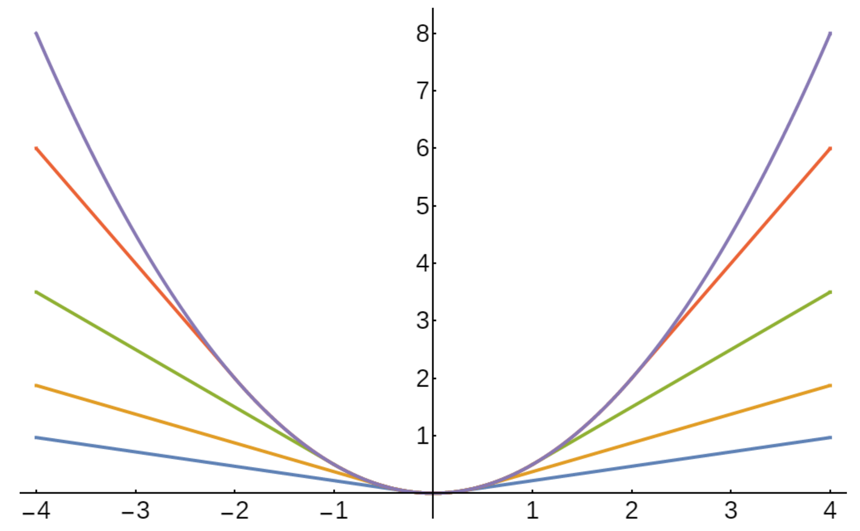


An Alternative Probabilistic Interpretation of the Huber Loss

Gregory P. Meyer
gregorpm@gmail.com

Introduction

- The Huber loss is a robust loss function used for a wide range of regression tasks.
- To utilize the Huber loss, a parameter that controls the transitions from a quadratic function to an absolute value function needs to be selected.
- In this work, we propose an alternative probabilistic interpretation of the Huber loss, which relates minimizing the loss to minimizing an upper-bound on the Kullback-Leibler divergence between Laplace distributions, where one distribution represents the noise in the ground-truth and the other represents the noise in the prediction.
- We demonstrate that the parameters of the Laplace distributions are directly related to the transition point of the Huber loss.
- As a result, our interpretation provides an intuitive way to identify well-suited hyper-parameters by approximating the amount of noise in the data.

Background

Huber Loss

- Loss functions commonly used for regression are $L_1(x) = |x|$ and $L_2(x) = \frac{1}{2}x^2$.
- Both of these functions have advantages and disadvantages:
  - L_1 is less sensitive to outliers in the data, but it is not differentiable at zero.
  - The L_2 is differentiable everywhere, but it is highly sensitive to outliers.
- Huber proposed the following loss as a compromise between the L_1 and L_2 losses:

$$H_\alpha(x) = \begin{cases} \frac{1}{2}x^2, & |x| \le \alpha \\ \alpha\left(|x| - \frac{1}{2}\alpha\right), & |x| > \alpha \end{cases}$$

  where $\alpha \in \mathbb{R}^+$ controls the transition from L_1 to L_2.
- The Huber loss is both differentiable everywhere and robust to outliers.
- A disadvantage is that the parameter α needs to be selected.

Maximum Likelihood Estimation

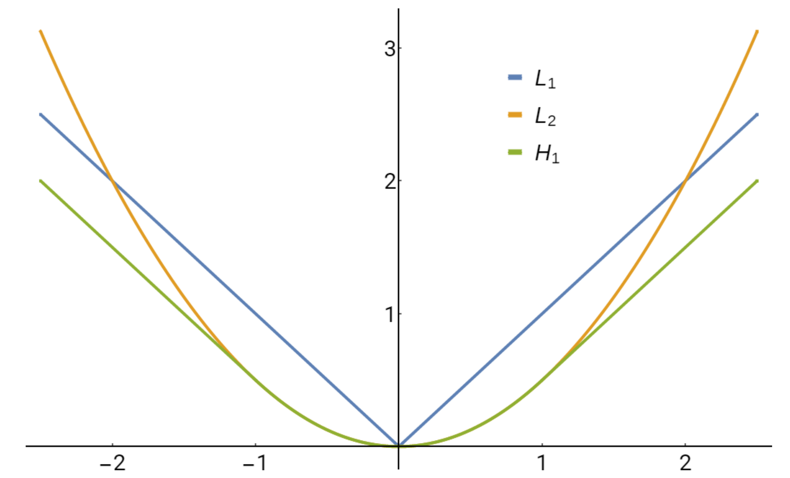
- Assume we have a dataset $\mathcal{D} = \{x_i, y_i\}_{i=0}^N$ drawn from an unknown distribution.
- Let us model the relationship between x_i and y_i as

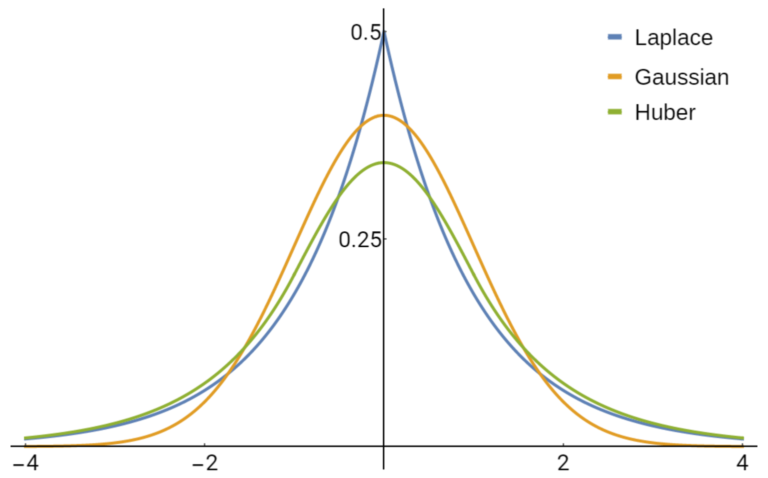
$$y_i = F_\theta(x_i) + \epsilon$$

  where F_θ is a function and ϵ is random noise drawn from some known distribution.
- The goal of maximum likelihood estimation is to identify $\hat{\theta}$ that maximizes the likelihood (or minimizes the negative log likelihood) of y_i given x_i across the dataset.

$$\hat{\theta} = \arg\max_\theta \prod_{i=0}^{N} p(y_i|x_i,\theta) = \arg\min_\theta -\sum_{i=0}^{N} \log p(y_i|x_i,\theta)$$

- Minimizing the Huber loss provides the maximum likelihood estimate of θ when $p(y_i|x_i,\theta) \propto \exp\left[-H_\alpha(y_i - F_\theta(x_i))\right]$, which is referred to as the Huber density.
- We believe this interpretation that relates the Huber loss to the Huber density fails to provide adequate intuition for identifying the transition point.





Proposal

- Assume we have a dataset $\mathcal{D} = \{x_i, y_i\}_{i=0}^N$, but consider the following relationships:

$$y_i^* = y_i + \epsilon_1 \qquad y_i^* = F_\theta(x_i) + \epsilon_2$$

  where y_i^* is an unknown value we would like to estimate with $F_\theta(x_i)$, y_i is a known estimate of y_i^*, and ϵ_1 and ϵ_2 are random noise drawn from separate distributions.
- In this case, we have two distributions: $p(y_i^*|y_i)$ which represents our uncertainty in the label, and $q(y_i^*|x_i,\theta)$ represents our uncertainty in the model's prediction.
- Assuming both the labels and the predictions are contaminated with outliers, i.e. both ϵ_1 and ϵ_2 are drawn from Laplace distributions, the probability densities become

$$p(y_i^*|y_i) = \frac{1}{2b_1}\exp\left(-\frac{|y_i^* - y_i|}{b_1}\right) \qquad q(y_i^*|x_i,\theta) = \frac{1}{2b_2}\exp\left(-\frac{|y_i^* - F_\theta(x_i)|}{b_2}\right)$$

  where $b_1 \in \mathbb{R}^+$ and $b_2 \in \mathbb{R}^+$ define the scale of the label and prediction uncertainty.
- We can identify $\hat{\theta}$ by minimizing the KL divergence between the distributions:

$$\hat{\theta} = \arg\min_\theta \sum_{i=0}^{N}\left(\frac{b_1\exp\left(-\frac{|y_i - F_\theta(x_i)|}{b_1}\right) + |y_i - F_\theta(x_i)|}{b_2} + \log\frac{b_2}{b_1} - 1\right)$$

- We propose the following loss function derived from the KL divergence:

$$D_{\alpha,\beta}(x) = \frac{\alpha\exp\left(-\frac{|x|}{\alpha}\right) + |x| - \alpha}{\beta}$$

- The variable x is equal to the difference in the means of the Laplace distributions. The parameter $\alpha \in \mathbb{R}^+$ directly corresponds to the scale of the noise in the label (b_1), and $\beta \in \mathbb{R}^+$ corresponds to the scale of the noise in the prediction (b_2).

Relationship to the Huber Loss

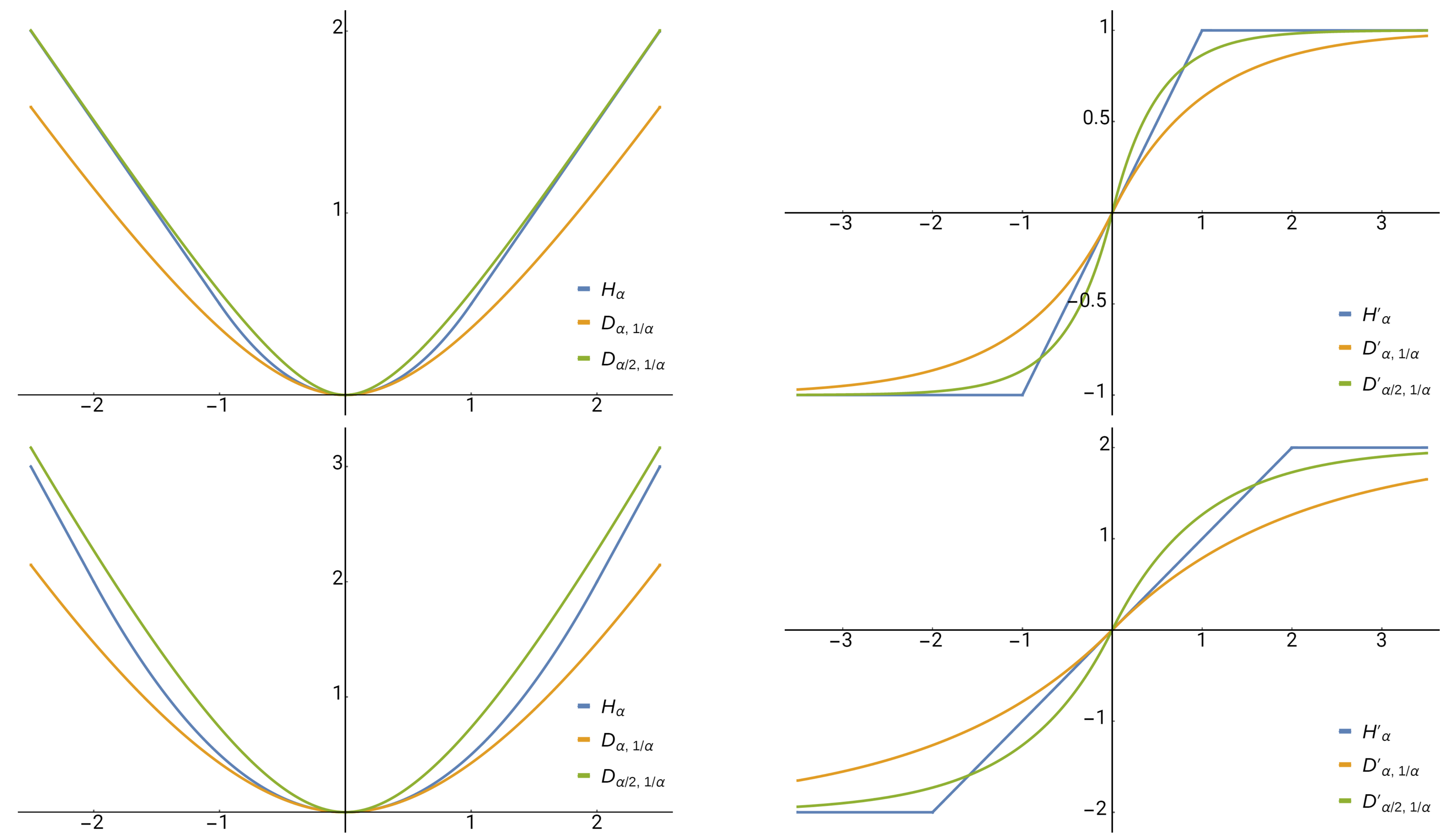
- Like the Huber loss, our proposed loss behaves quadratically when the residual is small and linearly when the residual is large.
- The following configurations tightly bound the Huber loss:

$$D_{\alpha,1/\alpha}(x) \le H_\alpha(x) \le D_{\alpha/2,1/\alpha}(x)$$

- Minimizing the Huber loss with parameter α is equivalent to minimizing an upper-bound on the KL divergence of two Laplace distributions when the scale of the label distribution $b_1 = \alpha$, and the scale of the prediction distribution $b_2 = 1/\alpha$.

$\alpha = 1$

$\alpha = 2$



Loss Functions | Derivatives

Case Study: Faster R-CNN

- With our interpretation, we analyze the loss functions used by the Faster R-CNN.
- The Faster R-CNN network architecture consists of two parts, a region proposal network and an object detection network. To regress a bounding box, both the proposal network and the detection network utilize the Huber loss.
- Let's analyze the center prediction; the target for the x-coordinate of the center is

$$t_x^* = \frac{x^* - x_a}{\sigma_x w_a}$$

  where x^* is the x-coordinate of the ground-truth center, x_a is the x-coordinate of the anchor, w_a is the width of the anchor, and $\sigma_x \in \mathbb{R}^+$ is a hyper-parameter.
- Faster R-CNN uses the loss $\frac{\lambda}{\alpha}H_\alpha(t_x - t_x^*)$ to penalize the model's prediction t_x.
- To interpret this loss, we re-write the residual in terms of the center displacement,

$$t_x - t_x^* = t_x - \frac{x^* - x_a}{\sigma_x w_a} = \frac{(t_x\sigma_x w_a + x_a) - x^*}{\sigma_x w_a} = \frac{x - x^*}{\sigma_x w_a}$$

  where $x = t_x\sigma_x w_a + x_a$ is the predicted x-coordinate of the center.
- Consider the relationship between their loss function and our proposed loss function:

$$\frac{\lambda}{\alpha}H_\alpha(t_x - t_x^*) \approx D_{\alpha\sigma_x w_a,\, \sigma_x w_a/\lambda}(x - x^*)$$

- With this formulation, the label and prediction noise can be independently changed.
- For the proposal network, the scale of the label noise is assumed to be $w_a/9$ and the prediction noise is w_a. For the detection network, the label and prediction noise is assumed to be $w_a/10$. Based on our interpretation, we believe the hyper-parameters could be improved upon, which we demonstrate through our experiments.

Experiments

- The goal of our experiments is to demonstrate that our interpretation of the Huber loss can lead to hyper-parameters better suited to the task of bounding box regression.
- Our aim is not to replace the Huber loss with our proposed loss; rather, we want to leverage the relationship between the losses to gain insight into the Huber loss.
- Therefore, we limit our modifications to Faster R-CNN to only the hyper-parameters of the Huber loss, and we propose three new sets of hyper-parameters.
- We were able to improve performance by reducing the assumed amount of noise in the labels and predictions. Specifically, we were able to raise performance at larger IoU thresholds, which requires more accurate bounding boxes.

Parameters	Label Noise		Prediction Noise		Mean Average Precision (mAP) @		
	Proposal	Detection	Proposal	Detection	0.5 IoU	0.75 IoU	0.5-0.95 IoU
Original	$w_a/9$	$w_a/10$	w_a	$w_a/10$	**44.7**	23.1	23.8
Experiment A	$w_a/20$	$w_a/20$	$w_a/5$	$w_a/10$	**44.7**	24.0	24.2
Experiment B	$w_a/20$	$w_a/20$	$w_a/10$	$w_a/20$	44.2	**25.0**	24.6
Experiment C	$w_a/20$	$w_a/20$	$w_a/5$	$w_a/20$	44.6	24.9	**24.7**

Conclusion

- In this work, we propose an alternative probabilistic interpretation of the Huber loss.
- We demonstrated that our interpretation can aid in hyper-parameter selection, and we were able to improve the performance of the Faster R-CNN object detector without needing to exhaustively search over hyper-parameters.
- The vast majority of recent papers that utilize the Huber loss use the same formulation as Faster R-CNN; therefore, these methods have the potential to be improved by leveraging our interpretation of the Huber loss to identify better suited hyper-parameters.