Efficient Mixture Modeling with RKHS Embeddings

A PAC-Bayesian Analysis

Matthew Higgs[1]

[1]Center for Computational Statistics and Machine Learning
University College London

Foundations and New Trends of PAC Bayesian Learning

Outline

Outline

Maximum Mean Discrepancy

- $P, Q \in \mathcal{M}_+^1(\mathbb{R}^d)$, probability measures on $\mathbb{R}^d$.
- $\mathcal{F} \subset \mathbb{R}^{\mathbb{R}^d}$, measurable.
- $f_Q := \int_{\mathbb{R}^d} f(x) dQ(x)$.

Maximum Mean Discrepancy

- $P, Q \in \mathcal{M}_+^1(\mathbb{R}^d)$, probability measures on $\mathbb{R}^d$.
- $\mathcal{F} \subset \mathbb{R}^{\mathbb{R}^d}$, measurable.
- $f_Q := \int_{\mathbb{R}^d} f(x) dQ(x)$.

Definition (Maximum Mean Discrepancy [Gretton et al., 2008])

$$MMD_{\mathcal{F}}(Q, P) := \sup_{f \in \mathcal{F}} |f_Q - f_P|. \tag{1}$$

Maximum Mean Discrepancy

- $P, Q \in \mathcal{M}_+^1(\mathbb{R}^d)$, probability measures on $\mathbb{R}^d$.
- $\mathcal{F} \subset \mathbb{R}^{\mathbb{R}^d}$, measurable.
- $f_Q := \int_{\mathbb{R}^d} f(x) dQ(x)$.

Definition (Maximum Mean Discrepancy [Gretton et al., 2008])

$$MMD_{\mathcal{F}}(Q, P) := \sup_{f \in \mathcal{F}} |f_Q - f_P|. \tag{1}$$

Examples

- $\mathcal{F} := C_b(\mathbb{R}^d)$
- $\mathcal{F} := \{f : ||f||_\infty \leq 1\}$
- $\mathcal{F} := \{\mathbf{1}_{(-\infty,t]}(\cdot) : t \in \mathbb{R}^d\}$,
- $\mathcal{F} := \left\{e^{i\langle\omega,\cdot\rangle} : \omega \in \mathbb{R}^d\right\}$,

Unit-ball in RHKS

- $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$, sym. p.d., $\sup_{x,x'} k(x, x') = C_k < \infty$.
- $\mathcal{H}$ s.t. $\langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{H}} = k(x, x')$, $|| \cdot ||_{\mathcal{H}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$.
- $k_Q(\cdot) := \int k(x, \cdot) dQ(x), \in \mathcal{H}$.
- $k_{Q,P} := \int \int k(x', x) dQ(x) dP(x'), \in \mathbb{R}$.

Unit-ball in RHKS

- $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$, sym. p.d., $\sup_{x,x'} k(x, x') = C_k < \infty$.
- $\mathcal{H}$ s.t. $\langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{H}} = k(x, x')$, $|| \cdot ||_{\mathcal{H}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$.
- $k_Q(\cdot) := \int k(x, \cdot) dQ(x), \in \mathcal{H}$.
- $k_{Q,P} := \int \int k(x', x) dQ(x) dP(x'), \in \mathbb{R}$.

Proposition ([Song et al., 2008])

For any $P, Q \in \mathcal{M}_+^1(\mathbb{R}^d)$, $\mathcal{F} = \{f : ||f||_{\mathcal{H}} \leq 1\}$,

$$MMD_{\mathcal{F}}(Q, P) = ||k_Q - k_P||_{\mathcal{H}}.$$

Unit-ball in RHKS

- $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$, sym. p.d., $\sup_{x,x'} k(x, x') = C_k < \infty$.
- $\mathcal{H}$ s.t. $\langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{H}} = k(x, x')$, $|| \cdot ||_{\mathcal{H}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$.
- $k_Q(\cdot) := \int k(x, \cdot) dQ(x), \in \mathcal{H}$.
- $k_{Q,P} := \int \int k(x', x) dQ(x) dP(x'), \in \mathbb{R}$.

Proposition ([Song et al., 2008])

For any $P, Q \in \mathcal{M}_+^1(\mathbb{R}^d)$, $\mathcal{F} = \{f : ||f||_{\mathcal{H}} \leq 1\}$,

$$MMD_{\mathcal{F}}(Q, P) = ||k_Q - k_P||_{\mathcal{H}}.$$

Corollary

$$MMD_{\mathcal{F}}^2(Q, P) = k_{Q,Q} - 2k_{Q,P} + k_{P,P}. \quad (2)$$

Outline

Reproducing KMM Mixture Model

- $S := (x_1, \ldots, x_n) \sim D^n$, $D_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$.
- $\Theta_{mult} := \{\theta \in \mathbb{R}^l : \theta^\top \mathbf{1} = 1, \theta \succ 0\}$.
- Mixture $Q_\theta = \sum_{i=1}^{l} \theta_i Q_i$, $\theta \in \Theta_{mult}$, $Q_i(\mathbb{R}^d) = 1$.
- $\mathbf{R}(i,j) := k_{Q_i,Q_j}$, $\mathbf{L}(i) := k_{Q_i,D_n}$.

Reproducing KMM Mixture Model

- $S := (x_1, \ldots, x_n) \sim D^n$, $D_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$.
- $\Theta_{mult} := \{\theta \in \mathbb{R}^l : \theta^\top \mathbf{1} = 1, \theta \succ 0\}$.
- Mixture $Q_\theta = \sum_{i=1}^{l} \theta_i Q_i$, $\theta \in \Theta_{mult}$, $Q_i(\mathbb{R}^d) = 1$.
- $\mathbf{R}(i,j) := k_{Q_i,Q_j}$, $\mathbf{L}(i) := k_{Q_i,D_n}$.

Corollary (KMM [Song et al., 2008])

Minimisation of $||k_{Q_\theta} - k_{D_n}||^2_{\mathcal{H}}$ *with reg.* $\lambda||\theta||^2$, *becomes QP*

$$\theta_{\min} := \arg\min_{\theta \in \Theta_{mult}} \quad \frac{1}{2}\theta^\top(\mathbf{R} + \lambda \mathbf{I})\theta - \mathbf{L}\theta \tag{3}$$

Reproducing KMM Mixture Model

- $S := (x_1, \ldots, x_n) \sim D^n$, $D_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$.
- $\Theta_{mult} := \{\theta \in \mathbb{R}^l : \theta^\top \mathbf{1} = 1, \theta \succ 0\}$.
- Mixture $Q_\theta = \sum_{i=1}^{l} \theta_i Q_i$, $\theta \in \Theta_{mult}$, $Q_i(\mathbb{R}^d) = 1$.
- $\mathbf{R}(i,j) := k_{Q_i,Q_j}$, $\mathbf{L}(i) := k_{Q_i,D_n}$.

Corollary (KMM [Song et al., 2008])

Minimisation of $||k_{Q_\theta} - k_{D_n}||^2_{\mathcal{H}}$ *with reg.* $\lambda||\theta||^2$, *becomes QP*

$$\theta_{\min} := \arg\min_{\theta \in \Theta_{mult}} \quad \frac{1}{2}\theta^\top(\mathbf{R} + \lambda \mathbf{I})\theta - \mathbf{L}\theta \tag{3}$$

Question

What PAC-Bayesian statistical guarantees do we have for $\theta_{\min}$*?*

Outline

U-Statistic PAC-Bayes Bound

- $S = (x_1, \ldots, x_n) \sim D^n$.
- $(\{h\})$ measurable, $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $a \leq h(x, x') \leq b$.
- $\bar{U}_S(h) := \frac{1}{n(n-1)} \sum_{i \neq j}^{n} \frac{h(x_i, x_j) - a}{b - a}$ with mean $\bar{U}(h)$.
- $\mathrm{kl}(q||p) := q \ln \frac{q}{p} + (1 - q) \ln (1 - q) \frac{1-q}{1-p}$.
- $\Phi_C(q, p) := \ln \frac{1}{1 - [1 - e^{-C}]p} - C \cdot q, \qquad C \in \mathbb{R}$.

U-Statistic PAC-Bayes Bound

- $S = (x_1, \ldots, x_n) \sim D^n$.
- $(\{h\})$ measurable, $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $a \leq h(x, x') \leq b$.
- $\bar{U}_S(h) := \frac{1}{n(n-1)} \sum_{i \neq j}^{n} \frac{h(x_i, x_j) - a}{b - a}$ with mean $\bar{U}(h)$.
- $\mathrm{kl}(q||p) := q \ln \frac{q}{p} + (1 - q) \ln (1 - q) \frac{1-q}{1-p}$.
- $\Phi_C(q, p) := \ln \frac{1}{1 - [1 - e^{-C}]p} - C \cdot q, \qquad C \in \mathbb{R}$.

Corollary

For any $\pi \in \mathcal{M}_+^1(\{h\})$, with probability $\geq 1 - \delta$ over the draw of S, for any $\rho \in \mathcal{M}_+^1(\{h\})$

$$\mathrm{kl}\Big(\mathbb{E}_{h \sim \rho} \bar{U}_S(h)) \Big|\Big| \mathbb{E}_{h \sim \rho} \bar{U}(h)\Big) \leq \frac{1}{\lfloor n/2 \rfloor} \left[\mathrm{KL}(\rho||\pi) + \ln \frac{\xi(\lfloor n/2 \rfloor)}{\delta} \right], \tag{4}$$

$$\Phi_C\Big(\mathbb{E}_{h \sim \rho} \bar{U}_S(h)), \mathbb{E}_{h \sim \rho} \bar{U}(h)\Big) \leq \frac{1}{\lfloor n/2 \rfloor} \left[\mathrm{KL}(\rho||\pi) + \ln \frac{1}{\delta} \right]. \tag{5}$$

Proof: iid blocks

- $\bar{h}(\cdot) = \frac{h(\cdot)-a}{b-a}$.
- Permutations σ on $\{1, \ldots, n\}$.
- $\bar{B}_{S,\sigma}(h) := \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1} \bar{h}(X_{\sigma(i)}, X_{\sigma(\lfloor n/2 \rfloor + i)})$.
- $\mathcal{D} : [0,1] \times [0,1] \to \mathbb{R}$ convex on $[0,1]^2$.

Proof: iid blocks

- $\bar{h}(\cdot) = \frac{h(\cdot)-a}{b-a}$.
- Permutations σ on $\{1, \ldots, n\}$.
- $\bar{B}_{S,\sigma}(h) := \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1} \bar{h}(X_{\sigma(i)}, X_{\sigma(\lfloor n/2 \rfloor + i)})$.
- $\mathcal{D} : [0,1] \times [0,1] \to \mathbb{R}$ convex on $[0,1]^2$.

Theorem

For any $\pi \in \mathcal{M}_+^1(\{h\})$, with probability $\geq 1-\delta$ over the draw of S, for any $\rho \in \mathcal{M}_+^1(\{h\})$

$$\mathcal{D}\Big(\mathbb{E}_{h\sim\rho}\bar{U}_S(h)), \mathbb{E}_{h\sim\rho}\bar{U}(h)\Big) \leq \frac{1}{\lfloor n/2 \rfloor}\left[\mathrm{KL}(\rho||\pi) + \ln\frac{\mathcal{L}_{\mathcal{D}}}{\delta}\right], \tag{6}$$

where $\mathcal{L}_{\mathcal{D}} = \mathbb{E}_{h\sim P}\mathbb{E}_S e^{\lfloor n/2 \rfloor\left[\mathcal{D}\left(\mathbb{E}_{h\sim\rho}\bar{U}_S(h)), \mathbb{E}_{h\sim\rho}\bar{U}(h)\right)\right]}$, and (by Jensen's)

$$\mathcal{L}_{\mathcal{D}} \leq \frac{1}{m!}\sum_{\sigma}\mathbb{E}_{h\sim P}\mathbb{E}_S e^{\lfloor n/2 \rfloor\left[\mathcal{D}\left(\bar{B}_{S,\sigma}(h), \mathbb{E}_S\bar{B}_{S,\sigma}(h)\right)\right]}. \tag{7}$$

Proof: Bounded to Bernoulli

[Maurer, 2004].

$X = (X_1, \ldots, X_n)$ iid, $X_i \in [0,1]$; $X' = (X'_1, \ldots, X'_n)$ iid, $X'_i \in \{0,1\}$, $\mathbb{E}X' = \mathbb{E}X$. Then $\phi : [0,1]^n \to \mathbb{R}$ convex, then

$$\mathbb{E}\phi(X) \leq \mathbb{E}\phi(X'). \tag{8}$$

Proof: Bounded to Bernoulli

[Maurer, 2004].

$X = (X_1, \ldots, X_n)$ iid, $X_i \in [0,1]$; $X' = (X'_1, \ldots, X'_n)$ iid, $X'_i \in \{0,1\}$, $\mathbb{E}X' = \mathbb{E}X$. Then $\phi : [0,1]^n \to \mathbb{R}$ convex, then

$$\mathbb{E}\phi(X) \leq \mathbb{E}\phi(X'). \tag{8}$$

□

Proposition

For $\mathcal{L}_{\mathcal{D}} \leq \frac{1}{m!} \sum_{\sigma} \mathbb{E}_{h \sim P} \mathbb{E}_S e^{\lfloor n/2 \rfloor \left[\mathcal{D}\left(\bar{B}_{S,\sigma}(h), \mathbb{E}_S \bar{B}_{S,\sigma}(h) \right) \right]}$, $S = (x_1, \ldots, x_n)$,

$$\mathcal{L}_{\mathrm{kl}} \leq \xi(\lfloor n/2 \rfloor) \tag{9}$$

$$\mathcal{L}_{\Phi_C} \leq 1, \qquad \forall C \in \mathbb{R}^+ \tag{10}$$

where $\xi(m) = \mathcal{O}(\sqrt{m})$.

Outline

Is $||k_{Q_{\widehat{\theta}}} - k_{D_n}||^2_{\mathcal{H}}$ a U-statistic?

Answer

No. But almost...

- $V(\theta) := \frac{1}{n}\sum_{k=1}^{n} ||k_{Q_\theta} - k(x_k, \cdot)||^2_{\mathcal{H}}$ (bias term).

Is $||k_{Q_{\widehat{\theta}}} - k_{D_n}||^2_{\mathcal{H}}$ a U-statistic?

Answer

No. But almost...

- $V(\theta) := \frac{1}{n}\sum_{k=1}^{n} ||k_{Q_\theta} - k(x_k, \cdot)||^2_{\mathcal{H}}$ (bias term).
- $\mathrm{kl}_{1/2}(q||p) := \mathrm{kl}\left(\frac{1}{2}q + \frac{1}{2}||\frac{1}{2}p + \frac{1}{2}\right)$.
- NB: $Q_\theta = \sum_{i=1}^{I} \theta_i Q_i$.

Is $||k_{Q_{\widehat{\theta}}} - k_{D_n}||^2_{\mathcal{H}}$ a *U*-statistic?

Answer

No. But almost...

- $V(\theta) := \frac{1}{n}\sum_{k=1}^{n} ||k_{Q_\theta} - k(x_k, \cdot)||^2_{\mathcal{H}}$ (bias term).
- $\mathrm{kl}_{1/2}(q||p) := \mathrm{kl}\left(\frac{1}{2}q + \frac{1}{2}||\frac{1}{2}p + \frac{1}{2}\right)$.
- NB: $Q_\theta = \sum_{i=1}^{I} \theta_i Q_i$.

Theorem

For any RKHS $\mathcal{H}$ with kernel k, $\mathcal{F} = \{f : ||f||_{\mathcal{H}} \leq 1\}$, any $\pi \in \mathcal{M}^1_+(\Theta_{mult})$, with probability $\geq 1 - \delta$ over the draw of S, for any $\widehat{\rho} \in \mathcal{M}^1_+(\Theta_{mult})$ with mean $\widehat{\theta} \in \Theta_{mult}$:

$$\mathrm{kl}_{1/2}\left(\frac{n||k_{Q_{\widehat{\theta}}} - k_{D_n}||^2_{\mathcal{H}} - V(\widehat{\theta})}{2C_k(n-1)}\middle|\middle|\frac{||k_{Q_{\widehat{\theta}}} - k_D||^2_{\mathcal{H}}}{2C_k}\right) \leq \frac{2\mathrm{KL}(\widehat{\rho}||\pi) + \ln\frac{\xi(\lfloor n/2\rfloor)}{\delta}}{\lfloor n/2\rfloor}.$$

Is $||k_{Q_{\widehat{\theta}}} - k_{D_n}||^2_{\mathcal{H}}$ a U-statistic?

Proof.

- $h_{(\theta,\theta')}(x, x') := \langle k_{Q_\theta} - k(x, \cdot), k_{Q_{\theta'}} - k(x', \cdot) \rangle_{\mathcal{H}} \leq 2C_k$.

Is $||k_{Q_{\widehat{\theta}}} - k_{D_n}||^2_{\mathcal{H}}$ a U-statistic?

Proof.

- $h_{(\theta,\theta')}(x, x') := \langle k_{Q_\theta} - k(x, \cdot), k_{Q_{\theta'}} - k(x', \cdot)\rangle_{\mathcal{H}} \leq 2C_k$.

For any $\rho^2, \pi^2 \in \mathcal{M}^1_+(\Theta^2_{mult})$, with probability $\geq 1 - \delta$

$$\mathrm{kl}\Big(\mathbb{E}_{(\theta,\theta')\sim\rho^2}\bar{U}_S\big(h_{(\theta,\theta')}\big)\Big|\Big|\mathbb{E}_{(\theta,\theta')\sim\rho^2}\mathbb{E}_S\bar{U}_S\big(h_{(\theta,\theta')}\big)\Big) \leq \frac{2\mathrm{KL}(\rho||\pi)}{\lfloor n/2\rfloor} + c \quad (11)$$

where $c = \frac{\ln \frac{\xi(\lfloor n/2\rfloor)}{\delta}}{\lfloor n/2\rfloor}$ and $\mathrm{KL}(\rho^2||\pi^2) = 2\mathrm{KL}(\rho||\pi)$.

Is $||k_{Q_{\hat{\theta}}} - k_{D_n}||^2_{\mathcal{H}}$ a *U*-statistic?

Proof.

- $h_{(\theta,\theta')}(x, x') := \langle k_{Q_\theta} - k(x, \cdot), k_{Q_{\theta'}} - k(x', \cdot)\rangle_{\mathcal{H}} \leq 2C_k$.

For any $\rho^2, \pi^2 \in \mathcal{M}^1_+(\Theta^2_{mult})$, with probability $\geq 1 - \delta$

$$\mathrm{kl}\left(\mathbb{E}_{(\theta,\theta')\sim\rho^2}\bar{U}_S\left(h_{(\theta,\theta')}\right)\middle|\middle|\mathbb{E}_{(\theta,\theta')\sim\rho^2}\mathbb{E}_S\bar{U}_S\left(h_{(\theta,\theta')}\right)\right) \leq \frac{2\mathrm{KL}(\rho||\pi)}{\lfloor n/2\rfloor} + c \quad (11)$$

where $c = \frac{\ln\frac{\xi(\lfloor n/2\rfloor)}{\delta}}{\lfloor n/2\rfloor}$ and $\mathrm{KL}(\rho^2||\pi^2) = 2\mathrm{KL}(\rho||\pi)$. Then

$$\frac{\mathbb{E}_{(\theta,\theta')\sim\rho^2}}{n(n-1)}\sum_{i\neq j}^{n}\langle k_{Q_\theta} - k(x_i, \cdot), k_{Q_{\theta'}} - k(x_j, \cdot)\rangle_{\mathcal{H}} = \frac{n||k_{Q_{\hat{\theta}}} - k_{D_n}||^2_{\mathcal{H}} - V(\widehat{\theta})}{(n-1)}$$

$$\frac{\mathbb{E}_{(\theta,\theta')\sim\rho^2}}{n(n-1)}\sum_{i\neq j}^{n}\mathbb{E}_S\langle k_{Q_\theta} - k(x_i, \cdot), k_{Q_{\theta'}} - k(x_j, \cdot)\rangle_{\mathcal{H}} = ||k_{Q_{\hat{\theta}}} - k_D||^2_{\mathcal{H}}.$$

Outline

Choosing $\mathrm{KL}(\widehat{\rho}||\pi)$ for $\widehat{\rho}, \pi \in \mathcal{M}_+^1(\Theta_{mult})$

Question

What are suitable choices for ρ and π such that $\mathrm{KL}(\widehat{\rho}||\pi)$ is manageable?

$$\mathrm{kl}_{1/2}\left(\frac{n||k_{Q_{\widehat{\theta}}} - k_{D_n}||_{\mathcal{H}}^2 - V(\widehat{\theta})}{2C_k(n-1)}\middle|\middle|\frac{||k_{Q_{\widehat{\theta}}} - k_D||_{\mathcal{H}}^2}{2C_k}\right) \leq \frac{2\mathrm{KL}(\widehat{\rho}||\pi) + \ln\frac{\xi(\lfloor n/2\rfloor)}{\delta}}{\lfloor n/2\rfloor}.$$

Choosing $\mathrm{KL}(\widehat{\rho}||\pi)$ for $\widehat{\rho}, \pi \in \mathcal{M}_+^1(\Theta_{mult})$

Question

What are suitable choices for ρ and π such that $\mathrm{KL}(\widehat{\rho}||\pi)$ is manageable?

$$\mathrm{kl}_{1/2}\left(\frac{n||k_{Q_{\widehat{\theta}}} - k_{D_n}||_{\mathcal{H}}^2 - V(\widehat{\theta})}{2C_k(n-1)}\middle|\middle|\frac{||k_{Q_{\widehat{\theta}}} - k_D||_{\mathcal{H}}^2}{2C_k}\right) \leq \frac{2\mathrm{KL}(\widehat{\rho}||\pi) + \ln\frac{\xi(\lfloor n/2\rfloor)}{\delta}}{\lfloor n/2\rfloor}.$$

Example (Dirichlet)

$\widehat{\rho}, \pi \in Dir(\Theta_{mult})$ with respective means $\widehat{\theta}$ and *uninformative* $\mathbf{1}/I$. Then

$$\mathrm{KL}_{Dir}(\widehat{\rho}||\pi) = \sum_{i=1}^{I} \log\frac{\Gamma(\widehat{\theta}_i)}{\Gamma(1/I)} + \sum_{i=1}^{I}[\widehat{\theta}_i - 1/I][\Psi(\widehat{\theta}_i) - \Psi(1/I)], \qquad (12)$$

where Γ and Ψ are the Gamma and Digamma functions.

Log-Normal Projection

- $\tau : \Theta_{mult} \rightarrow \mathbb{R}^l, \theta \mapsto (\ln(\frac{\theta_1}{g(\theta)}), \ldots, \ln(\frac{\theta_l}{g(\theta)}))^\top, g(\theta) = (\theta_1 \cdots \theta_l)^{1/l}.$

Log-Normal Projection

- $\tau : \Theta_{mult} \rightarrow \mathbb{R}^l$, $\theta \mapsto (\ln(\frac{\theta_1}{g(\theta)}), \ldots, \ln(\frac{\theta_l}{g(\theta)}))^\top$, $g(\theta) = (\theta_1 \cdots \theta_l)^{1/l}$.

Definition (Log-Normal Density)

Any $\rho \in LN(\Theta_{mult})$ with mean $\mu \in \Theta_{mult}$ and variance $1/\lambda$,

$$\frac{d\rho(\theta)}{d\theta} = |\nabla_\theta \tau| \left(\frac{2\pi}{\lambda}\right)^{-(l-1)/2} \exp\left(-\frac{\lambda}{2} ||\tau(\theta) - \tau(\mu)||^2 \right). \tag{13}$$

Log-Normal Projection

- $\tau : \Theta_{mult} \to \mathbb{R}^l$, $\theta \mapsto (\ln(\frac{\theta_1}{g(\theta)}), \ldots, \ln(\frac{\theta_l}{g(\theta)}))^\top$, $g(\theta) = (\theta_1 \cdots \theta_l)^{1/l}$.

Definition (Log-Normal Density)

Any $\rho \in LN(\Theta_{mult})$ with mean $\mu \in \Theta_{mult}$ and variance $1/\lambda$,

$$\frac{d\rho(\theta)}{d\theta} = |\nabla_\theta \tau| \left(\frac{2\pi}{\lambda}\right)^{-(l-1)/2} \exp\left(-\frac{\lambda}{2} ||\tau(\theta) - \tau(\mu)||^2 \right). \tag{13}$$

Example (Log-Normal KL)

$\widehat{\rho}, \pi \in LN(\Theta_{mult})$ with respective means $\widehat{\theta}$ and $\mathbf{1}/l$, and variance $1/\lambda$,

$$\mathrm{KL}_{LN}(\widehat{\rho}||\pi) = \frac{\lambda}{2} ||\tau(\widehat{\theta})||^2. \tag{14}$$

Outline

Conclusion

Summary:

- Derived a simple PAC-Bayes upper bound on the Maximum Mean Discrepancy between the input distribution and a mixture-based approximation.
- Given a class of prior and posterior with simple KL divergences.

Conclusion

Summary:

- Derived a simple PAC-Bayes upper bound on the Maximum Mean Discrepancy between the input distribution and a mixture-based approximation.
- Given a class of prior and posterior with simple KL divergences.

Future work:

- Find projections $\tau : \Theta_{mult} \rightarrow \mathbb{R}^l$ and distributions $\widehat{\rho}$ and π that provide efficient $\mathrm{KL}(\widehat{\rho}||\pi)$ for optimisation.
- Examine bounding $\ln \mathbb{E}_S \exp\left(\lambda\left[||k_{Q_{\widehat{\theta}}} - k_D||^2_{\mathcal{H}} - ||k_{Q_{\widehat{\theta}}} - k_{D_n}||^2_{\mathcal{H}}\right]\right)$.


Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. J. (2008).
CoRR abs/0805.2368.


Maurer, A. (2004).


Song, L., Zhang, X., Smola, A., Gretton, A. & Schölkopf, B. (2008).
In ICML '08: Proceedings of the 25th international conference on Machine learning pp. 992–999, ACM, New York, NY, USA.