Estimation Theory- Final Project

Maximum Likelihood Estimation for Learning Populations of Parameters

Main Results

Summary of the problem

- N coins with bias p_i , where the biases are from a distribution P
- X_i , i=1,2,...N represents the number of heads among t tosses for coins with bias p_i , hence X_i is a binomial distribution
- Aim is to estimate P using MLE, that is – $P_{MLE} = \arg\min_{Q \in D} KL(\mathbf{h}^{obs}, E_Q[\mathbf{h}])$
- Maximise the product of expectations of the fraction of coins with a particular number of heads

Wasserstein-1 Distance

- Measure of the accuracy of the estimated distribution
- Given by – $W_1(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} \int^1_{x=0} \int^1_{y=0} |x-y| d\gamma(x,y)$
- It can be shown that the Wasserstein-1 distance between an optimal solution to the MLE (P_{MLE}) and the true underlying distribution has a bound

Some definitions

- $O(\ f(x)\ )$ - in the worst case this quantity is of the order of $f(x)$
- $\Omega(\ f(x)\ )$ - in the best case this quantity is of the order of $f(x)$
- $\Theta(\ f(x)\ )$ - for large x, this quantity is bounded by $k_1 f(x)$ and $k_2 f(x)$, $k_1 < k_2$

Why MLE?

The following table enlists the orders of bounds on the W-1 distance or the EMD.

Estimators	Bound on EMD
Empirical	$\Theta\left(\frac{1}{\sqrt{t}}\right) + \Theta\left(\frac{1}{\sqrt{N}}\right)$ in all regimes
Moment Matching (Tian et al., 2017)	• $\Theta\left(\frac{1}{t}\right)$ when $t = \mathcal{O}(\log N)$ • Fails when $t = \Omega(\log N)$
MLE (this paper)	• $\Theta\left(\frac{1}{t}\right)$ when $t = \mathcal{O}(\log N)$ • $\Theta\left(\frac{1}{\sqrt{t}\ \log N}\right)$, when $t \in \left[\Omega(\log N), \mathcal{O}\left(N^{2/9-\epsilon}\right)\right]$

Why MLE?

- **Empricial Estimator :** simple ‘plug-in’ estimator which estimates the biases
- Has two error terms – one arising from the errors due to estimating the biases
- The second error term is due to estimating the underlying CDF
- In the sparse regime – large error due to first error term ($O(1/\sqrt{t})$), irrespective of how large N is
- Good estimator in the large regime

Why MLE?

- **Local Moment Matching Estimator :**
- expressing the population moments (i.e., the expected values of powers of the random variable under consideration) as functions of the parameters of interest
- Works well when $t = O(\log N)$ (same as MLE)
- Fails when $t = \Omega(\log N)$

Why MLE?

- It can be shown that MLE obtains optimal error bounds in sparse, medium and large regimes
- No hyperparameter tuning needed
- Therefore we use MLE

W_1 distance bound in small sample regime

- $t = O(\log N)$
- $W_1(P^*, P_{MLE}) <= O_\delta(1/t)$
- $\theta_\delta(1/t)$ is information theoretically optimal
- $F : \mathbf{X} \rightarrow f(\mathbf{X})$, then

  $\inf_f \sup_P E[W_1(P,f(X))] > 1/4t$ // explain

W_1 distance bound in medium sample regime

- t > Ω(logN)
- There exists Ɛ >0 s.t for t ε [Ω(logN) , O (N^(2/9 – Ɛ)], with probability atleast 1 – 2δ – W_1 (P^* , P_{MLE}) <= O_δ(1/ √tlogN)
- Θ (1/ √tlogN) lower bound on minimax rate for estimating P^*
- **MLE is minimax optimal upto a constant factor in both regimes**

- Explain minimax optimal

Proof Sketches

- Bounds on Wasserstein-1 distance
- W_1 (P,Q) = sup $_{f\,\varepsilon\,\mathrm{Lip}(1)}\int^1_{x=0} f(x)(p(x) - q(x))dx$ on P,Q supported on [0,1]
- Lip(1) -denotes Lipschitz functions
- Can be approximated by $\tilde{f}(x) = \Sigma^t_{j=0}\, b_j \binom{t}{j} x^j (1-x)^{(t-j)}$
- $\int^1_{x=0} f(x)(p(x) - q(x))dx = \int^1_{x=0} (f(x)-(\tilde{f}(x))(p(x) - q(x))dx + \int^1_{x=0} \tilde{f}(x)(p(x) - q(x))dx$, which can be bounded by
  - $2\|\, f - \tilde{f}\, \|_\infty + \int^1_{x=0} \Sigma^t_{j=0}\, b_j \binom{t}{j} x^j (1-x)^{(t-j)} (p(x) - q(x))dx$

    $= 2\|\, f - \tilde{f}\, \|_\infty + \Sigma^t_{j=0}\, b_j\, (E_P[h_j] - E_Q[h_j])$

    where $\|\, f - \tilde{f}\, \|_\infty = \max|\, f(x) - \tilde{f}(x)|$ is the approximate error

Proof Sketches

Therefore, the W_1 distance can be bounded as:

$W_1 (P^* , P_{MLE}) <= \sup \{2|| f – ƒ ||_\infty + \Sigma^t_{j=0} b_j (E_P[h_j] – h_j^{obs}) + \Sigma^t_{j=0} b_j (h_j^{obs} – E_{pmle}[h_j])$

- First term : approximation error for using Berstein polynomials
- Second term : error due to sampling
- Third term : error in matching fingerprints
- We can bound the second and third terms using the following lemmas

Lemma 1

- With probability of atleast 1-δ,

  $| \Sigma^{t}_{j=0} b_j (h_j - E[h_j])| <= O(\max_j |b_j| \sqrt{} (\log 1/\delta)/N)$

Lemma 2

- $| \Sigma^t_{j=0} b_j (h_j - E[h_j])| <= \max_j |b_j| \sqrt{2\ln 2}\sqrt{(t/2N \,.\, \log(4N/t) + \log(3e/\delta)/N)}$
- $\sqrt{t}$ dependence in the bound is unexpected
- This is because of the first inequality ?
- Hence we have to analyse the bound on the term $|b_j|$ to exactly analyse the bound of the EMD