Efficient Non-greedy Optimization of Decision Trees and Forests: Supplementary Material

Mohammad Norouzi[1] **Maxwell D. Collins**[2 *] **Matthew Johnson**[3]
David J. Fleet[4] **Pushmeet Kohli**[5]

[1,4] Department of Computer Science, University of Toronto
[2] Department of Computer Science, University of Wisconsin-Madison
[3,5] Microsoft Research

1 Proofs

Upper bound on loss. For any pair $(\mathbf{x}, y)$, the loss $\ell(\Theta^\mathsf{T} f(\operatorname{sgn}(W\mathbf{x})), y)$ is upper bounded by:

$$\ell(\Theta^\mathsf{T} f(\operatorname{sgn}(W\mathbf{x})), y) \;\le\; \max_{\mathbf{g}\in\mathcal{H}^m} \left\{ \mathbf{g}^\mathsf{T} W\mathbf{x} + \ell(\Theta^\mathsf{T} f(\mathbf{g}), y) \right\} - \max_{\mathbf{h}\in\mathcal{H}^m} \left\{ \mathbf{h}^\mathsf{T} W\mathbf{x} \right\} . \tag{1}$$

Proof.

$$\begin{aligned}
\text{RHS} &= \max_{\mathbf{g}\in\mathcal{H}^m} \left\{ \mathbf{g}^\mathsf{T} W\mathbf{x} + \ell(\Theta^\mathsf{T} f(\mathbf{g}), y) \right\} - \max_{\mathbf{h}\in\mathcal{H}^m} \left\{ \mathbf{h}^\mathsf{T} W\mathbf{x} \right\} \\
&= \max_{\mathbf{g}\in\mathcal{H}^m} \left\{ \mathbf{g}^\mathsf{T} W\mathbf{x} + \ell(\Theta^\mathsf{T} f(\mathbf{g}), y) \right\} - \operatorname{sgn}(W\mathbf{x})^T W\mathbf{x} \\
&\ge \max_{\mathbf{g}\in\{\operatorname{sgn}(W\mathbf{x})\}} \left\{ \mathbf{g}^\mathsf{T} W\mathbf{x} + \ell(\Theta^\mathsf{T} f(\mathbf{g}), y) \right\} - \operatorname{sgn}(W\mathbf{x})^T W\mathbf{x} \\
&= \operatorname{sgn}(W\mathbf{x})^T W\mathbf{x} + \ell(\Theta^\mathsf{T} f(\operatorname{sgn}(W\mathbf{x})), y) - \operatorname{sgn}(W\mathbf{x})^T W\mathbf{x} \\
&= \ell(\Theta^\mathsf{T} f(\operatorname{sgn}(W\mathbf{x})), y) \\
&= \text{LHS}
\end{aligned}$$

□

Proposition 1. *The upper bound on the loss becomes tighter as a constant multiple of W gets larger. More formally, for any $\alpha > \beta > 0$, we have:*

$$\max_{\mathbf{g}\in\mathcal{H}^m} \left\{ \alpha\mathbf{g}^\mathsf{T} W\mathbf{x} + \ell(\Theta^\mathsf{T} f(\mathbf{g}), y) \right\} - \max_{\mathbf{h}\in\mathcal{H}^m} \left\{ \alpha\mathbf{h}^\mathsf{T} W\mathbf{x} \right\} \;\le$$
$$\max_{\mathbf{g}'\in\mathcal{H}^m} \left\{ \beta\mathbf{g}'^\mathsf{T} W\mathbf{x} + \ell(\Theta^\mathsf{T} f(\mathbf{g}'), y) \right\} - \max_{\mathbf{h}'\in\mathcal{H}^m} \left\{ \beta\mathbf{h}'^\mathsf{T} W\mathbf{x} \right\} . \tag{2}$$

Proof. Let

$$\widehat{\mathbf{g}}_\alpha = \operatorname*{argmax}_{\mathbf{g}\in\mathcal{H}^m} \left\{ \alpha\mathbf{g}^\mathsf{T} W\mathbf{x} + \ell(\Theta^\mathsf{T} f(\mathbf{g}), y) \right\} , \qquad \widehat{\mathbf{g}}_\beta = \operatorname*{argmax}_{\mathbf{g}\in\mathcal{H}^m} \left\{ \beta\, \mathbf{g}^\mathsf{T} W\mathbf{x} + \ell(\Theta^\mathsf{T} f(\mathbf{g}), y) \right\} ,$$

then we have:

$$\beta\, \widehat{\mathbf{g}}_\alpha^\mathsf{T} W\mathbf{x} + \ell(\Theta^\mathsf{T} f(\widehat{\mathbf{g}}_\alpha), y) \;\le\; \beta\, \widehat{\mathbf{g}}_\beta^\mathsf{T} W\mathbf{x} + \ell(\Theta^\mathsf{T} f(\widehat{\mathbf{g}}_\beta), y) . \tag{3}$$

We also have:

$$\max_{\mathbf{h}\in\mathcal{H}^m} \left\{ \alpha\, \mathbf{h}^\mathsf{T} W\mathbf{x} \right\} \;=\; \alpha \operatorname{sgn}(W\mathbf{x})^\mathsf{T} W\mathbf{x} , \qquad \text{and} \qquad \max_{\mathbf{h}\in\mathcal{H}^m} \left\{ \beta\, \mathbf{h}^\mathsf{T} W\mathbf{x} \right\} \;=\; \beta \operatorname{sgn}(W\mathbf{x})^\mathsf{T} W\mathbf{x} . \tag{4}$$

Moreover,

$$\begin{aligned} \widehat{\mathbf{g}}_\alpha^\mathsf{T} W\mathbf{x} &\leq \operatorname{sgn}(W\mathbf{x})^\mathsf{T} W\mathbf{x} \quad \Longrightarrow \\ (\alpha-\beta)\,\widehat{\mathbf{g}}_\alpha^\mathsf{T} W\mathbf{x} &\leq (\alpha-\beta)\operatorname{sgn}(W\mathbf{x})^\mathsf{T} W\mathbf{x} \quad \Longrightarrow \\ (\alpha-\beta)\,\widehat{\mathbf{g}}_\alpha^\mathsf{T} W\mathbf{x} - \alpha\operatorname{sgn}(W\mathbf{x})^\mathsf{T} W\mathbf{x} &\leq -\beta\operatorname{sgn}(W\mathbf{x})^\mathsf{T} W\mathbf{x}\,. \end{aligned} \tag{5}$$

Now, summing the two sides of (3) and (5), and using (4), the inequality is proved. $\square$