A NBCS Hausdorff distance Approximation

In this section we show that the nested barycentric coordinate system (NBCS) can represent an arbitrarily close approximation to any convex body. As stated the NBCS produces a (not necessarily convex) piece-wise linear classifier. In fact, this method can approximate multiple convex bodies. For simplicity, we focus on the case of a single convex body, and demonstrate how our method approximates it. This will be done by placing split points at the barycenters of their containing simplices, where the barycenter of a simplex with vertices $p_0, \ldots, p_d$ is given by $(p_0 + \cdots + p_d)/(d+1)$.

In order to state our result formally, we introduce some notation: Given a point $p \in \mathbb{R}^d$ and a parameter $\varepsilon > 0$, let $B_\varepsilon(p) = \{q \in \mathbb{R}^d : \|q - p\|_2 \leq \varepsilon\}$ be the ball of radius ε centered at p. Given a set $X \subseteq \mathbb{R}^d$, let

$$X^{(-\varepsilon)} = \{p \in X : B_\varepsilon(p) \subseteq X\}$$

be the set of all points of X that are at distance at least ε from the boundary of X. Recall that S denotes the unit simplex.

Theorem A.1. *Let $P \subseteq S$ be a given convex body of diameter 1, and let $0 < \varepsilon < 1$ be given. Set $s = 2^{O(d)} \ln^2(1/\varepsilon)$. Then there exists a nested system B_t of $\min\{(d+1)^s, dn\}$ distinct simplices obtained by always placing split points at the barycenters of their containing simplices, and a corresponding set of weights w, such that*

$$\tilde{P} = \{x \in S : w \cdot \phi_t(x) \geq 0\} \tag{9}$$

satisfies the following:

1. $\operatorname{vol}(\tilde{P} \setminus P) < \varepsilon \operatorname{vol}(S)$.
2. $P^{(-\varepsilon)} \subseteq \tilde{P}^{(-\varepsilon)} \subseteq P \subseteq \tilde{P}$.

Further, this system may be computed in time

$$O(\min\{(d+1)^s, dns\}[d^2 n + e^{O(\sqrt{d \log d})}]).$$

The importance of Theorem A.1 is that a simple NBCS can closely approximate any convex body P. And if P has margin ε, NBCS can produce $(\tilde{P}^{(-\varepsilon)})$ which falls fully within the margin of P. As described in section 2, finding a consistent polytope to a convex body with margin is a problem of great interest which was widely investigated.

Proof. The construction proceeds in stages $i = 0, 1, \ldots, s$. At stage 0 the only points present are the vertices of S. At each stage i, $i \geq 1$, a new split point is placed at the barycenter of each existing simplex, and the final construction is called the s-stage *uniform subdivision* of S. Let A_i be the set of simplices present at stage i; as there is nothing to be gained by splitting an empty simplex, every empty simplex must have a sibling containing a point, and so clearly $|A_i| = \min\{(d+1)^i, dn\}$. Note that all simplices in A_i have the same volume.

The weights w_i are assigned as follows: Initially, vertices $q_0, \ldots, q_d$ of S are assigned weights $w_0 = \cdots = w_d = -1$. At each stage $i \geq 1$, each new split point is given the smallest possible weight that ensures $\tilde{P} \supseteq P$, where $\tilde{P}$ is given by (9). Once a weight is assigned to a point, it is never changed again. In other words, for those points of B_{i+1} that already belonged to B_i, their weights at B_{i+1} are the same as their weights at B_i. This completes the system construction.

We first derive the runtime of the construction: The difficult step is finding the weight of the new split point, specified to be the smallest weight ensuring that $\tilde{P} \supseteq P$. Now, the intersection between a hyperplane and a simplex has at most d^2 vertices, which lie along some of the edges of the simplex. Decreasing the weight of the new split point causes the hyperplane to shift, and equivalently, causes each intersection point to move along an edge of the simplex. The hyperplane intersects P when one of these points enters P, and so it suffices to compute for each edge its intersection with P. This can be done via linear programming (Eisenstat, 2014) in time $O(d^2 n + e^{O(\sqrt{d \log d})})$ (Gärtner, 1995). As each stage has $\min\{(d+1)^i, dn\}$ simplices, the runtime follows, and we proceed to prove the remaining bounds of the Theorem.

Let $S' \in A_i$ be a simplex with vertices $q_{i_0}, \ldots, q_{i_d}$ and weights $w_{i_0}, \ldots, w_{i_d}$, respectively. Let $q' = (q_{i_0} + \cdots + q_{i_d})/(d+1)$ be the barycenter of S'. By Theorem 3.2, if q' is assigned weight $w_{\text{avg}} = (w_{i_0} + \cdots w_{i_d})/(d+1)$, then $\tilde{P} \cap S'$ remains unchanged. Hence, the weight w' that will be assigned to q' by our construction will satisfy $w' \leq w_{\text{avg}}$. And therefore, at each stage, $\tilde{P}$ only shrinks. If at stage i a certain simplex $S' \in A_i$ satisfies $S' \cap P = \emptyset$, then at stage $i+1$ the barycenter of S' will be assigned weight $-\infty$, so that the interior of S' will lie completely outside of $\tilde{P}$.

Let us denote by $\tilde{P}_s$ the region $\tilde{P}$ produced by this construction after stage s. (See Figure 6 for an illustration in the plane.) We will now prove that, if s is made large enough, then $\tilde{P}_s$ approximates the given convex body P arbitrarily well, as stated in the theorem.

The *diameter* of a compact subset of $\mathbb{R}^d$ is the maximum distance between two points in the set. In particular, the diameter of a simplex is the largest distance between two vertices of the simplex.

Lemma A.2. *Let S' be a simplex with vertices*

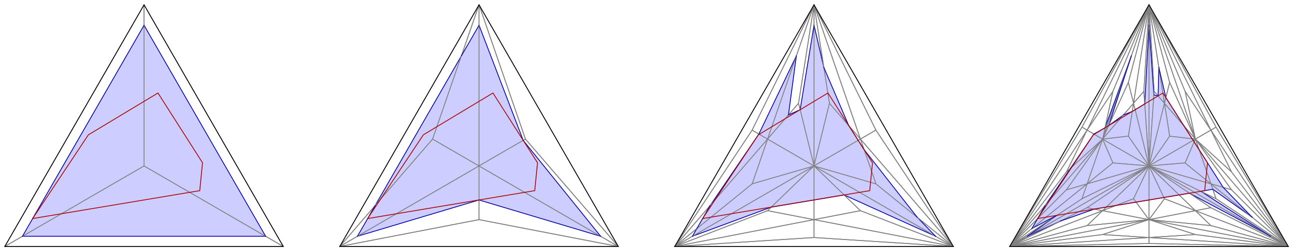


Figure 6: Four stages of the approximation of a given convex polygon in the plane.

$p_0, \dots, p_d$, let c be the diameter of S', and let q be the barycenter of S'. Then the distance between q and any vertex p_i is at most $cd/(d+1)$.

Proof. Fix $p_i = 0$ for concreteness. Then, under the constraints $\|p_j\|_2 \le c$ for $j \ne i$, the distance between q and p_i is maximized by the degenerate simplex that has $p_j = (c, 0, \dots 0)$ for all $j \ne i$, which yields the claimed distance. □

Lemma A.3. *Let S' be a simplex with diameter c. Let A be the collection of the $(d+1)^d$ simplices obtained by a d-stage uniform subdivision of S'. Then there are at least $(d+1)!$ simplices in A with diameter at most $cd/(d+1)$.*

Proof. By Lemma A.2, every simplex in A that contains at most one vertex of S' will have diameter at most $cd/(d+1)$. Each time a simplex S'' is subdivided into $d+1$ simplices by an interior point q, the new simplices share only d of their vertices with S''. Hence, at stage 1 of the subdivision of S', there are $d+1$ simplices that share only d vertices with S'; at stage 2, there are $(d+1)d$ simplices that share only $d-1$ vertices with S'; and so on, until at stage d there are $(d+1)d \cdots 2 = (d+1)!$ simplices that share only one vertex with S'. □

Recall that A_i denotes the collection of simplices present in the i-stage uniform subdivision of S.

Lemma A.4. *Let k, z be integers, and set $s = zkd$. Then at most a $\left(z(1-e^{-d})^k\right)$-fraction of the simplices in A_s have diameter larger than $(d/(d+1))^z$.*

Proof. By repeated application of Lemma A.3. After kd stages, at most an α-fraction of the simplices in A_{kd} have diameter larger than $d/(d+1)$, for $\alpha = \left(1 - \frac{(d+1)!}{(d+1)^d}\right)^k$. All the other simplices have diameter at most $d/(d+1)$. Of the latter simplices, after kd more stages, at most an α-fraction of their descendants have diameter larger than $(d/(d+1))^2$. Hence, in A_{2kd}, the fraction of simplices with diameter larger than $(d/(d+1))^2$ is at most $\alpha + (1-\alpha)\alpha < 2\alpha$. And so on. In A_{zkd}, the fraction of simplices with diameter larger than $(d/(d+1))^z$ is at most $z\alpha$. Since $(d+1)!/(d+1)^d > e^{-d}$ for all d, the lemma follows. □

We can now complete the proof of Theorem A.1. Given ε, let $\rho = \varepsilon/(2\sqrt{2}d^2)$. Choose z minimally so that $(d/(d+1))^z \le \rho$, and then choose k minimally so that $z(1-e^{-d})^k \le \varepsilon/2$. Let $s = zkd$. (Hence, we have $s \le c^d \ln^2(1/\varepsilon)$ for some c.) Let Z_1 be the region surrounding P that is at distance at most ρ from P, and let Z_2 be the union of all the simplices in A_s with diameter larger than ρ. By the choice of s, every point in $\tilde{P}_s \setminus P$ belongs to $Z_1 \cup Z_2$. Let us bound each of $\mathrm{vol}(Z_1)$ and $\mathrm{vol}(Z_2)$.

As $\rho \to 0$ (keeping P fixed) we have $\mathrm{vol}(Z_1) \le (1+o(1))\rho\,\mathrm{surf}(P)$, where surf denotes the $(d-1)$-dimensional surface volume.

Furthermore, P and S are both convex with $P \subseteq S$, so $\mathrm{surf}(P) \le \mathrm{surf}(S)$. Since $S = S_d$ where $S_d \subset \mathbb{R}^d$ is a regular simplex of unit side-length, we have $\mathrm{vol}(S_d) = \sqrt{d+1}/(d!\sqrt{2^d})$ and $\mathrm{surf}(S_d) = (d+1)\,\mathrm{vol}(S_{d-1}) \approx \sqrt{2}d^2\,\mathrm{vol}(S_d)$. Hence, by the choice of ρ, we have $\mathrm{vol}(Z_1) \le (\varepsilon/2)\,\mathrm{vol}(S)$. By Lemma A.4, we also have $\mathrm{vol}(Z_2) \le (\varepsilon/2)\,\mathrm{vol}(S)$. Hence, $\mathrm{vol}(\tilde{P}_s \setminus P) \le \varepsilon\,\mathrm{vol}(S)$, and the first item follows.

For the second item, by construction $P \subseteq \tilde{P}$. Now given a parameter $\varepsilon > 0$, apply the first part of the theorem with $\varepsilon' = \mathrm{vol}(B_\varepsilon)/(2\,\mathrm{vol}(S))$, where $\mathrm{vol}(B_\varepsilon) \approx \varepsilon^d \pi^{d/2}/(d/2)!$ is the volume of a d-dimensional ball of radius ε. (A calculation shows that $\varepsilon' \ge \varepsilon^d$, so it suffices to take $s = (c')^d \ln^2(1/\varepsilon)$ for an appropriate constant c'.) Suppose for a contradiction that there exists a point $p \in \tilde{P}_s^{(-\varepsilon)}$ that is outside of P. Then the ball $B = B_\varepsilon(p)$ is contained in $\tilde{P}$. But since P is convex, more than half of B is outside of P. Hence, $\mathrm{vol}(\tilde{P}_s \setminus P) > \mathrm{vol}(B)/2 = \varepsilon'\,\mathrm{vol}(S)$, contradicting the first part of the theorem. This implies that $\tilde{P}_s^{(-\varepsilon)} \subseteq P$. Finally, $P \subset \tilde{P}$ implies that $P^{(-\varepsilon)} \subset \tilde{P}^{(-\varepsilon)}$, concluding the second item and the proof of Theorem A.1. □

B NBCS MSE Function Approximation

In this section we show how to approximate any concave function $f(x)$ using a B_t NBCS. Our objective is to reduce the mean square error between the target function $f(x)$ and its piecewise linear approximation $\tilde{f}_t(x) = w \cdot \phi_t(x)$:

$$\mathrm{err}_t(x) = f(x) - \tilde{f}_t(x)$$
$$L(w,t) = \int_{\mathbf{V}} ||\,\mathrm{err}_t(x)||^2 dV$$

where $dV = dx_1 dx_2 \ldots$. For simplicity we will add a split point inside each simplex at stage $t+1$, while retaining all the same weights from stage t.

Lemma B.1. *Given a NBCS of rank t with a set of weights that minimizes $L(w,t)$, then for any choice of a new coordinate q_{t+1} inside simplex S', there exists a weight w_{t+1} such that:*

$$L(w, t+1) \leq L(w,t). \tag{10}$$

Proof. Let

$$\phi_t(x) = (\alpha_0, \ldots, \alpha_t),$$
$$\phi_{t+1}(x) = (\beta_0, \ldots, \beta_{t+1}),$$
$$\phi_t(q_{d+1}) = (\gamma_0, \ldots, \gamma_d).$$

As show in Section 3.2 $\alpha_i = \beta_i + \beta_{d+1}\gamma_i$, so it follows that,

$$\begin{aligned}
\mathrm{err}_{t+1}(x) &= \int_{\mathbf{V}} ||f(x) - \sum_{i=0}^{t+1} w_i\beta_i(x)||^2 dV \\
&= \int_{\mathbf{V}} ||f(x) - \sum_{i=0}^{t} w_i\beta_i(x) - w_{t+1}\beta_{t+1}(x)||^2 dV \\
&= \int_{\mathbf{V}} ||f(x) - \sum_{i=0}^{t} w_i\alpha_i(x) \\
&\qquad + \beta_{t+1}(x)\sum_{i=0}^{t} w_i\gamma_i - w_{t+1}\beta_{t+1}(x)||^2 dV \\
&= \int_{\mathbf{V}} ||f(x) - \sum_{i=0}^{t} w_i\alpha_i(x) \\
&\qquad - \beta_{t+1}(x)(w_{t+1} - \sum_{i=0}^{t} w_i\gamma_i)||^2 dV \\
&= \int_{\mathbf{V}} ||\,\mathrm{err}_t(x) - \beta_{t+1}(x)(w_{t+1} - \sum_{i=0}^{t} w_i\gamma_i)||^2 dV.
\end{aligned} \tag{11}$$

This above equation implies the following:

1. If point $x \notin S'$ then $\beta_{t+1}(x) = 0$ and the contribution of x remains the same as in the previous step. We will therefore only integrate over S'.
2. If we assign the weight $w_{t+1} = \sum_{i=0}^{t} w_i\gamma_i$ then $L(w, t+1) = L(w,t)$.
3. If
$$\sum_{i=0}^{t} w_i\gamma_i < w_{t+1} < \sum_{i=0}^{t} w_i\gamma_i + 2\frac{\mathrm{err}_t(x)}{\beta_{t+1}(x)} \forall x \in S' \tag{12}$$
then $L(w, t+1) < L(w,t)$ and the algorithm strictly reduces the objective function at step $t+1$.

In order to find the optimal weight w_{t+1}, we will differentiate with respect to w_{t+1}:

$$\begin{aligned}
&\frac{\partial L}{\partial w_{t+1}} = 0 \\
&\implies \int_{\mathbf{V}} -2err_t(x)\beta_{t+1}(x) + 2\beta_{t+1}^2(x)(w_{t+1} - \sum_{i=0}^{t} w_i\gamma_i) = 0 \\
&\implies w_{t+1} = \sum_{i=0}^{t} w_i\gamma_i + \int_{S'} \frac{\mathrm{err}_t(x)}{\beta_{t+1}(x)} dV.
\end{aligned} \tag{13}$$

This equation can be interpreted as follows: the new weight is the linear estimation of point q_{t+1}, since $\sum_{i=0}^{t} w_i\gamma_i = \tilde{f}(q_{t+1})$, plus a weighted sum over the rest of the points of their deviation from the previous estimated model weighted in some sence by their distance from point q_{t+1}. □

Lemma B.1 establishes that any new split point added to the hierarchical structure of NBCS can potentially reduce the target function. But it does not establish which candidate split point is the best, nor does it bound the number of times this procedure must be repeated in order to reduce the target function beneath an error estimate of ϵ. In order to find an upper bound on the number of splits necessary to achieve a given error estimate ϵ , and in order to describe a procedure to find the best split points, we need to assume additional properties on $f(\cdot)$ — specifically, continuity and concavity — and also make the trivial stipulation that we choose $w_{t+1} = f(q_{t+1})$; that is, the weight of point q_t must be the value of the function at that point.

Theorem B.2. *Let $f(\cdot)$ be a concave function and and let $\varepsilon > 0$ be a given number, then there exists a B_t NBCS and a corresponding set of weights w, such that the integrated mean square error $L(w,t) \leq \varepsilon$ (for $t = O(\ln\frac{1}{\epsilon})$).*

Proof. We shall prove this theorem by means of a constructive proof. The proposed construction proceeds in stages $i = 0, 1, \ldots, t$. At stage 0 the only points present are the vertices of the original d-dimensional simplex to which all data points are confined. At step zero the weight of coordinate q_i is given by $w_i = f(q_i)$. Let S_0 be the simplex $S_0 = \{x, y|y = w \cdot \phi_0(x)\}$ (Note , this a d dimensional simplex in the $d+1$ space). Likewise let the manifold $F = \{x, y \quad |y = f(x)\}$. Next, we perform the shearing transformation $f'(x) = f(x) - w \cdot \phi_0(x)$. Now the function $f'(x)$ evaluated at the vertices of S_0 is zero, and all subsequent stages will produce concave functions with this property. As the volume under the curve remains the same under the shearing transformation, and the concavity property is unaffected, then $f'(x)$ is a concave function defined inside S_0. Next, create the simplex S'_0 which is both parallel to S_0 and is also tangent to the manifold F at some point p' (see figure 7 for an illustration).

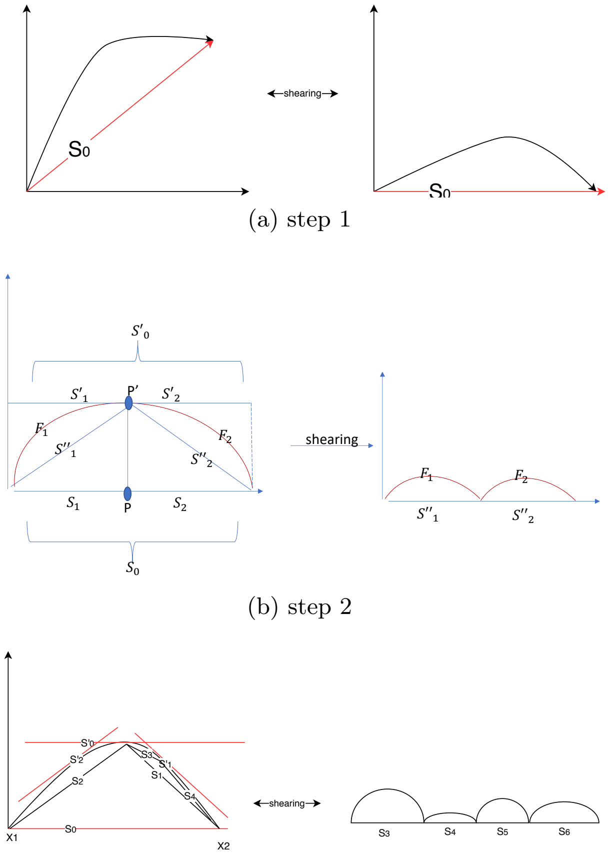
The concavity property implies that $\forall y \in S'_0 \quad y \geq f(x)$. We will say that manifold A is above B if $\forall y_1 \in A, y_2 \in B \quad y1 \geq y2$. Project the point p' back to S_0, and call the projection point p. The point p is the choice of our new knot (split point). The point p' splits S'_0 into $d+1$ simplices, let S'_k be the simplex where the point p' substitutes the vertex q'_k. Likewise the p split S_0 into $d+1$ simplices, let S_k be the simplex where the point p substitutes the vertex q_k. We call 2 simplices properly parallel if they share the same x coordinates and only their y coordinate differs in a constant value. Thus, we have constructed $d+1$ properly parallel simplices ($\{S_k, S'_k\}$). If we take the vertices of 2 properly parallel simplices, we can construct from the vertices a V-polytope which is hyper-parallelogram in the $d+1$ space. We have divided our space to $d+1$ such hyper-parallelograms termed Π_k, where each contains different parts of the manifold F the simplex S_0 and S'_0 and which all meet at line $p - p'$. $\forall k \quad F_k \in \Pi_k, S_k \in \Pi_k, S'_k \in \Pi_k$. Take the point p' and construct $d+1$ simplices where p' substitutes on of S_0 vertices, let S''_k be the simplex where the point p' substitues the vertex q_k. The integrated area under S''_k is $\frac{\text{vol}(\Pi_k)}{d}$, also we note that because of the concavity F_k is above S''_k and coincide with the vertices of S''_k. From the above we conclude that:

$$\text{vol}(S''_k) \leq \text{vol}(F_k) \leq \text{vol}(\Pi_k) \tag{14}$$

The integrated error therefore between F and our linear approximation is $\frac{d-1}{d}$ from the error in the previous stage.

For our next step we will perform the shearing transformation $f'(x) = f(x) - \sum w_i \phi_1(x)$, which results in $d+1$ new segments. These divide $f(\cdot)$ into $d+1$ different parts (see figure 7 for an illustration). After applying the transformation, each S''_k is a simplex taking the place of S_0 in the induction step, while $f'(x)$ is a concave function which evaluates to zero on the vertices and represents the integrated error as did $f(x)$ did in the previous stage. This allows the process to be inductively repeated. It follows that the error at stage $t+1$ holds $L(w, t+1) \leq \frac{d-1}{d} L(w, t) \leq (\frac{d-1}{d})^t L(w, 0)$. For any given positive number ε the error at stage $t \geq c \ln(\varepsilon)$ for some constant c is smaller than ε. □

Remark. As an aside, we note that another application for NBCS is multivariate numeric integration. Numeric integration for very large dimensions is difficult for precisely the same reason that piecewise linear function approximation is difficult in the multivariate case. In this case we do not only evaluate f by $\sum w_i \phi(x)$ but also its integral by the area under curve. We call this the *generalized Archimedes integration process*, which is reminiscent of Archimedes integration process for finding the area of a parabola by triangulation, except that this method generalizes the procedure to any concave function and to multiple dimensions. To the best of our knowledge, there is no known extension of Archimedes integration method to a general concave function in the multivariate case.



(a) step 1

(b) step 2

(c) step 3

Figure 7: Steps in function approximation method

C Hybrid PAC-compression bounds

In this section, we present a hybrid compression bound used in the derivation of Theorem 5.2.

General theory. It will be convenient to present our results in generality and then specialize. Our notation and terminology (*compression scheme*, etc) will be in line with Hanneke and Kontorovich (2019). Let P be a distribution on $\mathcal{Z}$. We write $Z_{[n]} = (Z_1, \dots, Z_n) \sim P^n$ and, for $f \in [0,1]^{\mathcal{Z}}$,

$$R(f,P) := \mathop{\mathbb{E}}_{Z \sim P} f(Z), \qquad \hat{R}(f, Z_{[n]}) := \frac{1}{n}\sum_{i=1}^n f(Z_i).$$

We write

$$\Delta_n(f) = \Delta_n(f, P, Z_{[n]}) := |R(f,P) - \hat{R}(f, Z_{[n]})|$$

and our main object of interest will be

$$\bar{\Delta}_n(\mathcal{F}) := \sup_{f \in \mathcal{F}} \Delta_n(f, P, Z_{[n]}), \tag{15}$$

for $\mathcal{F} \subset [0,1]^{\mathcal{Z}}$.

The catch is that $\mathcal{F}$ may itself be random, determined by the $Z_{[n]}$ via a compression scheme. For a fixed $k \in \mathbb{N}$, consider a fixed mapping $\rho : \mathcal{Z}^k \mapsto f \in \mathbb{R}^{\mathcal{Z}}$. In words, ρ maps k-tuples over $\mathcal{Z}$ into real-valued functions over $\mathcal{Z}$ and as such, is a *reconstruction* function in a sample compression scheme. Denote by $\mathcal{F}_\rho(Z_{[n]})$ the (random) collection of all functions constructable by ρ on a given $Z_{[n]}$:

$$\mathcal{F}_\rho(Z_{[n]}) = \left\{ \rho(Z_I) : I \in \binom{[n]}{k} \right\}, \tag{16}$$

where $\binom{[n]}{k}$ is the set of all k-subsets of $[n]$, and Z_I is the restriction of $Z_{[n]}$ to the index set I. [2]

A trivial application of the union bound yields

$$\mathbb{P}\left(\bar{\Delta}_n(\mathcal{F}_\rho(Z_{[n]})) \ge \varepsilon\right) \le \binom{n}{k} \max_{I \in \binom{[n]}{k}} \mathbb{P}\left(\Delta_n(\rho(Z_I)) \ge \varepsilon\right).$$

The key observation is that, conditioned on Z_I, the function $\rho(Z_I)$ becomes deterministic and independent of Z_J, where $J := [n] \setminus I$. Thus,

$$\mathbb{P}\left(\Delta_n(\rho(Z_I)) \ge \varepsilon\right) = \mathop{\mathbb{E}}_{Z_I}\left[\mathbb{P}\left(\Delta_n(\rho(Z_I)) \ge \varepsilon \,|\, Z_I\right)\right].$$

Conditional on Z_I, we have, for $f = \rho(Z_I)$,

$$\begin{aligned}
\Delta_n(f, P, Z_{[n]}) &= |R(f,P) - \hat{R}(f, Z_{[n]})| \\
&= \left| \frac{1}{n}\sum_{i=1}^n \mathbb{E} f(Z) - f(Z_i) \right| \\
&\le \left| \frac{1}{n}\sum_{i \in J} \mathbb{E} f(Z) - f(Z_i) \right| \\
&+ \left| \frac{1}{n}\sum_{i \in I} \mathbb{E} f(Z) - f(Z_i) \right| \\
&\le \frac{1}{n-k}\left| \sum_{i \in J} \mathbb{E} f(Z) - f(Z_i) \right| \\
&+ \frac{k}{n}\max_{i \in I} |\mathbb{E} f - f(Z_i)| \\
&= \Delta_{n-k}(f, P, Z_J) \\
&+ \frac{k}{n}\max_{i \in I} |\mathbb{E} f(Z) - f(Z_i)| \\
&\le \Delta_{n-k}(f, P, Z_J) + \frac{k}{n}.
\end{aligned}$$

Thus (conditioned on Z_I),

$$\Delta_n(\rho(Z_I), P, Z_{[n]}) \le \Delta_{n-k}(\rho(Z_I), P, Z_{[n]\setminus I}) + \frac{k}{n}$$

holds with probability 1.

We now state the main result of this section:

Theorem C.1.

$$\mathbb{P}\left(\bar{\Delta}_n(\mathcal{F}_\rho(Z_{[n]})) \ge \frac{k}{n} + \varepsilon\right) \le \binom{n}{k}\mathbb{P}\left(\bar{\Delta}_{n-k}(\rho(Z_{[k]}), P, Z_{[n]\setminus[k]}) \ge \varepsilon\right).$$

To apply this result to examples of interest, let us compute the right-hand side of the bound for some function classes.

Example: VC classes. In our first example, suppose that ρ maps k-tuples of $\mathcal{Z}$ to binary concept classes — which might well be different for each k-tuple — of VC-dimension at most d. More precisely, we take $\mathcal{Z} = \mathcal{X} \times \{0,1\}$, where $\mathcal{X}$ is an instance space. Let $\mathcal{H} = \mathcal{H}_z \subseteq \{0,1\}^{\mathcal{X}}$ be a concept class defined by the k-tuple $z \in \mathcal{Z}^k$, with VC-dimension d. Define $\mathcal{F} \subseteq \{0,1\}^{\mathcal{Z}}$ to be its associated loss class, where we write $\mathcal{F}^{\text{FIX}}$ to distinguish fixed (i.e., deterministic) vs. random function classes:

$$\mathcal{F}^{\text{FIX}} = \left\{ f_h : (x,y) \mapsto \mathbb{1}_{\{h(x) \ne y\}}; h \in \mathcal{H} \right\}.$$

We call this setting a *hybrid* (k,d) VC sample-compression scheme. It is well-known (see, e.g., (Anthony and Bartlett, 1999, Theorem 4.9)) that

$$\mathbb{E}[\bar{\Delta}_n(\mathcal{F}^{\text{FIX}})] \le c\sqrt{d/n}, \tag{17}$$

[2] We consider, for concreteness, permutation and repetition-invariant compression schemes; the extension to general ones is straightforward. The only requisite change consists of replacing $I \in \binom{[n]}{k}$ with $I \in [n]^k$ in (16).

where $c > 0$ is a universal constant. Further, $\bar{\Delta}_n(\mathcal{F}^{\mathrm{FIX}})$ is known to be concentrated about its mean (see, e.g., (Mohri et al., 2012, Theorem 3.1)):

$$\mathbb{P}\left(\bar{\Delta}_n(\mathcal{F}^{\mathrm{FIX}}) \geq \mathbb{E}[\bar{\Delta}_n(\mathcal{F}^{\mathrm{FIX}})] + \varepsilon\right) \leq \exp(-2n\varepsilon^2). \quad (18)$$

Combining Theorem C.1 with (17), and (18), we conclude:

Corollary C.1.1. *In a hybrid* (k, d) *VC sample compression scheme, on a sample of size* n*, a learner's sample error* $\widehat{\mathrm{err}}(\hat{h}_n)$ *and generalization error* $\mathrm{err}(\hat{h}_n)$ *satisfy*

$$\mathrm{err}(\hat{h}_n) \leq \widehat{\mathrm{err}}(\hat{h}_n) + c\sqrt{\frac{d}{n-k}} + \sqrt{\frac{\log[\delta^{-1}\binom{n}{k}]}{2(n-k)}} + \frac{k}{n}$$

with probability at least $1 - \delta$.

Example: Margin classes. Here, we take $\mathcal{X}$ to be an abstract set, $\mathcal{Y} = \{-1, 1\}$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and define

$$\tilde{\mathcal{H}} = \{h_w : \mathcal{X} \ni x \mapsto w \cdot \Psi(x); \|w\| \leq 1\},$$

where $\Psi(x) = \Psi_z(x)$ is a map from $\mathcal{X}$ to $\mathbb{R}^N$ determined by some k-tuple $z \in \mathcal{Z}$, with $\|\Psi_z(\cdot)\| \leq 1$. Associate to $\tilde{\mathcal{H}}$ the γ-margin loss class

$$\mathcal{F}_\gamma^{\mathrm{FIX}} = \left\{f_h : \mathcal{X} \times \{-1, 1\} \ni (x, y) \mapsto \Phi_\gamma(yh(x)); h \in \tilde{\mathcal{H}}\right\},$$

where $\Phi_\gamma(t) = \min(0, \max(1, 1 - t/\gamma))$. We refer to this setting as a *hybrid* (k, γ) margin sample compression scheme. It is a standard fact (see, e.g., (Mohri et al., 2012, Theorem 4.4)) that

$$\mathbb{P}\left(\bar{\Delta}_n(\mathcal{F}_\gamma^{\mathrm{FIX}}) \geq \frac{2}{\gamma\sqrt{n}} + \varepsilon\right) \leq \exp(-2n\varepsilon^2). \quad (19)$$

Combining Theorem C.1, (19), and a standard stratification argument (see (Mohri et al., 2012, Theorem 4.5)), we obtain the following result. Fix a map $\rho : \mathcal{Z}^k \to \Psi(\cdot)$. Given a sample $Z_{[n]} = (X_i, Y_i)_{i \in [n]}$ drawn iid, the learner chooses some k examples to define the random mapping $\Psi_z : \mathcal{X} \to \mathbb{R}^N$. Having mapped the sample to R^N, he runs SVM and obtains a hyperplane w.

Corollary C.1.2. *With probability at least* $1 - \delta$*, we have*

$$\begin{aligned}
\mathop{\mathbb{E}}_{(X,Y)}[\mathrm{sgn}(Yw \cdot \Psi(X)) \leq 0 \mid Z_{[n]}] &\leq \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - Y_i w \cdot \Psi(X_i)) \\
&+ \frac{4}{\|w\|\sqrt{n-k}} \\
&+ \sqrt{\frac{\log\log_2 \frac{2}{\|w\|}}{n-k}} \\
&+ \sqrt{\frac{\log(2\binom{n}{k}/\delta)}{2(n-k)}} + \frac{k}{n}.
\end{aligned}$$