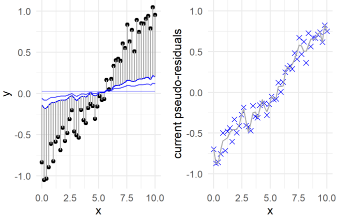
Introduction to Machine Learning

Boosting
Gradient Boosting: Illustration



Learning goals

- See simple visualizations of boosting in regression
- Understand impact of different losses and base learners

GRADIENT BOOSTING ILLUSTRATION - GAM

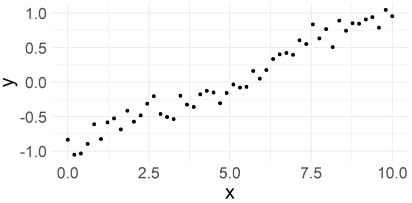
GAM / Splines as BL and compare *L*2 vs. *L*1 loss.

- L2: Init = optimal constant = mean(y); for L1 it's median(y)
- BLs are cubic *B*-splines with 40 knots.
- PRs $L2$: $\tilde{r}(f) = r(f) = y - f(\mathbf{x})$
- PRs $L1$: $\tilde{r}(f) = sign(y - f(\mathbf{x}))$
- Constant learning rate 0.2

Univariate toy data:

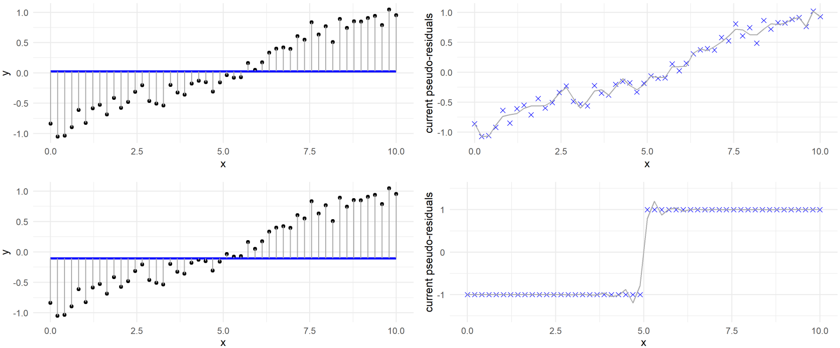
$$y^{(i)} = -1 + 0.2 \cdot x^{(i)} + 0.1 \cdot sin(x^{(i)}) + \epsilon^{(i)}$$

$$n = 50 \text{ ; } \epsilon^{(i)} \sim \mathcal{N}(0, 0.1)$$



GAM WITH $L2$ VS $L1$ LOSS

Top: $L2$ loss, bottom: $L1$ loss

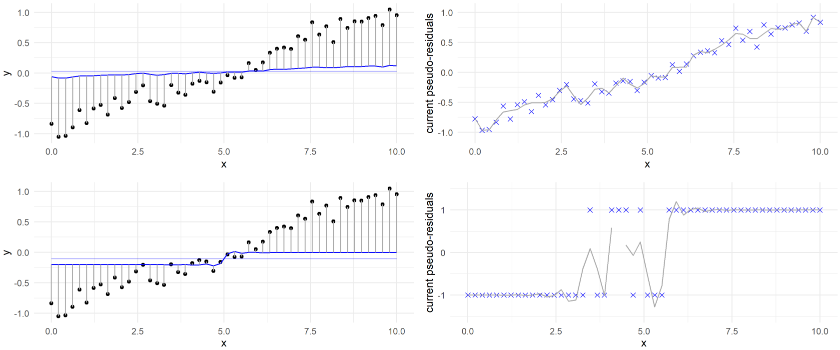


Iteration 1

Shape of PRs affects gradual model fit: $L1$ only sees resids' sign, BLs are not affected size of values as in $L2$ and hence lead to more moderate changes.

GAM WITH $L2$ VS $L1$ LOSS

Top: $L2$ loss, bottom: $L1$ loss

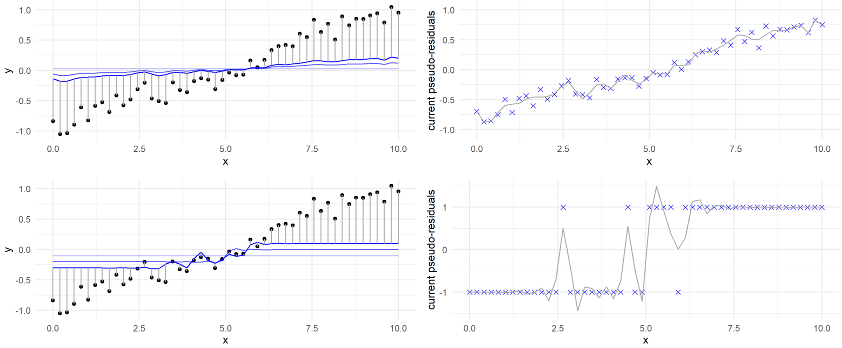


Iteration 2

Shape of PRs affects gradual model fit: $L1$ only sees resids’ sign, BLs are not affected size of values as in $L2$ and hence lead to more moderate changes.

GAM WITH $L2$ VS $L1$ LOSS

Top: $L2$ loss, bottom: $L1$ loss

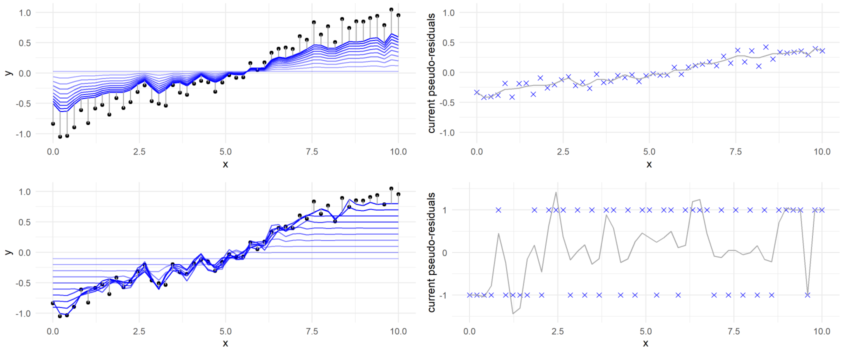


Iteration 3

Shape of PRs affects gradual model fit: $L1$ only sees resids' sign, BLs are not affected size of values as in $L2$ and hence lead to more moderate changes.

GAM WITH $L2$ VS $L1$ LOSS

Top: $L2$ loss, bottom: $L1$ loss

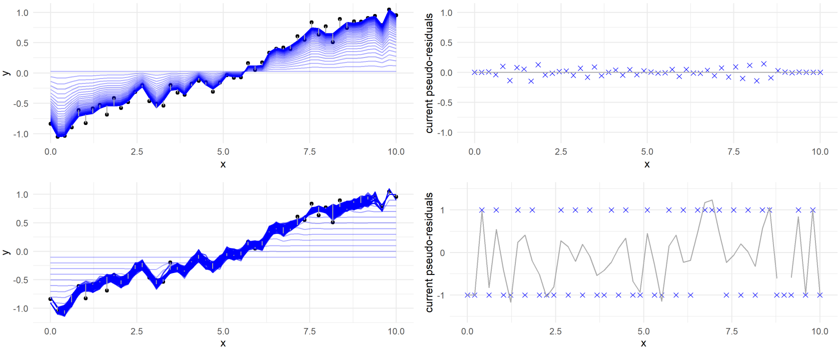


Iteration 10

Shape of PRs affects gradual model fit: $L1$ only sees resids' sign, BLs are not affected size of values as in $L2$ and hence lead to more moderate changes.

GAM WITH $L2$ VS $L1$ LOSS

Top: $L2$ loss, bottom: $L1$ loss

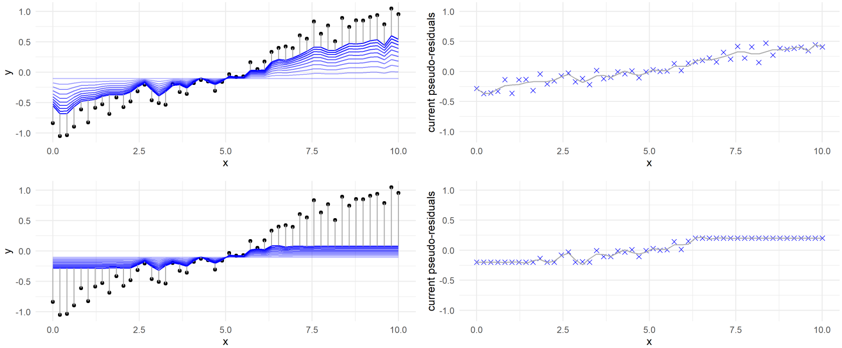


Iteration 100

Shape of PRs affects gradual model fit: $L1$ only sees resids' sign, BLs are not affected size of values as in $L2$ and hence lead to more moderate changes.

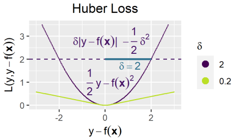
GAM WITH HUBER LOSS

Top: δ = 2, bottom: δ = 0.2.



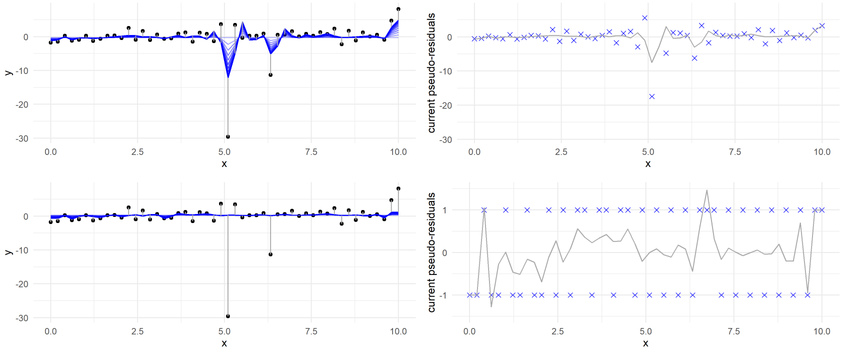
Iteration 10

For small δ, PRs are often bounded, resulting in $L1$-like behavior, while the upper plot more closely resembles $L2$ loss.



GAM WITH OUTLIERS

Instead of Gaussian noise, let's use t-distrib, that leads to outliers in y.
Top: $L2$, bottom: $L1$.

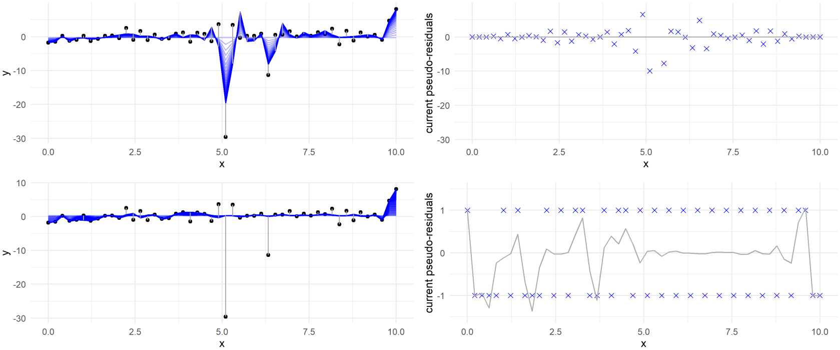


Iteration 10

$L2$ loss is affected by outliers rather strongly, whereas $L1$ solely considers residuals' sign and not their magnitude, resulting in a more robust model.

GAM WITH OUTLIERS

Instead of Gaussian noise, let's use *t*-distrib, that leads to outliers in *y*.
Top: *L*2, bottom: *L*1.

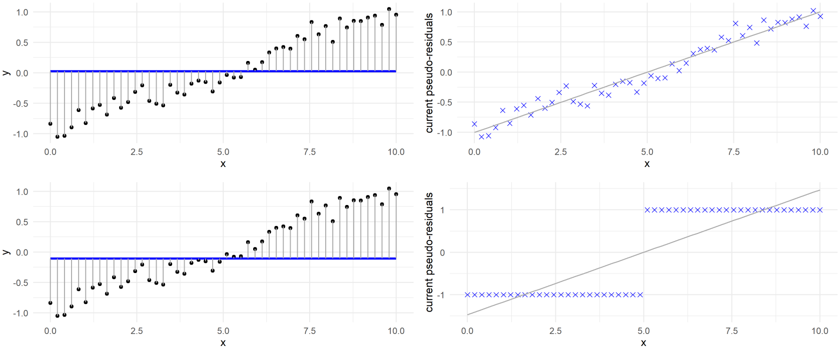


Iteration 100

*L*2 loss is affected by outliers rather strongly, whereas *L*1 solely considers residuals' sign and not their magnitude, resulting in a more robust model.

LM WITH $L2$ VS $L1$ LOSS

Top: $L2$, bottom: $L1$.

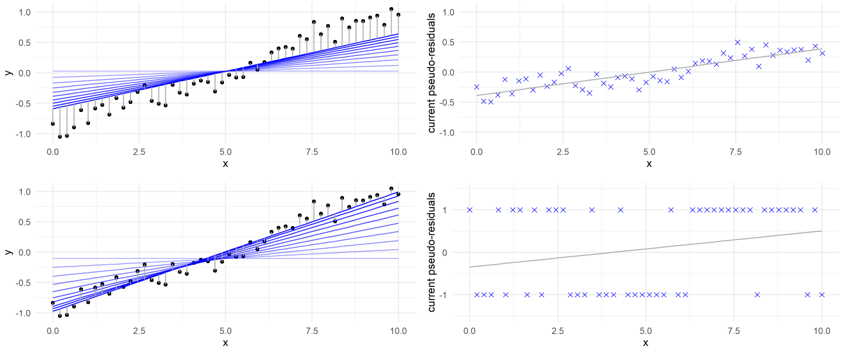


Iteration 1

$L2$: as $\tilde{r}(f) = r(f)$, BL of 1st iter already optimal; but learn rate slows us down.

LM WITH $L2$ VS $L1$ LOSS

Top: $L2$, bottom: $L1$.

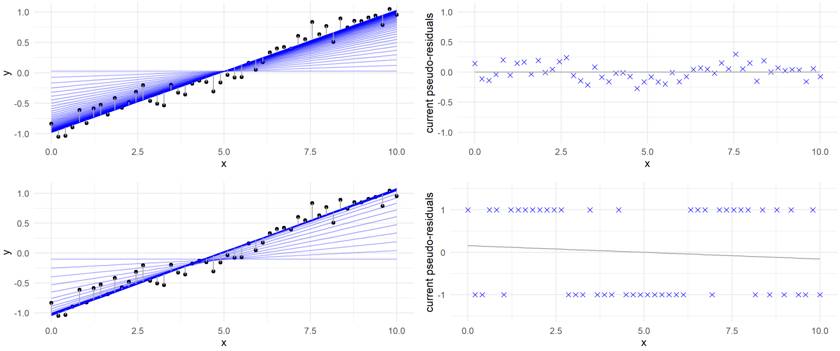


Iteration 10

$L2$: as $\tilde{r}(f) = r(f)$, BL of 1st iter already optimal; but learn rate slows us down.

LM WITH $L2$ VS $L1$ LOSS

Top: $L2$, bottom: $L1$.



Iteration 100

$L2$: as $\tilde{r}(f) = r(f)$, BL of 1st iter already optimal; but learn rate slows us down.