Compressed Maximum Likelihood

Yi Hao, Alon Orlitsky
UC San Diego

June 19, 2021

Learning as functional estimation

Maximum likelihood (ML)

ML's statistical guarantees

Compressed maximum likelihood (CML)

CML's statistical guarantees

Implications

Compressed maximum likelihood (CML) is a *unified and sample-efficient* ML approach for statistical learning

Notation

- $\mathcal{Z}$ domain space
- $\mathcal{P}$ distribution collection over $\mathcal{Z}$
- $\mathcal{Q}$ space equipped with pseudo-metric d (*loss*)
- $f : \mathcal{P} \to \mathcal{Q}$ functional
- $\hat{f} : \mathcal{Z} \to \mathcal{Q}$ estimator

Learning objective

- Observe $Z \sim p$
- Design $\hat{f}$ s.t. $d(\hat{f}(Z), f(p))$ is small

The most fundamental and general statistical estimation technique

ML plug-in estimator

- Every $z \in Z$
- ML mapping $p_z := \mathrm{argmax}_{p \in \mathcal{P}}\ p(z)$
- Given $Z \sim p$, ML (plug-in) estimate $f(p_Z)$

Example – parameter estimation

- $\mathcal{Q} = \Theta$ parameter space with loss d
- $\mathcal{P}$ collection of distributions indexed by Θ
- $f(p_\theta) = \theta$ for each $p_\theta \in \mathcal{P}$
- ML estimate $f(p_{\theta^\star}) = \theta^\star$, where $\theta^\star$ maximizes $p_\theta(Z)$

Classical *i.i.d.* sampling model

Sample space $\mathcal{X}$, domain $\mathcal{Z} = \mathcal{X}^*$

Sample size n, distribution $p_{\mathcal{X}} \in \Delta_{\mathcal{X}}$

$Z = (X_1, \ldots, X_n) \sim p = p_{\mathcal{X}}^n$, estimate $f(p_{\mathcal{X}}^n)$

While the ML principle is quite natural, showing its finite-sample efficiency (finite n) is often not easy

An argument bypassing such difficulty:

Lemma (Acharya et al. 2017)

Let $d_f(p,q) = d(f(p), f(q))$. *For any* $\mathcal{Z}$, $\mathcal{P}$, *and accuracy* $\varepsilon > 0$,

$$\max_{p \in \mathcal{P}} \Pr_{Z \sim p}\left(d_f(p, p_Z) > 2\varepsilon\right) \le |\mathcal{Z}| \cdot \max_{p \in \mathcal{P}} \Pr_{Z \sim p}\left(d(f(p), \hat{f}(Z)) > \varepsilon\right).$$

Definition

- Co-domain Φ
- Compressor $\varphi : \mathcal{Z} \to \Phi$
- CML mapping $p_{\varphi(z)} := \operatorname{argmax}_{p \in \mathcal{P}} p(\varphi(z)), \quad \forall z \in \mathcal{Z}$

Compressor's quality for statistical learning

- Typicality A compressor is (m, γ)-typical for $m \in \mathbb{N}$ and $\gamma \in (0, 1)$, if for every $p \in \mathcal{P}$, there is a size-m set $\mathcal{T} \subseteq \Phi$, s.t., $p(\mathcal{T}) \geq 1 - \gamma$
- Learnability Given error parameters ε, δ, a compressor enables (ε, δ)-learning if there is an algorithm $\mathcal{A} : \Phi \to \mathcal{Q}$,

$$\Pr_{Z \sim p} \left(d(f(p), \mathcal{A}(\varphi(Z))) > \varepsilon \right) \leq \delta, \ \forall p \in \mathcal{P}$$

Theorem (CML's Competitiveness)

For any compressor φ that is (m, γ)-typical and enables (ε, δ)-learning, distribution $p \in \mathcal{P}$, and $Z \sim p$,

$$\Pr\left(d_f(p, p_{\varphi(Z)}) > 2\varepsilon\right) \leq \gamma + m \cdot \delta$$

Corollary (Approximate CML (ACML))

For any $\beta \leq 1$, $z \in \mathcal{Z}$, and compressor φ, a distribution $\tilde{p}_{\varphi(z)} \in \mathcal{P}$ is a β-approximate CML if $\tilde{p}_{\varphi(z)}(\varphi(z)) \geq \beta \cdot p_{\varphi(z)}(\varphi(z))$

Under the conditions in Theorem, a β-approximate CML achieves

$$\Pr\left(d_f(p, \tilde{p}_{\varphi(Z)}) > 2\varepsilon\right) \leq \gamma + m \cdot \delta / \beta$$

Compressed maximum likelihood (CML) is a *unified and sample-efficient* ML approach for statistical learning:

- Continuous structured densities
  - E.g., Guassian mixtures
- Discrete structured distributions
  - E.g., Poisson/binomial mixtures
- Distribution probability multisets; profile maximum likelihood (PML), introduced in Orlitsky et al. 2004
  - E.g., Coin tosses – $\{p_{\mathsf{h}}, p_{\mathsf{t}}\}$
- Symmetric distribution functionals; variation of Acharya et al. 2017
  - E.g., Shannon entropy, support size

Thank you