van Ommen (2017). Since we have only weak supervision, we propose setting $\boldsymbol{\eta}$ by imposing user-specified properties on the final predictive distribution. Specifically, we are concerned with the prediction interval

$$C_{\alpha}(\boldsymbol{x}; \boldsymbol{\eta}) = \{\mathrm{y} : p(\mathrm{y} \mid \boldsymbol{x}, \widetilde{\boldsymbol{D}}; \boldsymbol{\eta}) \geq \epsilon_{\alpha}\}, \tag{4}$$

where ϵ_{α} is set such that the probability of $C_{\alpha}(\boldsymbol{x}; \boldsymbol{\eta})$ is $1 - \alpha$, for $\alpha \in (0, 1)$. We optimize $\boldsymbol{\eta}$ by defining objectives that encourage the predictive distribution and interval to have properties such as sharpness or diversity. We describe several objectives in Section 5, after discussing some modifications to increase the scalability of the framework.

4. Scaling via Local Posterior Combination

Unfortunately, the framework presented in Section 3 incurs substantial computational cost when the posterior does not have an analytical solution, which is all but always the case. Specifically, when optimizing the learning rates, every change to η_l requires re-computing the posterior $p(\theta \mid \widetilde{\boldsymbol{D}}; \boldsymbol{\eta})$. If computing the posterior requires running Markov chain Monte Carlo (MCMC), then we will have to run MCMC to convergence for every update to $\boldsymbol{\eta}$ – a costly inner loop. Below we describe how to sidestep this issue by decoupling the computations using optimal transport.

Local Posteriors. We propose decoupling the weak sources by computing their individual posterior and predictive distributions. We term these *local* distributions, meaning the distribution is specific to a weak source. We modify Equation 2 to be

$$p(\theta \mid \widetilde{\boldsymbol{D}}_l; \eta_l) \propto p(\theta) \cdot p(\widetilde{\boldsymbol{D}}_l \mid \theta)^{\eta_l} \tag{5}$$

and compute one posterior per weak source, resulting in L total posteriors. L local predictive distributions $p(\mathbf{y} \mid$

$\mathbf{X}, \widetilde{\boldsymbol{D}}_l; \eta_l)$ are then computed just as in Equation 3, except that $p(\theta \mid \widetilde{\boldsymbol{D}}_l; \eta_l)$ is used for the posterior.

With L local predictive distributions in hand, we next combine them to obtain an aggregated predictor. We perform the combination via optimal transport – specifically, by finding the Wasserstein barycenter of the local predictive distributions.

Definition 1. ***Barycenter Predictive Distribution***. *Let $\{\mu_1, \ldots, \mu_L\}$ denote the measures corresponding to the L local predictive distributions. Assuming the space of measures with finite second moment, the Wasserstein barycenter predictive distribution is:*

$$\pi(\mathrm{y}|\boldsymbol{x}, \widetilde{\boldsymbol{D}}; \boldsymbol{\eta}) = \arg\min_{\mu} \sum_{l=1}^{L} \mathcal{W}_2^2(\mu, \mu_l), \tag{6}$$

where W_2^2 is the Wasserstein-2 distance.

With this formulation, modifying η_l changes only the lth local posterior, which in turn changes only the corresponding local predictive distribution $p(\mathbf{y} \mid \mathbf{X}, \widetilde{\boldsymbol{D}}_l; \eta_l)$ (denoted μ_l). However, we will still need to re-compute the barycenter π.

Univariate Method. In one dimension, the computation becomes easier still. The optimal transport problem can be trivially solved by averaging the quantile functions of the predictive distributions (Li et al., 2017):

$$Q_\pi(\mathrm{u}|\boldsymbol{x}, \widetilde{\boldsymbol{D}}; \boldsymbol{\eta}) = \frac{1}{L} \cdot \sum_{l=1}^{L} Q_l(\mathrm{u}|\boldsymbol{x}, \widetilde{\boldsymbol{D}}_l; \eta_l), \tag{7}$$

where $Q_\pi(\mathrm{u}|\boldsymbol{x}, \widetilde{\boldsymbol{D}}; \boldsymbol{\eta})$ is the quantile function of $\pi(\mathrm{y}|\boldsymbol{x}, \widetilde{\boldsymbol{D}}; \boldsymbol{\eta})$, and $Q_l(\mathrm{u}|\boldsymbol{x}, \widetilde{\boldsymbol{D}}_l; \eta_l)$ is the quantile function of $p(\mathbf{y} \mid \mathbf{X}, \widetilde{\boldsymbol{D}}_l; \eta_l)$. Thus, for continuous univariate regression, the only computation that is possibly burdensome is computing a local posterior and/or predictive distribution.

5. Objective Functions

We now return to the issue of defining objectives to tune the final predictive distribution as a function of $\boldsymbol{\eta}$.

Simulating weak examples. In order to calculate our tuning objectives, we require samples from the weak sources ψ_l. However, we assume that we have access only to $\widetilde{\boldsymbol{D}}_l$, which we have already used to fit the local distributions. Instead, when we require samples from the lth weak source, we draw samples from $p(\mathbf{y} \mid \mathbf{X}, \widetilde{\boldsymbol{D}}_l; \eta_l)$, as this should be a good model of ψ_l and provides an unlimited number of samples.

5.1. Objective #1: Calibration from Weak Sources

The first objective we consider is *calibration* with respect to the weak sources. We first need to define a generative model of the weak supervision that unifies the multiple sources. We specifying the following hierarchical model that uniformly mixes over the L sources: $l \sim \mathrm{Uniform}([1, L]), \mathbf{x} \sim \mathrm{Uniform}(\mathcal{A}^{\mathcal{X}})$, $\widetilde{\mathrm{y}} \sim p(\mathrm{y} \mid \mathbf{x}, \widetilde{\boldsymbol{D}}_l; \eta_l)$, where $\mathcal{A}^{\mathcal{X}}$ is the subset of feature space $\mathcal{X}$ for which we are interested in making predictions. We define calibration w.r.t. this generative model:

Definition 2. ***Calibration of weak supervision.*** *A predictive interval $C_\alpha(\mathbf{x}; \boldsymbol{\eta})$ is calibrated at $\mathcal{A}^{\mathcal{X}}$ with respect to a uniform mixture over weak sources if $P\left(\widetilde{\mathrm{y}} \in C_\alpha(\mathbf{x}; \boldsymbol{\eta})\right) = 1 - \alpha$ for all $\mathbf{x} \in \mathcal{A}^{\mathcal{X}}$.*

This objective ensures that the resulting predictive model is roughly aligned with the variability of the weak sources. We define a minimization objective for calibration as follows:

$$T_{\text{cal}}(\boldsymbol{\eta}; \alpha) = \mathbb{E}\left[\left|P\left(\widetilde{\mathrm{y}} \in C_\alpha(\mathbf{x}; \boldsymbol{\eta})\right) - (1 - \alpha)\right|\right] \tag{8}$$

We seek to minimize the difference between the proportion of weak examples that are contained in the corresponding predictive interval and the target $(1 - \alpha)$.

5.2. Objective #2: Sharpness

The sharpness of a predictive distribution corresponds to its concentration (Gneiting et al., 2007) and is typically measured via its variance. A sharper predictive distribution implies that the distribution is more concentrated around the mean. We define a minimization tuning objective $T_{\text{sharp}}(\boldsymbol{\eta})$ as simply the standard deviation of y under the predictive distribution $p(\mathrm{y} \mid \boldsymbol{x}, \widetilde{\boldsymbol{D}}; \boldsymbol{\eta})$. Intuitively, a sharper distribution generally implies that $C_\alpha(\mathbf{x}; \boldsymbol{\eta})$ has a smaller width.

5.3. Objective #3: Diversity of Sources

We measure diversity in terms of the relative contribution from the weak sources. Inference is most diverse when it is drawing information equally from each source (i.e. all entries of $\boldsymbol{\eta}$ are equal), and it is least diverse when just one source contributes (i.e. all entries of $\boldsymbol{\eta}$ are zero except for one). We define the maximization diversity objective as:

$$T_{\text{div}}(\boldsymbol{\eta}) = \mathbb{H}\left[\mathtt{softmax}\left(\boldsymbol{\eta}\right)\right], \tag{9}$$

where $\mathbb{H}$ denotes the entropy of the categorical distribution parameterized by a softmax transformation of the learning rate vector.

6. Related work

Weak supervision. In the crowdsourcing setting, different model-based approaches have been proposed to estimate workers' error probabilities (Dawid and Skene, 1979; Khetan and Oh, 2016; Kleindessner and Awasthi, 2018) or jointly model the labels and worker quality (Khetan et al., 2017). Weak labels can come from different sources, for

example crowdsourcing (Gao et al., 2011; Krishna et al., 2017), distant supervision (Mintz et al., 2009; Niu et al., 2012), and labelling heuristics (Gupta and Manning, 2014; Bunescu and Mooney, 2007; Ratner et al., 2016).

Generalized Bayes. The Generalized Bayes framework exponentiates the likelihood with a non-negative real learning rate before combining it with the prior distribution via Bayes theorem (Zhang et al., 2006). Various learning rate selection procedures with differing goals have been proposed. When performing divide-and-conquer strategy for Bayesian inference (Srivastava et al., 2018), learning rate is used to make sure that the subset posteriors have variances of the same order of magnitude to the true posterior. To ensure that the credible intervals for parameters are calibrated, Syring and Martin (2019) proposed bootstrapping the observed data and applying stochastic approximation method to tune the learning rate such that calibration is achieved. To remedy posterior predictive inconsistency, the SafeBayes algorithm (Grünwald and Van Ommen, 2017) uses grid search to select a learning rate which minimizes the cumulative expected log-loss.

7. Experiments

Simulated data. We generated three weak sources so that two are in relative agreement and the third is in disagreement with the first two. The covariates are specified on a grid, ranging from $[0, 0.7)$, $[0.3, 1)$, and $[0.3, 0.7)$ respectively with a step size of 0.05. The weak labels are sampled from a linear model: $\mathrm{y} = x + c_l + \epsilon$, $\epsilon \sim \mathrm{N}(0, 1)$. The source-specific intercept c_l is set as $\{5, 5, 10\}$ respectively. Weak sources $\widetilde{D}_1$ and $\widetilde{D}_2$ intersect in $(0.3, 0.7)$ and are in agreement by nature of having the same intercept ($c_1 = c_2 = 5$). $\widetilde{D}_3$ is also defined in $(0.3, 0.7)$ but is in disagreement, due to its larger intercept ($c_1 = 5$ vs $c_3 = 10$).

Bayesian optimization. We optimize the tuning objectives by employing the q-Expected Hypervolume Improvement (qEHVI) multi-objective Bayesian optimization algorithm (Daulton et al., 2020), a scalable and performant gradient-based EHVI approach. For each run, we make use of a Pareto frontier to arrive at a final solution set.

Model. We use a feedforward Bayesian neural network with one hidden layer containing one node. We use $\mathrm{N}(0, 2)$ as prior for all parameters and NUTS (Hoffman and Gelman, 2014) to perform posterior inference.

Barycenter predictive distributions. We visualize the weak sources and the barycenter predictive distributions in Figure 1. Subfigure (a) shows the solution using traditional Bayesian inference ($\boldsymbol{\eta} = 1$). While having an equal contribution from all sources may at first seem like a good inductive bias, we see from Subfigure (a) that, by trying to satisfy all sources, the solution cannot faithfully model any of them. Subfigure (b) presents a solution that has reached some consensus—favoring $\widetilde{D}_1$ and $\widetilde{D}_2$—by tuning the learning rates to achieve both sharpness and calibration.

7.1. Use cases and results

In order to motivate our multi-objective optimization, we describe a few practical settings, where specific tuning objectives can be used to incorporate an inductive bias to the resulting predictive distribution. We show the Pareto frontier for every setting.

7.1.1. Low stakes

In low-stakes applications, a modeller may decide to use a predictor that generally has a peaked predictive distribution, in order to make certain predictions and minimize abstains. Example applications include cold-start regimes of ranking and recommender systems—situations in which it is more valuable to collect feedback from users than to not serve any items. We optimize T_{sharp} and T_{cal} (Figure 2a) and choose a solution $\boldsymbol{\eta} = (0.4, 0.8, 0.5)$, which is sharp but still calibrated (calibration estimate < 0.05). In turn, the optimized predictive intervals (Figure 1) are also narrow, even though the learning rates are strictly less than 1. This also implies that the prior information influences the final predictive distribution more strongly.

7.1.2. High stakes

In high-stakes applications, a modeller may decide to use a predictor that has moderate sharpness, so that the predictive distribution is uncertain when there is label noise, and sufficiently high diversity, so that more "opinions" from different weak sources are aggregated. Example applications include autonomous vehicles and medical applications. We optimize T_{sharp} and T_{div}, while constraining T_{cal} (Figure 2b) such that all the Pareto frontier solutions are generally calibrated. We choose a solution $\boldsymbol{\eta} = (0.4, 0.7, 0.5)$ which is highly similar to the low-stakes case, but slightly more diverse and less sharp. Qualitatively, the resulting predictive distribution is very similar to that in Figure 1.

7.1.3. Balanced information gathering

In balanced information gathering mode, a modeller may seek to enforce an inductive bias where all weak sources should contribute the same amount of information when building the final predictive distribution. Example applications include settings where the opinions from all the members of a committee should carry the same weight. We optimize T_{div} and T_{cal} (Figure 2c) and choose a solution $\boldsymbol{\eta} = (0.4, 0.6, 0.6)$. Note that this calibrated solution has learning rates that are almost identical, which satisfies the

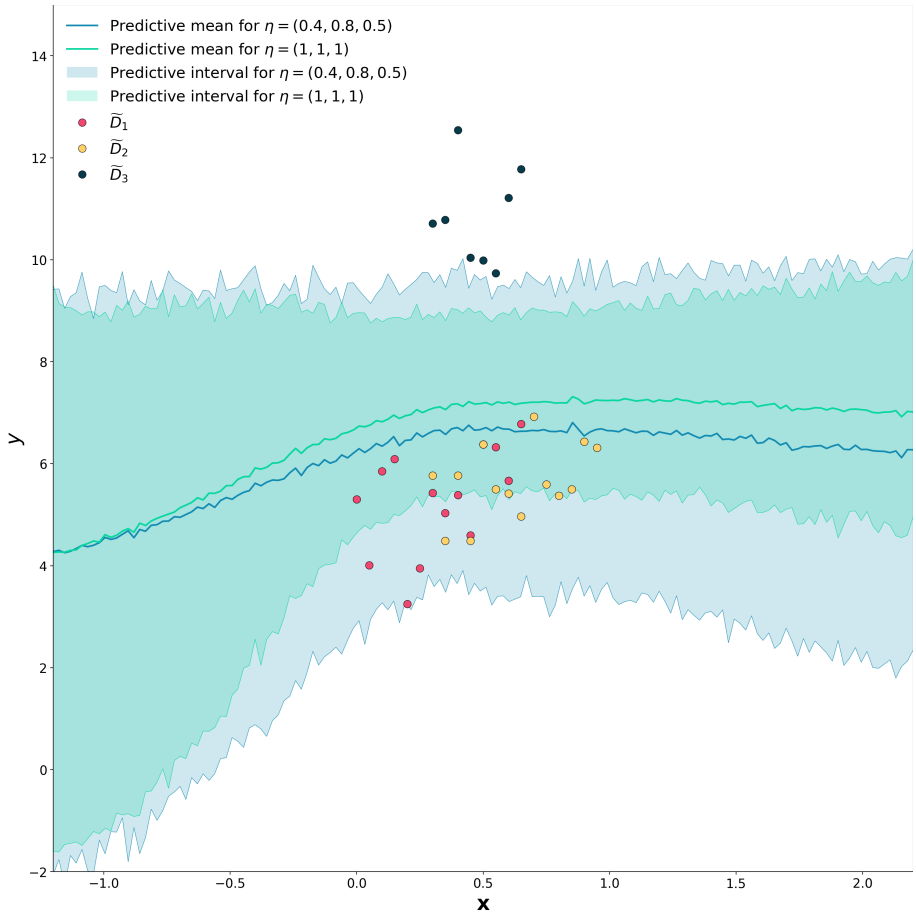


Figure 1. Barycenter Predictive Distributions. Green predictive mean and interval corresponds to performing traditional Bayesian inference ($\eta_l = 1$) for every $\widetilde{D}_l$ and then aggregating to the barycenter. Teal predictive mean and interval shows the solution obtained when simultaneously optimizing for calibration and sharpness. The prediction intervals are slightly wider than for traditional Bayesian inference, but this solution is calibrated with respect to the weak sources. Traditional Bayesian inference fails to accurately model the overall uncertainty from all the weak sources. It "overcounts" in the region where $\widetilde{D}_1$ and $\widetilde{D}_2$ overlap.

intended bias. Qualitatively the predictive distribution is very similar to that in Figure 1.

8. Conclusions, Limitations, and Future Work

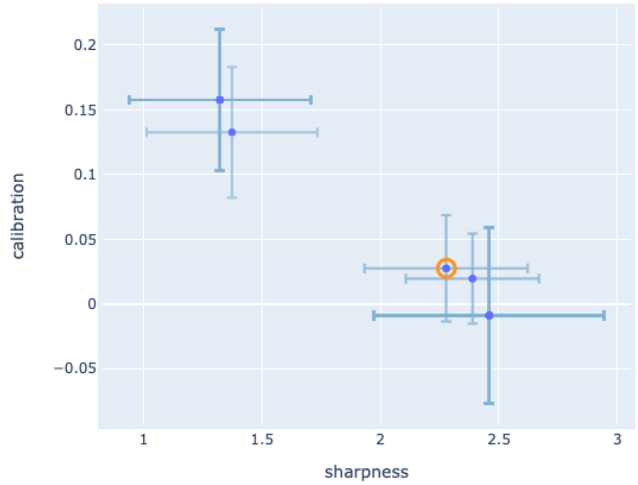
We have described a framework for uncertainty propagation in weakly supervised regression. Inspired by existing work on controlling for model misspecification, we employ learning rates to mitigate misspecification in the weak labels. We show how using the Wasserstein-2 Barycenter to aggregate the local distributions allows for a scalable, modular framework. Lastly, we use simulations to validate that our optimization framework can indeed produce learning rates whose Barycenter predictive distribution satisfy user-specified properties, such as calibration, sharpness, and diversity.

As for limitations, computing the local posteriors with high fidelity requires the user to be experienced in the Bayesian workflow (Gelman et al., 2020). Moreover, the posterior computations can be costly, as they need to be re-run when the corresponding learning rate is updated.

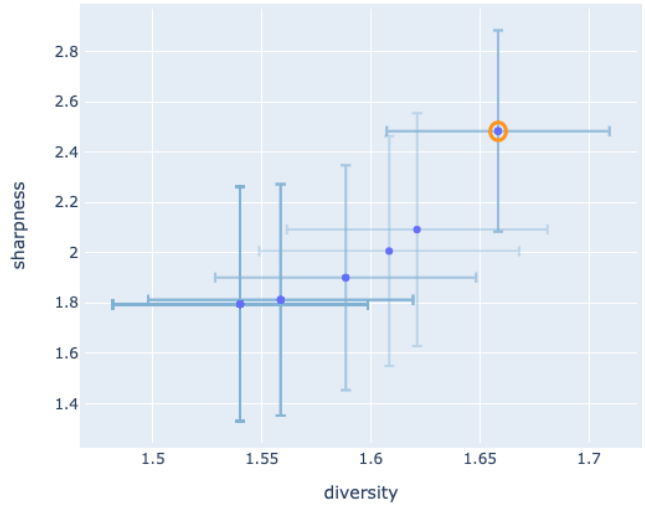
In future work, we would like to test our framework on larger regression problems and explore alternative formulations that improve computational tractability. This could involve using models that are expressive, while still admitting conjugacy (e.g. the neural linear model). We could also explore alternative aggregation strategies besides the barycenter.

9. Acknowledgments

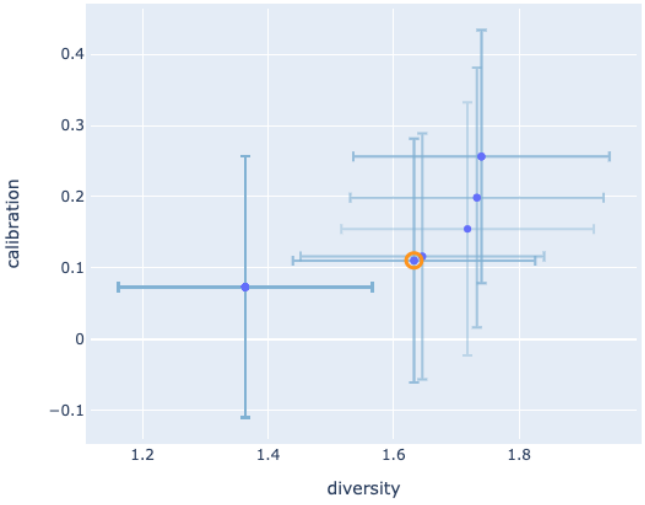
This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.



(a) Orange: $\boldsymbol{\eta} = (0.4, 0.8, 0.5)$



(b) Orange: $\boldsymbol{\eta} = (0.4, 0.7, 0.5)$



(c) Orange: $\boldsymbol{\eta} = (0.4, 0.6, 0.6)$

Figure 2. Pareto frontiers. Subfigure (a) is obtained from optimizing T_{cal} and T_{sharp}. Subfigure (b) is obtained from optimizing T_{sharp} and T_{div}. Subfigure (c) is obtained from optimizing T_{cal} and T_{div}. For each use case, we choose the solutions marked by the orange circles.

References

S. H. Bach, D. Rodriguez, Y. Liu, C. Luo, H. Shao, C. Xia, S. Sen, A. Ratner, B. Hancock, H. Alborzi, et al. Snorkel drybell: A case study in deploying weak supervision at industrial scale. In *Proceedings of the 2019 International Conference on Management of Data*, pages 362–375, 2019.

R. Bunescu and R. Mooney. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 576–583, 2007.

S. Daulton, M. Balandat, and E. Bakshy. Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization. *arXiv*, Jun 2020. URL https://arxiv.org/abs/2006.05078v3.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

H. Gao, G. Barbier, and R. Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3):10–14, 2011.

A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (2):243–268, 2007.

P. Grünwald and T. Van Ommen. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*, 12(4):1069–1103, 2017.

S. Gupta and C. D. Manning. Improved pattern learning for bootstrapped entity extraction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 98–108, 2014.

M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

A. Khetan and S. Oh. Reliable crowdsourcing under the generalized dawid-skene model. *arXiv preprint arXiv:1602.03481*, 2:28–56, 2016.

A. Khetan, Z. C. Lipton, and A. Anandkumar. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*, 2017.

M. Kleindessner and P. Awasthi. Crowdsourcing with arbitrary adversaries. In *International Conference on Machine Learning*, pages 2708–2717. PMLR, 2018.

R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

C. Li, S. Srivastava, and D. B. Dunson. Simple, scalable and accurate posterior interval estimation. *Biometrika*, 104(3):665–680, 2017.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

F. Niu, C. Zhang, C. Ré, and J. W. Shavlik. Deepdive: Web-scale knowledge-base construction using statistical learning and inference. *VLDS*, 12:25–28, 2012.

A. Ratner, C. De Sa, S. Wu, D. Selsam, and C. Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29: 3567, 2016.

S. Srivastava, C. Li, and D. B. Dunson. Scalable bayes via barycenter in wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

N. Syring and R. Martin. Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486, 2019.

T. Zhang et al. From e-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

Z.-H. Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.