Approximation by mixtures and geometric mixtures

W. D. Brinda, Jason M. Klusowski, and Andrew R. Barron

In a groundbreaking paper, Jones [1992] proved that the integrated squared error between a function in a Hilbert space and the best k-term linear combination greedily selected from a spanning set decays with order $1/k$ as long as a certain L^1-type norm is finite. Implications for neural network approximation of sigmoidal functions were worked out in detail by Barron [1993]; bounds for greedily estimating neural nets from data were given in Barron [1994]. These developments were significant for two main reasons: they showed that good approximation is possible without the number of nodes growing exponentially with the dimension of the function's domain, and they provided a more feasible optimization algorithm (greedily, one node at a time) for defining the nodes.

Under the advisement of Andrew Barron, Jonathan Li established analogous $1/k$ rates of approximation error and risk bounds for greedy k-component mixture *density estimation.* Their work is detailed in Li's doctoral thesis (Li [1999]) and summarized by Li and Barron [2000]. However, their inequality requires the family to have a uniformly bounded density ratio. As a result, it does not apply to familiar families, including Gaussian mixtures. In such cases, Li and Barron [2000] advocate truncating the distributions and restricting the parameter space to a compact subset of $\mathbb{R}^d$. Chapter 3 of Brinda [2018] indicated that the I-divergence approximation error result can hold without requiring a bounded density ratio if a certain complexity constant is finite. However, we will show that a much cleaner result can be established if a different divergence is used to quantify approximation error.

Barron and Li used I-divergence (relative entropy) for approximation error because it fits neatly with MDL penalized likelihood risk bounds. Instead, we will use *K-divergence*; it is closely related to relative entropy but is not readily applied to bounding statistical risk. In Section 1, we show that $1/k$ rates of approximation error hold for K-divergence with very little required of the target distribution or the family. This will imply an analogous result for Hellinger distance.

Section 2 introduces the notion of greedy geometric mixtures and bounds the J-divergence approximation error of such mixtures under appropriate conditions.

Proofs of lemmas and theorems are at the end of each section.

1 Hellinger approximation error of mixtures

Suppose $\Phi := \{\phi_\mu : \mu \in \Gamma\}$ is a family of probability densities on a measurable space $\mathcal{X}$ with respect to a σ-finite dominating measure. Let Q be a probability measure on Γ whose domain σ-algebra is fine enough that $(\mu, x) \mapsto \phi_\mu(x)$ is product-measurable.[1] Let $\bar{\phi}_Q$ denote the integral transform of Q defined by

$$\begin{aligned}\bar{\phi}_Q(x) &:= \int_\Gamma \phi_\mu(x) dQ(\mu) \\ &= \mathbb{E}_{\mu\sim Q}\phi_\mu(x).\end{aligned}$$

Tonelli's Theorem allows us to conclude that $\bar{\phi}_Q$ is measurable, and, by interchanging integrals, that $\bar{\phi}_Q$ must be a probability density as well. The corresponding probability measure on $\mathcal{X}$ is denoted $\bar{\Phi}_Q$ and is called the Q *mixture (over* Φ*)*.

We let $\mathcal{C}(\Phi)$ denote set of all such integral transforms of probability measures (each defined on a sufficiently fine σ-algebra of Γ); this set is convex. Notice that $\mathcal{C}(\Phi)$ includes all discrete mixtures from Φ. Importantly, $\mathcal{C}(\Phi)$ also includes all of the other well-defined "mixtures" such as *continuous mixtures*, as allowed by the nature of Γ.

Given any "target" probability measure P on $\mathcal{X}$, the greedy algorithm of Barron and Li constructs a sequence of approximating mixtures

$$p_{\theta_{k+1}^{(P)}} = (1-\alpha_{k+1})p_{\theta_k^{(P)}} + \alpha_{k+1}\phi_{\mu_{k+1}^{(P)}}.$$

The mixture components $\theta_k^{(P)} = \{\mu_1^{(P)}, \ldots, \mu_k^{(P)}\}$ are greedily chosen according to

$$\begin{aligned}\mu_1^{(P)} &:= \operatorname*{argmax}_{\mu\in\Gamma} \mathbb{E}_{X\sim P} \log \phi_\mu(X), \qquad \text{followed by} \\ \mu_{j+1}^{(P)} &:= \operatorname*{argmax}_{\mu\in\Gamma} \mathbb{E}_{X\sim P} \log[(1-\alpha_{j+1})p_{\theta_j^{(P)}}(X) + \alpha_{j+1}\phi_\mu(X)].\end{aligned}$$

We will assume throughout this paper that a maximizer exists at each step; it need not be unique.

We will use the term "Barron's weights" to refer to the sequence $\alpha_j = 2/(j+1)$. Barron and Li suggest using either these weights or finding the optimal weights at each step.[2] After k steps, the weight of component $j \in \{1, \ldots, k\}$ is $\alpha_j \prod_{i=j+1}^k (1-\alpha_i)$; with Barron's weights, this simplifies to $\frac{2j}{k(k+1)}$. We will provide results for this choice of weights and also for the choice $\alpha_j = 1/j$ which results in an equal-weighted mixture.

[1] By the theory of Carathéodory functions, if $\mathcal{X}$ is a separable metrizable space and each density $\phi_\mu : \mathcal{X} \to \mathbb{R}^+$ is continuous, then product-measurability is guaranteed as long as the domain σ-algebra is fine enough that $\mu \mapsto \phi_\mu(x)$ is measurable for every $x \in \mathcal{X}$ — see [Aliprantis and Border, 2006, Lem 4.51].

[2] Technically, Li presented the slightly different sequence $\alpha_2 = 2/3$ and $\alpha_j = 2/j$ thereafter. The sequence $2/(j+1)$ also works and is a bit simpler.

The work of Barron and Li proves that

$$D(P\|P_{\theta_k^{(P)}}) \leq D(P\|\bar{\Phi}_Q) + \frac{(1 + \log \sup_{x,\gamma_1,\gamma_2} \frac{\gamma_1(x)}{\gamma_2(x)})c_Q^2(P)}{k}$$

where

$$c_Q^2(P) := \mathbb{E}_{X\sim P} \frac{\mathbb{E}_{\mu\sim Q}\phi_\mu^2(X)}{\bar{\phi}_Q^2(X)}.$$

Section 3.2 of Li's dissertation discusses $c_Q^2(P)$, pointing out that it is 1 plus an expected coefficient of variation; his Lemma 3.1 shows that $c_Q^2(\bar{\Phi}_Q)$ is bounded by the number of components of $\bar{\Phi}_Q$ if it is a discrete mixture from the model.

Because the inequality holds for every Q, we can state an infimum version:

$$D(P\|P_{\theta_k^(P)}) \leq D(P\|\mathcal{C}(\Phi)) + \frac{(1 + \log \sup_{x,\gamma_1,\gamma_2} \gamma_1(x)/\gamma_2(x))c_\Phi^2(P)}{k}$$

where

$$c_\Phi^2(P) := \lim_{\epsilon\to 0} \inf \left\{ c_Q^2(P) : Q \text{ s.t. } D(P\|Q) \leq D(P\|\mathcal{C}(\Phi)) + \epsilon \right\}.$$

If equal weights are used rather than Barron's weights, then Brinda [2018] verifies that the results holds with $(1 + \log k)/k$ replacing $1/k$.

A popular alternative to I-divergence is *Jensen-Shannon divergence.*

$$D_{\text{JS}}(P, Q) := \tfrac{1}{2}D(P\|\tfrac{1}{2}Q + \tfrac{1}{2}P) + \tfrac{1}{2}D(Q\|\tfrac{1}{2}P + \tfrac{1}{2}Q)$$

It is defined in terms of I-divergence, but unlike I-divergence it is symmetric and finite. The Jensen-Shannon divergence is the average of two quantities that are called *K-divergences* by Lin [1991].

$$D_{\text{K}}(P\|Q) := D(P\|\tfrac{1}{2}Q + \tfrac{1}{2}P)$$

The K-divergence was generalized to the family of *skewed K-divergences* by Nielsen [2010] as

$$D_{\text{K},\lambda}(P\|Q) := D(P\|[1-\lambda]Q + \lambda P)$$

for any $\lambda \in [0, 1)$. Notice that $D_{\text{K},\lambda}$ becomes increasingly similar to I-divergence as λ approaches zero. As pointed out by Nielsen, $D_{\text{K},\lambda}$ is the f-divergence defined by $f(t) = -t\log(\lambda + \frac{1-\lambda}{t})$.

If instead of I-divergence, one uses skewed K-divergence to quantify approximation error, a remarkably clean statement about greedy mixtures holds. In this context, we apply a new greedy algorithm to create $\hat{\theta}_k = (\hat{\gamma}_1, \ldots, \hat{\gamma}_k)$ by maximizing the expected log of a modified density.

The key to our skewed K-divergence approximation error result is the following inequality which is analogous to [Li, 1999, Lem 5.9].

Theorem 1.1. *Let $\Phi := \{\phi_\gamma : \gamma \in \Gamma\}$ be a family of probability densities with respect to a σ-finite dominating measure, and let Q be a probability measure on Γ for which $(\gamma, x) \mapsto \phi_\gamma(x)$ is product-measurable. Given $\lambda \in [0,1]$, let $P_{\tilde{\theta}_1}$, $P_{\tilde{\theta}_2}$, ... be the sequence of mixtures that greedily maximize $\mathbb{E}_{X\sim P} \log\left([1-\lambda]p_{\theta_1}(X) + \lambda\bar{\phi}_Q(X)\right)$, $\mathbb{E}_{X\sim P} \log\left([1-\lambda]p_{\theta_2}(X) + \lambda\bar{\phi}_Q(X)\right)$, ... If either Barron's weights or optimal weights were used, then*

$$\mathbb{E}_{X\sim P} \log \frac{\bar{\phi}_Q(X)}{[1-\lambda]p_{\tilde{\theta}_k}(X) + \lambda\bar{\phi}_Q(X)} \leq \frac{\log(3\sqrt{e}/\lambda)c_Q^2(P)}{k}.$$

Alternatively, if equal weights were used, then

$$\mathbb{E}_{X\sim P} \log \frac{\bar{\phi}_Q(X)}{[1-\lambda]p_{\tilde{\theta}_k}(X) + \lambda\bar{\phi}_Q(X)} \leq \frac{(1+\log k)\log(2\sqrt{e}/\lambda)c_Q^2(P)}{k}.$$

When the target P is itself a mixture over Φ, a direct application of Theorem 1.1 gives approximation error rates for the greedy mixtures. To simplify, we specialize to ordinary K-divergence.

Corollary 1.2. *Let $\Phi := \{\phi_\gamma : \gamma \in \Gamma\}$ be a family of probability densities with respect to a σ-finite dominating measure, and let Q be a probability measure on Γ for which $(\gamma, x) \mapsto \phi_\gamma(x)$ is product-measurable. Let $P_{\tilde{\theta}_1}$, $P_{\tilde{\theta}_2}$, ... be the sequence of mixtures that greedily maximize $\mathbb{E}_{X\sim P} \log\left(\frac{1}{2}p_{\theta_1}(X) + \frac{1}{2}\bar{\phi}_Q(X)\right)$, $\mathbb{E}_{X\sim P} \log\left(\frac{1}{2}p_{\theta_2}(X) + \frac{1}{2}\bar{\phi}_Q(X)\right)$, ... If either Barron's weights or optimal weights were used, then*

$$D_K(\bar{\Phi}_Q \| P_{\tilde{\theta}_k}) \leq \frac{\log(6\sqrt{e})\, c_Q^2(P)}{k}.$$

Alternatively, if equal weights were used, then

$$D_K(\bar{\Phi}_Q \| P_{\tilde{\theta}_k}) \leq \frac{(1+\log k)\ \log(4\sqrt{e})\, c_Q^2(P)}{k}.$$

K-divergence can be used to bound Hellinger distance d_{H} using a Pinsker's-type inequality.

Lemma 1.3. *For any probability measures P and Q on a measurable space,*

$$d_H^2(P,Q) \leq 6D_K(P\|Q).$$

We use the triangle inequality to state a Hellinger approximation error bound that works for arbitrary distributions.

Corollary 1.4. *Let $\Phi := \{\phi_\gamma : \gamma \in \Gamma\}$ be a family of probability densities with respect to a σ-finite dominating measure, and let Q be a probability measure on Γ for which $(\gamma, x) \mapsto \phi_\gamma(x)$ is product-measurable. Let $P_{\tilde{\theta}_1}$, $P_{\tilde{\theta}_2}$, ... be the sequence of mixtures that greedily maximize $\mathbb{E}_{X\sim P} \log\left(\frac{1}{2}p_{\theta_1}(X) +\right.$*

$\frac{1}{2}\bar{\phi}_Q(X))$, $\mathbb{E}_{X\sim P}\log\left(\frac{1}{2}p_{\theta_2}(X) + \frac{1}{2}\bar{\phi}_Q(X)\right)$, $\ldots$. *If either Barron's weights or optimal weights were used, then*

$$d_H(P, P_{\tilde{\theta}_k}) < d_H(P, \bar{\Phi}_Q) + \frac{4\, c_Q(\bar{\Phi}_Q)}{\sqrt{k}}.$$

Alternatively, if equal weights were used, then

$$d_H(P, P_{\tilde{\theta}_k}) < d_H(P, \bar{\Phi}_Q) + \frac{(1+\log k)\, 4\, c_Q(\bar{\Phi}_Q)}{\sqrt{k}}.$$

Finally, define

$$\widetilde{c}_\Phi(P) := \lim_{\epsilon\to 0}\inf\left\{c_Q(\bar{\Phi}_Q) : Q \text{ s.t. } d_{\mathrm{H}}(P, \bar{\Phi}_Q) \le d_{\mathrm{H}}(P, \mathcal{C}(\Phi)) + \epsilon\right\}.$$

Because the modified greedy algorithm depends on Q, Corollary 1.5 only claims that a mixture with the specified approximation error exists and does not explicitly say how to construct it.

Corollary 1.5. *Let $\Phi := \{\phi_\gamma : \gamma \in \Gamma\}$ be a family of probability densities with respect to a σ-finite dominating measure, and let P be a probability measure defined on the same measurable space as the dominating measure. There exists a Barron-weighted k-component mixture P_{θ_k} from Φ such that*

$$d_H(P, P_{\theta_k}) < d_H(P, \mathcal{C}(\Phi)) + \frac{4\, \widetilde{c}_\Phi(P)}{\sqrt{k}}.$$

Additionally, there exists an equally-weighted k-component mixture $P_{\theta'_k}$ from Φ such that

$$d_H(P, P_{\theta'_k}) < d_H(P, \mathcal{C}(\Phi)) + \frac{(1+\log k)\, 4\, \widetilde{c}_\Phi(P)}{\sqrt{k}}.$$

Unlike the I-divergence mixture approximation bounds, these Hellinger results do not require awkward conditions on the family or on the distribution to be approximated. Note that an approximation result for total variation distance also follows using the fact that it is bounded by $\sqrt{2}$ times Hellinger distance.

Proofs

Proof of Theorem 1.1. Define the family $\Phi_{\lambda\bar{\Phi}_Q} := \{[1-\lambda]\phi_\gamma + \lambda\bar{\phi}_Q : \gamma \in \Gamma\}$. Observe that $\bar{\phi}_Q$ is in $\mathcal{C}(\Phi_{\lambda\bar{\phi}_Q})$ as well; specifically, $\mathbb{E}_{\gamma\sim Q}[(1-\lambda)\phi_\gamma + \lambda\bar{\phi}_Q] = \bar{\phi}_Q$. Importantly, the modified greedy algorithm on Φ is identical to the ordinary greedy algorithm on $\Phi_{\lambda\bar{\Phi}_Q}$ as we now verify by induction. For the first step, this is clear. Next, assume that $(1-\lambda)p_{\hat{\theta}_k} + \lambda\bar{\phi}_Q$ is the greedily optimal choice from $\Phi_{\lambda\bar{\Phi}_Q}$. The next greedy step (with weight α on the new component) optimizes an expected log of

$$(1-\alpha)[(1-\lambda)p_{\hat{\theta}_k} + \lambda\bar{\phi}_Q] + \alpha[(1-\lambda)\phi_\gamma + \lambda\bar{\phi}_Q] = (1-\lambda)[(1-\alpha)p_{\hat{\theta}_k} + \lambda\phi_\gamma] + \lambda\bar{\phi}_Q$$

where the second representation makes it clear that this maximization is the same as a step of the modified greedy algorithm.

Thus we are justified in applying the ordinary greedy algorithm results to $\Phi_{\lambda\bar{\Phi}_Q}$. Compare our situation to the proofs of [Li, 1999, Lem 5.8 and Lem 5.9]. Two discrepancies arise in the term labeled "3" on page 57 of Li.

First, using the decreasing property of ζ shown in [Li, 1999, Lem 5.3],

$$\begin{aligned}\zeta\left([1-\alpha]\frac{(1-\lambda)p_{\hat{\theta}_{k-1}}+\lambda\bar{\phi}_Q}{\bar{\phi}_Q}\right) &\leq \zeta\left([1-\alpha]\frac{\lambda\bar{\phi}_Q}{\bar{\phi}_Q}\right)\\ &\leq \zeta([1-\alpha]\lambda).\end{aligned}$$

Secondly, this term's expectation, which arises in the proof of [Li, 1999, Lem 5.9], will involve

$$\begin{aligned}&\mathbb{E}_{X\sim P}\frac{\mathbb{E}_{\gamma\sim Q}[[1-\lambda]\phi_\gamma+\lambda\bar{\phi}_Q]^2(X)}{[\mathbb{E}_{\gamma\sim Q}[[1-\lambda]\phi_\gamma+\lambda\bar{\phi}_Q](X)]^2}\\ &\qquad= \mathbb{E}_{X\sim P}\frac{\mathbb{E}_{\gamma\sim Q}[[1-\lambda]\phi_\gamma+\lambda\bar{\phi}_Q]^2(X)}{\bar{\phi}_Q^2(X)}\\ &\qquad= \mathbb{E}_{X\sim P}\frac{[1-\lambda]^2\mathbb{E}_{\gamma\sim Q}\phi_\gamma^2(X)+2\lambda(1-\lambda)\bar{\phi}_Q^2(X)+\lambda^2\bar{\phi}_Q^2(X)}{\bar{\phi}_Q^2(X)}\\ &\qquad= (1-\lambda)^2c_Q^2(P)+\lambda(2-\lambda)\\ &\qquad\leq (1-\lambda)^2c_Q^2(P)+\lambda(2-\lambda)c_Q^2(P)\\ &\qquad= c_Q^2(P);\end{aligned}$$

the inequality follows from the fact that $c_Q^2(P)\geq 1$.

In light of these observations, Li's proofs establish that for a step with weight α on the new component

$$\begin{aligned}P\log\frac{\bar{\phi}_Q}{[1-\lambda]p_{\hat{\theta}_{k+1}}+\lambda\bar{\phi}_Q} &\leq (1-\alpha)P\log\frac{\bar{\phi}_Q}{[1-\lambda]p_{\hat{\theta}_k}+\lambda\bar{\phi}_Q}+\alpha^2\zeta((1-\alpha)\lambda)(1-\lambda^2)c_Q^2(P)\\ &\leq (1-\alpha)P\log\frac{\bar{\phi}_Q}{[1-\lambda]p_{\hat{\theta}_k}+\lambda\bar{\phi}_Q}\\ &\qquad+\alpha^2[1/2+\log\tfrac{1}{1-\alpha}+\log\tfrac{1}{\lambda}]c_Q^2(P) \qquad (1)\end{aligned}$$

using [Li, 1999, Lem 5.4]. The initial term is

$$P\log\frac{\bar{\phi}_Q}{[1-\lambda]\phi_{\hat{\gamma}_1}+\lambda\bar{\phi}_Q}\leq\log(1/\lambda).$$

The factor multiplying α^2 in (1) is at least $1/2+\log\frac{1}{\lambda}$, which is at least $\log\frac{1}{\lambda}$, so it bounds the initial term.

With Barron's weights, apply [Li, 1999, Lem 5.6] and use $\alpha\leq 2/3$; with equal weights, apply [Brinda, 2018, Lem 3.1.1] and use $\alpha\leq 1/2$. □

Proof of Lemma 1.3. We noted that K-divergence is the f-divergence defined by $f_K(t) := t \log \frac{2t}{t+1}$. It can also be expressed by $\tilde{f}_K(t) := t \log \frac{2t}{t+1} + \frac{1-t}{2}$ since the additional term has Q-integral zero when the density dP/dQ is substituted for t. We will show that $f_H(t) := (\sqrt{t} - 1)^2$, which defines the squared Hellinger divergence is bounded point-wise by $6\tilde{f}_K$. This proves that

$$\begin{aligned} D_{\mathrm{K}}(P \| Q) &= \int \tilde{f}_K(dP/dQ)dQ \\ &\leq 6 \int f_H(dP/dQ)dQ \\ &= 6d_{\mathrm{H}}^2(P, Q). \end{aligned}$$

To verify the point-wise inequality, first consider the region $t \geq 38$. One can confirm that $\log \frac{2t}{t+1} \geq 2/3$ when t is this large.

$$\begin{aligned} f_H(t) = (\sqrt{t} - 1)^2 & \\ &\leq t + 1 \\ &\leq 6\left((\tfrac{2}{3} - \tfrac{1}{2})t + \tfrac{1}{2}\right) \\ &\leq 6\left((\log \frac{2t}{t+1} - \tfrac{1}{2})t + \tfrac{1}{2}\right) \\ &= 6\tilde{f}_K \end{aligned}$$

Next, we will consider the region $[.9, 1.1]$. Note that $f_H(1) = f'_H(1) = f_K(1) = f'_K(1) = 0$. By Taylor expansion at 1 with Lagrange remainder,

$$f_H(t) = (t-1)^2 \left(\frac{1}{2t_1^{3/2}} \right)$$

and

$$6\tilde{f}_K(t) = (t-1)^2 \left(\frac{6}{t_2(t_2+1)^2} \right)$$

for some t_1 and t_2 between 1 and t. A plot shows that the second derivative of $6\tilde{f}_K$ is uniformly larger than the second derivative of f_H on $[.9, 1.1]$, so our expression for $6\tilde{f}_K(t)$ is uniformly larger than our expression for $f_H(t)$ on that interval regardless of what t_1 and t_2 are.

Finally, for the remaining regions $[0, .9)$ and $(1.1, 38)$, the point-wise inequality is easy to confirm with a plot.

□

2 Jeffreys approximation error of geometric mixtures

LET JASON WRITE THIS section on our overleaf document then copy and paste it here.

Proofs

PROOFS AT THE END

References

Charalambos D Aliprantis and Kim Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer Science & Business Media, 2006.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

W. D. Brinda. *Adaptive Estimation with Gaussian Radial Basis Mixtures*. PhD thesis, Yale University, 2018.

Lee K Jones. A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *The annals of Statistics*, pages 608–613, 1992.

Jonathan Q Li. *Estimation of Mixture Models*. PhD thesis, Yale University, 1999.

Jonathan Q Li and Andrew R Barron. Mixture density estimation. In *Advances in Neural Information Processing Systems 12*, pages 279–285. The MIT Press, 2000.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Frank Nielsen. A family of statistical symmetric divergences based on jensen's inequality. *arXiv preprint arXiv:1009.4004*, 2010.