Bayesian Weak Supervision via an Optimal Transport Approach

Putra Manggala[1] **Holger Hoos**[2] **Eric Nalisnick**[1]

[1]University of Amsterdam,
[2]RWTH Aachen University

Abstract

Large-scale machine learning is often impeded by a lack of labeled training data. To address this problem, the paradigm of weak supervision aims to collect and then aggregate multiple noisy labels. We propose a Bayesian probabilistic model that employs a tractable Sinkhorn-based optimal transport formulation to derive a ground-truth label. The translation between true and weak labels is cast as a transport problem with an inferred cost structure. Our approach achieves strong performance on the WRENCH weak supervision benchmark. Moreover, the posterior distribution over cost matrices allows for exploratory analysis of the weak sources.

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]—like crowdsourcing Raykar et al. [2010]—seeks to solve this problem by learning from an abundance of cheap but low-quality labels. Two-stage weak supervision is a popular paradigm: a model infers a latent ground-truth label from the weak sources, and the latent labels are used to train a downstream predictive model [Zhang et al., 2022]. The weak sources can range from being human experts to automated heuristics. In all cases but especially when the sources are humans, we wish to have an interpretable model that allows for easy identification of poorly- and well-performing labelers.

In this paper, we model the relationship between the latent label and observed weak label as an optimal transport problem Villani [2009]. The cost matrix then encodes the mislabeling tendencies of the weak source. This allows practitioners to identify weak sources that should be rebuilt or excluded from the current system. Iterative improvement of weak supervision systems via source analysis is an essential property for successful deployment.[1] Our contributions are:

1. We propose a novel generative model that re-casts weak supervision as an optimal transport problem between latent and observed (weak) labels. This is the first time optimal transport is used in the context of weak supervision.
2. We propose a post-hoc cost analysis that identifies poorly performing weak sources and their mislabeling tendencies. The latter can be used to create interpretable feedback to train and improve human labelers.
3. We empirically validate our model in two ways. Using simulated data, we show that we can identify poorly performing weak sources and improve the performance of a practical baseline method after pruning them. We also show that our method is highly competitive in the WRENCH benchmark [Zhang et al., 2021].

2 BACKGROUND

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). We use italics to differentiate observations and constants (e.g. y, $\boldsymbol{y}$) from random variables (e.g. y, $\mathbf{y}$).

We consider the task of weakly supervised classification: predicting labels $\mathrm{y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_N\}$ where the true label y_i for data point i takes a value $c \in \mathcal{C} = \{1, 2, \ldots, d\}$. We specify weak supervision using L sets of weak labels (each from a weak source): $\widetilde{\boldsymbol{y}} = \{\widetilde{\boldsymbol{y}}^1, \ldots, \widetilde{\boldsymbol{y}}^L\}$. Weak label $\widetilde{\boldsymbol{y}}_i^l$ for example i from weak source l takes a value from the set $\mathcal{C}^l$. The core assumption of weak supervision is that we have access to $\widetilde{\boldsymbol{y}}$ but not realizations of y. Hence our goal is to estimate a 'strong' latent label y from weak labels $\widetilde{\boldsymbol{y}}$.

[1]See for example: `https://snorkel.ai/programmatic-labeling/`

Accepted for the 5th Workshop on Tractable Probabilistic Modeling at UAI (TPM 2022).

In practice, due to a lack of expertise, we assume a weak source l can also abstain from providing a label ($\perp \in \mathcal{C}^l$) and a weak source may only be able to label certain values $\mathcal{C}^l \subseteq \mathcal{C}$). We also assume that weak sources are always more accurate than random guessing [Ratner et al., 2016].

Optimal Transport with Entropic Regularization We next define the optimal transport problem for discrete probability measures. Let $\boldsymbol{\nu}$ and $\boldsymbol{y}$ (a latent label) be discrete probability measures of the forms $\sum_{c \in \mathcal{C}^l} \boldsymbol{\nu}_c \delta_c$ and $\sum_{c \in \mathcal{C}} \boldsymbol{y}_c \delta_c$ respectively, where δ_c is the Dirac at position c, and $\boldsymbol{\nu}$ and $\boldsymbol{y}$ are probability simplex weights. The cost matrix $\mathbf{C} \in \mathbb{R}_+^{d \times d^l}$ stores all pairwise costs between values in $\mathcal{C}$ and $\mathcal{C}^l$. The entropic optimal transport problem obtains the optimal coupling matrix $\boldsymbol{P}^\lambda$ by solving:

$$\boldsymbol{P}^\lambda = \underset{\boldsymbol{P} \in \boldsymbol{U}(\boldsymbol{y}, \boldsymbol{\nu})}{\arg\min} \langle \boldsymbol{P}, \boldsymbol{C} \rangle - \lambda \mathbb{H}(\boldsymbol{P}), \tag{1}$$

where $\mathbb{H}(\boldsymbol{P}) = -\sum_{c,c'} \boldsymbol{P}_{c,c'}(\log \boldsymbol{P}_{c,c'} - 1)$ is the entropy of coupling $\boldsymbol{P}$, $\lambda > 0$ controls the regularization, and $\boldsymbol{U}(\boldsymbol{y}, \boldsymbol{\nu}) = \{\boldsymbol{P} \in \mathbb{R}^{d \times d^l} : \boldsymbol{P}\mathbf{1}_{d^l} = \boldsymbol{y} \text{ and } \boldsymbol{P}^T\mathbf{1}_d = \boldsymbol{\nu}\}$ is the set of all admissible couplings between $\boldsymbol{y}$ and $\boldsymbol{\nu}$. The optimal coupling matrix entry $\boldsymbol{P}^\lambda_{c,c'}$ represents the amount of mass transferred from $c \in \mathcal{C}$ to $c' \in \mathcal{C}^l$. As $\lambda \to 0$, $\boldsymbol{P}^\lambda$ converges to the Kantorovich optimal transport, which tends to admit a sparse optimal coupling matrix. In our problem, we favor the use of entropic regularization which admits a denser solution. This allows conflicting labels for a data point to be probabilistically combined without incurring a large transport cost during optimization. This bias is similar to that in minimax entropy approaches in crowdsourcing Zhou et al. [2012]. The entropic regularization in Equation 1 renders the problem efficiently solvable by Sinkhorn scaling [Cuturi, 2013].

3 AN OPTIMAL TRANSPORT MODEL

We propose the following hierarchical generative model that captures how strong latent labels are 'corrupted' by the weak sources to produce the weak labels. Specifically, this 'corruption' is encoded as a transportation problem between discrete measures. The weak sources move probability mass away from the latent label according to a per-source cost structure. Inferring a lower cost means that a weak labeler is more likely to flip the latent (presumably true) label to the (presumably incorrect) observed label.

Generative Model We write this model formally as follows. Each weak source $l \in [1, L]$ has a latent cost matrix $\boldsymbol{C}^l \sim p(\boldsymbol{C}^l; \gamma)$. We will discuss specification of $p(\boldsymbol{C}^l; \gamma)$ below. Each data point i for $i \in [1, N]$, is associated with a strong latent label $\mathrm{y}_i \sim \text{Dirichlet}(\boldsymbol{\rho_0})$ and a weak label

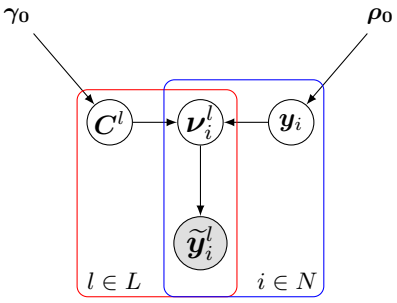

Figure 1: Directed graphical model with plates over weak sources L (red) and data points N (blue).

$\widetilde{\boldsymbol{y}}_i^l \sim \text{Categorical}(\boldsymbol{\nu}_i^l)$ for source l. We model the strong label using the Dirichlet distribution since the label model is expected to output probabilistic (soft) labels for downstream end models. This is different from other truth inference models which require a Categorical latent variable Dawid and Skene [1979], Kim and Ghahramani [2012] and have expensive posterior inference due to discrete enumeration. We now come to our key insight of using optimal transport to relate the weak and strong labels. We model the weak label's parameters $\boldsymbol{\nu}_i^l$ as a quantity proportional to the OT cost of 'break up' the mass of y_i and re-arranging it as $\boldsymbol{\nu}_i^l$ according to costs $\boldsymbol{C}^l$:

$$p\left(\boldsymbol{\nu}_i^l \mid \mathrm{y}_i, \boldsymbol{C}^l; \lambda, T, \psi\right) \propto \exp\left\{-\frac{\max\left(0, < \boldsymbol{P}^\lambda_{i,l}, \boldsymbol{C}^l > -\psi\, \mathbb{H}(\boldsymbol{C}^l)\right)}{T}\right\}, \tag{2}$$

where $\boldsymbol{P}^\lambda_{i,l}$ is the coupling, T is a temperature that controls the spread of the distribution, and ψ controls the amount of regularization due to the entropy of the cost matrix. This distribution over $\boldsymbol{\nu}_i^l$ is realistic since weak sources may not select the corresponding $\widetilde{\boldsymbol{y}}$ with the smallest optimal transportation cost. Informally, the mode of $\prod_l p(\boldsymbol{\nu}_i^l \mid \mathrm{y}_i, \boldsymbol{C}^l)$ can be thought of as a (regularized) barycenter across the weak sources. The plate diagram for the complete model is shown in Figure 1.

Specification Details As the optimal transport cost decreases to 0, the probability in Eqn 2 increases. In order to instill a mislabeling structure into the cost matrix, we set $\boldsymbol{C}^l \sim \text{Dirichlet}(\boldsymbol{\gamma_0})$, where $\boldsymbol{\gamma_0}$ is a $|\mathcal{C}| \times |\mathcal{C}^l|$ matrix where each row corresponds to a vector concentration parameter. Thus, each row of $\boldsymbol{C}^l$ is a Dirichlet random variable which sums up to one. In order to break the symmetry during cost inference, we set the diagonal concentration parameter value to be smaller than the off-diagonals. This matches the practical setting, as weak sources tend to label examples correctly. This prior choice can however be washed out by the observed weak labels. We set diagonal and off-diagonal concentration parameters to $\mathbf{1}$ and $\mathbf{2}$, respectively. Lastly,

we set $\boldsymbol{\gamma_0} = \mathbf{1}$. In practice, we find the use of lower temperature T and small cost-based regularization ψ to be favorable (see discussion in Figure 2).

Posterior Inference The Sinkhorn scaling procedure used to evaluate the RHS of Eqn 2 is differentiable and we can use Hamiltonian Monte Carlo (HMC) to perform sampling Neal [2011]. We use the expectation of y under the marginal posterior distribution as a prediction for the true label.

4 RELATED WORK

Weak Supervision and Crowdsourcing. Weak labels can come from different sources, for example crowd annotations [Krishna et al., 2017], distant supervision [Mintz et al., 2009], and labelling heuristics [Gupta and Manning, 2014, Bunescu and Mooney, 2007]. Various two-stage models have been proposed and they differ in the way they model the joint distribution of observed weak labels and latent true label [Ratner et al., 2016, 2019, Fu et al., 2020]. 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]. Bayesian aggregation methods have been proposed to allow for prior specification and uncertainty quantification [Kim and Ghahramani, 2012, Raykar et al., 2010, Paun et al., 2018].

Inverse and Belief Transport The inverse optimal transport problem seeks to recover the cost matrix from observed couplings [Chiu et al., 2021, Stuart and Wolfram, 2020, Cuturi and Avis, 2014]. The idea of transporting human beliefs using forward optimal transport has been proposed in the context of modeling cooperation [Shafto et al., 2021].

5 EXPERIMENTS

5.1 COST ANALYSIS

We describe how to use the the posterior cost matrices $\{\hat{\boldsymbol{C}}^l\}_{l\in L}$, i.e., $\hat{\boldsymbol{C}}^l = \mathbb{E}_{\boldsymbol{C}^l|\widetilde{\boldsymbol{y}}}\left[\boldsymbol{C}^l\right]$ to analyze the quality of weak sources. Without loss of generality we assume $\mathcal{C} = \mathcal{C}^l$. The diagonal entry $\hat{\boldsymbol{C}}_{c,c}$ corresponds to the cost of correctly labeling class c, while the off-diagonal entries $\hat{\boldsymbol{C}}_{c,c'}$'s correspond to the costs of incorrectly labeling class c as c'. To determine the quality of weak source l, we compute its health score:

$$\texttt{HS}(l) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\left(\frac{1}{|\mathcal{C}|-1}\sum_{c'\in\mathcal{C}\setminus c}\hat{\boldsymbol{C}}^l_{c,c'}\right) - \hat{\boldsymbol{C}}^l_{c,c}, \quad (3)$$

which is the expected difference between the average off-diagonal cost and diagonal cost. Intuitively, when the health score is high, the weak source is good at labeling that class since the diagonal has a generally lower cost than the off-diagonal.

Synthetic Data We consider a binary classification problem with cost matrix of the form $\left(\begin{smallmatrix}\alpha & 1-\alpha\\ 1-\alpha & \alpha\end{smallmatrix}\right)$. We consider a weak source to be good when $\alpha \in [0, 0.2]$ and bad when $\alpha \in [0.8, 1]$. This cost matrix structure has a very predictable impact on simple aggregation methods like majority voting – a performance increase can be isolated to the addition of good sources and/or deletion of bad sources. If a source is deleted and the performance of majority voting increases, then the deleted source is a bad weak source. Using the health scores, we can rank the weak sources and delete the lowest ranked ones.

We create $L = 12$ weak sources by fixing 8 good and 4 bad weak sources by sampling the appropriate α's. We create $N = 1000$ true labels by sampling uniformly from a binary $\mathcal{C}$. Using the generative model in Figure 1, we perform forward sampling to obtain the weak labels by conditioning on the the costs and true labels. Conditioning on these weak labels, we perform posterior inference using Hamiltonian Monte Carlo (HMC) to obtain estimates of the true labels and compute the health scores. Sampling diagnostics show that HMC has been successful in exploring the posteriors.

Method	Accuracy	Brier score
EOT	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.12 \pm 0.08}$
Majority	0.75	0.25
Majority, Pruned	0.81	0.19

Table 1: *Predictive performance against simulated data.* We compare accuracy and Brier score across three methods: EOT (our model), Majority voting, and Majority (pru) voting after EOT cost-based pruning. Credible intervals for EOT are obtained from y's posterior simulation sets. There is no uncertainty for Majority and Majority (pruned) because they are only computed once from the simulated weak labels.

To measure the accuracy of our true label estimates, we compare the accuracy and Brier score of our approach (EOT) to majority voting (Majority). From Table 1, we see that EOT is able to improve upon Majority, as it is able to infer the right cost structure. We then delete two weak sources with the lowest health scores and run majority voting on the remaining weak sources (Majority (pru)). We choose to only delete two weak sources, since heuristically in practice we may not want to delete more than 20% of the available weak sources. We find that the two deleted weak sources are bad weak sources. In general, we can use a small amount of ground truth development set, typically available in practice [Ratner et al., 2019], to determine the number of weak sources to prune. We see that this improves upon Majority,

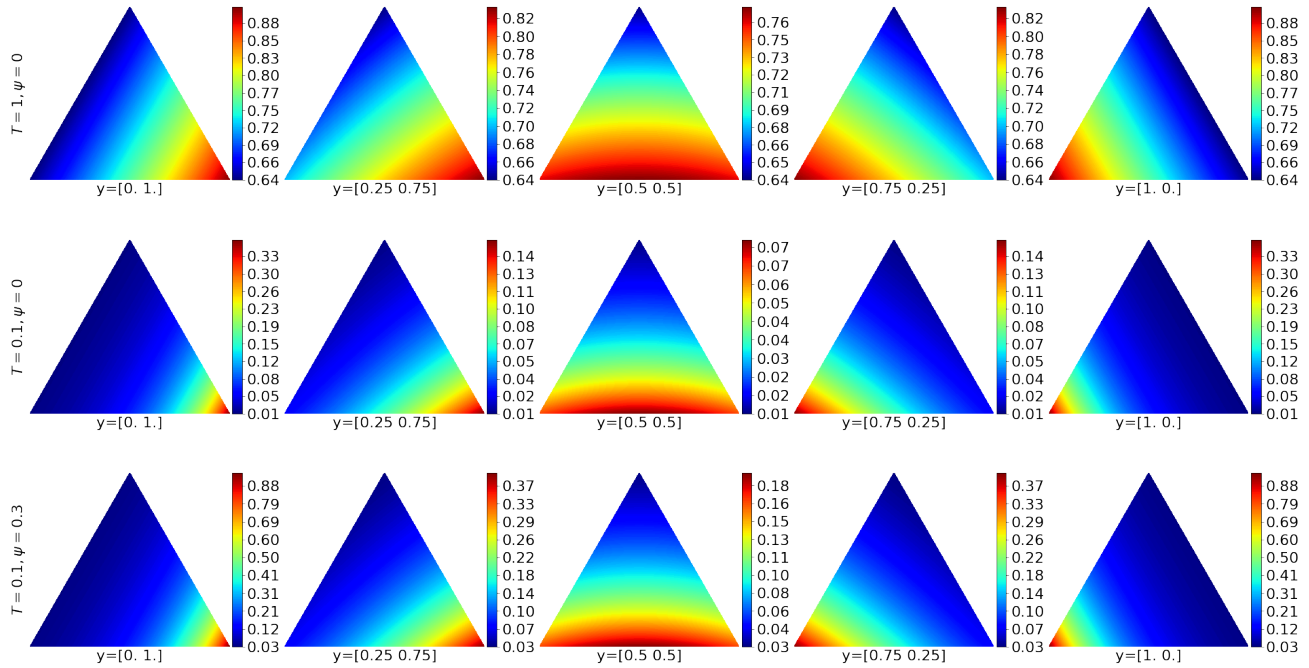

Figure 2: *Unnormalized probability density from Eqn 2*: Each row corresponds to different values of T and ψ. Each column corresponds to the density of 3-dimensional ν given fixed cost matrix $\boldsymbol{C} = \left(\begin{smallmatrix} 0.1 & 0.45 & 0.45 \\ 0.45 & 0.1 & 0.45 \end{smallmatrix}\right)$ and 2-dimensional $\boldsymbol{y}$. This cost matrix prefers correct weak label assignment (with 0.1 cost) over incorrect assignment and abstaining. First row shows the density when $T = 1, \psi = 0$, which corresponds to no temperature scaling and cost-based regularization. Second row sets $T = 0.1, \psi = 0$, which transforms the density to be more peaked and the downstream categorical choice to be more concentrated. This tends to occur in practice, as weak sources tend to be confident at labeling certain classes. Third row sets $T = 0.1, \psi = 0.3$, which increases the absolute magnitude and widens the range of density according to the entropy of $\boldsymbol{C}$. This causes density levels across different cost matrix entropy levels to be more similar and more posterior mass is assigned to high entropy cost matrices.

as bad weak sources have been removed. This is impactful in practice, as we can use EOT to prune weak sources without using it as a long-running label model and instead use simpler baseline methods.

5.2 WRENCH BENCHMARK

The WRENCH benchmark includes various datasets and label model implementations for evaluating weakly supervised learning [Zhang et al., 2021]. We evaluate our method (EOT) against other two-stage model baselines from weak supervision: majority voting (Majority), Snorkel [Ratner et al., 2016] and Flying Squid [Fu et al., 2020] and crowdsourcing: Dawid Skene (DS) [Dawid and Skene, 1979], IBCC [Kim and Ghahramani, 2012], and Hierarchical Dawid Skene (HDS) [Paun et al., 2018]. Snorkel and Flying Squid are popular label models in the weak supervision community. DS is fitted using a closed-form Expectation-Maximization algorithm, whereas IBCC and HDS received a fully Bayesian treatment and are fitted using Hamiltonian Monte Carlo. Similar sampler settings are used for IBCC, HDS and EOT and sampling diagnostics show that HMC has been successful in exploring the posteriors.

Method	Accuracy
EOT	**0.796** $\pm$.02
Majority	0.761 $\pm$.04
Snorkel	0.747 $\pm$.03
Flying Squid	0.752 $\pm$.03
DS	0.547 $\pm$.02
IBCC	0.611 $\pm$.04
HDS	0.690 $\pm$.15

Table 2: *Accuracy results on IMDB dataset.*

We perform our experiments on the `imdb` dataset, which is a dataset for binary sentiment classification with 20,000 movie reviews for training, 2,500 for validation and 2,500 for testing. The weak sources are 4 heuristics rules on keywords and 1 heuristics rule on expressions. This dataset and its weak sources are generated in Ren et al. [2020]. Since the features are textual, we use BERT as the end model [Devlin et al., 2018]. We measure accuracy as a performance

metric. From Table 2, we observe that our model (EOT) is competitive as a label model.

6 CONCLUSION

We have proposed a novel generative model that exploits optimal transport to translate strong labels into weak ones. A latent cost matrix is estimated for each weak source, giving us insight into the source's misspecification w.r.t. the latent label. We demonstrated in the experiments how inspecting the posterior cost matrices can be used to prune poor-performing weak sources. Moreover, our empirical results suggest that our model is competitive with state-of-the-art methods for weak supervision (e.g. Snorkel) and crowdsourcing. In future work, we plan to explore alternative cost priors, to analyze scenarios where weak sources are only able to label certain values ($\mathcal{C}^l \subset \mathcal{C}$), and to scale the work to larger data sets.

7 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`.

References

Razvan Bunescu and Raymond 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.

Wei-Ting Chiu, Pei Wang, and Patrick Shafto. Probabilistic inverse optimal transport. *arXiv preprint arXiv:2112.09754*, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Marco Cuturi and David Avis. Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564, 2014.

Alexander Philip Dawid and Allan 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Daniel Fu, Mayee Chen, Frederic Sala, Sarah Hooper, Kayvon Fatahalian, and Christopher Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. In *International Conference on Machine Learning*, pages 3280–3291. PMLR, 2020.

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

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

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

Hyun-Chul Kim and Zoubin Ghahramani. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, pages 619–627. PMLR, 2012.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David 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.

Mike Mintz, Steven Bills, Rion Snow, and Dan 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.

Radford M Neal. MCMC Using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585, 2018.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29:3567, 2016.

Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4763–4771, 2019.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11(4), 2010.

Wendi Ren, Yinghao Li, Hanting Su, David Kartchner, Cassie Mitchell, and Chao Zhang. Denoising multi-source weak supervision for neural text classification. *arXiv preprint arXiv:2010.04582*, 2020.

Patrick Shafto, Junqi Wang, and Pei Wang. Cooperative communication as belief transport. *Trends in cognitive sciences*, 25(10):826–828, 2021.

Andrew M Stuart and Marie-Therese Wolfram. Inverse optimal transport. *SIAM Journal on Applied Mathematics*, 80(1):599–619, 2020.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. Wrench: A comprehensive benchmark for weak supervision. *arXiv preprint arXiv:2109.11377*, 2021.

Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. A survey on programmatic weak supervision. *arXiv preprint arXiv:2202.05433*, 2022.

Dengyong Zhou, Sumit Basu, Yi Mao, and John Platt. Learning from the wisdom of crowds by minimax entropy. *Advances in neural information processing systems*, 25, 2012.

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