Notes on 'Remember the Curse of Dimensionality: The Case of Goodness-of-Fit Testing in Arbitrary Dimension'

Alden Green

February 13, 2019

Define the *worst-case risk* of a test ϕ (Borel-measurable function to $[0,1]$) to be

$$R_\epsilon^{(m)}(\phi; f_0; \mathcal{H}) = \mathbb{E}_{f_0}^{(m)} \phi + \sup_{\substack{f \in \mathcal{H} \\ \delta(f, f_0) \leq \epsilon}} \left\{ \mathbb{E}_f^{(m)} (1 - \phi) \right\}$$

where $\delta(f,g) = \int (f-g)^2$.

The *minimax risk* is then

$$R_\epsilon^{(m)}(f_0; \mathcal{H}) = \inf_\phi R_\epsilon^{(m)}(\phi; f_0; \mathcal{H}).$$

1 One-sample goodness of fit problem for Holder class

Let $\mathcal{H}_s^d(L)$ be the *Holder class* of functions $f : [0,1]^d \to \mathbb{R}$ such that

$$\left| f^{\lfloor s \rfloor}(x) - f^{\lfloor s \rfloor}(y) \right| \leq L \left\| x - y \right\|^{s - \lfloor s \rfloor} \qquad \text{(for all } x, y \in [0,1]^d\text{)}$$

and additionally

$$\left\| f^{(s')} \right\|_\infty \leq L, \quad \forall s' \in \{0, \cdots, \lfloor s \rfloor\}$$

Theorem 1 (One-sample lower bound over Holder densities)**.** *For the one-sample problem*

$$f_0 = \textit{uniform distribution over } [0,1]^d \textit{ known}$$

under known Holder regularity, there is a constant $c > 0$ depending only on (s, d, L) such that

$$\mathcal{R}_\epsilon^{(m)}(f_0; \mathcal{H}_s^d(L)) \geq 1/2 \textit{ for all } \epsilon \leq c m^{-\frac{2s}{4s+d}}$$

1.1 One-sample chi-squared test

Let the *one sample chi-squared statistic* be given by

$$\Gamma_\kappa^{one} = \sum_{k \in [\kappa]^d} \left| M_{k,\kappa} - m\kappa^{-d} \right|^2$$

where

$$M_{k,\kappa} = \# \left\{ x_i : x_i \in \left(\frac{k-1}{\kappa}, \frac{k}{\kappa} \right] \right\}$$

For now, we treat s as known, and set

$$\kappa = \kappa(s,d) := \left\lfloor m^{\frac{2}{4s+d}} \right\rfloor$$

Theorem 2 (One-sample chi-squared test). *In the one-sample problem under known Holder regularity, consider the chi-squared test $\phi_{\kappa,\tau} = \mathbb{I}\{\Gamma_\kappa^{one} > \tau\}$. There are constants c_1 depending on only (s,L) and c_2 depending on only (s,d,L) such that for $\tau = m + am\kappa^{-d/2}$ with $a \geq 1$*

$$R_\epsilon^{(m)}(\phi_{\kappa,\tau}; f_0, \mathcal{H}_s^d(L)) \leq c_1/a^2$$

for any $\epsilon \geq c_2 a m^{-\frac{2s}{4s+d}}$.

2 Two-Sample Testing

In the two sample case, we observe data $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$. For a given test (Borel measurable function of the data) ϕ, define the worst case risk to be

$$R_\epsilon^{(m,n)}(\phi; \mathcal{H}) = \sup_{f \in \mathcal{H}} \mathbb{E}_{f,f}(\phi) + \sup_{\substack{f,g \in \mathcal{H} \\ \delta(f,g) \geq \epsilon^2}} \mathbb{E}_{f,g}(1-\phi)$$

and the minimax risk to be

$$R_\epsilon^{(m,n)}(\mathcal{H}) = \inf_\phi R_\epsilon^{(m,n)}(\phi; \mathcal{H}).$$

It is intuitively clear that the two-sample problem is at least as hard as the one-sample problem.

Lemma 1 (Two-sample is harder than one-sample.). *For any class $\mathcal{H}$, pseudo-metric δ, $\epsilon > 0$, and $f_0 \in \mathcal{H}$,*

$$R_\epsilon^{(m)}(f_0; \mathcal{H}) \leq R_\epsilon^{(m,n)}(\mathcal{H})$$

Theorem 3 (Two-sample lower bound over Holder densities). *For the two-sample problem under known Holder regularity, there exists constant c depending only on (s,d,L) such that*

$$R_\epsilon^{(m,n)}(\mathcal{H}_s^d(L)) \geq 1/2$$

for any $\epsilon \leq c(m \wedge n)^{-\frac{2s}{4s+d}}$.

2.1 Two-sample chi-squared test.

Define the *two-sample chi-squared* test statistic

$$\Gamma_\kappa = \sum_{k \in [\kappa]^d} (M_{k,\kappa} - N_{k,\kappa})^2$$

where

$$M_{k,\kappa} = \# \left\{ x_i : x_i \in \left(\frac{k-1}{\kappa}, \frac{k}{\kappa} \right] \right\}, \quad N_{k,\kappa} = \# \left\{ y_i : y_i \in \left(\frac{k-1}{\kappa}, \frac{k}{\kappa} \right] \right\}$$

are bin counts, and for simplicity we let $m = n$ so no normalization is needed.

Theorem 4 (Two-sample chi-squared test.)**.** *For the two-sample testing problem under known Holder regularity, let $\phi_{k,\kappa} = \mathbb{I}(\Gamma_\kappa > \tau)$. There are constants c_1 depending only on L, and c_2 depending only on (s, d, L) such that for $\tau = 2m + am\kappa^{-d/2}$*

$$R_\epsilon^{(m,m)}(\phi_{k,\kappa}) \le c_1/a^2$$

whenever $\epsilon \ge ac_2 m^{-\frac{2s}{4s+d}}$.

3 Proofs

Proof of Theorem 1. Let $h : \mathbb{R}^d \to \mathbb{R}$ be infinitely differentiable with support on $[0,1]^d$ such that $\int h = 0$, $\int h^2 = 1$. Let $\kappa \ge 1$ be an integer, and for $j \in \mathbb{Z}^d$, define

$$h_{j,\kappa}(x) = \kappa^{d/2} h(\kappa x - j + 1)$$

supported on $[\frac{(j-1)}{\kappa}, \frac{j}{\kappa}]$. (Note that $\|h_{j,\kappa}\| = 1$)

For $\eta = (\eta_1, \ldots, \eta_{\kappa^d}) \in \{-1, +1\}^{\kappa^d}$ and $\rho > 0$ to be defined later, define

$$f_\eta = f_0 + \rho \sum_{j \in [\kappa]^d} \eta_j h_{j,\kappa}(x)$$

and note that since $\int h_{j,\kappa} = 0$, $\int f_\eta = 1$. Additionally, because the $h_{j,\kappa}$'s have disjoint support,

- If $\rho\kappa^{d/2} \|h\|_\infty \le 1$,
$$f_\eta \ge 0$$
- For $C := 4 \left\| h^{(\lfloor s \rfloor)} \right\|_\infty \vee 2 \left\| h^{(\lfloor s \rfloor + 1)} \right\|_\infty$, if $\rho\kappa^{d/2+s} C \le L$, then
$$h_{j,\kappa}^{(\lfloor s \rfloor)} \in \mathcal{H}_s^d(L)$$

To see the second point, for arbitrary $x, y \in [0,1]^d$ let $x \in [\frac{k-1}{\kappa}, \frac{k}{\kappa}]$, $y \in [\frac{l-1}{\kappa}, \frac{l}{\kappa}]$. Then

$$\begin{aligned}
\left|f_\eta^{(\lfloor s\rfloor)}(x) - f_\eta^{(\lfloor s\rfloor)}(y)\right| &\leq 2\rho\kappa^{d/2+\lfloor s\rfloor}\Big(\left|h^{(\lfloor s\rfloor)}(\kappa x - k + 1) - h^{(\lfloor s\rfloor)}(\kappa y - k + 1)\right| + \\
&\quad \left|h^{(\lfloor s\rfloor)}(\kappa x - l + 1) - h^{(\lfloor s\rfloor)}(\kappa y - k + 1)\right|\Big) \\
&\leq 2\rho\kappa^{d/2+\lfloor s\rfloor}\left(2\kappa \left\|h^{(\lfloor s\rfloor+1)}\right\|_\infty \|x-y\| \wedge 4\left\|h^{(\lfloor s\rfloor)}\right\|_\infty\right) \\
&\leq 2\rho\kappa^{d/2+\lfloor s\rfloor}\left(\left[2\left\|h^{(\lfloor s\rfloor+1)}\right\|_\infty \vee 4\left\|h^{(\lfloor s\rfloor)}\right\|_\infty\right][1 \wedge \kappa\|x-y\|]\right) \\
&\leq L\|x-y\|^{s-\lfloor s\rfloor}
\end{aligned}$$

where the step follows from $(1 \wedge u) \leq u^a$ for all $u > 0, 0 < a \leq 1$.

Take ρ small enough to satisfy the above conditions, let $\epsilon = \rho\kappa^{d/2}$. Note that

$$\|f_0 - f_\eta\|_2^2 \leq \rho^2 \sum_{j\in[\kappa]^d} \|h_{j,\kappa}\|_2^2 = \rho^2\kappa^d = \epsilon^2.$$

The minimax risk is lower bounded by the Bayes risk; in this case, consider the uniform prior distribution over $\left\{f_\eta : \eta \in \{-1,1\}^{[\kappa]^d}\right\}$. The Bayes risk is achieved by the likelihood ratio test $\{W > 1\}$, where

$$W = \frac{1}{2^{\kappa^d}} \sum_{\eta\in\{-1,1\}^{\kappa^d}} \prod_{i=1}^m f_\eta(x_i)$$

and its known that the risk of the likelihood ratio test is lower bounded by $1 - \sqrt{\frac{1}{2}\mathrm{Var}_{f_0}(W)}$.[1]

We can compute $\mathrm{Var}_{f_0}(W) \leq \exp\{(\rho^2 m)^2\kappa_d\}$ for $rho^2 m \leq 1$. Choosing $\kappa = \lfloor m^{2/(4s+d)}\rfloor$ and $\rho = cm^{-(2s+d)/(4s+d)}$, we have that the upper bound on $\mathrm{Var}_{f_0}(W)$ is 1, and so the minimax risk is at least 1/2. $\epsilon = \kappa^{d/2}\rho = cm^{\frac{-2s}{4s+d}}$. It can be verified that the choice of ρ satisfies the necessary conditions imposed above. $\square$

3.1 Proof of Theorem 4

Test error for chi-squared test over discrete distributions Let p and q be discrete distributions over $\mathcal{K}$, and let

$$T = \sum_{k\in\mathcal{K}} (M_k - N_k)^2$$

[1]Compare the χ^2-divergence and TV-distance, and use the lower bound on risk $R \geq 1/2 - 1/2\|P_0 - P_1\|_{TV}$.

where

$$M_k = \#\{A_i = k\},\ N_k = \#\{B_j = k\}$$

for $A_1, \ldots, A_m \sim p, B_1, \ldots, B_m \sim q$ independent.

Lemma 2 (Moment bounds for T)**.** *We have*

$$\begin{aligned}\mathbb{E}(T) &= 2m + m^2\langle (p-q)^2\rangle - m(\langle p^2\rangle + \langle q^2\rangle)\\ \mathrm{Var}(T) &= 2m^2\langle (p+q)^2\rangle + 4m^3(\langle (p+q)(p-q)^2\rangle + 2\langle pq\rangle\langle (p-q^2)\rangle)\end{aligned}$$

Corollary 1. *Consider testing within the class of probability distributions r on $\mathcal{K}$ such that $\|r\|_\infty \leq \eta$ for some $\eta > 0$. There are universal constants ν_1 and ν_2 such that for any $a > 0$, the test with rejection region*

$$\{T - 2m \geq am\sqrt{\eta}\}$$

has size at most ν_1/a^2 and power at least $1-\nu_1/a^2$ against alternatives satisfying

$$\|p-q\|^2 \geq \nu_2(a\sqrt{\eta} \vee a^2\eta \vee \eta)/m$$

Approximation Error.

Lemma 3 (Approximation error of binning)**.** *For a continuous function $h : [0,1]^d \to \mathbb{R}$ and an integer $\kappa \geq 2$, define*

$$W_\kappa[h] = \sum_{k\in[\kappa^d]} \kappa^d \int_{H_k} h(x)dx\mathbb{I}(H_k)$$

Then there are constants $b_1, b_2 > 0$ depending only on (s, d, L) such that

$$\|W_k[h]\|_2 \geq b_1 \|h\|_2 - b_2\kappa^{-s},\ \forall h \in H_s^d(L)$$

Proof of Lemma 3.

Lemma 4 (Taylor expansion)**.** *Fix any $h \in \mathcal{H}_s^d(L)$ and any $x_0 \in [0,1]^d$. Let u denote the $\lfloor s \rfloor$-th order Taylor expansion of h around x_0. Then, there is a constant L' depending only on (s, d, L) such that*

$$|h(x) - u(x)| \leq L' \|x - x_0\|^s \qquad (\forall u \in [0,1]^d)$$

Fix $h \in \mathcal{H}_s^d(L)$, and let u_j be the $\lfloor s \rfloor$-th order Taylor expansion around $(j-1)r/\kappa$, for some $j = (j_1, \ldots, j_d)$, $r \geq 1$ is an integer. Then, define

$$u = \sum_j u_j \mathbb{I}_{\widetilde{\mathcal{H}}_j},\ \widetilde{\mathcal{H}}_j = \prod_{l=1}^m \widetilde{\mathcal{H}}_{j_l},\ \widetilde{\mathcal{H}}_{j_l} = \left(\frac{(j_l-1)r}{\kappa}, \frac{j_l r}{\kappa}\right]$$

Then,

$$|u(x) - h(x)| \leq L' \left(\frac{2\sqrt{d}r}{\kappa}\right)^s =: c_1\kappa^{-s}$$

and as a result

$$\begin{aligned}\|W_\kappa[h]\|_2 &\geq \|W_\kappa[u]\|_2 - \|W_\kappa[h] - W_\kappa[u]\|_2 \\ &\geq \|W_\kappa[u]\|_2 - \|h-u\|_2 \\ &\geq \|W_\kappa[u]\|_2 - c_1\kappa^{-s}\end{aligned}$$

Lemma 5 (Averaging of polynomial preserves norm)**.** *Let $\mathcal{P}_m^d$ denote the class of polynomials on $\mathbb{R}^d$ of degree at most m. For a given partition $\mathcal{Q} = (Q_i)$, define*

$$W_{\mathcal{Q}}[v] = \sum_{Q_i \in \mathcal{Q}} \frac{1}{|Q_i|}\left(\int_{Q_j} v(x)dx\right)\mathbb{I}_{Q_i}.$$

Then there exists constant $c_2 > 0$ depending only on (d, m) such that, when $\max_j \operatorname{diam}(Q_j) \leq c_2$*,*

$$\|W_{\mathcal{Q}}[v]\| \geq c_2 \|v\|_2\,, \textit{ for all } v \in \mathcal{P}_m^d.$$

□

Proof of Theorem 4. Define

$$H_k = \left(\frac{k-1}{\kappa}, \frac{k}{\kappa}\right],\ p_k = \int_{H_k} f(x)dx,\ q_k = \int_{H_k} g(x)dx$$

and let $p = (p_k)_{k\in[\kappa]^d}$, $q = (q_k)_{k\in[\kappa]^d}$.

Application of Lemma 3 to $f - g$ yields

$$\|p-q\|^2 \geq \kappa^{-d}\left(b_1^* \|f-g\|_2 - b_2^*\kappa^{-s}\right)^2.$$

Recall that $\|g - f\|_2 \geq \epsilon \geq c_1 a m^{-\frac{2s}{4s+d}}$. Since $\kappa^{-s} = m^{-\frac{2s}{4s+d}}$, an appropriate choice of c_1, independent of κ, yields

$$\|p-q\|^2 \geq \kappa^{-d}\epsilon^2$$

□

4 Relevant Citations

1. A distribution free version of the smirnov two-sample test in the p-variate case (Bickel 69)
2. Optimal kernel choice for large-scale two-sample tests. (Gretton 12)
3. Permutation tests for equality of distributions in high-dimensional settings (Hall 02)