# Loss Math

## Algorithm

We let $w$ be the weight of a given parent leaf, and $w_L$, $w_R$ be the weights of the left and right children of the parent leaf. For ease, denote $\Delta_L = w - w_L$ and $\Delta_R = w - w_R$. Our prediction $\hat{y}^{(t-1)} = o_i + \eta w$. We increment our left leaf by some value $\Delta_L$ to get leaf value $w_L$. Our updated prediction is thus $\hat{y}^{(t)} = o_i + \eta w + \eta\Delta_L$. We will actually set $\hat{y}^{(t)} = o_i + \eta w + \epsilon\eta\Delta_L$, but that is not needed for this calculation.

By Taylor Expansion, we have that

$$\begin{aligned}
l(y_i, \hat{y_i}^{(t)}) &= \\
l(y_i, \hat{y_i}^{(t-1)} + \eta\Delta_L) &= \\
l(y_i, \hat{y_i}^{(t-1)}) + \eta\Delta_L \frac{\partial l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}} + \frac{\eta^2\Delta_L^2}{2}\frac{\partial^2 l(y_i, \hat{y_i}^{(t-1)})}{\partial (\hat{y_i}^{(t-1)})^2} &= \\
l(y_i, \hat{y_i}^{(t-1)}) + \eta\Delta_L g_i + \frac{\eta^2\Delta_L^2}{2} h_i &
\end{aligned}$$

We also add a regularization term of $\frac{1}{2}\lambda(w+\Delta_L)^2 = \frac{1}{2}\lambda(w^2 + 2w\Delta_L + \Delta_L^2)$ Combining terms and removing constants with respect to $\Delta_L$, we wish to minimize

$$\eta\Delta_L g_i + \frac{\eta^2\Delta_L^2}{2} h_i + \frac{1}{2}\lambda(2w\Delta_L + \Delta_L^2)$$

Taking the derivative with respect to $\Delta_L$ and adding in all values on the left side, we have that

$$\begin{aligned}
\eta G_L + \eta^2\Delta_L H_L + \lambda w + \lambda\Delta_L = 0 &\implies \\
\Delta_L\left(\eta^2 H_L + \lambda\right) = -\eta G_L - \lambda w &\implies \\
\Delta_L = -\frac{\eta G_L + \lambda w}{\eta^2 H_L + \lambda} &
\end{aligned}$$

We know that for a quadratic function in the form of $\frac{1}{2}bx^2 + ax$, the minimum value is $-\frac{a^2}{2b}$. We have that $b = \eta^2 H_L + \lambda$, and $a = \eta G_L + w\lambda$. So, the gain at the leaf is

$$\frac{1}{2}\frac{(\eta G_L + w\lambda)^2}{\eta^2 H_L + \lambda}$$

The determination, then, is whether to begin with w or the mean of y as the initial value.

# Old algorithm

For a given leaf and split, we have that minimum loss is given by

$$\min_{\gamma} \sum_{i=1}^{n} \mathcal{L}(y_i, \frac{\alpha o_i + \gamma}{\alpha + 1})$$

We can set the "plain optimum" to be

$$\gamma^* = \arg\min_{\gamma} \sum_{i \in \text{leaf}} \mathcal{L}(y_i, \gamma)$$

And the "ensemble optimum" to be

$$\tilde{\gamma}^* = \arg\min_{\gamma} \sum_{i \in \text{leaf}} \mathcal{L}(y_i, \frac{\alpha o_i + \gamma}{\alpha + 1})$$

Letting a star denote the optimal value that combines ensemble predictions, I split at $A^*$ and define the the left leaf value as

$$B_{\text{leaf}} = (1 - \delta)(\gamma^*) + \delta \tilde{\gamma}^*$$

With a split at $A^*$

The split is defined at:

$$\arg\min_{A} \sum_{i=1}^{n} \mathcal{L}(y_i, \frac{\alpha o_i + (1 - \delta) \sum_{i=1}^{n} \arg\min_{\gamma} \mathcal{L}(y_i, \gamma) + \delta \sum_{i=1}^{n} \arg\min_{\gamma} \mathcal{L}(y_i, \frac{\alpha o_i + \gamma}{\alpha + 1})}{\alpha + 1})$$

Where we define

$$\sum_{i=1}^{n} (1 - \delta) \arg\min_{\gamma} \mathcal{L}(y_i, \gamma) + \sum_{i=1}^{n} \delta \arg\min_{\gamma} \mathcal{L}(y_i, \frac{\alpha o_i + \gamma}{\alpha + 1}) = B_n$$

# Generic Algorithm

Begin with initial value

$$\gamma_0 = \arg\min_{\gamma} \sum_{i=1}^{n} \mathcal{L}(y_i, \gamma)$$

Find error at split A

$$\gamma^* = \gamma_0 - \eta \frac{G}{H + \lambda}$$

Given

$$\mathcal{L}(y_i, \frac{\alpha o_i + \gamma_0}{\alpha + 1})$$

We have that $g_i =$

$$\frac{1}{\alpha + 1} \frac{\partial \mathcal{L}(y_i, z)}{\partial z}\Big|_{z=\frac{\alpha o_i + \gamma_0}{\alpha + 1}}$$

And $h_i =$

$$\frac{1}{(\alpha + 1)^2} \frac{\partial^2 \mathcal{L}(y_i, z)}{\partial z^2}\Big|_{z=\frac{\alpha o_i + \gamma_0}{\alpha + 1}}$$

So that $G = \sum_{i \in \text{leaf}} g_i$ and $H = \sum_{i \in \text{leaf}} h_i$

So, we have that

$$\frac{G}{H} =$$

$$\frac{\sum_{i \in \text{leaf}} g_i}{\sum_{i \in \text{leaf}} h_i} =$$

$$\frac{\sum_{i \in \text{leaf}} \frac{1}{\alpha+1} \frac{\partial \mathcal{L}(y_i, z)}{\partial z}\Big|_{z=\frac{\alpha o_i + \gamma_0}{\alpha+1}}}{\sum_{i \in \text{leaf}} \frac{1}{(\alpha+1)^2} \frac{\partial^2 \mathcal{L}(y_i, z)}{\partial z^2}\Big|_{z=\frac{\alpha o_i + \gamma_0}{\alpha+1}}} =$$

$$(\alpha + 1) \frac{\sum_{i \in \text{leaf}} \frac{\partial \mathcal{L}(y_i, z)}{\partial z}}{\sum_{i \in \text{leaf}} \frac{\partial^2 \mathcal{L}(y_i, z)}{\partial z^2}}$$

# MSE

The gradient we have as

$$\frac{\partial \mathcal{L}(y_i, z)}{\partial z}\Big|_{z=\frac{\alpha o_i + \gamma_0}{\alpha + 1}} = -2 \left( y_i - \frac{\alpha o_i + \gamma_0}{\alpha + 1} \right)$$

And hessian we have as

$$\frac{\partial^2 \mathcal{L}(y_i, z)}{\partial z^2}\Big|_{z=\frac{\alpha o_i + \gamma_0}{\alpha + 1}} = 2$$

So that

$$
\begin{aligned}
-\frac{G}{H} &= \\
\sum_{i=1}^{n} (\alpha+1) y_i - \alpha o_i - \gamma_0 &= \\
\sum_{i=1}^{n} (\alpha+1) y_i - \sum_{i=1}^{n} \alpha o_i - \sum_{i=1}^{n} \sum_{i=1}^{n} \frac{y_i}{n} &= \\
\alpha \sum_{i=1}^{n} (y_i - o_i)
\end{aligned}
$$

# New version of loss-agnostic algorithm

The algorithm can be desribed as following:

$$
\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) = \sum_{k=1}^{K} w^k_{q_k(x_i)}
$$

We define the loss as

$$
\mathcal{L}(\phi) = \sum_{i=1}^{n} \mathcal{L}(y_i, \phi(x_i)) + \sum_{i=1}^{k} \sum_{j=1}^{T_i} 1 + \frac{1}{2}\lambda (w^k_j)^2
$$

Where $T_i$ represents the number of leaves in the $i$th tree and $w^k_j$ represents the value of the $j$th leaf in the $k$th tree.

We begin by creating all trees to the warmup depth, and then update each one at a time. Suppose we begin with tree m WLOG, updating at leaf $j$. Denote $I_j = \{i : q_k(x_i) = j\}$. Similarly, denote $I_{j,l}$ to be the points split to the left of leaf j in tree k after iteration t - 1.

Then,

ONLY WORK WITH ONE SPLIT LEAF AT A TIME, I THINK?

$$
\mathcal{L}^{(t)} = \begin{array}{c}
\left( \sum_{i \notin I_j} l(y_i, \hat{y}_i^{(t-1)}) + \sum_{k \neq m} \Omega(f_k) \right) + 1 + \frac{1}{2}\lambda w_{j,l}^2 + \frac{1}{2}\lambda w_{j,r}^2 - \frac{1}{2}\lambda w_j^2 + \\
\sum_{i \in I_{j,l}} l \left( y_i, w_{j,l} + \sum_{k \neq m} f_k(x_i) \right) + \sum_{i \in I_{j,r}} l \left( y_i, w_{j,r} + \sum_{k \neq m} f_k(x_i) \right)
\end{array} =
$$

$$
\frac{1}{2}
$$