Borrowing strength in hierarchical Bayes: convergence of the Dirichlet base measure

Long Nguyen

Department of Statistics
University of Michigan

9th Bayesian Nonparametrics Conference
Amsterdam, June 2013

Inference of the Dirichlet process base measure

- let $Q_1, \ldots, Q_m$ be m random measures drawn from $DP_{\alpha G}$, where $G = G_0$, how to infer about G_0 on the basis of Q_i's?
  - studied by Korwar and Hollander (Ann. Prob., 1973)

Inference of the Dirichlet process base measure

- let $Q_1, \ldots, Q_m$ be m random measures drawn from $DP_{\alpha G}$, where $G = G_0$, how to infer about G_0 on the basis of Q_i's?
  - studied by Korwar and Hollander (Ann. Prob., 1973)

- A realistic elaboration: assume that we have *no* direct observations of Q_i's, only iid observations from mixture models $Q_i * f$

- Moreover, base measure G is endowed with a prior distribution, namely another Dirichlet process prior

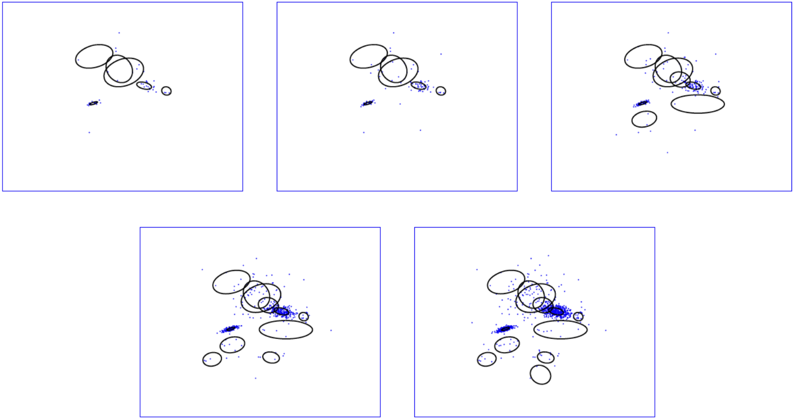
Inference of the Dirichlet process base measure

- let $Q_1, \ldots, Q_m$ be m random measures drawn from $DP_{\alpha G}$, where $G = G_0$, how to infer about G_0 on the basis of Q_i's?
  - studied by Korwar and Hollander (Ann. Prob., 1973)

- A realistic elaboration: assume that we have *no* direct observations of Q_i's, only iid observations from mixture models $Q_i * f$

- Moreover, base measure G is endowed with a prior distribution, namely another Dirichlet process prior
  - this is the Hierarchial Dirichlet Process (Teh, Jordan, Blei and Beal, JASA, 2006)
  - we ask: what is the posterior concentration behavior of G, given the observed data?

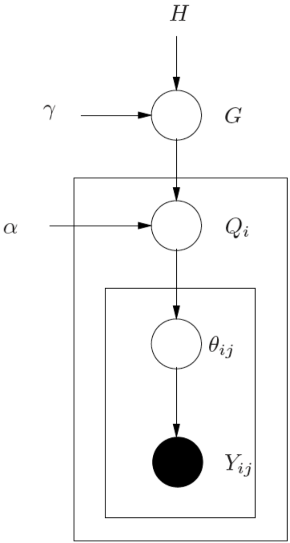
Modeling of exchangeable groups of exchangeable data



motivated by De Finetti's, each group can be modeled by a mixture model, while the mixture models are coupled by a nonparametric Bayesian hierarchy

Hierarchical Dirichlet process mixture

(Teh et al, JASA 2006)

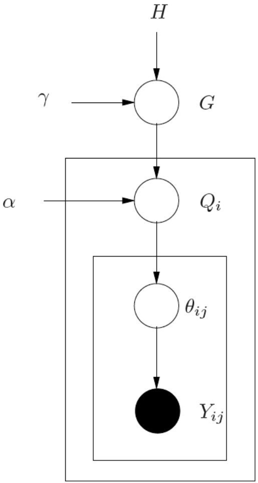


$$G \sim DP_{\gamma H}$$

$$Q_1, \ldots, Q_m | G \overset{iid}{\sim} DP_{\alpha G}$$

$$Y_{i1}, \ldots, Y_{in} | Q_i \overset{iid}{\sim} Q_i * f$$

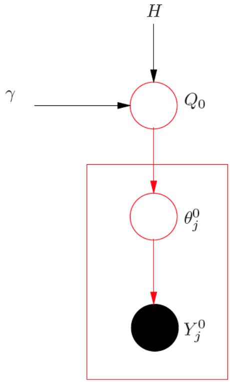
Posterior concentration of "tables" and "dishes" in Chinese restaurants:



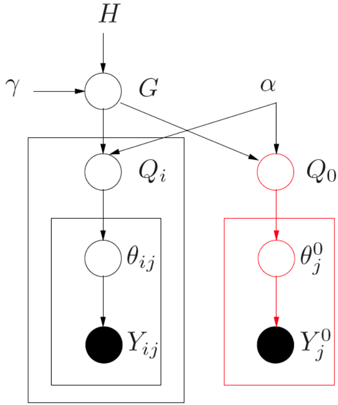
- posterior concentration behavior of latent G?
- posterior concentration behavior of Q_i's
- quantifying benefits of "borrowing of strength": hierarchical model vs treating groups separately?

Benefits of "borrowing strength"

given $\tilde{n}$-sample $(Y_1^0, \ldots, Y_{\tilde{n}}^0)$ from mixture distribution $Q_0 * f$
Q_0 is assumed to share the same atoms as Q_i's



versus



Stand-alone DP mixture

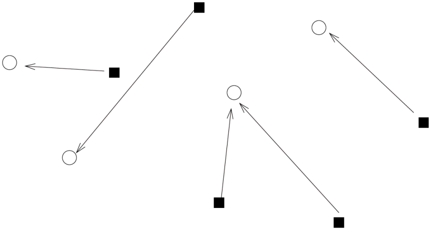
Hierarchical DP mixture

Talk outline

- tools from optimal transportation theory
  - Wasserstein metrics for nonparametric Bayesian hierarchies
- two main theorems
  - posterior concentration rate of Dirichlet base measure
  - benefits of "borrowing strength": improvement from nonparametric to parametric rate of convergence
- main ingredients of proof
  - concentration of Dirichlet measure
  - concentration of measure along the boundary between two Dirichlet processes

Optimal transport problem (Monge-Kantorovich)

- goods are transported from producers to customers in the optimal way (given that transportation cost is proportional to distance)
- the optimal transportation cost defines a distance between "production density" and "consumption density"



squares: locations of producers; circles: locations of consumers

Wasserstein distance

Let $G, G' \in \mathcal{P}(\Theta)$, the space of Borel probability measures on Θ, $\mathcal{T}(G, G')$ set of all couplings of G, G', i.e., all joint distributions on $\Theta \times \Theta$ which project to marginals G, G'

Definition

Let ρ be a distance function on Θ, the Wasserstein distance is defined by:

$$d_\rho(G, G') = \inf_{\kappa \in \mathcal{T}(G,G')} \int \rho(\theta, \theta') d\kappa.$$

Wasserstein distance

Let $G, G' \in \mathcal{P}(\Theta)$, the space of Borel probability measures on Θ, $\mathcal{T}(G, G')$ set of all couplings of G, G', i.e., all joint distributions on $\Theta \times \Theta$ which project to marginals G, G'

Definition

Let ρ be a distance function on Θ, the Wasserstein distance is defined by:

$$d_\rho(G, G') = \inf_{\kappa \in \mathcal{T}(G,G')} \int \rho(\theta, \theta') d\kappa.$$

When $\Theta \subset \mathbb{R}^d$, for $r \geq 1$, use $\| \cdot \|^r$ as a distance function on $\mathbb{R}^d$ to obtain L_r Wasserstein metric:

$$W_r(G, G') := \left[\inf_{\kappa \in \mathcal{T}(G,G')} \int \|\theta - \theta'\|^r d\kappa \right]^{1/r}.$$

Facts and Examples

Wasserstein distance W_r metrizes weak convergence in the space of probability measures on Θ.

Facts and Examples

Wasserstein distance W_r metrizes weak convergence in the space of probability measures on Θ.

If $\Theta = \mathbb{R}$, then $W_1(G, G') = \|CDF(G) - CDF(G')\|_1$.

Facts and Examples

Wasserstein distance W_r metrizes weak convergence in the space of probability measures on Θ.

If $\Theta = \mathbb{R}$, then $W_1(G, G') = \|CDF(G) - CDF(G')\|_1$.

If $G_0 = \delta_{\theta_0}$ and $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$, then

$$W_1(G_0, G) = \sum_{i=1}^{k} p_i \|\theta_0 - \theta_i\|.$$

Facts and Examples

Wasserstein distance W_r metrizes weak convergence in the space of probability measures on Θ.

If $\Theta = \mathbb{R}$, then $W_1(G, G') = \|CDF(G) - CDF(G')\|_1$.

If $G_0 = \delta_{\theta_0}$ and $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$, then

$$W_1(G_0, G) = \sum_{i=1}^{k} p_i \|\theta_0 - \theta_i\|.$$

If $G = \sum_{i=1}^{k} \frac{1}{k} \delta_{\theta_i}$, $G' = \sum_{j=1}^{k} \frac{1}{k} \delta_{\theta'_j}$, then

$$W_1(G, G') = \inf_{\pi} \sum_{i=1}^{k} \frac{1}{k} \|\theta_i - \theta'_{\pi(i)}\|,$$

where π ranges over the set of permutations on $(1, \dots, k)$.

Distance of nonparametric Bayesian hierarchies

Recall that $W_r(G, G')$ is Wasserstein metric on $\mathcal{P}(\Theta)$

Further up in the Bayesian hierarchy, again using Wasserstein-type distance

Distance of nonparametric Bayesian hierarchies

Recall that $W_r(G, G')$ is Wasserstein metric on $\mathcal{P}(\Theta)$

Further up in the Bayesian hierarchy, again using Wasserstein-type distance

Distance on measures of measures

Let $\mathcal{D}, \mathcal{D}' \in \mathcal{P}(\mathcal{P}(\Theta))$ (the space of Borel probability measures on $\mathcal{P}(\Theta)$). Define Wasserstein distance between $\mathcal{D}, \mathcal{D}'$

$$W_r(\mathcal{D}, \mathcal{D}') := \inf_{\mathcal{K} \in \mathcal{T}(\mathcal{D}, \mathcal{D}')} \left[\int W_r^r(G, G')\, d\mathcal{K}(G, G') \right]^{1/r}.$$

$\mathcal{T}(\mathcal{D}, \mathcal{D}')$ is the space of all couplings of $\mathcal{D}, \mathcal{D}' \in \mathcal{P}(\mathcal{P}(\Theta))$

Distance between two Dirichlet processes (Nguyen, 2013)

Let $\mathcal{D} = DP_{\alpha G}$ and $\mathcal{D}' = DP_{\alpha' G'}$. Then

$$W_r(\mathcal{D}, \mathcal{D}') \geq W_r(G, G').$$

Moreover, if $\alpha = \alpha'$ then $W_r(\mathcal{D}, \mathcal{D}') = W_r(G, G')$.

Set-up: posterior concentration of Dirichlet base measure

Let $Q_1, \ldots, Q_m$ be iid from $DP_{\alpha G}$, where $G = G_0$ (fixed non-random)

G is endowed with another Dirichlet prior $G \sim DP_{\gamma H}$, where H non-atomic

Set-up: posterior concentration of Dirichlet base measure

Let $Q_1, \ldots, Q_m$ be iid from $DP_{\alpha G}$, where $G = G_0$ (fixed non-random)

G is endowed with another Dirichlet prior $G \sim DP_{\gamma H}$, where H non-atomic

Each Q_i gives a mixture distribution $Q_i * f$, of which an n-iid sample is given

Set-up: posterior concentration of Dirichlet base measure

Let $Q_1, \ldots, Q_m$ be iid from $DP_{\alpha G}$, where $G = G_0$ (fixed non-random)

G is endowed with another Dirichlet prior $G \sim DP_{\gamma H}$, where H non-atomic

Each Q_i gives a mixture distribution $Q_i * f$, of which an n-iid sample is given

We will show that

As $m \to \infty$ and $n = n(m) \to \infty$ at a suitable rate, there is $\epsilon_{m,n} \to 0$ such that

$$\Pi_G\left(W_1(G, G_0) \geq C\epsilon_{m,n} \middle| m \times n \,\mathrm{Data}\, Y_{[n]}^{[m]} \right) \longrightarrow 0$$

in probability.

Assumptions

On kernel density f, and base probability measure H of the Dirichlet prior for G

(A1) For some $r \geq 1, C_1 > 0$, $h(f(\cdot|\theta), f(\cdot|\theta')) \leq C_1 \|\theta - \theta'\|^r$ and $K(f(\cdot|\theta), f(\cdot|\theta')) \leq C_1 \|\theta - \theta'\|^r \; \forall \theta, \theta' \in \Theta$.

(A2) There holds $M = \sup_{\theta, \theta' \in \Theta} \chi(f(\cdot|\theta), f(\cdot|\theta')) < \infty$.

(A3) $H \in \mathcal{P}(\Theta)$ is non-atomic, and for some constant $c_0 > 0$, $H(B) \geq c_0 \epsilon^d$ for any closed ball B of radius ϵ.

Main Theorems

Let Θ be a bounded subset of $\mathbb{R}^d$. Suppose that

(a) Assumptions (A1–A3) hold.

(b) G_0 has a finite number of support points in Θ.

(c) The Dirichlet parameters $\alpha \in (0, 1], \gamma > 0$, and $H \in \mathcal{P}(\Theta)$ non-atomic.

Theorem 1 (Nguyen, 2013)

As $m \to \infty$ and $n \to \infty$ such that $n_1(m) \leq n \leq n_2(m)$ for some sequences $n_2(m)$ and $n_1(m) \to \infty$, there holds

$$\Pi_G \left(W_1(G, G_0) \geq C \left(\frac{n^{3d} \log m}{m} \right)^{1/(2d+2)} \Big| m \times n \,\mathrm{Data}\, Y_{[n]}^{[m]} \right) \longrightarrow 0$$

in probability for a large constant C.

Remarks

The details of $n_1(m)$ and $n_2(m)$ depend on additional conditions of f. Define

$$\alpha^* := \min_{\theta \in \text{spt } G_0} \alpha G_0(\{\theta\}).$$

(i) If f is ordinary smooth with parameter β, then it suffices to set

$$n_1(m) \asymp m^{\frac{4+(2\beta+1)d'}{3d(4+(2\beta+1)d')+(2d+2)\alpha^*}}$$

and $n_2(m) \asymp (m/\log m)^{1/3d}$, for any $d' > d$. In particular, if n is allowed to grow at the rate $n \asymp n_1(m)$ then the posterior concentration rate is

$$\epsilon_{m,n} \asymp n^{-\frac{\alpha^*}{4+(2\beta+1)d'}} (\log n)^{1/(2d+2)} \asymp m^{-\gamma} (\log m)^{1/(2d+2)},$$

where

$$\gamma = \frac{\alpha^*}{3d(4+(2\beta+1)d')+(2d+2)\alpha^*} < \frac{1}{2d+2}.$$

(ii) If f is supersmooth with parameter β, then it suffices to set

$$\frac{m}{\log m (\log n)^{\alpha^*(2d+2)/\beta}} \lesssim n^{3d} \lesssim \frac{m}{\log m}.$$

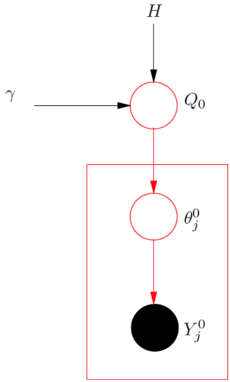
In particular, if n satisfies $n^{3d}(\log n)^{\alpha^*(2d+2)/\beta} \asymp \frac{m}{\log m}$, then we obtain the concentration rate $\epsilon_{m,n} \asymp (\log n)^{-\alpha^*/\beta} \asymp (\log m)^{-\alpha^*/\beta}$.

(iii) Requirements of the type $n_1(m) \leq n \leq n_2(m)$ appear crucial in deriving posterior concentration rates in hierarchical models. Beyond this range, we do not know the rates

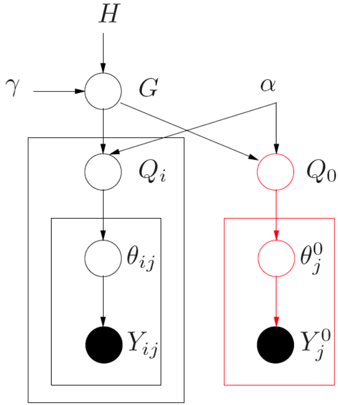
(iv) If G_0 has infinite support, we conjecture that polynomial rate is no longer possible.

Effects of "borrowing strength"

given $\tilde{n}$-sample $(Y_1^0, \ldots, Y_{\tilde{n}}^0)$ from mixture distribution $Q_0 * f$
Q_0 is assumed to share the same atoms as Q_i's



versus



Stand-alone DP mixture

Hierarchical DP mixture

Stand-alone setting

Suppose that an iid $\tilde{n}$-sample $Y^0_{[\tilde{n}]}$ drawn from a mixture model $Q_0 * f$ is available, where $Q_0 = Q_0^* \in \mathcal{P}(\Theta)$ is unknown:

$$Y^0_{[\tilde{n}]} | Q_0 \overset{iid}{\sim} Q_0 * f.$$

In a stand-alone setting Q_0 is endowed with a Dirichlet prior: $Q_0 \sim DP_{\alpha_0 H_0}$ for some known $\alpha_0 > 0$ and non-atomic base measure $H_0 \in \mathcal{P}(\Theta)$.

Stand-alone setting

Suppose that an iid $\tilde{n}$-sample $Y^0_{[\tilde{n}]}$ drawn from a mixture model $Q_0 * f$ is available, where $Q_0 = Q_0^* \in \mathcal{P}(\Theta)$ is unknown:

$$Y^0_{[\tilde{n}]} | Q_0 \stackrel{iid}{\sim} Q_0 * f.$$

In a stand-alone setting Q_0 is endowed with a Dirichlet prior: $Q_0 \sim DP_{\alpha_0 H_0}$ for some known $\alpha_0 > 0$ and non-atomic base measure $H_0 \in \mathcal{P}(\Theta)$.

(Nguyen, Ann Stat (2013))

Then

$$\Pi_Q\left(h(Q_0 * f, Q_0^* * f) \geq (\log \tilde{n}/\tilde{n})^{\frac{1}{d+2}} \,\middle|\, Y^0_{[\tilde{n}]} \right) \longrightarrow 0$$

in $P_{Y^0_{[\tilde{n}]}|Q_0^*}$-probability.

Alternatively, in hierarchical DP setting

suppose Q_0 is attached to the hierarchical Dirichlet process in the same way as the $Q_1, \ldots, Q_m$, i.e.:

$$G \sim DP_{\gamma H}, \quad Q_0, Q_1, \ldots, Q_m | G \overset{iid}{\sim} DP_{\alpha G}.$$

- implicitly Q_0 is assumed to share the same set of supporting atoms as $Q_1, \ldots, Q_m$, as they share with the (latent) discrete base measure G.

Alternatively, in hierarchical DP setting

suppose Q_0 is attached to the hierarchical Dirichlet process in the same way as the $Q_1, \ldots, Q_m$, i.e.:

$$G \sim DP_{\gamma H}, \quad Q_0, Q_1, \ldots, Q_m | G \stackrel{iid}{\sim} DP_{\alpha G}.$$

- implicitly Q_0 is assumed to share the same set of supporting atoms as $Q_1, \ldots, Q_m$, as they share with the (latent) discrete base measure G.

Then, as $\tilde{n} \to \infty$ and $m, n \to \infty$ at suitable rates, there is $\delta_{m,n,\tilde{n}} \downarrow 0$ such that

$$\Pi_Q \left(h(Q_0 * f, Q_0^* * f) \geq \delta_{m,n,\tilde{n}} \middle| Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]} \right) \longrightarrow 0$$

in $P_{Y_{[\tilde{n}]}^0 | Q_0^*} \times P_{G_0}^m$-probability, where

$$\delta_{m,n,\tilde{n}} \asymp (\log \tilde{n}/\tilde{n})^{1/(d+2)} + \epsilon_{m,n}^{r_0/2} \log(1/\epsilon_{m,n}),$$

Here, $\epsilon_{m,n}$ is an assumed concentration rate for the posterior of G.

- extra term $\epsilon_{m,n}^{r_0/2} \log(1/\epsilon_{m,n})$ suggests decreased efficiency due to the maintainance of the latent hierarchy

- extra term $\epsilon_{m,n}^{r_0/2} \log(1/\epsilon_{m,n})$ suggests decreased efficiency due to the maintainance of the latent hierarchy

- if m and n grow sufficiently fast relatively to $\tilde{n}$ so that $\epsilon_{m,n}$ is suitably small, then the impact of "borrowing of strength" from the $m \times n$ data set $Y_{[n]}^{[m]}$ on the inference about the data set $Y_{[\tilde{n}]}^{0}$ is quite striking:

- extra term $\epsilon_{m,n}^{r_0/2} \log(1/\epsilon_{m,n})$ suggests decreased efficiency due to the maintainance of the latent hierarchy

- if m and n grow sufficiently fast relatively to $\tilde{n}$ so that $\epsilon_{m,n}$ is suitably small, then the impact of "borrowing of strength" from the $m \times n$ data set $Y_{[n]}^{[m]}$ on the inference about the data set $Y_{[\tilde{n}]}^{0}$ is quite striking:

Theorem 2 (Nguyen, 2013)

1. if f is an ordinary smooth kernel density, then $\delta_{m,n,\tilde{n}} \asymp (\log \tilde{n}/\tilde{n})^{1/2}$.
2. if f is a supersmooth kernel density with smoothness $\beta > 0$, then $\delta_{m,n,\tilde{n}} \asymp (1/\tilde{n})^{1/(\beta+2)}$.

- the above theorem shows the improved efficiency for groups with small size $\tilde{n}$ — recall nonparametrate rate if using stand-alone mixture model, $(\log \tilde{n}/\tilde{n})^{1/(d+2)}$

Proof ingredients

- Existence of test argument: a subset in $\mathcal{P}(\Theta)$ that can be used to discriminate a pair of Dirichlet processes
- Existence of a point-estimate for mixing measures in a mixture model that admits finite-sample probability bounds
  - implying a lower bound of Hellinger distance of HDP data densities in terms of Wasserstein distance of Dirichlet processes
- Posterior concentration under a perturbation of base measure
  - requiring concentration of Dirichlet measure
- The rest are standard Bayesian asymptotics techniques (e.g., Ghosal, Ghosh and van der Vaart (2000))

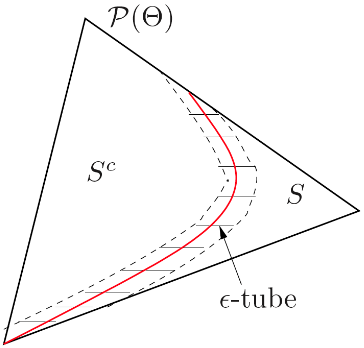
Existence of test sets

Consider a test $DP_{\alpha G}$ against $DP_{\alpha G'}$, we need to show existence of test set $S \subset \mathcal{P}(\Theta)$ the difference of measures on S is sufficiently large

Existence of test sets

Consider a test $DP_{\alpha G}$ against $DP_{\alpha G'}$, we need to show existence of test set $S \subset \mathcal{P}(\Theta)$ the difference of measures on S is sufficiently large

Boundary of S is "regular": the Dirichlet measure on the ϵ-tube defined along the boundary of S (in Wasserstein metric) needs to go to 0 at certain rate as $\epsilon \to 0$



Point estimate of mixing measures with finite-sample bounds

Given the assumption on kernel density f, with constants $C_1 > 0, r \geq 1$. Given n-sample from a mixture distribution $Q_0 * f$, there exists a point estimate $\hat{Q}_n$ of Q_0 and $\hat{f}_n = \hat{Q}_n * f$ such that for any $Q_0 \in \mathcal{Q}$: under $Q_0 * f$-measure,

$$\mathbb{P}(h(\hat{f}_n, Q_0 * f) \geq \epsilon_n) \leq 5 \exp(-c_2 n \epsilon_n^2),$$

$$\mathbb{P}(W_2(\hat{Q}_n, Q_0) \geq \delta_n) \leq 5 \exp(-c_2 n \epsilon_n^2),$$

where c_1, c_2 are some universal positive constants.

Point estimate of mixing measures with finite-sample bounds

Given the assumption on kernel density f, with constants $C_1 > 0, r \geq 1$. Given n-sample from a mixture distribution $Q_0 * f$, there exists a point estimate $\hat{Q}_n$ of Q_0 and $\hat{f}_n = \hat{Q}_n * f$ such that for any $Q_0 \in \mathcal{Q}$: under $Q_0 * f$-measure,

$$\mathbb{P}(h(\hat{f}_n, Q_0 * f) \geq \epsilon_n) \leq 5\exp(-c_2 n\epsilon_n^2),$$
$$\mathbb{P}(W_2(\hat{Q}_n, Q_0) \geq \delta_n) \leq 5\exp(-c_2 n\epsilon_n^2),$$

where c_1, c_2 are some universal positive constants. And:

(a) $\epsilon_n = C_2(\log n/n)^{r/2d}$, if $d > 2r$; $\epsilon_n = C_2(\log n/n)^{r/(d+2r)}$ if $d < 2r$, and $\epsilon_n = (\log n)^{3/4}/n^{1/4}$ if $d = 2r$.

(b) If f is ordinary smooth with parameter $\beta > 0$, then $\delta_n = C_3\epsilon_n^{\frac{2}{4+(2\beta+1)d'}}$ for any $d' > d$. If f is supersmooth with parameter $\beta > 0$, then $\delta_n = C_3[-\log \epsilon_n]^{-1/\beta}$.

Here, C_2, C_3 are different constants in each case. C_2 depends only on d, r, Θ and C_1, while C_3 depends only d, β, Θ and C_2.

Posterior concentration under perturbation

Suppose that spt $Q_0 \subset$ spt G_0, and we use a Dirichlet prior $Q \sim DP_{\alpha G}$ such that $W_r(G, G_0)$ is "small", then the posterior of Q given the data concentrates on the true Q_0 at a suitably fast rate

Posterior concentration under perturbation

Suppose that spt $Q_0 \subset$ spt G_0, and we use a Dirichlet prior $Q \sim DP_{\alpha G}$ such that $W_r(G, G_0)$ is "small", then the posterior of Q given the data concentrates on the true Q_0 at a suitably fast rate

To prove the above statement, one needs to carefully construct suitable sieves that occupy most of the probability mass, while requiring small entropy

Posterior concentration under perturbation

Suppose that spt $Q_0 \subset$ spt G_0, and we use a Dirichlet prior $Q \sim DP_{\alpha G}$ such that $W_r(G, G_0)$ is "small", then the posterior of Q given the data concentrates on the true Q_0 at a suitably fast rate

To prove the above statement, one needs to carefully construct suitable sieves that occupy most of the probability mass, while requiring small entropy

This requires new facts about the concentration of the Dirichlet process

Summary

- posterior concentration of latent hierarchies in the hierarchical Dirichlet process
  - convergence of the Dirichlet mean measure from mixture data
  - asymptotic gain of borrowing information in the Bayes hierarchy

- for details see
  - Nguyen, X. *Borrowing strength in hierarchical Bayes: convergence of the Dirichlet base measure.* arxiv.org/abs/1301.0802