IT IS LIKELY THAT YOUR LOSS SHOULD BE A LIKELIHOOD

Anonymous authors
Paper under double-blind review

ABSTRACT

Many common loss functions such as mean-squared-error, cross-entropy, and reconstruction loss are unnecessarily rigid. Under a probabilistic interpretation, these common losses correspond to distributions with fixed shapes and scales. We instead argue for optimizing full likelihoods that include parameters like the normal variance and softmax temperature. Joint optimization of these "likelihood parameters" with model parameters can adaptively tune the scales and shapes of losses in addition to the strength of regularization. We explore and systematically evaluate how to parameterize and apply likelihood parameters for robust modeling, outlier-detection, and re-calibration. Additionally, we propose adaptively tuning L_2 and L_1 weights by fitting the scale parameters of normal and Laplace priors and introduce more flexible element-wise regularizers.

1 INTRODUCTION

Choosing the right loss matters. Many common losses arise from likelihoods, such as the squared error loss from the normal distribution , absolute error from the Laplace distribution, and the cross entropy loss from the softmax distribution. The same is true of regularizers, where L_2 arises from a normal prior and L_1 from a Laplace prior.

Deriving losses from likelihoods recasts the problem as a choice of distribution which allows data-dependant adaptation. Standard losses and regularizers implicitly fix key distribution parameters, limiting flexibility. For instance, the squared error corresponds to fixing the normal variance at a constant. The full normal likelihood retains its scale parameter and allows optimization over a parametrized set of distributions. This work examines how to jointly optimize distribution and model parameters to select losses and regularizers that encourage generalization, calibration, and robustness to outliers. We explore three key likelihoods the normal, softmax, and the robust regressor likelihood ρ of Barron (2019). Additionally, we cast adaptive *priors* in the same light and introduce adaptive regularizers. In summary:

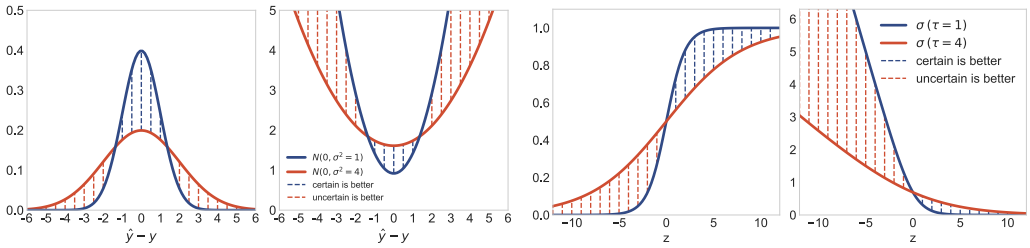
1. We systematically survey and evaluate global, data, and predicted likelihood parameters and introduce a new self-tuning variant of the robust adaptive loss ρ
2. We apply likelihood parameters to create new classes of robust models, outlier detectors, and re-calibrators.
3. We propose adaptive versions of $L1$ and $L2$ regularization using parametrized normal and Laplace priors on model parameters.

2 BACKGROUND

Notation We consider a dataset $\mathcal{D}$ of points x_i and targets y_i indexed by $i \in \{1, \ldots, N\}$. Targets for regression are real numbers and targets for classification are one-hot vectors. The model f with parameters θ makes predictions $\hat{y}_i = f_\theta(x)$. A loss $L(\hat{y}, y)$ measures the quality of the prediction given the target. To learn model parameters we solve the following loss optimization:

$$\min_\theta \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} L(\hat{y} = f_\theta(x), y) \tag{1}$$

A likelihood $\mathcal{L}(\hat{y}|y, \phi)$ measures the quality of the prediction as a distribution over $\hat{y}$ given the target y and likelihood parameters ϕ. We use to the negative log-likelihood ℓ (NLL), and the likelihood



(a) The Normal PDF and NLL (b) The Softmax CDF and NLL

Figure 1: Optimizing likelihood parameters adapts the loss without manual hyperparameter tuning to balance accuracy and certainty.

interchangeably since both have the same optima. We define the full likelihood optimization:

$$\min_{\theta,\phi} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \ell(\hat{y} = f_\theta(x)|y, \phi) \tag{2}$$

to jointly learn model and likelihood parameters. "Full" indicates the inclusion of ϕ, which control the distribution and induced NLL loss. We focus on full likelihood optimization in this work. We note that the target, y, is the only supervision needed to optimize model and likelihood parameters, θ and ϕ respectively. Additionally, though the shape and scale varies with ϕ, reducing the error $\hat{y} - y$ always reduces the NLL for our distributions.

Distributions Under Investigation This work considers the normal likelihood with variance σ (Bishop et al., 2006; Hastie et al., 2009), the softmax likelihood with temperature τ (Hinton et al., 2015), and the robust likelihood ρ (Barron, 2019) with shape α and scale σ that control the scale and shape of the likelihood. We note that changing the scale and shape of the likelihood distribution is not "cheating" as there is a trade-off between uncertainty and credit. Figure 1 shows how this trade-off affects the Normal and softmax distributions and their NLLs.

The normal likelihood has terms for the residual $\hat{y} - y$ and the variance σ as

$$\mathcal{N}(\hat{y}|y, \sigma) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(\hat{y} - y)^2}{\sigma^2}\right), \tag{3}$$

with $\sigma \in (0, \infty)$ scaling the distribution. The normal NLL can be written $\ell_{\mathcal{N}} = \frac{1}{2\sigma^2}(\hat{y} - y)^2 + \log\sigma$, after simplifying and omitting constants that do not affect minimization. We recover the squared error by substituting $\sigma = 1$.

The softmax defines a categorical distribution defined by scores z for each class c as

$$\text{softmax}(\hat{y} = y|z, \tau) = \frac{e^{z_y \tau}}{\sum_c e^{z_c \tau}}, \tag{4}$$

with the temperature, $\tau \in (0, \infty)$, adjusting the entropy of the distribution. The softmax NLL is We recover the classification cross-entropy loss, $-\log p(\hat{y} = y)$, by substituting $\tau = 1$ in the respective NLL. We state the gradients of these likelihoods with respect to their σ and τ in Section A of the supplement.

The robust loss ρ and its likelihood are

$$\rho(x, \alpha, \sigma) = \frac{|\alpha - 2|}{\alpha}\left(\left(\frac{(x/\sigma)^2}{|\alpha - 2|} + 1\right)^{\alpha/2} - 1\right) \text{ and} \tag{5}$$

$$p(\hat{y} \mid y, \alpha, \sigma) = \frac{1}{\sigma Z(\alpha)} \exp\left(-\rho(\hat{y} - y, \alpha, \sigma)\right), \tag{6}$$

with shape $\alpha \in [0, \infty)$, scale $\sigma \in (0, \infty)$, and normalization function $Z(\alpha)$. This likelihood generalizes the normal, Cauchy, and Student's t distributions.

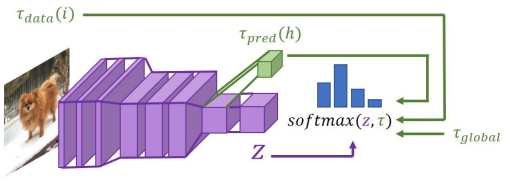


Figure 2: An image classifier with three different types of temperature conditioning: global, model, and data.

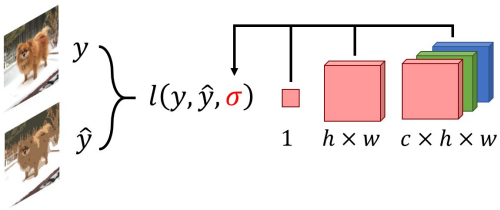


Figure 3: An image loss function with three different likelihood parameter dimensionalities

3 Related Work

Likelihood optimization follows from maximum likelihood estimation (Hastie et al., 2009; Bishop et al., 2006), yet is uncommon in practice for fitting deep regressors and classifiers for discriminative tasks. However Kendall & Gal (2017); Kendall et al. (2018); Barron (2019); Saxena et al. (2019) optimize likelihood parameters to their advantage yet differ in their tasks, likelihoods, and parameterizations. In this work we aim to systematically experiment, clarify usage, and encourage their wider adoption.

Early work on regressing means and variances (Nix & Weigend, 1994) had the key insight that optimizing the full likelihood can fit these parameters and adapt the loss. Some recent works use likelihoods for loss adaptation, and interpret their parameters as the uncertainty (Kendall & Gal, 2017; Kendall et al., 2018), robustness (Kendall & Gal, 2017; Barron, 2019; Saxena et al., 2019), and curricula (Saxena et al., 2019) of losses. MacKay (1992) uses Bayesian evidence to select hyper-parameters and losses based on proper likelihood normalization. Barron (2019) define a generalized robust regression loss, ρ, to jointly optimize the type and degree of robustness with global, data-independent, parameters. Kendall & Gal (2017) predict variances for regression and classification to handle data-dependent uncertainty. Kendall et al. (2018) balance multi-task loss weights by optimizing variances for regression and temperatures for classification. These global parameters depend on the task but not the data, and are interpreted as inherent task uncertainty. Saxena et al. (2019) define a differentiable curriculum for classification by assigning each training point its own temperature. These data parameters depend on the index of the data but not its value. We compare these different likelihood parametizations across tasks and distributions.

In the calibration literature, Guo et al. (2017) have found that deep networks are often miscalibrated, but they can be re-calibrated by cross-validating the temperature of the softmax. In this work we explore several generalizations of this concept. Alternatively, Platt scaling (Platt, 1999) fits a sigmoid regressor to model predictions to calibrate probabilities. Kuleshov et al. (2018) re-calibrate regressors by fitting an Isotonic regressor to the empirical cumulative distribution function.

4 Likelihood Parameter Types

We explore the space of likelihood parameter representations for model optimization and inference. Though we note that some losses, like adversarial losses, are difficult to represent as likelihoods, many different losses in the community have a natural probabilistic interpretation. Often, these probabilistic interpretations can be parametrized in a variety of ways. We explore two key axes of generality when building these loss functions: conditioning and dimensionality.

Conditioning We represent the likelihood parameters by three functional classes: global, data, and predicted. *Global* parameters, $\phi = c$, are independent of the data and model and define the same likelihood distribution for all points. *Data* parameters, ϕ_i, are conditioned on the index, i, of the data, x_i, but not its value. Every training point is assigned an independent likelihood parameter, ϕ_i that define different likelihoods for each training point. *Predicted* parameters, $\phi(x) = g_\eta(x)$, are determined by a model, g, with parameters η (not to be confused with the task model parameters θ). Global and predicted parameters can be used during training and testing, but data parameters are only assigned to each training point and are undefined for testing. We show a simple example of predicted temperature in Figure 4, and an illustration of the parameter types in Figure 2.

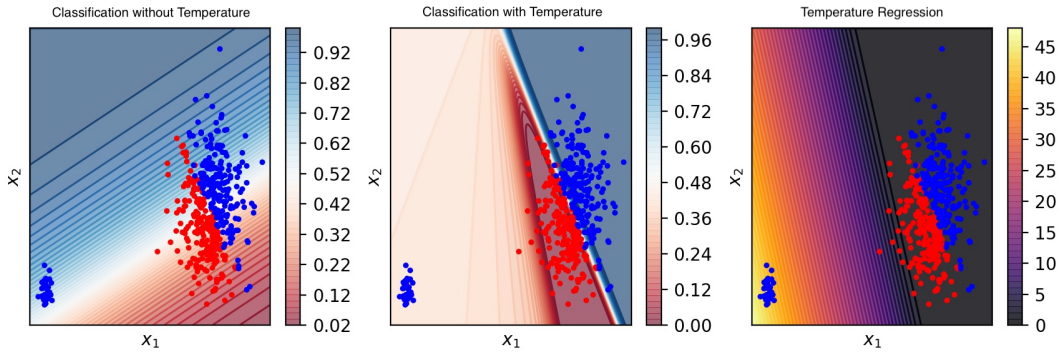


Figure 4: A synthetic logistic regression experiment. Regressing softmax temperature reduces the influence of outliers (blue, bottom-left), by locally raising temperature. The jointly optimized model achieves a more accurate classification.

We note that for certain global parameters like a learned Normal scale, changing the scale does not affect the optima, but does change the probabilistic interpretation. This invariance has led many authors to drop the scale from their formulations. However, when models can predict these scale parameters they can naturally remain calibrated in the presence of heteroskedasticity and outliers. Additionally we note that for the shape parameter of the robust likelihood, ρ, changing global parameters does affect model fitting. Previous works have adapted a global softmax temperature for model distillation (Hinton et al., 2015), and recalibration (Guo et al., 2017).

Dimensionality The dimensionality, $|\phi|$, of likelihood parameters can vary with the dimension of the task prediction, $\hat{y}$. For example, image regressors can use a single likelihood parameter for each image $|\phi| = 1$, RGB image channel $|\phi| = C$, or even every pixel $|\phi| = W \times H \times C$ as in Figure 3. These choices correspond to different likelihood distribution classes. Dimensionality and Conditioning of likelihood parameters can interact. For example, data parameters with $|\phi| = W \times H \times C$ would result in $N \times W \times H \times C$ additional parameters, where N is the size of the dataset. This can complicate implementations and slow down optimization due to disk I/O when their size exceeds memory. Table 7 in the appendix contrasts the computational requirements of different likelihood parameter types.

5 Applications

5.1 Robust Modeling

Data in the wild is noisy, and machine learning methods should be robust to noise, heteroskedasticity, and corruption. The standard mean squared error (MSE) loss is highly susceptible to outliers due to its fixed variance (Huber, 2004). Likelihood parameters naturally transform standard methods such as regressors, classifiers, and manifold learners into robust variants without expensive outer-loop of model fitting such as RANSAC (Fischler & Bolles, 1981) and Theil-Sen (Theil, 1992). Figure 4 demonstrates this effect with a simple classification dataset, and we point readers to Figures 9 and 10 of the Supplement for similar examples for regression and manifold learning. By mitigating the harmful effect of outliers, regressing variance can often improve generalization ability in terms of MSE, accuracy, and NLL on an unseen test set. Furthermore, by allowing models to adjust their certainty in a principled manner, this also reduces the miscalibration (CAL) of regressors. Table 2 shows this effect for deep models on the datasets used in (Kuleshov et al., 2018).

Often the assumption of Gaussianity is too restrictive for regression models and one must consider more robust approaches. This has led many to investigate broader classes of likelihoods such as Generalized Linear Models (GLMs) (Nelder & Wedderburn, 1972) or the more recent general robust loss, ρ, of Barron (2019). To systematically explore how likelihood parameter dimension and conditioning affect model robustness and quality, we reproduce Barron (2019)'s variational auto-encoding (Kingma & Ba, 2015) (VAE) experiments on faces from the CelebA dataset (Liu et al., 2015). We compare global, predicted, and data parameters (Saxena et al., 2019) as well as two natural choices of parameter dimensionality: a single set of parameters for the whole image, and a

Table 1: MSE, Time, and Memory increase (compared to standard normal likelihood) for reconstruction by variational auto-encoders with different parameterizations of the robust loss, ρ. Predicted likelihood parameters yield more accurate reconstruction models.

Param.	Dim	MSE	Time	Mem
Global	1×1×1	225.8	**1.04**×	$<$1**KB**
Data	...	244.2	2.70×	0.6GB
Pred.	...	228.5	**1.04**×	$<$1MB
Global	H×W×C	231.1	1.08×	$<$1MB
Data	...	252.6	9.42×	4.4GB
Pred.	...	**222.3**	1.08×	$<$1MB

Table 2: Effect of predicted likelihood parameters on Calibration (CAL) (Kuleshov et al., 2018), MSE, and NLL evaluated on test data for deep regressors. Linear results are in Table 6 of the supplement.

	CAL		MSE		NLL	
DATASET	BASE	TEMP	BASE	TEMP	BASE	TEMP
CRIME	**0.018**	0.146	0.028	**0.088**	**220.1**	478.4
KIN.	0.122	**0.001**	0.064	**0.006**	0.061	**-0.471**
BANK	0.417	**0.001**	0.127	**0.008**	1.361	**-1.464**
WINE	0.023	**0.003**	0.032	**0.011**	1.614	**0.302**
MPG	0.208	**0.006**	0.083	**0.020**	**0.307**	5.050
CPU	0.554	**0.021**	0.150	**0.022**	**-0.102**	12.218
SOIL	0.602	**0.307**	0.160	**0.100**	-0.131	**-4.078**
FRIED	0.472	**0.000**	0.129	**0.002**	0.301	**-1.039**

set of parameters for each pixel and channel. We note that predicted parameters achieve the best performance while maintaining speed and a small memory footprint. We visualize these parameters in Section E of the Appendix, and show that they demarcate challenging areas of target images. More details on experimental conditions, datasets, and models are provided in Sections I and J in the appendix.

5.2 OUTLIER DETECTION

Likelihood parameter prediction gives models a direct channel to express their "uncertainty" for each data-point with respect to the task. This allows models to naturally down-weight and clean outliers from the dataset which can improve model robustness. Consequently, one can harness this effect to create outlier detectors from *any* underlying model architecture by using learned scales or temperatures as an outlier score function. Furthermore, predicted likelihood parameters allow these methods to detect outliers in unseen data. In Figure 5 we show how auditing temperature or noise parameters can help practitioners spot erroneous labels and poor quality examples. In particular, the temperature of an image classifier correlates strongly with blurry, dark, and difficult examples on the Street View House Number (SVHN) dataset.

One can also leverage auto-encoders (Kramer, 1991) or self-supervision (Zhang et al., 2016; Tian et al., 2019) to yield label-free measures of uncertainty. We use this approach to create a simple outlier detection algorithms by considering deep (AE+S) and linear (PCA+S) auto-encoders with a data-conditioned scale parameters as outlier scores. In Table 3 we quantitatively demonstrate the quality of these simple likelihood parameter approaches across 22 datasets from the Outlier Detection Datasets (ODDS), a standard outlier detection benchmark (Rayana, 2016). The ODDS benchmark supplies ground truth outlier labels for each dataset, which allows one to treat outlier detection as an unsupervised classification problem. We compare against a variety of established outlier detection approaches including: One-Class SVMs (OCSVM) (Schölkopf et al., 2000), Local Outlier Fraction (LOF) (Breunig et al., 2000), Angle Based Outlier Detection (ABOD) (Kriegel et al., 2008), Feature Bagging (FB) (Lazarevic & Kumar, 2005), Auto Encoder Distance (AE) (Aggarwal, 2015), K-Nearest Neighbors (KNN) (Ramaswamy et al., 2000; Angiulli & Pizzuti, 2002), Copula Based Outlier Detection (COPOD) (Li et al., 2020), Variational Auto Encoders (VAE)

Figure 5: The data with the lowest (top) and highest (bottom) predicted temperatures in the SVHN dataset. High temperature entries are blurry, cropped poorly, and generally difficult to classify.

Table 3: Median outlier detection performance of several methods across 22 benchmark datasets from ODDS.

Method	Median AUC
LOF	.669
FB	.702
ABOD	.727
AE	.737
VAE	.792
COPOD	.799
PCA	.808
OCSVM	.814
MCD	.820
KNN	.822
HBOS	.822
IF	.823
CBLOF	.836
AE+S (Ours)	.846
PCA+S (Ours)	**.868**

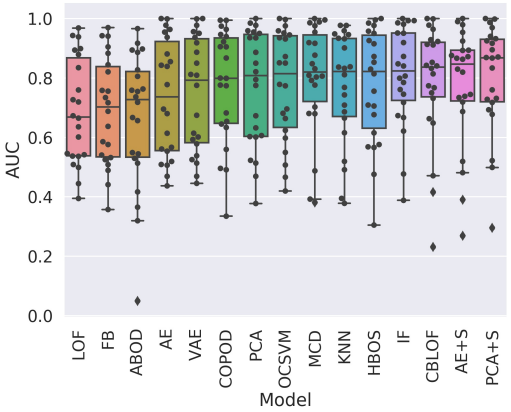


Figure 6: Distribution of Outlier Detection AUC across the ODDS Benchmark. Our approaches, PCA+S and AE+S, are competitive with other Outlier Detection systems.

(Kingma & Welling, 2013), Minimum Covariance Determinants with Mahlanohbis Distance (MCD) (Rousseeuw & Driessen, 1999; Hardin & Rocke, 2004), Histogram-based Outlier Scores (HBOS) (Goldstein & Dengel, 2012), Principal Component Analysis (PCA) (Shyu et al., 2003), Isolation Forests (IF) (Liu et al., 2008; 2012), and the Clustering-Based Local Outlier Factor (CBLOF) (He et al., 2003). For experimental details please see Sections I and J of the Appendix.

Our methods (PCA+S and AE+S) use a similar principle as isolation-based approaches that determine outliers based on how difficult they are to model. In existing approaches, outliers influence and skew the isolation model which causes the model to exhibit less confidence on the whole. This hurts a model's ability to distinguish between inliers and outliers. In contrast, our approach allows the underlying model to down-weight outliers. This yields a more consistent model with a clearer decision boundary between outliers and inliers as shown in Figure 4. As a future direction of investigation we note that our approach is model-architecture agnostic, and can be combined with domain-specific architectures to create outlier detection methods tailored to images, text, and audio.

5.3 Adaptive Regularization with Prior Parameters

In addition to optimizing the shape and scale of the likelihood distribution we can use the same approach to optimize a model's *prior* distribution. More specifically, we propose adaptive regularizers for a model's parameters, θ. This approach optimizes the distribution parameters of the prior, ϕ_{prior}, to naturally tune the degree of regularization. In particular, the Normal (Ridge, L2) and Laplace (LASSO, L1) priors, with scale parameters σ and b, regularize model parameters for small magnitude and sparsity respectively (Hastie et al., 2009). The degree of regularization, $\lambda \in [0, \infty)$, is

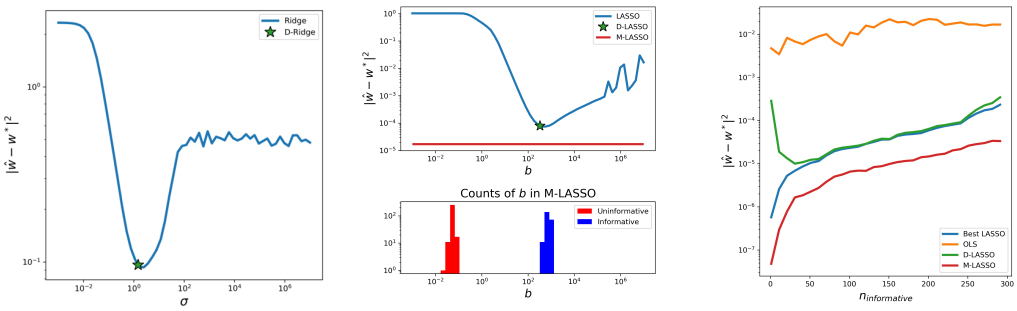


Figure 7: Performance of $L2$ (left) and $L1$ (middle) regularized linear regression on a 500 dimensional synthetic dataset where the true parameters, w^*, are known. Dynamic Ridge (D-Ridge) and D-LASSO regression find the regularization strength that best estimates the true parameters. M-LASSO outperforms any single global regularization strength and does not shrink informative weights. (right) Performance of adaptive $L1$ regularization methods as a function of true model sparsity. In all cases, Multi-LASSO outperforms other methods by orders of magnitude.

conventionally a hyperparameter of the regularized loss function:

$$\min_{\theta} \sum_{i}^{N} (\hat{y}_i := f_\theta(x_i) - y_i)^2 + \lambda \sum_{j}^{P} |\theta_j|. \tag{7}$$

We note that we cannot choose λ by direct minimization because it admits a trivial minimum at $\lambda = 0$. In the linear case, one can select this weight efficiently using Least Angle Regression (Efron et al., 2004). However, in general λ is usually learned through expensive cross validation methods. Instead, we retain the prior with its scale parameter, and jointly optimize over the full likelihood:

$$\min_{\theta,\sigma,b} \sum_{i}^{N} \left(\frac{1}{2\sigma^2} (\hat{y}_i - y_i)^2 + \log \sigma \right) + \sum_{j}^{P} \left(\frac{|\theta_j|}{b} + \log b \right) \tag{8}$$

This approach, the Dynamic Lasso (D-LASSO), admits no trivial solution for the prior parameter b, and must balance the effective regularization strength, $\frac{1}{b}$, with the normalization factor, $\log b$. D-LASSO selects the degree of regularization by gradient descent, rather than expensive black-box search. In Figure 7 (left) and (middle) we show that this approach, and its Ridge equivalent, yield ideal settings of the regularization strength. Figure 7 (right) shows D-LASSO converges to the best LASSO regularization strength for a variety of true-model sparsities. As a further extension, we replace the global σ or b with a σ_j or b_j for each model parameter, θ_j, to locally adapt regularization to each model weight (**M**ulti-Lasso). This consistently outperforms any global setting of the regularization strength and shields important weights from undue shrinkage 7 (middle).

5.4 Re-calibration

The work of (Guo et al., 2017) shows that modern networks are accurate, yet systematically overconfident, a phenomenon called mis-calibration. We investigate the role of optimizing likelihood parameters to re-calibrate models. More specifically, we can fit likelihood parameter regressors on a validation set to modify an existing model's confidence to better align with the validation set. This approach is a generalization of Guo et al. (2017)'s Temperature Scaling method, which we refer to as Global Scaling (GS) for notational consistency. Global Scaling re-calibrates classifiers with a learned global parameter, τ in the loss function: $\sigma(\vec{x}, \tau)$.

Fitting model-conditioned likelihood parameters to to a validation set defines a broad class of re-calibration strategies. From these we introduce three new re-calibration methods. Linear Scaling (LS) learns a linear mapping, l, to transform logits to a softmax temperature: $\sigma(\vec{x}, l(\vec{x}))$. Linear Feature Scaling (LFS) learns a linear mapping, l, to transform the *features* prior to the logits, $\vec{f}$, to a softmax temperature: $\sigma(\vec{x}, l(\vec{f}))$. Finally, we introduce Deep Scaling (DS) for regressors which learns a nonlinear network, N, to transform features, $\vec{f}$, into a temperature: $\sigma(\vec{x}, N(\vec{f}))$.

Table 4: Comparison of calibration methods by ECE for ResNet-50 (RN50) and DenseNet-121 (DN121) architectures on test data. Our predicted likelihood parameter methods: Linear Scaling (LS) and Linear Feature Scaling (LFS) outperform other approaches. In all cases our methods reduce miscalibration with comparable computation time as GS.

Model	Dataset	Uncalibrated	Platt	Isotonic	GS	VS	LS	LFS
RN50	CIFAR-10	.250	.034	.053	.046	.037	**.018**	**.018**
RN50	CIFAR-100	.642	.061	.072	.035	.044	**.030**	.173
RN50	SVHN	.072	.053	.010	.029	.022	**.009**	**.009**
RN50	ImageNet	.430	.018	.070	.019	.023	.026	**.015**
DN121	CIFAR-10	.253	.048	.042	.039	.034	**.028**	**.028**
DN121	CIFAR-100	.537	.049	.067	.024	.024	**.014**	.031
DN121	SVHN	.079	.018	**.010**	.022	.017	.011	**.010**
DN121	ImageNet	.229	.028	.095	.021	**.019**	.043	**.019**

In Table 4 we compare our recalibration approaches to the previous state of the art: Global Scaling. We note that (Guo et al., 2017) have already shown that Global Scaling outperform Bayesian Binning into Quantiles (Naeini et al., 2015), Histogram binning (Zadrozny & Elkan, 2001), and Isotonic Regression. We recalibrate both ResNet50 (He et al., 2016) and DenseNet121 (Huang et al., 2017) on a variety of vision datasets. We measure classifier miscalibration using the Expected Calibration Error (ECE) (Guo et al., 2017) to align with prior art. We additionally evaluate Isotonic recalibration, Platt Scaling (Platt, 1999), and Vector Scaling (VS) (Guo et al., 2017), which learns a vector, $\vec{v}$, to re-weight logits: $\sigma(\vec{v}\vec{x}, 1)$. LS and LFS tend to outperform other approaches like GS and VS, which demonstrates that richer likelihood parametrizations can improve calibration akin to how richer models can improve prediction.

For recalibrating regressors, we compare against the previous state of the art, Kuleshov et al. (2018), who use an Isotonic regressor to correct a regressors' confidence. We use the same datasets, and regressor calibration metric (CAL) as Kuleshov et al. (2018). Table 5 shows that our approaches can outperform this baseline as well as the regression equivalent of Global Scaling.

Table 5: Comparison of regression calibration methods as evaluated by their calibration error as defined in (Kuleshov et al., 2018). Predicted likelihood parameters often outperform other methods.

Dataset	Uncalibrated	Isotonic	GS	LS	DS
crime	0.3624	0.3499	0.0693	**0.0125**	0.0310
kinematics	0.0164	0.0103	0.0022	**0.0021**	0.0032
bank	0.0122	0.0056	0.0027	0.0024	**0.0020**
wine	0.0091	0.0108	0.0152	0.0131	**0.0064**
mpg	0.2153	0.2200	0.1964	0.1483	**0.0233**
cpu	0.0862	**0.0340**	0.3018	0.2078	0.1740
soil	0.3083	**0.3000**	0.3130	0.3175	0.3137
fried	0.0006	**0.0002**	**0.0002**	**0.0002**	**0.0002**

6 Conclusion

Optimizing the full likelihood can improve model quality by adapting losses and regularizers. Full likelihoods are agnostic to the architecture, optimizer, and task, which makes them simple substitutes for standard losses. Global, data, and predicted likelihood parameters offer different degrees of expressivity and efficiency. In particular, predicted parameters adapt the likelihood to each data point during training and testing without significant time and space overhead. By including these parameters in a loss function one can improve a model's robustness and generalization ability and create new classes of outlier detectors and recalibrators that outperform baselines. More generally, we hope this work encourages joint optimization of model and likelihood parameters, and argue it is likely that your loss should be a likelihood.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `http://tensorflow.org/`. Software available from tensorflow.org.

Charu C Aggarwal. Outlier analysis. In *Data mining*, pp. 237–263. Springer, 2015.

Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *European conference on principles of data mining and knowledge discovery*, pp. 15–27. Springer, 2002.

Jonathan T Barron. A general and adaptive robust loss function. In *CVPR*, 2019.

Christopher M Bishop. Mixture density networks. 1994.

Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

William S Cleveland. *Visualizing data*. Hobart Press, 1993.

George E Dahl, Tara N Sainath, and Geoffrey E Hinton. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *ICASSP*, pp. 8609–8613. IEEE, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. Ieee, 2009.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, pp. 249–256, 2010a.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, pp. 249–256, 2010b.

Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, pp. 59–63, 2012.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

Johanna Hardin and David M Rocke. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, 44(4): 625–638, 2004.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650, 2003.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, pp. 5574–5584, 2017.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pp. 7482–7491, 2018.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.

Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 444–452, 2008.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Jennifer Dy and Andreas Krause (eds.), *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2801–2809, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 157–166, 2005.

Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. Copod: copula-based outlier detection. *arXiv preprint arXiv:2009.09463*, 2020.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39, 2012.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, December 2015.

David JC MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *ICNN*, volume 1, pp. 55–60. IEEE, 1994.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

John C Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 427–438, 2000.

Shebuti Rayana. Odds library. *Stony Brook,-2016*, (2017), 2016.

Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

Shreyas Saxena, Oncel Tuzel, and Dennis DeCoste. Data parameters: A new family of parameters for learning a differentiable curriculum. In *NeurIPS*, pp. 11093–11103, 2019.

Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In *Advances in neural information processing systems*, pp. 582–588, 2000.

Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. Technical report, MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING, 2003.

Henri Theil. A rank-invariant method of linear and polynomial regression analysis. In *Henri Theil's contributions to economics and econometrics*, pp. 345–381. Springer, 1992.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, volume 1, pp. 609–616. Citeseer, 2001.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019. URL `http://jmlr.org/papers/v20/19-011.html`.