Choosing a Penalty Function

J. Wren*

1 Overview

First, let's make the notation clear.

- denote a generic loss function, $\ell(\hat{d}, d^{\text{tgt}})$, where $\hat{d}$ is the predicted value and d^{tgt} is the target value
- we will consider loss functions of the form $p(\hat{d} - d^{\text{tgt}}) = p(r)$
- p is the penalty function, and r denotes the prediction error (residual)
- average loss (a measure of performance) is $\frac{1}{S}\sum_{s=1}^{S} \ell(\hat{d}_s, d^{\text{tgt}})$

Our penalty function has been $p^{\text{sqr}}(r) = r^2$, with $\widehat{\text{MSE}}$ as the average loss. $p^{\text{sqr}}(r)$ is perhaps the most ubiquitous penalty function because of its two essential qualities:

1. punishment is more severe for large residuals
2. p is symmetric, i.e., $p(-r) = p(r)$

Symmetry implies indifference between over-estimates ($r > 0$), and under-estimates ($r < 0$). But since an under-estimate leads to bounds that are conservative, whereas an over-estimate leads to bounds that are *incorrect*, we should punish over-estimates more than under-estimates.

*Becker Friedman Institute for Research in Economics, University of Chicago.

2 Penalty Functions

2.1 A natural $p^{\mathbf{sqr}}(r)$ extension

To impose asymmetrical penalization, we can use the right-tilted square penalty function:

$$p_\alpha^{\text{rts}}(r) \equiv \begin{cases} \alpha r^2, & \text{if } r \geq 0 \\ (1-\alpha) r^2, & \text{if } r < 0, \end{cases} \tag{1}$$

where $\alpha \in \left[\frac{1}{2}, 1\right)$.[1] $\alpha = \frac{1}{2}$ gives us a symmetrical punishment, while the penalization for over-estimating increases as $\alpha \to 1$. Since $\alpha = 1$ results in no punishment for underestimates, α should be strictly less than 1.

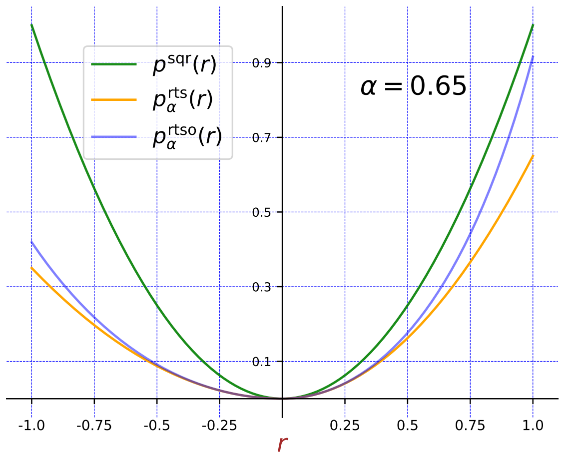
2.2 A large error emphasis

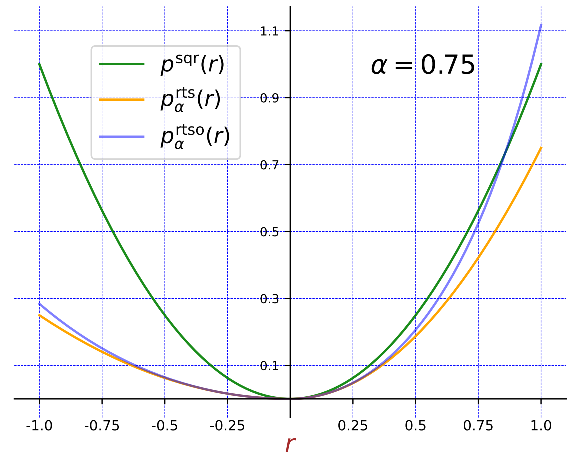
Let's also consider one more penalty function that more heavily punishes large residuals (outliers). Define:

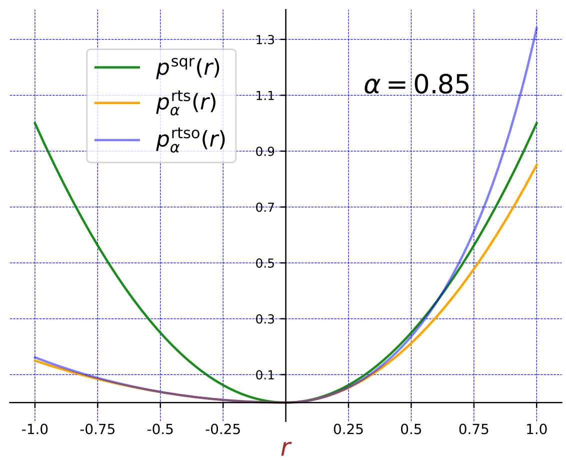
$$p_\alpha^{\text{rtso}}(r) \equiv \begin{cases} \frac{1}{\exp\left(-\alpha r^2\right)} - 1, & \text{if } r \geq 0 \\ \frac{1}{\exp\left(-(1-\alpha) r^2\right)} - 1, & \text{if } r < 0, \end{cases} \tag{2}$$

where again $\alpha \in \left[\frac{1}{2}, 1\right)$. As $\alpha \to 1$, the additional punishment (relative to p_α^{rts}) for large residuals increases.

[1] While a more general definition would allow for $\alpha \in [0, 1]$, since we want to impose a higher cost for over-estimating, we should only consider values of $\alpha \geq \frac{1}{2}$.


$p^{\text{sqr}}(r)$
$p_\alpha^{\text{rts}}(r)$
$p_\alpha^{\text{rtso}}(r)$
$\alpha = 0.65$
0.9
0.7
0.5
0.3
0.1
-1.0
-0.75
-0.5
-0.25
0.25
0.5
0.75
1.0
r


$p^{\text{sqr}}(r)$
$p_\alpha^{\text{rts}}(r)$
$p_\alpha^{\text{rtso}}(r)$
$\alpha = 0.75$
1.1
0.9
0.7
0.5
0.3
0.1
-1.0
-0.75
-0.5
-0.25
0.25
0.5
0.75
1.0
r


$p^{\text{sqr}}(r)$
$p_\alpha^{\text{rts}}(r)$
$p_\alpha^{\text{rtso}}(r)$
$\alpha = 0.85$
1.3
1.1
0.9
0.7
0.5
0.3
0.1
-1.0
-0.75
-0.5
-0.25
0.25
0.5
0.75
1.0
r

References

Boyd, S. (2022): "EE104/CME107: Introduction to Machine Learning," https://ee104.stanford.edu, accessed: 2023–1-18.