

Concentration of tempered posteriors and of their variational approximations

Pierre Alquier



Workshop “Approximating high dimensional functions”
Dec. 19, 2017

Talk based on the preprint :



P. Alquier & J. Ridgway (2017). Concentration of tempered posteriors and of their variational approximations. *Preprint arxiv :1706.09293*.



Outline of the talk

- 1 Introduction : tempered posteriors & variational approx.
 - Tempered posteriors
 - Variational approximations
- 2 Main results
 - Concentration of the tempered posterior
 - A result in expectation
 - The misspecified case
- 3 Applications
 - Gaussian VB
 - Matrix completion

Notations

Assume that we observe $X_1, \dots, X_n$ i.i.d from P_{θ_0} in a model $\{P_\theta, \theta \in \Theta\}$ dominated by $Q : \frac{dP_\theta}{dQ} = p_\theta$. Prior π on Θ .

Notations

Assume that we observe $X_1, \dots, X_n$ i.i.d from P_{θ_0} in a model $\{P_\theta, \theta \in \Theta\}$ dominated by $Q : \frac{dP_\theta}{dQ} = p_\theta$. Prior π on Θ .

The likelihood

$$L_n(\theta) = \prod_{i=1}^n p_\theta(X_i)$$

Notations

Assume that we observe $X_1, \dots, X_n$ i.i.d from P_{θ_0} in a model $\{P_{\theta}, \theta \in \Theta\}$ dominated by $Q : \frac{dP_{\theta}}{dQ} = p_{\theta}$. Prior π on Θ .

The likelihood

$$L_n(\theta) = \prod_{i=1}^n p_{\theta}(X_i)$$

The posterior

$$\pi_n(d\theta) \propto L_n(\theta)\pi(d\theta).$$

Notations

Assume that we observe $X_1, \dots, X_n$ i.i.d from P_{θ_0} in a model $\{P_{\theta}, \theta \in \Theta\}$ dominated by $Q : \frac{dP_{\theta}}{dQ} = p_{\theta}$. Prior π on Θ .

The likelihood

$$L_n(\theta) = \prod_{i=1}^n p_{\theta}(X_i)$$

The posterior

$$\pi_n(d\theta) \propto L_n(\theta)\pi(d\theta).$$

The tempered posterior - $0 < \alpha < 1$

$$\pi_{n,\alpha}(d\theta) \propto [L_n(\theta)]^{\alpha}\pi(d\theta).$$

Various reasons to use a tempered posterior

- easier to sample from



G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

Various reasons to use a tempered posterior

- easier to sample from



G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

- robust to model misspecification



P. Grünwald and T. van Ommen (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*.

Various reasons to use a tempered posterior

- easier to sample from



G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

- robust to model misspecification



P. Grünwald and T. van Ommen (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*.

- theoretical analysis easier



A. Bhattacharya, D. Pati & Y. Yang (2016). Bayesian fractional posteriors. *Preprint arxiv :1611.01125*.

Bhattacharya, Pati & Yang's approach (1/2)

The α -Rényi divergence for $\alpha \in (0, 1)$

$$D_{\alpha}(P, R) = \begin{cases} \frac{1}{\alpha-1} \log \int \left(\frac{dP}{dR}\right)^{\alpha-1} dP & \text{if } P \ll R \\ +\infty & \text{otherwise.} \end{cases}$$

Bhattacharya, Pati & Yang's approach (1/2)

The α -Rényi divergence for $\alpha \in (0, 1)$

$$D_\alpha(P, R) = \begin{cases} \frac{1}{\alpha-1} \log \int \left(\frac{dP}{dR}\right)^{\alpha-1} dP & \text{if } P \ll R \\ +\infty & \text{otherwise.} \end{cases}$$

All the properties derived in :



T. Van Erven & P. Harremos (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*.

Among others, for $1/2 \leq \alpha$, link with Hellinger and Kullback :

$$\mathcal{H}^2(P, R) \leq D_\alpha(P, R) \xrightarrow[\alpha \nearrow 1]{} \mathcal{K}(P, R).$$

Bhattacharya, Pati & Yang's approach (2/2)

$$\mathcal{B}(r) = \left\{ \theta \in \Theta : \mathcal{K}(P_{\theta_0}, P_{\theta}) \leq r \text{ and } \text{Var} \left[\log \frac{p_{\theta}(X_i)}{p_{\theta_0}(X_i)} \right] \leq r. \right\}$$

Theorem (Bhattacharya, Pati & Yang)

For any sequence (r_n) such that

$$-\log \pi[B(r_n)] \leq nr_n$$

we have

$$\mathbb{P} \left[\int D_{\alpha}(P_{\theta}, P_{\theta_0}) \pi_{n,\alpha}(\mathrm{d}\theta) \leq \frac{2(1+\alpha)}{1-\alpha} r_n \right] \geq 1 - \frac{2}{nr_n}.$$

Computational issues

Popular methods to compute / sample from the (tempered) posterior :

Computational issues

Popular methods to compute / sample from the (tempered) posterior :

- Monte-Carlo methods : MCMC (Gibbs Sampler, Metropolis-Hastings), SMC, Langevin Monte-Carlo etc.

Computational issues

Popular methods to compute / sample from the (tempered) posterior :

- Monte-Carlo methods : MCMC (Gibbs Sampler, Metropolis-Hastings), SMC, Langevin Monte-Carlo etc.
- optimization methods : **variational Bayes (VB)** and expectation-propagation (EP).

Computational issues

Popular methods to compute / sample from the (tempered) posterior :

- Monte-Carlo methods : MCMC (Gibbs Sampler, Metropolis-Hastings), SMC, Langevin Monte-Carlo etc.
- optimization methods : **variational Bayes (VB)** and expectation-propagation (EP).

Principle of VB : chose a family $\mathcal{F}$ of probability distributions on Θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}$:

Computational issues

Popular methods to compute / sample from the (tempered) posterior :

- Monte-Carlo methods : MCMC (Gibbs Sampler, Metropolis-Hastings), SMC, Langevin Monte-Carlo etc.
- optimization methods : **variational Bayes (VB)** and expectation-propagation (EP).

Principle of VB : chose a family $\mathcal{F}$ of probability distributions on Θ and approximate $\pi_{n,\alpha}$ by a distribution in $\mathcal{F}$:

$$\tilde{\pi}_{n,\alpha} := \arg \min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}).$$

Variational approximations

$$\begin{aligned}\tilde{\pi}_{n,\alpha} &= \arg \min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}) \\ &= \arg \min_{\rho \in \mathcal{F}} \left\{ -\alpha \int \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i) \rho(d\theta) + \mathcal{K}(\rho, \pi) \right\}.\end{aligned}$$

Examples :

Variational approximations

$$\begin{aligned}\tilde{\pi}_{n,\alpha} &= \arg \min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}) \\ &= \arg \min_{\rho \in \mathcal{F}} \left\{ -\alpha \int \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i) \rho(d\theta) + \mathcal{K}(\rho, \pi) \right\}.\end{aligned}$$

Examples :

- parametric approximation

$$\mathcal{F} = \{ \mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d^+ \}.$$

Variational approximations

$$\begin{aligned}\tilde{\pi}_{n,\alpha} &= \arg \min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}) \\ &= \arg \min_{\rho \in \mathcal{F}} \left\{ -\alpha \int \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i) \rho(d\theta) + \mathcal{K}(\rho, \pi) \right\}.\end{aligned}$$

Examples :

- parametric approximation

$$\mathcal{F} = \{ \mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d^+ \}.$$

- mean-field approximation, $\Theta = \Theta_1 \times \Theta_2$ and

$$\mathcal{F} : \{ \rho : \rho(d\theta) = \rho_1(d\theta_1) \times \rho_2(d\theta_2) \}.$$

Extension of previous result to VB

Theorem

Assume that (r_n) is such that there is a distribution $\rho_n \in \mathcal{F}$ with

$$\int \mathcal{K}(P_{\theta_0}, P_{\theta}) \rho_n(d\theta) \leq r_n, \quad \int \mathbb{E} \left[\log^2 \left(\frac{p_{\theta}(X_i)}{p_{\theta_0}(X_i)} \right) \right] \rho_n(d\theta) \leq r_n$$

and

$$\mathcal{K}(\rho_n, \pi) \leq nr_n.$$

Then, for any $\alpha \in (0, 1)$,

$$\mathbb{P} \left[\int D_{\alpha}(P_{\theta}, P_{\theta_0}) \tilde{\pi}_{n,\alpha}(d\theta) \leq \frac{2(\alpha + 1)}{1 - \alpha} r_n \right] \geq 1 - \frac{2}{nr_n}.$$

A simpler result in expectation

Theorem

If we only require that there is $\rho_n \in \mathcal{F}$ such that

$$\int \mathcal{K}(P_{\theta_0}, P_{\theta}) \rho_n(d\theta) \leq r_n$$

and

$$\mathcal{K}(\rho_n, \pi) \leq nr_n,$$

then, for any $\alpha \in (0, 1)$,

$$\mathbb{E} \left[\int D_{\alpha}(P_{\theta}, P_{\theta_0}) \tilde{\pi}_{n,\alpha}(d\theta) \right] \leq \frac{1 + \alpha}{1 - \alpha} r_n.$$

Misspecified case

Assume now that $X_1, \dots, X_n$ i.i.d from $Q \notin \{P_\theta, \theta \in \Theta\}$.

Put :

$$\theta^* := \arg \min_{\theta \in \Theta} \mathcal{K}(Q, P_\theta).$$

Theorem

Assume that there is $\rho_n \in \mathcal{F}$ such that

$$\int \mathbb{E} \left[\log \frac{dP_{\theta^*}}{dP_\theta} \right] \rho_n(d\theta) \leq r_n \text{ and } \mathcal{K}(\rho_n, \pi) \leq nr_n,$$

then, for any $\alpha \in (0, 1)$,

$$\mathbb{E} \left[\int D_\alpha(P_\theta, Q) \tilde{\pi}_{n,\alpha}(d\theta) \right] \leq \frac{\alpha}{1-\alpha} \mathcal{K}(Q, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha} r_n.$$

- 1 Introduction : tempered posteriors & variational approx.
 - Tempered posteriors
 - Variational approximations
- 2 Main results
 - Concentration of the tempered posterior
 - A result in expectation
 - The misspecified case
- 3 Applications
 - Gaussian VB
 - Matrix completion

Gaussian VB

- Let $\Theta = \mathbb{R}^p$.

Gaussian VB

- Let $\Theta = \mathbb{R}^p$.
- We start with the family of approximations

$$\mathcal{F}_{\mathcal{G}}^{\Phi} := \left\{ \Phi(d\theta; m, \Sigma), \quad m \in \mathbb{R}^d, \Sigma \in \mathcal{G} \subset \mathcal{S}_+^d(\mathbb{R}) \right\},$$

Gaussian VB

- Let $\Theta = \mathbb{R}^p$.
- We start with the family of approximations

$$\mathcal{F}_{\mathcal{G}}^{\Phi} := \left\{ \Phi(d\theta; m, \Sigma), \quad m \in \mathbb{R}^d, \Sigma \in \mathcal{G} \subset \mathcal{S}_+^d(\mathbb{R}) \right\},$$

- We assume that for a model $\{p_{\theta}, \theta \in \Theta\}$ there exists a measurable real valued function $M(\cdot)$ and $p \in \mathbb{N}^* \cup \{\frac{1}{2}\}$

$$|\log p_{\theta}(X_1) - \log p_{\theta'}(X_1)| \leq M(X_1) \|\theta - \theta'\|_2^{2p}$$

Furthermore we assume that

$$\mathbb{E}M(X_1) =: B_1, \quad \mathbb{E}M^2(X_1) =: B_2 < \infty.$$

Application of the result

Theorem

Let the family of approximation be $\mathcal{F}$ with $\mathcal{F}_{\sigma^2 I}^\Phi \subset \mathcal{F}$ as defined above. We put

$$r_n = \frac{B_1}{n} \vee \frac{B_2}{n^2} \vee C \frac{d}{n} \log n$$

Then for any $\alpha \in (0, 1)$,

$$\mathbb{E} \left[\int D_\alpha(P_\theta, P_{\theta_0}) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n) \right] \leq \frac{1 + \alpha}{1 - \alpha} r_n.$$

Stochastic Variational Bayes

- To implement the idea we write

$$\mathcal{F}_B^\Phi = \left\{ \Phi(d\theta; m, CC^t), \quad (m, C) \in \mathbb{B} \cap \mathbb{R}^d \times \mathcal{S}_+^d \right\}.$$

$$F : x = (m, C) \in \mathbb{R}^d \times \mathbb{R}^{d \times d} \mapsto \mathbb{E} [f(x, \xi)] = \mathcal{K}(\rho_{m,C}, \pi_n)$$

where $\xi \sim \mathcal{N}(0, I_d)$

Stochastic Variational Bayes

- To implement the idea we write

$$\mathcal{F}_B^\Phi = \{ \Phi(d\theta; m, CC^t), \quad (m, C) \in \mathbb{B} \cap \mathbb{R}^d \times \mathcal{S}_+^d \}.$$

$$F : x = (m, C) \in \mathbb{R}^d \times \mathbb{R}^{d \times d} \mapsto \mathbb{E} [f(x, \xi)] = \mathcal{K}(\rho_{m,C}, \pi_n)$$

where $\xi \sim \mathcal{N}(0, I_d)$

- The optimization problem can be written

$$\min_{x \in \mathbb{B} \cap \mathbb{R}^d \times \mathcal{S}_+^d} \mathbb{E} [f(x, \xi)],$$

where

$$f((m, C), \xi) := \log p_{m+ C\xi}(Y_1^n) + \log \frac{d\Phi_{m, CC^t}}{d\pi}(m + C\xi)$$

We can use stochastic gradient descent

Algorithm 1 Stochastic VB

Input : x_0, X_1^n, γ_T

For $i \in \{1, \dots, T\},$

a. Sample $\xi_t \sim \mathcal{N}(0, I_d)$

b. Update

$$x_t \leftarrow \mathcal{P}_{\mathbb{B}}(x_{t-1} - \gamma_T \nabla f(x_{t-1}, \xi_t))$$

End For .

Output : $\bar{x}_T = \frac{1}{T} \sum_{t=1}^T x_t$

where ∇f is the gradient of the integrand in the objective function

- Assume that f is convex in its first component x and that it has L -Lipschitz gradients.
- Define $\tilde{\pi}_{n,\alpha}^k(d\theta|X_1^n)$ to be the k -th iterate of the algorithm

Theorem

For some C ,

$$r_n = \frac{B_1}{n} \vee \frac{B_2}{n^2} \vee \left\{ \frac{d}{n} \left[\frac{1}{2} \log(\vartheta^2 n^2 C) + \frac{1}{n\vartheta^2} \right] + \frac{\|\theta_0\|^2}{n\vartheta^2} - \frac{d}{2n} \right\}$$

with $\gamma_k = \frac{B}{L\sqrt{2k}}$, we get

$$\mathbb{E} \left[\int D_\alpha(P_\theta, P_{\theta_0}) \tilde{\pi}_{n,\alpha}^k(d\theta|X_1^n) \right] \leq \frac{1+\alpha}{1-\alpha} r_n + \frac{1}{n(1-\alpha)} \sqrt{\frac{2BL}{k}}.$$

Matrix completion : notations

The parameter θ is a matrix $M^0 \in \mathbb{R}^{m \times p}$, with $m, p \geq 1$.
Under P_M , the observations are random entries of this matrix with possible noise :

$$Y_i = M_{i_k, j_k}^0 + \varepsilon_k$$

where the (i_k, j_k) are i.i.d $\mathcal{U}(\{1, \dots, m\} \times \{1, \dots, p\})$. Assume that the ε_k are i.i.d $\mathcal{N}(0, \sigma^2)$, σ^2 known. We have

$$\mathcal{K}(P_M, P_N) = \frac{1}{mp} \sum_{i=1}^m \sum_{j=1}^p \frac{(M_{i,j} - N_{i,j})^2}{2\sigma^2} = \frac{\|M - N\|_F^2}{2\sigma^2 mp}.$$

Matrix completion : notations

The parameter θ is a matrix $M^0 \in \mathbb{R}^{m \times p}$, with $m, p \geq 1$.
Under P_M , the observations are random entries of this matrix with possible noise :

$$Y_i = M^0_{i_k, j_k} + \varepsilon_k$$

where the (i_k, j_k) are i.i.d $\mathcal{U}(\{1, \dots, m\} \times \{1, \dots, p\})$. Assume that the ε_k are i.i.d $\mathcal{N}(0, \sigma^2)$, σ^2 known. We have

$$\mathcal{K}(P_M, P_N) = \frac{1}{mp} \sum_{i=1}^m \sum_{j=1}^p \frac{(M_{i,j} - N_{i,j})^2}{2\sigma^2} = \frac{\|M - N\|_F^2}{2\sigma^2 mp}.$$

Usual assumption : M^0 is low-rank.

Prior specification - main idea

Define :

$$\underbrace{M}_{p \times m} = \underbrace{U}_{p \times k} \underbrace{V^T}_{k \times m}.$$

Prior specification - main idea

Define :

$$\underbrace{M}_{p \times m} = \underbrace{U}_{p \times k} \underbrace{V^T}_{k \times m}.$$

Let $U_{\cdot,\ell} \sim \mathcal{N}(0, \gamma I)$ denote the ℓ -th column of M , we have :

$$M = \sum_{\ell=1}^k U_{\cdot,\ell} (V_{\cdot,\ell})^T \quad \Rightarrow \quad \text{rank}(M) \leq k.$$

Prior specification - adaptation



R. Salakhutdinov & A. Mnih (2008). Bayesian probabilistic matrix factorization using MCMC. *Proceedings of ICML'08*.

Prior specification - adaptation



R. Salakhutdinov & A. Mnih (2008). Bayesian probabilistic matrix factorization using MCMC. *Proceedings of ICML'08*.

$$M = \sum_{\ell=1}^k U_{\cdot,\ell} (V_{\cdot,\ell})^T$$

with k large - e.g. $k = \min(p, m)$.

Prior specification - adaptation



R. Salakhutdinov & A. Mnih (2008). Bayesian probabilistic matrix factorization using MCMC. *Proceedings of ICML'08*.

$$M = \sum_{\ell=1}^k U_{\cdot,\ell} (V_{\cdot,\ell})^T$$

with k large - e.g. $k = \min(p, m)$.

Definition of π :

- $U_{\cdot,\ell}, V_{\cdot,\ell} \sim \mathcal{N}(0, \gamma_{\ell} I)$,
- γ_{ℓ} is itself random, such that most of the $\gamma_{\ell} \simeq 0$

$$\frac{1}{\gamma_{\ell}} \sim \text{Gamma}(a, b).$$

Known results



T. Suzuki (2015). Convergence rate of Bayesian tensor estimator and its minimax optimality. *ICML*.

(truncation of the support of π : remove large values of $M_{i,j}$).

Known results



T. Suzuki (2015). Convergence rate of Bayesian tensor estimator and its minimax optimality. *ICML*.

(truncation of the support of π : remove large values of $M_{i,j}$).



T. T. Mai & P. Alquier (2014). A Bayesian Approach for Noisy Matrix Completion : Optimal Rate under General Sampling Distribution. *Electronic Journal of Statistics*.

(truncation of the support of π : remove large values of $U_{i,k}$ and $V_{j,k}$).

Known results



T. Suzuki (2015). Convergence rate of Bayesian tensor estimator and its minimax optimality. *ICML*.

(truncation of the support of π : remove large values of $M_{i,j}$).



T. T. Mai & P. Alquier (2014). A Bayesian Approach for Noisy Matrix Completion : Optimal Rate under General Sampling Distribution. *Electronic Journal of Statistics*.

(truncation of the support of π : remove large values of $U_{i,k}$ and $V_{j,k}$).

In both cases, (in expectation or with large probability),

$$\int \frac{\|M - M^0\|_F^2}{2\sigma^2 mp} \hat{\pi}_{n,\alpha}(\mathrm{d}M) \lesssim \frac{\mathrm{rank}(M^0) \max(m, p) \log(\dots)}{n}.$$

Variational approximation



Y. J. Lim & Y. W. Teh (2007). Variational Bayesian approach to movie rating prediction.
Proceedings of KDD cup and workshop.

Mean-field approximation, $\mathcal{F}$ given by :

$$\rho(dU, dV, d\gamma) = \bigotimes_{i=1}^m \rho_{U_i}(dU_{i,\cdot}) \bigotimes_{j=1}^p \rho_{V_j}(dV_{j,\cdot}) \bigotimes_{k=1}^K \rho_{\gamma_k}(\gamma_k).$$

Variational approximation



Y. J. Lim & Y. W. Teh (2007). Variational Bayesian approach to movie rating prediction.
Proceedings of KDD cup and workshop.

Mean-field approximation, $\mathcal{F}$ given by :

$$\rho(dU, dV, d\gamma) = \bigotimes_{i=1}^m \rho_{U_i}(dU_{i,\cdot}) \bigotimes_{j=1}^p \rho_{V_j}(dV_{j,\cdot}) \bigotimes_{k=1}^K \rho_{\gamma_k}(\gamma_k).$$

It can be shown that

- ❶ ρ_{U_i} is $\mathcal{N}(\mathbf{m}_{i,\cdot}^T, \mathcal{V}_i)$,
- ❷ ρ_{V_j} is $\mathcal{N}(\mathbf{n}_{j,\cdot}^T, \mathcal{W}_j)$,
- ❸ ρ_{γ_k} is $\Gamma(a + (m_1 + m_2)/2, \beta_k)$,

for some $m \times K$ matrix $\mathbf{m}$ whose rows are denoted by $\mathbf{m}_{i,\cdot}$,
some $p \times K$ matrix $\mathbf{n}$ and some vector $\beta = (\beta_1, \dots, \beta_K)$.

The VB algorithm

The parameters are updated iteratively through the formulae

1 moments of U :

$$\mathbf{m}_{i,\cdot}^T := \frac{2\alpha}{n} \mathcal{V}_i \sum_{k:j_k=i} Y_{i_k,j_k} \mathbf{n}_{j_k,\cdot}^T,$$

$$\mathcal{V}_i^{-1} := \frac{2\alpha}{n} \sum_{k:j_k=i} [\mathcal{W}_{j_k} + \mathbf{n}_{j_k,\cdot} \mathbf{n}_{j_k,\cdot}^T] + \left(a + \frac{m_1 + m_2}{2} \right) \text{diag}(\beta)^{-1}$$

2 moments of V :

$$\mathbf{n}_{j,\cdot}^T := \frac{2\alpha}{n} \mathcal{W}_j \sum_{k:j_k=j} Y_{i_k,j_k} \mathbf{m}_{i_k,\cdot}^T,$$

$$\mathcal{W}_j^{-1} := \frac{2\alpha}{n} \sum_{k:j_k=j} [\mathcal{V}_{i_k} + \mathbf{m}_{i_k,\cdot} \mathbf{m}_{i_k,\cdot}^T] + \left(a + \frac{m_1 + m_2}{2} \right) \text{diag}(\beta)^{-1}$$

3 moments of γ :

$$\beta_k := \frac{1}{2} \left[\sum_{i=1}^{m_1} \left(\mathbf{m}_{i,k}^2 + (\mathcal{V}_i)_{k,k} \right) + \sum_{j=1}^{m_2} \left(\mathbf{n}_{j,k}^2 + (\mathcal{V}_j)_{k,k} \right) \right].$$

Application of our theorem

Theorem

Assume $M = \bar{U}\bar{V}^T$ where

$$\bar{U} = (\bar{U}_{1,\cdot} | \dots | \bar{U}_{r,\cdot} | 0 | \dots | 0) \text{ and } \bar{V} = (\bar{V}_{1,\cdot} | \dots | \bar{V}_{r,\cdot} | 0 | \dots | 0)$$

and $\sup_{i,k} |U_{i,k}|, \sup_{j,k} |V_{j,k}| \leq B$. Take $a > 0$ as any constant and $b = \frac{B^2}{512(nmp)^4[(m \vee p)K]^2}$. Then

$$\mathbb{P} \left[\int D_\alpha(P_M, P_{M^0}) \tilde{\pi}_{n,\alpha}(\mathrm{d}M) \leq \frac{2(\alpha + 1)}{1 - \alpha} r_n \right] \geq 1 - \frac{2}{nr_n}$$

$$\text{where } r_n = \frac{\mathcal{C}(a, \sigma^2, B) r \max(m, p) \log(nmp)}{n}.$$

Thank you !