Differential Geometric Regularization for Supervised Learning of Classifiers Supplemental Materials

A. Proof of Theorem 1

Proof. For $\boldsymbol{f}: \mathbb{R}^N \to \Delta^{L-1} \subset \mathbb{R}^L$,

$$\{r_j = r_j(\boldsymbol{x}) = (0, \ldots, \overset{j}{1}, \ldots, 0, f_j^1, \ldots, f_j^L) : j = 1, \ldots N\}$$

is a basis of the tangent space $T_{\boldsymbol{x}}\mathrm{gr}(\boldsymbol{f})$ to $\mathrm{gr}(\boldsymbol{f})$. Here $f_j^i = \partial_{x^j} f^i$. Let $\{e_i\}$ be an orthonormal frame of $T_{\boldsymbol{x}}\mathrm{gr}(\boldsymbol{f})$. We have

$$e_i = B_i^j r_j$$

for some invertible matrix B_i^j.

Define the metric matrix g for the basis $\{r_j\}$ by

$$g = (g_{kj}) \text{ with } g_{kj} = r_k \cdot r_j = \delta_{kj} + f_k^i f_j^i.$$

Then

$$\begin{aligned}\delta_{ij} = e_i \cdot e_j = B_i^k B_j^t r_k \cdot r_t = B_i^k B_j^t g_{kt} \\ \Rightarrow I = (BB^T)g \Rightarrow BB^T = g^{-1}.\end{aligned}$$

Thus BB^T is computable in terms of derivatives of $\boldsymbol{f}$.

Let $D_u w$ be the $\mathbb{R}^{N+L}$ directional derivative of w in the direction u. Then

$$\begin{aligned}\text{Tr II} &= P^\nu D_{e_i} e_i = P^\nu D_{B_i^j r_j} B_i^k r_k = B_i^j P^\nu D_{r_j} B_i^k r_k \\ &= B_i^j P^\nu [(D_{r_j} B_i^k) r_k] + B_i^j B_i^k D_{r_j} r_k \\ &= B_i^j B_i^k P^\nu D_{r_j} r_k \\ &= (g^{-1})^{jk} P^\nu D_{r_j} r_k,\end{aligned}$$

since $P^\nu r_k = 0$.

We have

$$\begin{aligned}r_k &= (0, \ldots, 1, \ldots, f_k^1(x^1, \ldots, x^N), \ldots, f_k^L(x^1, \ldots, x^N)) \\ &= \partial_k^{\mathbb{R}^{N+L}} + \sum_{i=1}^{L} f_k^i \partial_{N+i}^{\mathbb{R}^{N+L}},\end{aligned}$$

so in particular, $\partial_\ell^{\mathbb{R}^{N+L}} r_k = 0$ if $\ell > N$. Thus

$$D_{r_j} r_k = (0, \ldots, \overset{N}{0}, f_{kj}^1, \ldots, f_{kj}^L).$$

So far, we have

$$\text{Tr II} = (g^{-1})^{jk} P^\nu (0, \ldots, \overset{N}{0}, f_{kj}^1, \ldots, f_{kj}^L).$$

Since g is given in terms of derivatives of $\boldsymbol{f}$, we need to write $P^\nu = I - P^T$ in terms of derivatives of $\boldsymbol{f}$. For any $u \in \mathbb{R}^{N+L}$, we have

$$\begin{aligned}P^T u &= (P^T u \cdot e_i) e_i = (u \cdot B_i^j r_j) B_i^k r_k \\ &= B_i^j B_i^k (u \cdot r_j) r_k \\ &= (g^{-1})^{jk} (u \cdot r_j) r_k.\end{aligned}$$

Thus

$$\begin{aligned}&\text{Tr II} && (1)\\ =\ & (g^{-1})^{jk} P^\nu (0, \ldots, \overset{N}{0}, f_{kj}^1, \ldots, f_{kj}^L) && (2)\\ =\ & (g^{-1})^{jk} (0, \ldots, \overset{N}{0}, f_{kj}^1, \ldots, f_{kj}^L) \\ & - P^T [(g^{-1})^{jk} (0, \ldots, \overset{N}{0}, f_{kj}^1, \ldots, f_{kj}^L)] \\ =\ & (g^{-1})^{jk} (0, \ldots, \overset{N}{0}, f_{kj}^1, \ldots, f_{kj}^L) \\ & - (g^{-1})^{jk} [(g^{-1})^{rs} (0, \ldots, \overset{N}{0}, f_{rs}^1, \ldots, f_{rs}^L) \cdot r_j] r_k \\ =\ & (g^{-1})^{jk} (0, \ldots, \overset{N}{0}, f_{kj}^1, \ldots, f_{kj}^L) \\ & - (g^{-1})^{jk} (g^{-1})^{rs} \left(f_{rs}^i f_j^i\right) r_k \\ =\ & (g^{-1})^{ij} \Big(0, \ldots, -(g^{-1})^{rs} \overset{j}{f_{rs}^a} f_i^a, \ldots, 0, && (3)\\ & f_{ji}^1 - (g^{-1})^{rs} f_{rs}^a f_i^a f_j^1, \ldots, f_{ji}^L - (g^{-1})^{rs} f_{rs}^a f_i^a f_j^L \Big),\end{aligned}$$

after a relabeling of indices. Therefore, the last L component of Tr II are given by

$$\begin{aligned}\text{Tr II}^L &= (g^{-1})^{ij} \Big(f_{ji}^1 - (g^{-1})^{rs} f_{rs}^a f_i^a f_j^1, \ldots, \\ & \qquad f_{ji}^L - (g^{-1})^{rs} f_{rs}^a f_i^a f_j^L \Big).\end{aligned}$$

□

B. An Easy Example with Bayes Consistency

We now give an example with a loss function that enables easy Bayes consistency proof under some mild initialization assumption. Related notation is summerized in §C.

For ease of reading, we change the notation for empirical penalty $\mathcal{P}_{\mathcal{T}_m}$ in this supplemental material to $\mathcal{P}_D$, i.e., $\mathcal{P} = \mathcal{P}_D + \lambda\mathcal{P}_G$. $\mathcal{P}_D$ measures the deviation of $\mathrm{gr}(\boldsymbol{f})$ from the mapped training points, a natural geometric distance penalty term is an L^2 distance in $\mathbb{R}^L$ from $\boldsymbol{f}(\boldsymbol{x})$ to the averaged $\boldsymbol{z}$ component of the k-nearest training points:

$$\mathcal{P}_D(\boldsymbol{f}) = R_{D,\mathcal{T}_m,k}(\boldsymbol{f}) = \int_{\mathcal{X}} d^2\left(\boldsymbol{f}(\boldsymbol{x}), \frac{1}{k}\sum_{i=1}^{k}\tilde{\boldsymbol{z}}_i\right) d\boldsymbol{x}, \tag{4}$$

where d is the Euclidean distance in $\mathbb{R}^L$, $\tilde{\boldsymbol{z}}_i$ is the vector of the last L components of $(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{z}}_i) = (\tilde{\boldsymbol{x}}_i^1, \ldots, \tilde{\boldsymbol{x}}_i^N, \tilde{\boldsymbol{z}}_i^1, \ldots, \tilde{\boldsymbol{z}}_i^L)$, with $\tilde{\boldsymbol{x}}_i$ the i^{th} nearest neighbor of $\boldsymbol{x}$ in $\mathcal{T}_m$, and $d\boldsymbol{x}$ is the Lebesgue measure. The gradient vector field is

$$\nabla(R_{D,\mathcal{T}_m,k})_{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) = \frac{2}{k}\sum_{i=1}^{k}(\boldsymbol{f}(\boldsymbol{x}) - \tilde{\boldsymbol{z}}_i).$$

However, $\nabla(R_{D,\mathcal{T}_m,k})_{\boldsymbol{f}}$ is discontinuous on the set $\mathcal{D}$ of points $\boldsymbol{x}$ such that $\boldsymbol{x}$ has equidistant training points among its k nearest neighbors. $\mathcal{D}$ is the union of $(N-1)$-dimensional hyperplanes in $\mathcal{X}$, so $\mathcal{D}$ has measure zero. Such points will necessarily exist unless the last L components of the mapped training points are all 1 or all 0. To rectify this, we can smooth out $\nabla(R_{D,\mathcal{T}_m,k})_{\boldsymbol{f}}$ to a vector field

$$V_{D,\boldsymbol{f},\phi} = \frac{2\phi(\boldsymbol{x})}{k}\sum_{i=1}^{k}(\boldsymbol{f}(\boldsymbol{x}) - \tilde{\boldsymbol{z}}_i). \tag{5}$$

Here $\phi(\boldsymbol{x})$ is a smooth damping function close to the singular function $\delta_{\mathcal{D}}$, which has $\delta_{\mathcal{D}}(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \mathcal{D}$ and $\delta_{\mathcal{D}}(\boldsymbol{x}) = 1$ for $\boldsymbol{x} \notin \mathcal{D}$. Outside any open neighborhood of $\mathcal{D}$, $\nabla R_{D,\mathcal{T}_m,k} = V_{D,\boldsymbol{f},\phi}$ for ϕ close enough to $\delta_{\mathcal{D}}$.

Recall the geometric penalty from the submission, i.e., $\mathcal{P}_G(\boldsymbol{f}) = \int_{\mathrm{gr}(\boldsymbol{f})} \mathrm{dvol}$, with the geometric gradient vector field being $V_{G,\boldsymbol{f}} = -\mathrm{Tr}\ \mathrm{II}^L$.

Then the gradient vector field $V_{tot,\lambda,m,\boldsymbol{f},\phi}$ of this example penalty $\mathcal{P}$ is,

$$\begin{aligned} V_{tot,\lambda,m,\boldsymbol{f},\phi} &= \nabla\mathcal{P}_f = V_{D,\boldsymbol{f},\phi} + \lambda V_{G,\boldsymbol{f}} \\ &= \frac{2\phi(\boldsymbol{x})}{k}\sum_{i=1}^{k}(\boldsymbol{f}(\boldsymbol{x}) - \tilde{\boldsymbol{z}}_i) - \lambda\mathrm{Tr}\ \mathrm{II}^L \end{aligned} \tag{6}$$

B.1. Consistency analysis

For a training set $\mathcal{T}_m$, we let $\boldsymbol{f}_{\mathcal{T}_m} = (f^1_{\mathcal{T}_m}, \ldots, f^L_{\mathcal{T}_m})$ be the class probability estimator given by our approach. We denote the generalization risk of the corresponding plug-in classifier $h_{\boldsymbol{f}_{\mathcal{T}_m}}$ by $R_P(\boldsymbol{f}_{\mathcal{T}_m}) = \mathbb{E}_P[\mathbb{1}_{h_{\boldsymbol{f}_{\mathcal{T}_m}}(\boldsymbol{x})\neq y}]$. The Bayes risk is defined by $R_P^* = \inf_{h:\mathcal{X}\to\mathcal{Y}} R_P(h) = \mathbb{E}_P[\mathbb{1}_{h_{\boldsymbol{\eta}}(\boldsymbol{x})\neq y}]$. Our algorithm is Bayes consistent if $\lim_{m\to\infty} R_P(\boldsymbol{f}_{\mathcal{T}_m}) = R_P^*$ holds in probability for all distributions P on $\mathcal{X}\times\mathcal{Y}$. Usually, gradient flow methods are applied to a convex functional, so that a flow line approaches the unique global minimum. If the domain of the functional is an infinite dimensional manifold of (e.g. smooth) functions, we always assume that flow lines exist and that the actual minimum exists in this manifold.

Because our functionals are not convex, we can only hope to prove Bayes consistency for the set of initial estimators in the stable manifold of a stable fixed point (or sink) of the vector field (Guckenheimer & Worfolk, 1993). Recall that a stable fixed point $\boldsymbol{f}_0$ has a maximal open neighborhood, the stable manifold $\mathcal{S}_{\boldsymbol{f}_0}$, on which flow lines tend towards $\boldsymbol{f}_0$. For the manifold $\mathcal{M}$, the stable manifold for a stable critical point of the vector field $V_{tot,\lambda,m,\boldsymbol{f},\phi}$ is infinite dimensional.

The proof of Bayes consistency for multiclass (including binary) classification follows these steps:

Step 1: $\lim_{\lambda\to 0} R^*_{D,P,\lambda} = 0$.

Step 2: $\lim_{n\to\infty} R_{D,P}(\boldsymbol{f}_n) = 0 \Rightarrow \lim_{n\to\infty} R_P(\boldsymbol{f}_n) = R_P^*$.

Step 3: For all $\boldsymbol{f} \in \mathcal{M} = \mathrm{Maps}(\mathcal{X}, \Delta^{L-1})$, $|R_{D,\mathcal{T}_m}(\boldsymbol{f}) - R_{D,P}(\boldsymbol{f})| \xrightarrow{m\to\infty} 0$ in probability.

Proofs of these steps are provided in following subsections. For the notation see §C. $R^*_{D,P,\lambda}$ is the minimum of the regularized D risk $R_{D,P,\lambda}(\boldsymbol{f})$ for $\boldsymbol{f}$: $R_{D,P,\lambda}(\boldsymbol{f}) = R_{D,P}(\boldsymbol{f}) + \lambda\mathcal{P}_G(\boldsymbol{f})$, with $R_{D,P}(\boldsymbol{f}) = \int_{\mathcal{X}} d^2(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x}))d\boldsymbol{x}$ the D-risk. Also, $R_{D,\mathcal{T}_m,\lambda}(\boldsymbol{f}) = R_{D,\mathcal{T}_m}(\boldsymbol{f}) + \lambda\mathcal{P}_G(\boldsymbol{f})$, with $R_{D,\mathcal{T}_m}(\boldsymbol{f}) = \int_{\mathcal{X}} d^2\left(\boldsymbol{f}(\boldsymbol{x}), \frac{1}{k}\sum_{i=1}^{k}\tilde{\boldsymbol{z}}_i\right) d\boldsymbol{x}$ the empirical D-risk.

Theorem 2 (Bayes Consistency)**.** *Let m be the size of the training data set. Let $\boldsymbol{f}_{1,\lambda,m} \in \mathcal{S}_{\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}}$, the stable manifold for the global minimum $\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}$ of $R_{D,\mathcal{T}_m,\lambda}$, and let $\boldsymbol{f}_{n,\lambda,m,\phi}$ be a sequence of functions on the flow line of $V_{tot,\lambda,m,\boldsymbol{f},\phi}$ starting with $\boldsymbol{f}_{1,\lambda,m}$ with the flow time $t_n \to \infty$ as $n \to \infty$. Then $R_P(\boldsymbol{f}_{n,\lambda,m,\phi}) \xrightarrow[\lambda\to 0,\phi\to\delta_{\mathcal{D}}]{m,n\to\infty} R_P^*$ in probability for all distributions P on $\mathcal{X}\times\mathcal{Y}$, if $k/m \to 0$ as $m \to \infty$.*

Proof. In the notation of §C, if $\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}$ is a global minimum for $R_{D,\mathcal{T}_m,\lambda}$, then outside of $\mathcal{D}$, $\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}$ is the limit of critical points for the negative flow of $V_{tot,\lambda,m,\boldsymbol{f},\phi}$ as $\phi \to \delta_{\mathcal{D}}$. To see this, fix an ϵ_i neighborhood $\mathcal{D}_{\epsilon_i}$ of $\mathcal{D}$. For a sequence $\phi_j \to \delta_{\mathcal{D}}$, $V_{tot,\lambda,m,f,\phi_j}$ is independent of $j \geq j(\epsilon_i)$ on $\mathcal{X}\setminus\mathcal{D}_{\epsilon_i}$, so we find a function $\boldsymbol{f}_i$, a critical point of $V_{tot,\lambda,m,\boldsymbol{f},\phi_{j(\epsilon_i)}}$, equal to $\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}$ on $\mathcal{X}\setminus\mathcal{D}_{\epsilon_i}$. Since any $\boldsymbol{x} \notin \mathcal{D}$ lies outside some $\mathcal{D}_{\epsilon_i}$, the sequence $\boldsymbol{f}_i$

converges at $\boldsymbol{x}$ if we let $\epsilon_i \to 0$. Thus we can ignore the choice of ϕ in our proof, and drop ϕ from the notation.

For our algorithm, for fixed λ, m, we have as above $\lim_{n\to\infty} \boldsymbol{f}_{n,\lambda,m} = \boldsymbol{f}_{D,\mathcal{T}_m,\lambda}$, so

$$\lim_{n\to\infty} R_{D,\mathcal{T}_m,\lambda}(\boldsymbol{f}_{n,\lambda,m}) = R_{D,\mathcal{T}_m,\lambda}(\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}),$$

for $\boldsymbol{f}_1 \in \mathcal{S}_{\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}}$. By Step 2, it suffices to show $R_{D,P}(\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}) \xrightarrow[\lambda\to 0]{m\to\infty} 0$. In probability, we have $\forall \delta > 0, \exists m > 0$ such that

$$\begin{aligned}
0 &\leq R_{D,P}(\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}) \\
&\leq R_{D,P}(\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}) + \lambda\mathcal{P}_G(\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}) \\
&\leq R_{D,\mathcal{T}_m}(\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}) + \lambda\mathcal{P}_G(\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}) + \frac{\delta}{3} \quad \text{(Step 3)} \\
&= R_{D,\mathcal{T}_m,\lambda}(\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}) + \frac{\delta}{3} \\
&\leq R_{D,\mathcal{T}_m,\lambda}(\boldsymbol{f}_{D,P,\lambda}) + \frac{\delta}{3} \quad \text{(minimality of } \boldsymbol{f}_{D,\mathcal{T}_m,\lambda}) \\
&= R_{D,\mathcal{T}_m}(\boldsymbol{f}_{D,P,\lambda}) + \lambda\mathcal{P}_G(\boldsymbol{f}_{D,P,\lambda}) + \frac{\delta}{3} \\
&\leq R_{D,P}(\boldsymbol{f}_{D,P,\lambda}) + \lambda\mathcal{P}_G(\boldsymbol{f}_{D,P,\lambda}) + \frac{2\delta}{3} \text{ (Step 3)} \\
&= R_{D,P,\lambda}(\boldsymbol{f}_{D,P,\lambda}) + \frac{2\delta}{3} = R^*_{D,P,\lambda} + \frac{2\delta}{3} \\
&\leq \delta, \quad \text{(Step 1)}
\end{aligned}$$

for λ close to zero. □

B.2. Step 1

Lemma 1. *(Step 1)* $\lim_{\lambda\to 0} R^*_{D,P,\lambda} = 0$.

Proof. After the smoothing procedure in §3.1 for the distance penalty term, the function $R_{D,P,\lambda} : \mathcal{M} \to \mathbb{R}$ is continuous in the Fréchet topology on $\mathcal{M}$. We check that the functions $R_{D,P,\lambda} : \mathcal{M} \to \mathbb{R}$ are equicontinuous in λ: for fixed $\boldsymbol{f}_0 \in \mathcal{M}$ and $\epsilon > 0$, there exists $\delta = \delta(\boldsymbol{f}_0, \epsilon)$ such that $|\lambda - \lambda'| < \delta \Rightarrow |R_{D,P,\lambda}(\boldsymbol{f}_0) - R_{D,P,\lambda'}(\boldsymbol{f}_0)| < \epsilon$. This is immediate:

$$|R_{D,P,\lambda}(\boldsymbol{f}_0) - R_{D,P,\lambda'}(\boldsymbol{f}_0)| = |(\lambda - \lambda')\mathcal{P}_G(\boldsymbol{f}_0)| < \epsilon,$$

if $\delta < \epsilon/|\mathcal{P}_G(\boldsymbol{f}_0)|$. It is standard that the infimum $\inf R_\lambda$ of an equicontinuous family of functions is continuous in λ, so $\lim_{\lambda\to 0} R^*_{D,P,\lambda} = R^*_{D,P,\lambda=0} = R_{D,P}(\boldsymbol{\eta}) = 0$. □

B.3. Step 2

We assume that the class probability function $\boldsymbol{\eta}(\boldsymbol{x}) : \mathbb{R}^N \to \mathbb{R}^L$ is smooth, and that the marginal distribution $\mu(\boldsymbol{x})$ is continuous. We also let μ denote the corresponding measure on $\mathcal{X}$.

Denote: $h_{\boldsymbol{f}}(\boldsymbol{x}) = \text{argmax}\{f^\ell(\boldsymbol{x}), \ell \in \mathcal{Y}\}$, and,

$$\mathbb{1}_{h_{\boldsymbol{f}}(\boldsymbol{x})\neq y} = \begin{cases} 1, & h_{\boldsymbol{f}}(\boldsymbol{x}) \neq y, \\ 0, & h_{\boldsymbol{f}}(\boldsymbol{x}) = y. \end{cases}$$

Lemma 2. *(Step 2 for a subsequence)*

$$\lim_{n\to\infty} R_{D,P}(\boldsymbol{f}_n) = 0 \Rightarrow \lim_{i\to\infty} R_P(\boldsymbol{f}_{n_i}) = R^*_P$$

for some subsequence $\{\boldsymbol{f}_{n_i}\}_{i=1}^\infty$ *of* $\{\boldsymbol{f}_n\}$.

Proof. The left hand side of the Lemma is

$$\int_{\mathcal{X}} d^2(\boldsymbol{f}_n(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x}))d\boldsymbol{x} \to 0,$$

which is equivalent to

$$\int_{\mathcal{X}} d^2(\boldsymbol{f}_n(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x}))\mu(\boldsymbol{x})d\boldsymbol{x} \to 0, \tag{7}$$

since $\mathcal{X}$ is compact and μ is continuous. Therefore, it suffices to show

$$\begin{aligned}
&\int_{\mathcal{X}} d^2(\boldsymbol{f}_n(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x}))\mu(\boldsymbol{x})d\boldsymbol{x} \to 0 \qquad (8) \\
&\quad\Longrightarrow\quad \mathbb{E}_P[\mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}] \to \mathbb{E}_P[\mathbb{1}_{h_{\boldsymbol{\eta}}(\boldsymbol{x})\neq y}].
\end{aligned}$$

We recall that L^2 convergence implies pointwise convergence a.e, so (7) implies that a subsequence of $\boldsymbol{f}_n$, also denoted $\boldsymbol{f}_n$, has $\boldsymbol{f}_n \to \boldsymbol{\eta}(\boldsymbol{x})$ pointwise a.e. on $\mathcal{X}$. (By our assumption on $\mu(\boldsymbol{x})$, these statements hold for either μ or Lebesgue measure.) By Egorov's theorem, for any $\epsilon > 0$, there exists a set $B_\epsilon \subset \mathcal{X}$ with $\mu(B_\epsilon) < \epsilon$ such that $\boldsymbol{f}_n \to \boldsymbol{\eta}(\boldsymbol{x})$ uniformly on $\mathcal{X} \setminus B_\epsilon$.

Fix $\delta > 0$ and set

$$\begin{aligned}
Z_\delta \;=\; &\{\boldsymbol{x} \in \mathcal{X} : \#\{\text{argmax}_{\ell\in\mathcal{Y}}\, \eta^\ell(\boldsymbol{x})\} = 1, \\
&|\max_{\ell\in\mathcal{Y}} \eta^\ell(\boldsymbol{x}) - \text{submax}_{\ell\in\mathcal{Y}}\, \eta^\ell(\boldsymbol{x})| < \delta\},
\end{aligned}$$

where $\text{submax}_{\ell\in\mathcal{Y}}$ denotes the second largest element in $\{\eta^1(\boldsymbol{x}), \ldots, \eta^L(\boldsymbol{x})\}$. For the moment, assume that $Z_0 = \{\boldsymbol{x} \in \mathcal{X} : \#\{\text{argmax}_{\ell\in\mathcal{Y}}\, \eta^\ell(\boldsymbol{x})\} > 1\}$ has $\mu(Z_0) = 0$.

It follows easily[1] that $\mu(Z_\delta) \to 0$ as $\delta \to 0$. On $\mathcal{X} \setminus (Z_\delta \cup B_\epsilon)$, we have $\mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y} = \mathbb{1}_{h_{\boldsymbol{\eta}}(\boldsymbol{x})\neq y}$ for $n > N_\delta$. Thus

$$\mathbb{E}_P[\mathbb{1}_{\mathcal{X}\setminus(Z_\delta\cup B_\epsilon)}\mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}] = \mathbb{E}_P[\mathbb{1}_{\mathcal{X}\setminus(Z_\delta\cup B_\epsilon)}\mathbb{1}_{h_{\boldsymbol{\eta}}(\boldsymbol{x})\neq y}].$$

[1] Let A_k be sets with $A_{k+1} \subset A_k$ and with $\mu(\cap_{k=1}^\infty A_k) = 0$. If $\mu(A_k) \not\to 0$, then there exists a subsequence, also called A_k, with $\mu(A_k) > K > 0$ for some K. We claim $\mu(\cap A_k) \geq K$, a contradiction. For the claim, let $Z = \cap A_k$. If $\mu(Z) \geq \mu(A_k)$ for all k, we are done. If not, since the A_k are nested, we can replace A_k by a set, also called A_k, of measure K and such that the new A_k are still nested. For the relabeled $Z = \cap A_k$, $Z \subset A_k$ for all k, and we may assume $\mu(Z) < K$. Thus there exists $Z' \subset A_1$ with $Z' \cap Z = \emptyset$ and $\mu(Z') > 0$. Since $\mu(A_i) = K$, we must have $A_i \cap Z' \neq \emptyset$ for all i. Thus $\cap A_i$ is strictly larger than Z, a contradiction. In summary, the claim must hold, so we get a contradiction to assuming $\mu(A_k) \not\to 0$.

(Here $\mathbb{1}_A$ is the characteristic function of a set A.)

As $\delta \to 0$,

$$\mathbb{E}_P[\mathbb{1}_{\mathcal{X}\setminus(Z_\delta \cup B_\epsilon)} \mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}] \to \mathbb{E}_P[\mathbb{1}_{\mathcal{X}\setminus B_\epsilon} \mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}].$$

and similarly for $\boldsymbol{f}_n$ replaced by $\boldsymbol{\eta}(\boldsymbol{x})$. During this process, N_δ presumably goes to ∞, but that precisely means

$$\lim_{n\to\infty} \mathbb{E}_P[\mathbb{1}_{\mathcal{X}\setminus B_\epsilon} \mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}] = \mathbb{E}_P[\mathbb{1}_{\mathcal{X}\setminus B_\epsilon} \mathbb{1}_{h_{\boldsymbol{\eta}}(\boldsymbol{x})\neq y}].$$

Since

$$\left|\mathbb{E}_P[\mathbb{1}_{\mathcal{X}\setminus B_\epsilon} \mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}] - \mathbb{E}_P[\mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}]\right| < \epsilon,$$

and similarly for $\boldsymbol{\eta}(\boldsymbol{x})$, we get

$$\begin{aligned}
&\left|\lim_{n\to\infty} \mathbb{E}_P[\mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}] - \mathbb{E}_P[\mathbb{1}_{h_{\boldsymbol{\eta}}(\boldsymbol{x})\neq y}]\right| \\
\leq\ & \left|\lim_{n\to\infty} \mathbb{E}_P[\mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}] - \lim_{n\to\infty} \mathbb{E}_P[\mathbb{1}_{\mathcal{X}\setminus B_\epsilon} \mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}]\right| \\
&\quad + \left|\lim_{n\to\infty} \mathbb{E}_P[\mathbb{1}_{\mathcal{X}\setminus B_\epsilon} \mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}] - \mathbb{E}_P[\mathbb{1}_{\mathcal{X}\setminus B_\epsilon} \mathbb{1}_{h_{\boldsymbol{\eta}}(\boldsymbol{x})\neq y}]\right| \\
&\quad + \left|\lim_{n\to\infty} \mathbb{E}_P[\mathbb{1}_{\mathcal{X}\setminus B_\epsilon} \mathbb{1}_{h_{\boldsymbol{\eta}}(\boldsymbol{x})\neq y}] - \mathbb{E}_P[\mathbb{1}_{h_{\boldsymbol{\eta}}(\boldsymbol{x})\neq y}]\right| \\
\leq\ & 3\epsilon.
\end{aligned}$$

(Strictly speaking, $\lim_{n\to\infty} \mathbb{E}_P[\mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y}]$ is first lim sup and then lim inf to show that the limit exists.) Since ϵ is arbitrary, the proof is complete if $\mu(Z_0) = 0$.

If $\mu(Z_0) > 0$, we rerun the proof with $\mathcal{X}$ replaced by Z_0. As above, $\boldsymbol{f}_n|_{Z_0}$ converges uniformly to $\boldsymbol{\eta}(\boldsymbol{x})$ off a set of measure ϵ. The argument above, without the set Z_δ, gives

$$\int_{Z_0} \mathbb{1}_{h_{\boldsymbol{f}_n}(\boldsymbol{x})\neq y} \mu(\boldsymbol{x}) d\boldsymbol{x} \to \int_{Z_0} \mathbb{1}_{h_{\boldsymbol{\eta}}(\boldsymbol{x})\neq y} \mu(\boldsymbol{x}) d\boldsymbol{x}.$$

We then proceed with the proof above on $\mathcal{X} \setminus Z_0$. □

Corollary 3. *(Step 2 in general) For our algorithm,* $\lim_{n\to\infty} R_{D,P}(\boldsymbol{f}_{n,\lambda,m}) = 0 \Rightarrow \lim_{i\to\infty} R_P(\boldsymbol{f}_{n,\lambda,m}) = R_P^*$.

Proof. Choose $\boldsymbol{f}_{1,\lambda,m}$ as in Theorem 2. Since $V_{tot,\lambda,m,\boldsymbol{f}_{n,\lambda,m}}$ has pointwise length going to zero as $n \to \infty$, $\{\boldsymbol{f}_{n,\lambda,m}(\boldsymbol{x})\}$ is a Cauchy sequence for all $\boldsymbol{x}$. This implies that $\boldsymbol{f}_{n,\lambda,m}$, and not just a subsequence, converges pointwise to $\boldsymbol{\eta}$. □

B.4. Step 3

Lemma 4. *(Step 3) If $k \to \infty$ and $k/m \to 0$ as $m \to \infty$, then for $\boldsymbol{f} \in \mathrm{Maps}(\mathcal{X}, \Delta^{L-1})$,*

$$|R_{D,\mathcal{T}_m}(\boldsymbol{f}) - R_{D,P}(\boldsymbol{f})| \xrightarrow{m\to\infty} 0 \ \textit{in probability},$$

for all distributions P that generate $\mathcal{T}_m$.

Proof. Since $R_{D,P}(\boldsymbol{f})$ is a constant for fixed $\boldsymbol{f}$ and P, convergence in probability will follow from weak convergence, i.e.,

$$\mathbb{E}_{\mathcal{T}_m}[|R_{D,\mathcal{T}_m}(\boldsymbol{f}) - R_{D,P}(\boldsymbol{f})|] \xrightarrow{m\to\infty} 0.$$

We have

$$\begin{aligned}
&|R_{D,\mathcal{T}_m}(\boldsymbol{f}) - R_{D,P}(\boldsymbol{f})| \\
=\ & \left|\int_{\mathcal{X}} \left[d^2\left(\boldsymbol{f}(\boldsymbol{x}), \frac{1}{k}\sum_{i=1}^k \tilde{\boldsymbol{z}}_i\right) - d^2(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x}))\right] d\boldsymbol{x}\right| \\
\leq\ & \int_{\mathcal{X}} \left|d^2\left(\boldsymbol{f}(\boldsymbol{x}), \frac{1}{k}\sum_{i=1}^k \tilde{\boldsymbol{z}}_i\right) - d^2(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x}))\right| d\boldsymbol{x}.
\end{aligned}$$

Set $\boldsymbol{a} = \boldsymbol{f}(\boldsymbol{x}) - \frac{1}{k}\sum_{i=1}^k \tilde{\boldsymbol{z}}_i$, $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{\eta}(\boldsymbol{x})$. Then

$$\begin{aligned}
&\left|\|\boldsymbol{a}\|_2^2 - \|\boldsymbol{b}\|_2^2\right| \\
=\ & \left|\sum_{\ell=1}^L a_\ell^2 - \sum_{\ell=1}^L b_\ell^2\right| = \left|\sum_{\ell=1}^L (a_\ell^2 - b_\ell^2)\right| \\
\leq\ & \sum_{\ell=1}^L |a_\ell^2 - b_\ell^2| \leq 2\sum_{\ell=1}^L |a_\ell - b_\ell| \max\{|a_\ell|, |b_\ell|\} \\
\leq\ & 2\sum_{\ell=1}^L |a_\ell - b_\ell|,
\end{aligned}$$

since $f^\ell(\boldsymbol{x}), \frac{1}{k}\sum_i^k \tilde{z}_i^\ell, \eta^\ell(\boldsymbol{x}) \in [0,1]$. Therefore, it suffices to show that

$$\sum_{\ell=1}^L \mathbb{E}_{\mathcal{T}_m}\left[\int_{\mathcal{X}} \left|((f^\ell(\boldsymbol{x}) - \frac{1}{k}\sum_i^k \tilde{z}_i^\ell) - (f^\ell(\boldsymbol{x}) - \eta^\ell(\boldsymbol{x}))\right| d\boldsymbol{x}\right] \xrightarrow{m\to\infty} 0,$$

so the result follows if

$$\lim_{m\to\infty} \mathbb{E}_{\mathcal{T}_m, \boldsymbol{x}}\left[\left|\eta^\ell(\boldsymbol{x}) - \frac{1}{k}\sum_i^k \tilde{z}_i^\ell\right|\right] = 0 \text{ for all } \ell. \quad (9)$$

By Jensen's inequality $(\mathbb{E}[f])^2 \leq \mathbb{E}(f^2)$, (9) follows if

$$\lim_{m\to\infty} \mathbb{E}_{\mathcal{T}_m, \boldsymbol{x}}\left[\left(\eta^\ell(\boldsymbol{x}) - \frac{1}{k}\sum_i^k \tilde{z}_i^\ell\right)^2\right] = 0 \text{ for all } \ell. \quad (10)$$

Let $\eta_{k,m}^\ell(\boldsymbol{x}) = \frac{1}{k}\sum_i^k \tilde{z}_i^\ell$. Then $\eta_{k,m}^\ell$ is actually an estimate of the class probability $\eta^\ell(\boldsymbol{x})$ by the k-Nearest Neighbor rule. Following the proof of Stone's Theorem (Stone, 1977; Devroye et al., 1996), if $k \xrightarrow{m\to\infty} \infty$ and $k/m \xrightarrow{m\to\infty} 0$, (10) holds for all distributions P. □

C. Notation

$$
\begin{aligned}
h_{\boldsymbol{f}}(\boldsymbol{x}) = \operatorname{argmax}\{f^{\ell}(\boldsymbol{x}), \ell \in \mathcal{Y}\} &: \text{plug-in classifier of estimator } \boldsymbol{f} : \mathcal{X} \to \Delta^{L-1} \\
\mathbb{1}_{h_{\boldsymbol{f}}(\boldsymbol{x}) \neq y} &= \begin{cases} 1, & h_{\boldsymbol{f}}(\boldsymbol{x}) \neq y, \\ 0, & h_{\boldsymbol{f}}(\boldsymbol{x}) = y. \end{cases} \\
R_P(\boldsymbol{f}) = \mathbb{E}_P[\mathbb{1}_{h_{\boldsymbol{f}}(\boldsymbol{x}) \neq y}] &: \text{generalization risk for the estimator } \boldsymbol{f} \\
\boldsymbol{\eta}(\boldsymbol{x}) = (\eta^1(\boldsymbol{x}), \ldots, \eta^L(\boldsymbol{x})) &: \text{class probability function: } \eta^{\ell}(\boldsymbol{x}) = P(y = \ell | \boldsymbol{x}) \\
R_P^* = R_P(\boldsymbol{\eta}) &: \text{Bayes risk} \\
(D\text{-risk for our } \mathcal{P}_D)\ R_{D,P}(\boldsymbol{f}) &= \int_{\mathcal{X}} d^2(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{\eta}(\boldsymbol{x})) d\boldsymbol{x} \\
(\text{empirical } D\text{-risk}) R_{D,\mathcal{T}_m}(\boldsymbol{f}) = R_{D,\mathcal{T}_m,k}(\boldsymbol{f}) &= \int_{\mathcal{X}} d^2 \left(\boldsymbol{f}(\boldsymbol{x}), \frac{1}{k} \sum_{i=1}^{k} \tilde{\boldsymbol{z}}_i \right) d\boldsymbol{x} \\
&\quad \text{where } \tilde{\boldsymbol{z}}_i \text{ is the vector of the last } L \text{ components of} \\
&\quad (\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{z}}_i), \text{ with } \tilde{\boldsymbol{x}}_i \text{ the } i^{\text{th}} \text{ nearest neighbor of } \boldsymbol{x} \text{ in } \mathcal{T}_m \\
(\text{volume penalty term})\ \mathcal{P}_G(\boldsymbol{f}) &= \int_{\operatorname{gr}(\boldsymbol{f})} \operatorname{dvol} \\
R_{D,P,\lambda}(\boldsymbol{f}) = R_{D,P}(\boldsymbol{f}) + \lambda \mathcal{P}_G(\boldsymbol{f}) &: \text{regularized } D\text{-risk for estimator } \boldsymbol{f} \\
R_{D,\mathcal{T}_m,\lambda}(\boldsymbol{f}) = R_{D,\mathcal{T}_m}(\boldsymbol{f}) + \lambda \mathcal{P}_G(\boldsymbol{f}) &: \text{regularized empirical } D\text{-risk for estimator } \boldsymbol{f} \\
\boldsymbol{f}_{D,P,\lambda} &= \text{function attaining the global minimum for } R_{D,P,\lambda} \\
R_{D,P,\lambda}^* = R_{D,P,\lambda}(\boldsymbol{f}_{D,P,\lambda}) &: \text{minimum value for } R_{D,P,\lambda} \\
\boldsymbol{f}_{D,\mathcal{T}_m,\lambda} = \boldsymbol{f}_{D,\mathcal{T}_m,k,\lambda} &: \text{function attaining the global minimum for } R_{D,\mathcal{T}_m,\lambda}(\boldsymbol{f})
\end{aligned}
$$

Note that we assume $\boldsymbol{f}_{D,P,\lambda}$ and $\boldsymbol{f}_{D,\mathcal{T}_m,\lambda}$ exist.

References

Devroye, Luc, Györfi, László, and Lugosi, Gábor. *A probabilistic theory of pattern recognition*. Springer, 1996.

Guckenheimer, John and Worfolk, Patrick. Dynamical systems: some computational problems. In *Bifurcations and periodic orbits of vector fields (Montreal, PQ, 1992)*, volume 408 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pp. 241–277. Kluwer Acad. Publ., Dordrecht, 1993.

Stone, Charles. Consistent nonparametric regression. *Annals of Statistics*, pp. 595–620, 1977.