Hedging Your Bets: Optimizing Accuracy-Specificity Trade-offs in Large Scale Visual Recognition Supplementary Material

Jia Deng[1,2], Jonathan Krause[1], Alexander C. Berg[3], Li Fei-Fei[1]
Stanford University[1], Princeton University[2], Stony Brook University[3]

1. An example of discontinuous $\Phi(f_\lambda)$

We give a concrete example of discontinuous $\Phi(f_\lambda)$ with respect to λ, *i.e.*, case(2) in Sec.4.2 of the main paper.

Consider the simplest hierarchy: two leaf nodes a and b, plus a root node c. Let the rewards $r_a = r_b = 1$ and $r_c = 0$. Let $p(a|x) = \Pr(Y = a|X = x)$ be the posterior probability of node a, which completely determines the posterior distribution.

Assuming that ties are broken alphabetically, using Eqn.6 in the main paper, it is easy to verify the following.

For $0 \leq \lambda \leq 1$,

$$f_\lambda(x) = \begin{cases} a, & p(a|x) \geq 0.5 \\ b, & p(a|x) < 0.5 \end{cases} . \quad (1)$$

For $\lambda > 1$,

$$f_\lambda(x) = \begin{cases} a, & p(a|x) \geq \frac{\lambda}{\lambda+1} \\ c, & \frac{1}{\lambda+1} < p(a|x) < \frac{\lambda}{\lambda+1} \\ b, & p(a|x) \leq \frac{1}{\lambda+1} \end{cases} . \quad (2)$$

Suppose $p(a|x)$ only takes two discrete values, $p_1 = 0.6$ and $p_2 = 0.4$. Let $\mu(p)$ be the percentage of examples such that $p(a|x) = p$. Here let $\mu(p_1) = \mu(p_2) = 0.5$, that is, half of examples have a posterior probability of 0.6, the other half 0.4. Then it is simple to confirm that for $0 \leq \lambda < 1.5$, $\Phi(f_\lambda) = 0.6$, *i.e.*, all examples are predicted to the leaf nodes. For $\lambda \geq 1.5$, $\Phi(f_\lambda) = 1$, *i.e.*, all examples are predicted to the root node. Therefore $\Phi(f_\lambda)$ is discontinuous at $\lambda = 1.5$ with a gap between 0.6 and 1. Thus for $1-\epsilon \in (0.6, 1)$, there exists no $\lambda^\dagger$ such that $\Phi(f_{\lambda^\dagger}) = 1-\epsilon$.

Note that in this example, the distribution of $p(a|x)$ concentrates on 0.4 and 0.6, violating our optimality condition stated in Theorem 4.1 of the main paper.

2. Our condition versus the one in [1]

In Theorem 4.1 of the main paper we established one condition under which DARTS converges to an optimal solution.

In [1], a different condition is established for strong duality in a general class selective rejection framework, *i.e.*, there exists a continuous density function $p(x|y)$ for each $y \in \mathcal{Y}$.

Here we show that their condition is insufficient to establish the optimality of DARTS by giving an example where their condition is satisfied but DARTS cannot converge to an optimal solution for certain $1 - \epsilon$.

Consider the same hierarchy as in Sec. 1, *i.e.*, a hierarchy of two leaf nodes a and b, plus a root node c. Let X be a one dimensional feature, specifically a real number on $[0, 1]$. Let $p(x|a)$ be the density function of X given $Y = a$ and we set

$$p(x|a) = \begin{cases} 8x, & x \in [0, \frac{1}{4}) \\ \frac{8}{3}(1-x), & x \in [\frac{1}{4}, 1] \end{cases} .$$

Let $p(x|b)$ be the density function of X given $Y = b$ and we set

$$p(x|b) = \begin{cases} \frac{8}{3}x, & x \in [0, \frac{3}{4}) \\ 8(1-x), & x \in [\frac{3}{4}, 1] \end{cases} .$$

Then $p(x|a)$ and $p(x|b)$ are both continuous and thus satisfy the condition in [1]. We plot $p(x|a)$ and $p(x|b)$ in Fig. 1.

Further assume that $\Pr(Y = a) = \Pr(Y = b) = \frac{1}{2}$. By Bayes' law,

$$p(a|x) = \begin{cases} 3/4, & x \in [0, 1/4] \\ 1-x, & x \in (1/4, 3/4) \\ 1/4, & x \in [3/4, 1] \end{cases} .$$

We plot $p(a|x)$ in Fig. 2.

Now consider f_λ. Eqn. 2 implies that if $\lambda > 3$, $\Phi(f_\lambda) = 1$, *i.e.*, every example is predicted to the root node. If $\lambda = 3$, all examples x in $[0, 1/4]$ are predicted to node a with $3/4$ of them being correct and all examples x in $[3/4, 1]$ are predicted to node b with $3/4$ of them being correct. The rest, *i.e.*, $x \in (1/4, 3/4)$ is predicted to the root node c with all of them being correct. Therefore with $\lambda = 3$, $\Phi(f_\lambda) = 3/4 \times 1/2 + 1 \times 1/2 = 7/8$.

Thus $\Phi(f_\lambda)$ is discontinuous at $\lambda = 3$ and for $1 - \epsilon \in (7/8, 1)$ there exists no $\lambda^\dagger$ such that $\Phi(f_{\lambda^\dagger}) = 1 - \epsilon$.

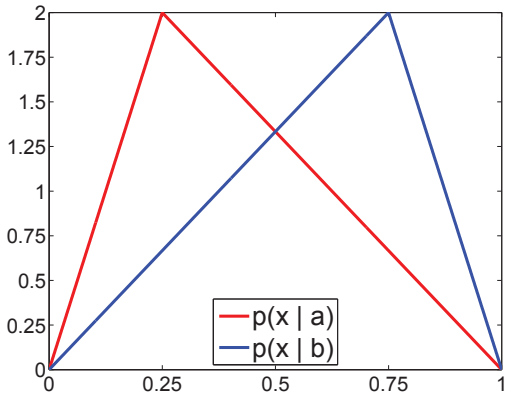

Figure 1. Probability density functions $p(x|a)$ and $p(x|b)$ for the example in Sec. 2.

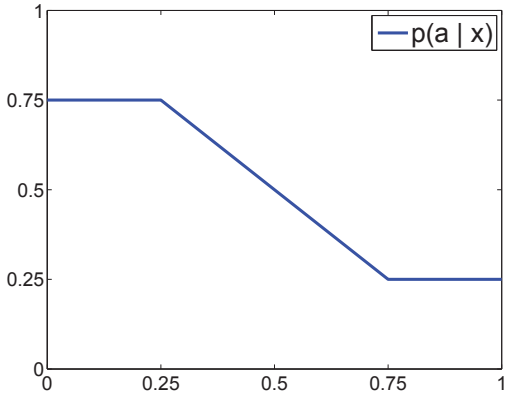

Figure 2. Posterior probability $p(a|x)$ for the example in Sec. 2.

Now we can show that DARTS fails to converge to an optimal solution. Let's set $1-\epsilon = 15/16$. The binary search in DARTS returns $\lambda' = 3+\delta$ where $\delta > 0$ is a small number. We then have $\Phi(f_{\lambda'}) = 1$ and $R(f_{\lambda'}) = 0$ since all examples are predicted to the root node. Consider a classifier g:

$$g(x) = \begin{cases} a, & x \in [0, 1/16] \\ c, & x \in (1/16, 1] \end{cases}.$$

Then $\Phi(g) \geq 15/16$ and $R(g) > 0$. Thus g is a better solution to OP1 than $f_{\lambda'}$.

The condition in [1] guarantees strong duality, *i.e.*, an optimal solution (to OP1 of the main paper) maximizes the Lagrange function that has a dual variable that minimizes the dual function. Thus under strong duality, if we can find all maximizers of the Lagrange function for any given dual variable, we then have a dual algorithm guaranteed to converge to an optimal solution. However, the maximizer is not necessarily unique and it can be impractical to find all of them. In fact, there can be infinitely many of them. DARTS finds a maximizer of the Lagrange function by only using posterior probabilities. This classifier is not necessarily optimal without certain conditions, as the example shows.

This example also shows that their condition does not imply ours, because otherwise by Theorem 4.1 of the main paper DARTS would converge to an optimal solution under their condition. On the other hand, our condition does not imply theirs either, because $p(x|b)$ and $p(x|a)$ do not need to be continuous to satisfy our condition.

3. Proof about $\bar{\lambda}$ in DARTS

We show that $\bar{\lambda}$ as defined in Line 5 of DARTS (see the main paper) can be an upper-bound of the binary search interval in DARTS. To that end, we first prove a lemma stating that given a $1-\epsilon$, if the transformed root reward $r_{\hat{v}} + \lambda$ is large enough, then the classifier f_λ that maximizes the Lagrange function $L(f, \lambda)$ is at least $1-\epsilon$ accurate.

Let $r_{max} = \max_{v \in V} r_v$ and we assume that $r_{max} > 0$ because otherwise all rewards are zero and OP1 is not meaningful.

Lemma 3.1. *For any $\lambda \geq 0$, if $(r_{\hat{v}}+\lambda)/(r_{max}+\lambda) \geq 1-\epsilon$, where $\hat{v}$ is the root node, then $\Phi(f_\lambda) \geq 1-\epsilon$.*

Proof. Assume to the contrary that $\Phi(f_\lambda) < 1-\epsilon$. Then

$$\begin{aligned} L(f_\lambda, \lambda) =& \mathbb{E}\,(r_{f_\lambda(X)} + \lambda)[f_\lambda(X) \in \pi(Y)] + \lambda(\epsilon - 1) \\ \leq& \mathbb{E}\,(r_{max} + \lambda)[f_\lambda(X) \in \pi(Y)] + \lambda(\epsilon - 1) \\ =& (r_{max} + \lambda)\Phi(f_\lambda) + \lambda(\epsilon - 1) \\ <& (r_{max} + \lambda)(1-\epsilon) + \lambda(\epsilon - 1) \\ \leq& r_{\hat{v}} + \lambda + \lambda(\epsilon - 1) = L(\hat{f}, \lambda), \end{aligned}$$

where $\hat{f}$ is the trivial solution that maps all examples to the root node. This contradicts that f_λ maximizes $L(f, \lambda)$. □

Now we are ready to prove the claim about $\bar{\lambda}$.

Lemma 3.2. *Let $\bar{\lambda} = (r_{max}(1-\epsilon) - r_{\hat{v}})/\epsilon$, where $\hat{v}$ is the root node and $r_{max} = \max_{v \in V} r_v$. If $\Phi(f_0) < 1-\epsilon$, then $\bar{\lambda} > 0$ and $\Phi(f_{\bar{\lambda}}) \geq 1-\epsilon$.*

Proof. Note that $\bar{\lambda} > 0$ because otherwise Lemma 3.1 implies that $\Phi(f_0) \geq 1-\epsilon$. It is easy to verify that $(r_{\hat{v}} + \bar{\lambda})/(r_{max} + \bar{\lambda}) \geq 1-\epsilon$. Lemma 3.1 then implies that $\Phi(f_{\bar{\lambda}}) \geq 1-\epsilon$. □

4. Proof of the main theorem

Finally we give the complete proof of Theorem 4.1 in the main paper. For completeness, we repeat the formal definitions here.

Let $\Delta = \{q \in \mathbb{R}^{|\mathcal{Y}|-1} : q \succeq 0, \|q\|_1 \leq 1\}$ be the set of possible posterior probabilities over $|\mathcal{Y}|-1$ leaf nodes. Note that for $|\mathcal{Y}|$ leaf nodes there are only $|\mathcal{Y}|-1$ degrees of freedom. With slight abuse of notation, we write q_y as the

probability of any leaf node $y \in \mathcal{Y}$, with the understanding that $q_{y_1} = 1 - \|q\|_1$, where y_1 is the leaf node not in the $|\mathcal{Y}| - 1$ leaf nodes. We also use $q_v = \sum_{y \in \mathcal{Y}} [v \in \pi(y)] q_y$ to mean the posterior probability at a node $v \in V$.

Let $\Delta^\ddagger = \{q \in \Delta : \|q\|_\infty = 1 \vee q = 0\}$ be the set of posterior probabilities at the vertices of Δ, where one of the leaf nodes takes probability 1.

Let $\vec{p}_{Y|X} : \mathcal{X} \to \Delta$ be a Borel measurable function that maps an example x to its posterior probabilities on the leaf nodes. Let $Q = \vec{p}_{Y|X}(X)$ be the posterior probabilities on the leaf nodes for the random variable X. As a function of X, Q is also a random variable.

Theorem 4.1. *If $\Pr(Q \in \Delta^\ddagger) = 1$, or Q has a probability density function with respect to the Lebesgue measure on $\mathbb{R}^{|\mathcal{Y}|-1}$ conditioned on $Q \notin \Delta^\ddagger$, then DARTS converges to an optimal solution of OP1.*

Proof. Per our analysis in Sec.4 of the main paper, we only need to show the continuity of $\Phi(f_\lambda)$ with respect to $\lambda \geq 0$.

We first have

$$\begin{aligned}\Phi(f_\lambda) =& p^\ddagger \mathbb{E}_{X,Y|Q\in\Delta^\ddagger}[f_\lambda(X) \in \pi(Y)] \\ &+ (1 - p^\ddagger)\mathbb{E}_{X,Y|Q\notin\Delta^\ddagger}[f_\lambda(X) \in \pi(Y)],\end{aligned} \tag{3}$$

where $p^\ddagger = \Pr(Q \in \Delta^\ddagger)$.

Consider the first expectation in Eqn. 3. Let $y^\ddagger$ be the leaf node such that $Q_{y^\ddagger} = 1$, *i.e.*, $Y = y^\ddagger$ with probability 1. Then $p_{Y|X}(v|X) = 1$ for any $v \in \pi(y^\ddagger)$ and $p_{Y|X}(v|X) = 0$ otherwise. Therefore $f_\lambda(X) = \operatorname{argmax}_{v\in V} r_v p_{Y|X}(v|X) = \operatorname{argmax}_{v\in\pi(y^\ddagger)} r_v$. Thus $f_\lambda(X) \in \pi(Y)$ with probability 1, *i.e.*, the first expectation is the constant 1.

Therefore we only need to show the continuity of the second expectation with respect to λ and we do this in the rest of the proof. For simplicity of notation, we drop $Q \notin \Delta^\ddagger$ hereafter and simply write the second expectation in Eqn. 3 as $\mathbb{E}_{X,Y}[f_\lambda(X) \in \pi(Y)]$.

Define $\tilde{f}_\lambda : \Delta \to V$ as

$$\tilde{f}_\lambda(q) = \operatorname{argmax}_{v\in V}(r_v + \lambda)q_v,$$

breaking ties the same way as Eqn.6 of the main paper. Then it follows that $\forall x \in \mathcal{X}$, $f_\lambda(x) = \tilde{f}_\lambda(\vec{p}_{Y|X}(x))$. We also define

$$\Gamma_v(\lambda) = \{q \in \Delta : (r_v + \lambda)q_v > (r_{v'} + \lambda)q_{v'}, \forall v' \neq v\}.$$

to be the open polyhedron in Δ such that $\tilde{f}_\lambda(q) = v, \forall q \in \Gamma_v(\lambda)$, *i.e.*, the set of posterior probabilities that lead to a prediction at node v given λ. Also let

$$\begin{aligned}\bar{\Gamma}(\lambda) =& \{q \in \Delta : \exists v', v'', v' \neq v'', \forall u \neq v', u \neq v'', \\ & (r_{v'} + \lambda)q_{v'} = (r_{v''} + \lambda)q_{v''} \geq (r_u + \lambda)q_u\},\end{aligned}$$

i.e., the set of probabilities that lie on a decision boundary. It is then a simple exercise to check that $\bar{\Gamma}(\lambda)$ and $\Gamma_v(\lambda), \forall v \in V$ partition Δ.

Let $p_Q(q)$ be the (conditional) density function of Q given $Q \notin \Delta^\ddagger$, then

$$\begin{aligned}&\mathbb{E}_{X,Y}[f_\lambda(X) \in \pi(Y)] \\ &= \mathbb{E}_X \mathbb{E}_{Y|X}[f_\lambda(X) \in \pi(Y)] \\ &= \mathbb{E}_Q \mathbb{E}_{X|Q} \mathbb{E}_{Y|X,Q}[f_\lambda(X) \in \pi(Y)] \\ &= \mathbb{E}_Q \mathbb{E}_{X|Q} \sum_{y\in\mathcal{Y}} [f_\lambda(X) \in \pi(y)] p_{Y|X}(y|X) \\ &= \mathbb{E}_Q \mathbb{E}_{X|Q} \sum_{y\in\mathcal{Y}} [\tilde{f}_\lambda(Q) \in \pi(y)] Q_y \\ &= \mathbb{E}_Q \sum_{y\in\mathcal{Y}} [\tilde{f}_\lambda(Q) \in \pi(y)] Q_y \\ &= \int_\Delta \sum_{y\in\mathcal{Y}} [\tilde{f}_\lambda(q) \in \pi(y)] q_y \, p_Q(q) \, \mathrm{d}q \\ &= \left(\int_{\bar{\Gamma}(\lambda)} + \sum_{v\in V} \int_{\Gamma_v(\lambda)}\right) \sum_{y\in\mathcal{Y}} [\tilde{f}_\lambda(q) \in \pi(y)] q_y \, p_Q(q) \, \mathrm{d}q \\ &= \sum_{v\in V} \int_{\Gamma_v(\lambda)} q_v p_Q(q) \, \mathrm{d}q \\ &= \sum_{v\in V} \int [q \in \Gamma_v(\lambda)] q_v p_Q(q) \, \mathrm{d}q.\end{aligned}$$

Note that the first two equalities are by iterated expectations. Also we can drop $\int_{\bar{\Gamma}(\lambda)}$ at the second to last step because $\bar{\Gamma}(\lambda)$ has fewer than $|\mathcal{Y}| - 1$ dimensions and therefore has zero measure.

Let $\phi_v(\lambda, q) = [q \in \Gamma_v(\lambda)] q_v p_Q(q)$. To prove continuity, it suffices to show that for each v, $\int \phi_v(\lambda, q) \, \mathrm{d}q$ is continuous with respect to λ, *i.e.*, for sequences $\{\lambda_n\}$, if $\lim_{n\to\infty} \lambda_n = \lambda$, then $\lim_{n\to\infty} \int \phi_v(\lambda_n, q) \, \mathrm{d}q = \int \phi_v(\lambda, q) \, \mathrm{d}q$. This is directly implied by Lebesgue's dominated convergence theorem if we can show (1)$\lim_{n\to\infty} \phi_v(\lambda_n, q) = \phi_v(\lambda, q)$ almost everywhere and (2) for all n and every q, $|\phi_v(\lambda_n, q)| \leq \psi(q)$ for some integrable ψ.

Note that condition(2) is trivial as $\phi_v(\lambda_n, q) \leq p_Q(q)$ and thus we only need to check condition(1). First note that condition(1) trivially holds for any $q \notin \Delta$. For $q \in \Delta$, there are three possibilities: (i)$q \in \Gamma_v(\lambda)$, (ii)$q \in \Gamma_u(\lambda)$ for some $u \neq v$, or (iii)$q \in \bar{\Gamma}(\lambda)$. We only need to check (i) and (ii) because $\bar{\Gamma}(\lambda)$ has zero measure.

For (i), let $\gamma_v(\lambda, q) = (r_v + \lambda)q_v - \max_{v'\neq v}(r_{v'} + \lambda)q_{v'}$. For any $q \in \Gamma_v(\lambda)$, as $n \to \infty$, $\gamma_v(\lambda_n, q) \to \gamma_v(\lambda, q) > 0$. Therefore there exists n' such that $\forall n > n', \gamma_v(\lambda_n, q) > 0$ and thus $\forall n > n', [q \in \Gamma_v(\lambda_n)] = 1$. Therefore $[q \in \Gamma_v(\lambda_n)] \to 1 = [q \in \Gamma_v(\lambda)]$.

For (ii), since $q \in \Gamma_u(\lambda)$ for some $u \neq v$, as $n \to \infty$,

$\gamma_v(\lambda_n, q) \to \gamma_v(\lambda, q) < 0$ and therefore $[q \in \Gamma_v(\lambda_n)] \to 0 = [q \in \Gamma_v(\lambda)]$.

□

References

[1] E. Grall-Maes and P. Beauseroy. Optimal decision rule with class-selective rejection and performance constraints. *PAMI*, 31(11):2073 –2082, 2009.