Online likelihood-ratio estimation

Alejandro de la Concha, Nicolas Vayatis, Argyris Kalogeratos

Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, France

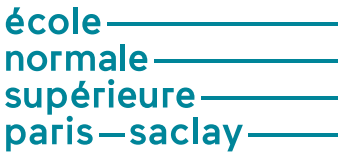


Table of contents

Table of contents

Likelihood-ratio and statistics

Definition

Consider a feature space $\mathcal{X} \subset \mathbb{R}^n$ and two probability distributions P and Q ($Q \ll P$) such that they admit density functions $p(x)$ and $q(x)$ with respect to dx, then the **likelihood-ratio** is defined as:

$$r(x) = \frac{q(x)}{p(x)} \quad x \in \mathcal{X}$$

Applications

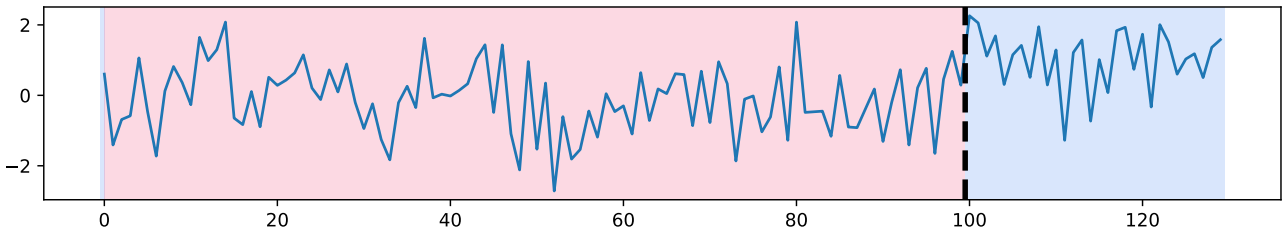
- *Hypothesis Testing* (Neyman-Pearson Lemma [Neyman et al., 1933])
- *Sequential Change-point Detection* [Page, 1954, Shiryaev, 1963]
- *Transfer Learning (Importance Sampling [Fishman, 1996])*

Sequential change-point detection

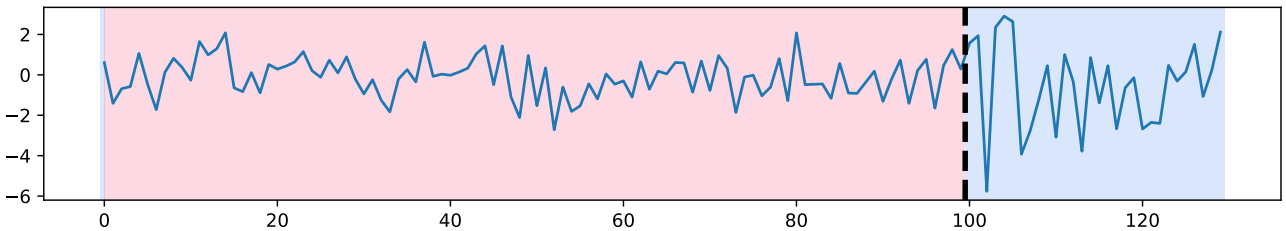
Hypothesis: Given a set of incoming observations $x_1, x_2, ...$, there exists a timestamp τ such that $\{x_t\}_{t=1}^{\tau-1} \sim p(x)$ and $\{x_t\}_{t=\tau}^{\infty} \sim q(x)$

Goal: Identify τ as soon as possible while minimizing the risk of false alarm

Change in the mean



Change in the variance



Change-point detectors

Shewart Chart

$$\tau_{Sh} = \inf\{t \geq 1 : r(x_t) \geq b)$$

Cumulative Sum Procedure (CUSUM)

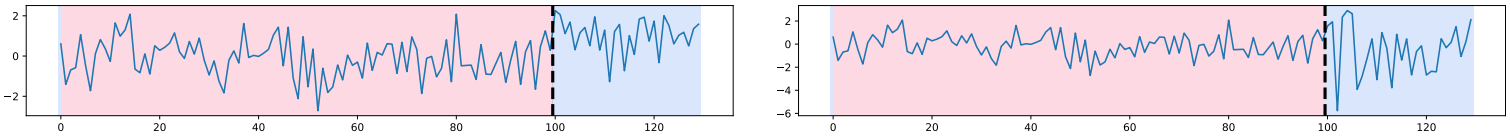
$$W_t = (W_{t-1} + \log(r(x_t)))^+, \quad W_0 = 0$$
$$\tau_{CS} = \inf\{t \geq 1 : W_t \geq b)$$

Shiryaev-Roberts Procedure

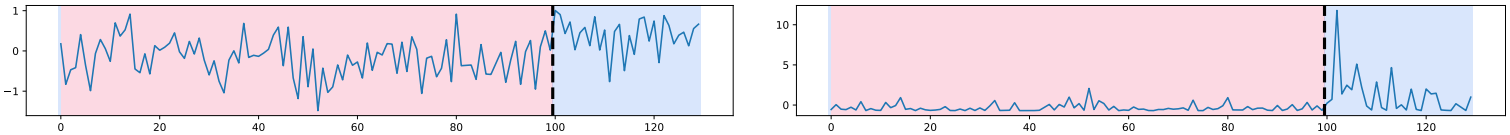
$$T_t = (1 + T_{t-1})r(x_t), \quad T_0 = 0$$
$$\tau_{SR} = \inf\{t \geq 1 : T_t \geq b)$$

Back to our example

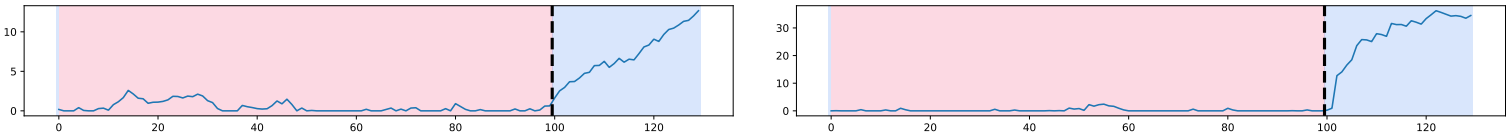
Original time series



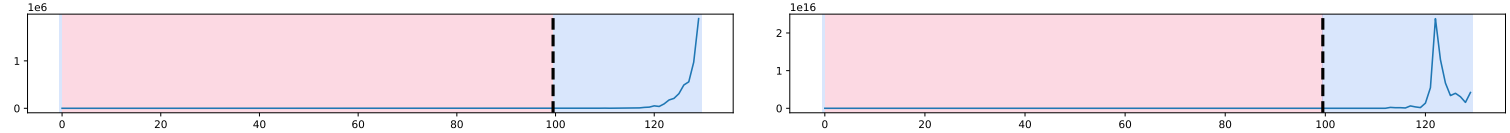
Shewart Chart



CUSUM



Shiryaev-Roberts Procedure



Limitations of existing methods

Take away message: The likelihood-ratio is a fundamental component in change-point detection
However:

1. $r(x)$ is rarely known in practice (q is by definition unknown)
2. Inference techniques assume the type of change is known in advance
3. Parametric models assume the distribution family before and after the change will be the same

Question

How can we estimate $r(x)$ without restrictive hypotheses on p and q?

Table of contents

ϕ-divergences

Definition

Let P and Q be two probability measures defined over the input space $\mathcal{X}$. A ϕ-divergence is a positive measure that quantifies the dissimilarity between P and Q:

$$\mathcal{D}_\phi(P\|Q) = \begin{cases} \int \phi\left(\frac{dQ}{dP}\right)(x)dP(x), & \text{if } Q \ll P; \\ +\infty, & \text{if } Q \not\ll P, \end{cases} \tag{1}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a convex and lower semi-continuous function with $\phi(1) = 0$ Csiszár [1967], Birrell et al. [2022].

If holds, $\mathcal{D}_\phi(P\|Q) \geq 0$, and if ϕ is strictly convex at 1, then $\mathcal{D}_\phi(P\|Q) = 0$ iff $P = Q$

ϕ-divergences and likelihood-ratio estimation

Lemma

Lemma 1 in Nguyen et al. [2008]. *Consider two probability distributions P and Q, both assumed to be absolutely continuous with respect to Lebesgue measure dx, with densities p and q. We assume Q is absolutely continuous with respect to P ($Q \ll P$). Then, for any class of measurable functions $\mathcal{F}$:*

$$\mathcal{D}_\phi(P\|Q) \geq \sup_{g\in\mathcal{F}} \int g(x')q(x')dx' - \int \phi^*(g)(x)\, p(x)dx \tag{2}$$

where ϕ^* denotes the convex conjugate of $\phi : \mathbb{R} \to \mathbb{R}$. Eq. 2 holds iff the subdifferential $\nabla\phi(r)$ contains an element of $\mathcal{F}$.

We consider $P^\alpha = (1-\alpha)P + \alpha Q$, then $Q \ll P^\alpha$ and $\mathcal{D}_\phi(P^\alpha\|Q) < \infty$.

Lemma 1 will refer to $r^\alpha(x) = \frac{q(x)}{p^\alpha(x)} \leq \frac{1}{\alpha}$

ϕ-divergences and likelihood-ratio estimation

How to solve the functional optimization problem in practice?

$$\sup_{g \in \mathcal{F}} \int g(x')q(x')dx' - \int \phi^*(g)(x)\, p^\alpha(x)dx$$

1. **The choice of the functional space $\mathcal{F}$**
   - Mutilple possible options: linear models, RKHS, neural-networks ...
   - $\mathcal{F}$ should be big enough to approximate r^α
   - Possibility to derive guarantees for convergence rates
2. **The choice of the ϕ-divergence**
   - The form of ϕ affects the numerical implementation of LRE
   - The computational complexity varies if the problem is quadratic, convex, strongly convex...

The choice of the functional space $\mathcal{F}$

Definition

Reproducing Kernel Hilbert Space. Let $\mathcal{X} \subset \mathbb{R}^n$ be a set, and $\mathbb{H} \subset \mathbb{R}^{\mathcal{X}}$ a class of functions forming a real Hilbert space with inner-product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$.
$\mathrm{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel function of $\mathbb{H}$ if:

1. $\mathbb{H}$ contains all functions of the form: $\forall x \in \mathcal{X}$, $\mathrm{K}(x, \cdot) : t \to \mathrm{K}(x, t)$.
2. $\langle K(x, \cdot), f \rangle_{\mathbb{H}} = f(x) \in \mathbb{R}$, for any $f \in \mathbb{H}$
3. $\mathbb{H} = \overline{\mathrm{span}}(\{K(x, \cdot) : \forall x \in \mathcal{X}\})$

- Notice that $r^\alpha \in L^p(p^\alpha)$, $p \in [1, \infty)$ since $r^\alpha \leq \frac{1}{\alpha}$
- **Theorem:** The space of continuous functions with compact support C_K is dense in $L^p(p^\alpha)$
- **Universal Kernels:** They are dense in the space C_K with respect to the maximum norm

The choice of ϕ

If $\phi(z) = \frac{1}{2}(z-1)^2$ recovers Pearson's χ^2-divergence [Pearson, 1900]:

$$\text{Æ}(P^\alpha \| Q) = \int \frac{1}{2}(r^\alpha(x) - 1)^2 p^\alpha(x) dx$$

Then the likelihood-ratio estimation problem takes the form:

$$\min_{f \in \mathbb{H}} (1-\alpha) \int \frac{f^2(x)}{2} dP(x) + \alpha \int \frac{f^2(x')}{2} dQ(x') - \int f(x')\, dQ(x'),$$

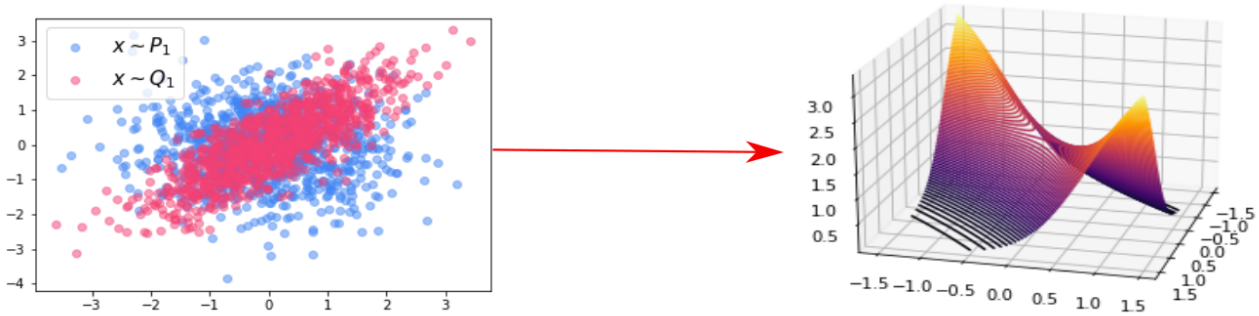
where f is an approximation of r^α

LRE problem is a Quadratic Problem in $\mathbb{H}$

Classical approach

Setting: Two datasets are available $\mathbf{X} = \{x_t \sim p\}_{t=1}^{n}$, $\mathbf{X}' = \{x'_t \sim q\}_{t=1}^{n'}$

$$r(\cdot)$$



Empirical Risk Minimizarion [Nguyen et al., 2008, Sugiyama et al., 2012].

$$\hat{f} = \min_{f \in \mathbb{H}} \frac{(1-\alpha)}{2n} \sum_{i=1}^{n} f(x_i) + \frac{\alpha}{2n'} \sum_{i=1}^{n'} f(x'_i) - \frac{1}{n'} \sum_{i=1}^{n'} f(x'_i) + \frac{\lambda_{n,n'}}{2} \|f\|_{\mathbb{H}}^2$$

Classical approach

The Representer Theorem implies: $\hat{f}(\cdot) = \sum_{i=1}^{n+n'} \hat{\theta}_i K(x_i, \cdot)$, meaning LRE takes the form:

$$\hat{\theta} = \min_{\theta \in \mathbb{R}^{n+n'}} \theta^\top \left[\frac{(1-\alpha)}{2n} \sum_{i=1}^{n} \phi(x_i)\phi(x_i)^\top + \frac{\alpha}{2} \sum_{i=1}^{n'} \phi(x'_i)\phi(x'_i)^\top + \frac{\lambda_n}{2} \mathcal{K} \right] \theta$$

$$- \theta^\top \left(\frac{1}{n'} \sum_{i=1}^{n'} \phi(x'_i) \right)$$

where $\phi(\cdot) = (K(x_1, \cdot), ..., K(x_{n+n'}, \cdot)$ and $\mathcal{K}_{ij} = K(x_i, x_j)$

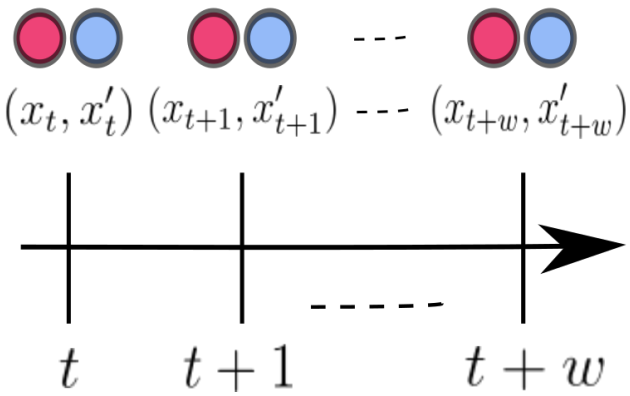
Drawbacks

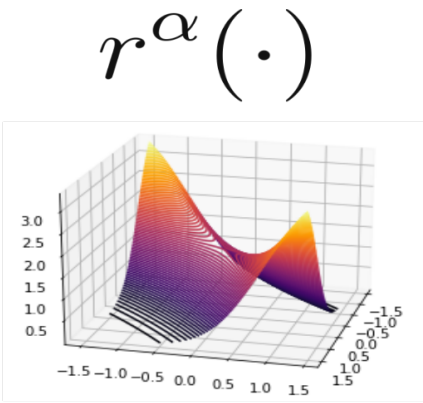
1. We require to know the number of observations in advance
2. Empirical risk minimization is prone to overfitting

Table of contents

Online likelihood-ratio estimation

What about the online case where data arrive on the fly?





Where $x_t, x_{t+1}, ..., x_{t+w}, ... \sim p$ and $x'_t, x'_{t+1}, ..., x'_{t+w}, ... \sim q$

LRE as a linear problem in $\mathbb{H}$

Covariance operator

$$C_\alpha = \mathbb{E}_{p^\alpha(y)}[K(y,\cdot) \otimes K(y,\cdot)]; \quad f, g \in \mathbb{H} \;\; \langle f, C_\alpha g \rangle_{\mathbb{H}} = \mathbb{E}_{p^\alpha(y)}[f(y)g(y)]$$

Mean embedding

$$\mu_q = \mathbb{E}_{q(x)}\left[K(x',\cdot)\right], \quad f \in \mathbb{H} \;\; \langle f, \mu_q \rangle_{\mathbb{H}} = \mathbb{E}_{q(x)}\left[f(x)\right]$$

LRE problem is a Linear Problem in $\mathbb{H}$

$$\operatorname*{argmin}_{f \in \mathbb{H}} \int \frac{f^2(y)}{2} dP^\alpha(y) - \int f(x')\, dQ(x') = \operatorname*{argmin}_{f \in \mathbb{H}} \frac{\langle f, C_\alpha f \rangle_{\mathbb{H}}}{2} - \langle f, \mu_q \rangle_{\mathbb{H}} \qquad (3)$$

Problem 3 is equivalent to find $f \in \mathbb{H}$ such that:

$$C_\alpha f = \mu_q$$

Gradients in a RKHS

Definition

Fréchet derivative. Let V and W be normed vector spaces, and $U \subset V$ be an open subset of V. A function $\ell : U \to W$ is called Fréchet differentiable at $f \in U$ if there exists a bounded linear operator $A : V \to W$ such that:

$$\lim_{\|h\|_V \to 0} \frac{\|\ell(f+h) - \ell(f) - Ah\|_W}{\|h\|_V} = 0$$

If A exits it is unique and $\nabla_f(\ell(f)) = A$ is called the Fréchet derivative.

If $\ell : \mathbb{H} \to \mathbb{R}$ by the reproducing property:

$$\begin{aligned} \nabla_f \ell(f(x)) &= \nabla_f \ell(\langle K(x,\cdot), f\rangle_{\mathbb{H}}) \\ &= \nabla_f [\ell \circ \langle K(x,\cdot), \cdot\rangle_{\mathbb{H}})(f)] \\ &= \ell'(f(x)) K(x,\cdot) \end{aligned} \tag{4}$$

Stochastic appoximation and regularization paths

How to solve the linear system $C_\alpha f = \mu_q$ as observations arrive and C_α has an unbounded inverse?

Given the instantaneous loss-function:

$$\ell_t^{\mathsf{PE}}(f) = (1-\alpha)\frac{f^2(x_t)}{2} + \alpha\frac{f^2(x_t')}{2} - f(x_t') + \frac{\lambda_t}{2}\|f\|_{\mathbb{H}}^2 .$$

Compute the Fréchet derivative $\nabla_f(\ell_t^{\mathsf{PE}}(f))(\cdot) \in \mathbb{H}$:

$$\nabla_f(\ell_t^{\mathsf{PE}}(f))(\cdot) = (1-\alpha)f(x_t)K(x_t,\cdot) + (\alpha f(x_t') - 1)K(x_t',\cdot) + \lambda_t f(\cdot).$$

A line of code :

$$\begin{aligned} f_t(\cdot) &= f_{t-1}(\cdot) - \eta_t \nabla_f(\ell_t^{\mathsf{PE}}(f_{t-1}))(\cdot) \\ &= (1-\eta_t\lambda_t)f_{t-1}(\cdot) - \eta_t[(1-\alpha)f_{t-1}(x_t)K(x_t,\cdot) + (\alpha f_{t-1}(x_t') - 1)K(x_t',\cdot)] \end{aligned}$$

where $\lambda_t, \eta_t \to 0$ as $t \to \infty$ such that $f_t \to r^\alpha$

Online likelihood-ratio estimation

Algorithm: – OLRE

Input: $\{x_t \sim p, x'_t \sim q\}_{t=1,\dots}$: stream of observation pairs;
t_0: size of the warm-up period;
$a \geq 4$, $\frac{1}{2} \leq \beta \leq 1$: fixed constants;
$0 < \alpha < 1$: prefixed regularization parameter;
K: predefined kernel function

Initialize $f_0(\cdot) = 0$; $D_0 = [\,]$; $\theta_0 = [\,]$
for $t = 1, 2, \dots$ **do**
    Get the incoming i.i.d pair of observations (x_t, x'_t)
    Compute the step-size and the penalization parameter:

$$\eta_t = a\left(\frac{1}{t_0 + t}\right)^{\frac{2\beta}{2\beta+1}}, \qquad \lambda_t = \frac{1}{a}\left(\frac{1}{t_0 + t}\right)^{\frac{1}{2\beta+1}}$$

    Update the dictionary: $D_t = D_{t-1} \cup \{x_t, x'_t\}$
    Update the weights: $\theta_t = [(1-\eta_t\lambda_t)\theta_{t-1}, \eta_t(\alpha - 1)f_{t-1}(x_t), \eta_t(1-\alpha f_{t-1}(x'_t))]$
    Update the relative likelihood-ratio estimate: $f_t(\cdot) = K(D_t, \cdot)\theta_t$

return $\{f_t\}_{t=1}^{T}$

General hypotheses

- **Assumption 1.** The pairs of observations $(x_t, x'_t), t = 1, 2, ...$ are iid in time and satisfy $x_t \sim p$ and $x'_t \sim q$
- **Assumption 2.** The reproducing kernel map can be upper-bounded by a constant $C > 0$: $\sup_{x \in \mathcal{X}} \sqrt{K(x, x)} \leq C < \infty$
- **Assumption 3.** p^α has full support on the feature space $\mathcal{X}$
- **Assumption 4.1.** $r^\alpha \in (C_\alpha)^\beta(\mathcal{L}^2_{p^\alpha})$ for $\frac{1}{2} \leq \beta \leq 1$
- **Assumption 4.2.** $r^\alpha \in (C_\alpha)^\beta(\mathcal{L}^2_{p^\alpha})$ for $\frac{1}{2} < \beta \leq \frac{3}{2}$

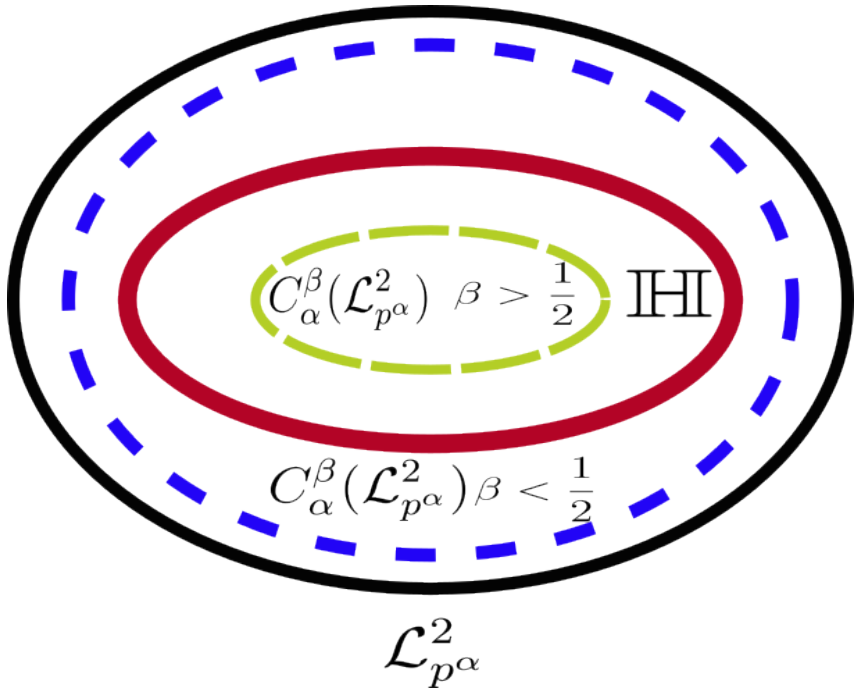
Source condition

The previous hypothesis implies C_α is a compact operator, then it is a self-adjoint, semi-definite positive operator in $\mathcal{L}^2_{p^\alpha}$:

$$C_\alpha = \sum_{i \in I} \mu_i \psi_i \otimes \psi_i$$

$$(C_\alpha)^\beta = \sum_{i \in I} \mu_i^\beta \psi_i \otimes \psi_i$$

where $\{\psi_i\}_{i \in I}$ is a basis of $\mathbb{H}$, and the eigenvalues $\{\mu_i\}_{i \in I}$ are strictly positive and arranged in decreasing order



Convergence rates

$$\begin{aligned}\|f_t - r^\alpha\|_{\mathcal{L}^2_{p^\alpha}} &\leq \left\|\bar{\Pi}_1^t(f_0 - f_{\lambda_0})\right\|_{\mathcal{L}^2_{p^\alpha}} + \left\|f_{\lambda_t} - r^\alpha\right\|_{\mathcal{L}^2_{p^\alpha}} \\ &+ \left\|\sum_{j=1}^t \bar{\Pi}_j^t(f_{\lambda_j} - f_{\lambda_{j-1}})\right\|_{\mathcal{L}^2_{p^\alpha}} + \left\|\sum_{j=1}^t \eta_j \bar{\Pi}_{j+1}^t \epsilon_j\right\|_{\mathcal{L}^2_{p^\alpha}} \\ &\leq \mathcal{E}_{\text{init}}(t) + \mathcal{E}_{\text{approx}}(t) + \mathcal{E}_{\text{drift}}(t) + \mathcal{E}_{\text{sample}}(t)\end{aligned}$$

Theorem

Under 1,2,3,4.1, we have with probabilty $1-\delta$:

$$\begin{aligned}\|f_t - r^\alpha\|_{\mathcal{L}^2_{p^\alpha}} &\leq \frac{C_1}{\bar{t}} + \left(C_2 a^{(-\beta)} + C_3\sqrt{a}\log(\frac{2}{\delta})\right)\left(\frac{1}{\bar{t}}\right)^{\frac{\beta}{2\beta+1}} \\ &+ \left(C_4 a^{\frac{5}{2}} + C_5 a^{\frac{7}{2}}\sqrt{log(\bar{t})}\right)(\log^2(\frac{2}{\delta}))\left(\frac{1}{\bar{t}}\right)^{\frac{4\beta-1}{4\beta+2}},\end{aligned}$$

where:

$$C_1 = \frac{2t_0}{\alpha}, C_2 = \frac{5\beta+1}{\beta(1+\beta)}\left\|L_K^{(-\beta)} r^\alpha\right\|_{\mathcal{L}^2_{p^\alpha}}, C_3 = \frac{16C}{\alpha}, C_4 = \frac{32C^3}{\alpha}, C_5 = \frac{8C^3(10C+3)}{\alpha}.$$

Convergence rates

$$\begin{aligned}\|f_t - r^\alpha\|_{\mathbb{H}} &\leq \left\|\Pi_1^t(f_0 - f_{\lambda_0})\right\|_{\mathbb{H}} + \left\|f_{\lambda_t} - r^\alpha\right\|_{\mathbb{H}} \\ &+ \left\|\sum_{j=1}^{t} \Pi_j^t(f_{\lambda_j} - f_{\lambda_{j-1}})\right\|_{\mathbb{H}} + \left\|\sum_{j=1}^{t} \eta_j \Pi_{j+1}^t \epsilon_j\right\|_{\mathbb{H}} \\ &\leq \mathcal{E}'_{\text{init}}(t) + \mathcal{E}'_{\text{approx}}(t) + \mathcal{E}'_{\text{drift}}(t) + \mathcal{E}'_{\text{sample}}(t)\end{aligned}$$

Theorem

Under 1,2,3,4.2, we have with probabilty $1-\delta$:

$$\|f_t - r^\alpha\|_{\mathbb{H}} \leq \frac{C'_1}{\bar{t}} + \left(C'_2 a^{\frac{1}{2}-\beta} + C'_3 a \log\left(\frac{2}{\delta}\right)\right)\left(\frac{1}{\bar{t}}\right)^{\frac{2\beta-1}{4\beta+2}},$$

where:

$$C'_1 = \frac{2\sqrt{a}t_0^{\frac{4\beta+1}{4\beta+2}}}{\alpha}, C'_2 = \frac{20\beta-2}{(2\beta-1)(2\beta+3)}\left\|L_K^{(-\beta)} r^\alpha\right\|_{\mathcal{L}^2_{p^\alpha}}, C'_3 = 6\left(\frac{(C+1)^2}{C\alpha}\right)$$

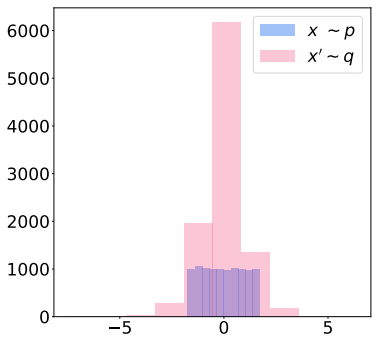
Table of contents

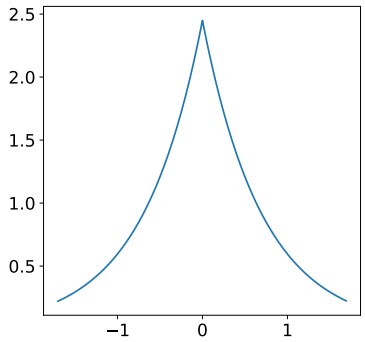
Competitors

Method	**Reference**	**Estimate**	ϕ-**divergence**	**Cost per iteration**		**Total cost**	
				#MA	#KE	#MA	#KE
KLIEP	Sugiyama et al. [2007]	l.-r.	KL-divergence	(T^2)	(T^2)	$(i_{\text{KLIEP}}(T)T^2)$	(T^2)
				(M^2)	(MT)	$(i_{\text{KLIEP}}(T)MT)$	(MT)
RULSIF	Yamada et al. [2011]	relative l.-r.	χ^2-divergence	(T^3)	(T^2)	(T^3)	(T^2)
				(M^3)	(MT)	(M^3)	(MT)
OLRE	this work	relative l.-r.	χ^2-divergence	(T)	(T)	(T^2)	(T^2)

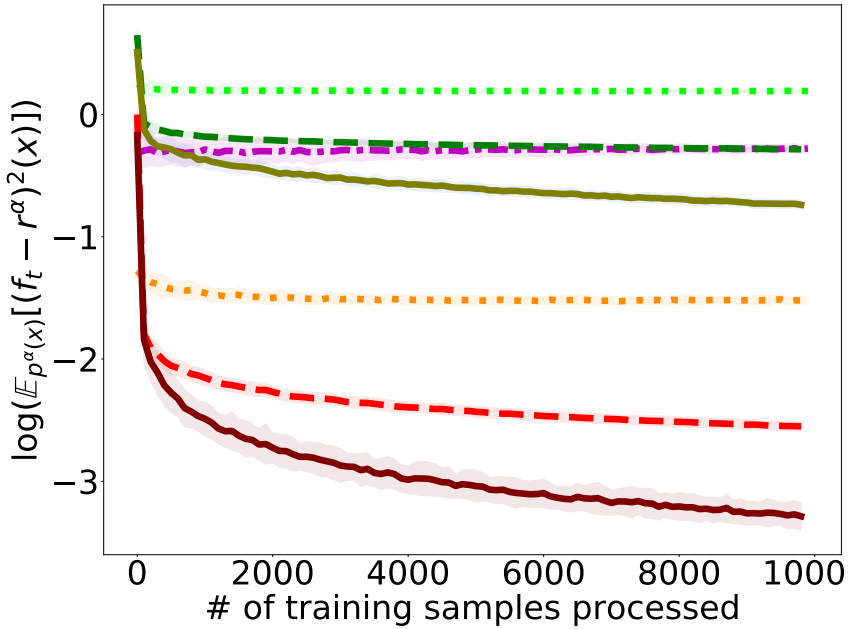
Experiments

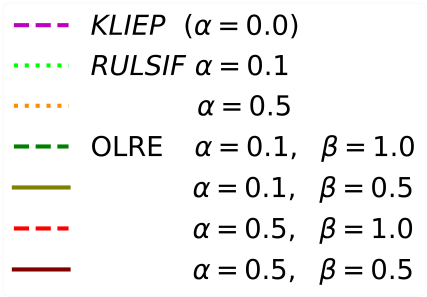
Experiment I: p is a uniform continuous distribution with zero mean and unit variance ($p = \mathcal{U}(-\sqrt{3}, \sqrt{3})$); q is a Laplace distribution with zero mean and unit variance.





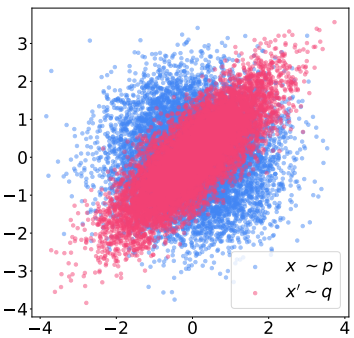
Results:

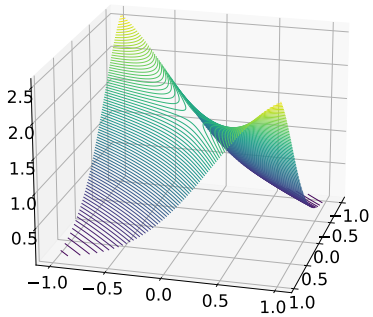




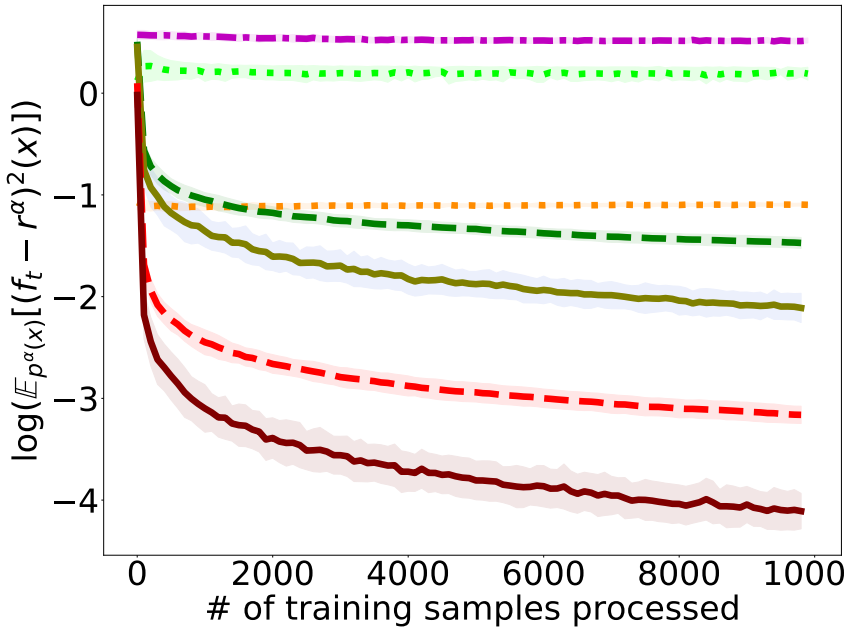
Experiments

Experiment II: p is a bivariate Gaussian distribution with zero mean, and a covariance matrix equal to the identity matrix; q has zero mean and covariance matrix such that $\Sigma_{1,1} = \Sigma_{2,2} = 1$, $\Sigma_{1,2} = \frac{4}{5}$.



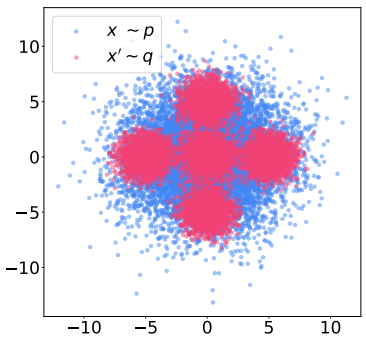


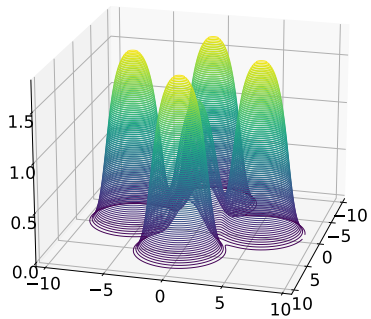
Results:



Experiments

Experiment III: p is bivariate Gaussian distribution with mean vector μ and covariance matrix $\Sigma_1 = 10 \times_{2 \times 2}$, and q is a mixture of five bivariate Gaussian distributions with the same covariance matrix.





Results:

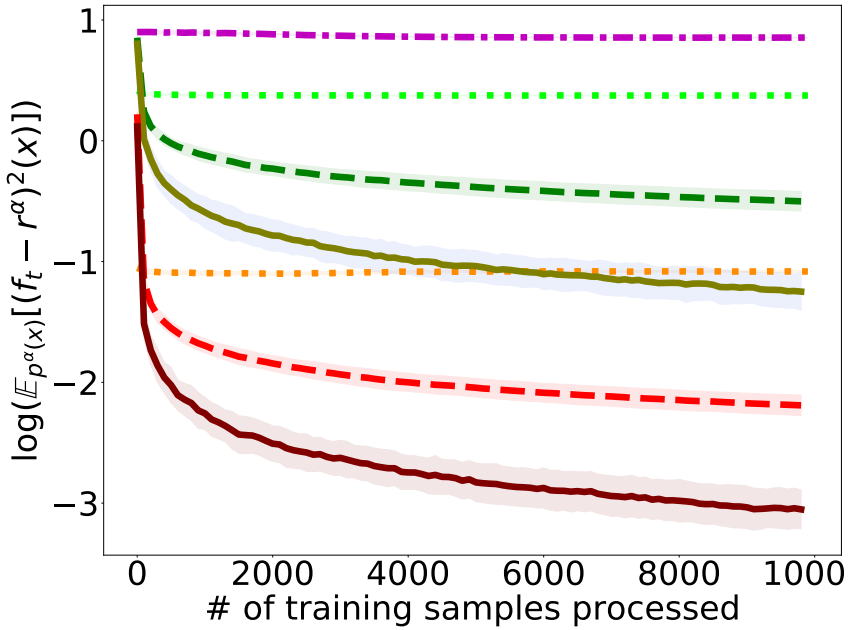


Table of contents

Conclusions

- OLRE does not require knowing in advance the sample size, which can be even infinite
- Our stochastic approximation aims at minimizing the generalization error directly avoiding over-fitting
- The cost of the iteration at time t is $O(t)$, hence in total $O(t^2)$ for up to time t
- Our convergence results provide guidelines on how to select the OLRE's hyperparameters
- OLRE outperforms other likelihood-ratio estimators

More details in the paper [de la Concha et al., 2023]

Thank you

Acknowledgments

This work was supported by the Industrial Data Analytics and Machine Learning Chair hosted at ENS Paris-Saclay, University Paris-Saclay, and grants from Région Ile-de-France.

References

Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, γ)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 23(39):1–70, 2022.

Imre Csiszár. On topological properties of f-divergences. *Studia Scientiarum Mathematicarum Hungarica*, 2:329—-339, 1967.

Alejandro de la Concha, Nicolas Vayatis, and Argyris Kalogeratos. Online non-parametric likelihood-ratio estimation by pearson-divergence functional minimization, 2023.

George S. Fishman. *Monte Carlo*. Springer New York, 1996.

Jerzy Neyman, Egon Sharpe Pearson, and Karl Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231 (694-706):289–337, 1933.

XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, 2008.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954.

Karl Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

A. N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1):22–46, 1963.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, 2007.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems*, 2011.