A Bayesian view to online sparse regression

CEA Postdoc Seminar

Kostas Themelis
Bat. 709, p. 277

Saclay, October 24, 2017

1. About me

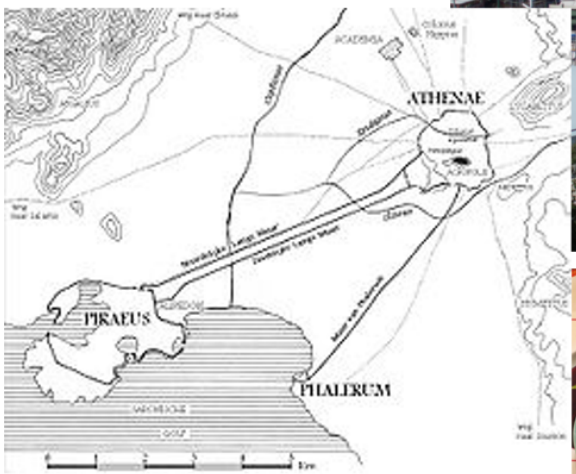
2. Introduction to Bayesian data analysis
   - Bayesian models
   - Variational Bayes inference

3. Variational Bayesian online parameter estimation
   - Online sparse regression
   - Proposed method
   - Experimental results

4. Work at CEA
   - Weak gravitational lensing

About me

born in Piraeus (Pireás), Greece



About me

- I am a postdoctoral researcher at the Cosmostat lab in CEA
- before that I was a postdoctoral researcher at the National Observatory of Athens in 2012-2017
- my area of expertise is in statistical signal processing
- my PhD thesis was on *Bayesian signal processing techniques for hyperspectral image unmixing*, completed at the University of Athens in 2012
- my research interests are in variational algorithms for approximate Bayesian inference with application to image processing
- I don't need a website, themelis.github.io

Bayesian data analysis

Let us assume a given set $\mathbf{y}$ of data points, and a suitable structural parametric model $\mathcal{M}$ that describes these data.

In the Bayesian framework,

- we treat the model parameters $\boldsymbol{\theta}$ as random variables, i.e., we assume some suitable $p(\boldsymbol{\theta})$
- we use Bayes' theorem to compute the posterior distribution of our model parameters

Bayes' theorem

- is mathematically expressed as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

- it can be seen as an inversion procedure expressed in a probabilistic way

Bayesian data analysis

Advantages

- Bayes' theorem provides a direct way to infer model parameters
- we are able to extract confidence intervals for our inferred parameter estimates

Challenges

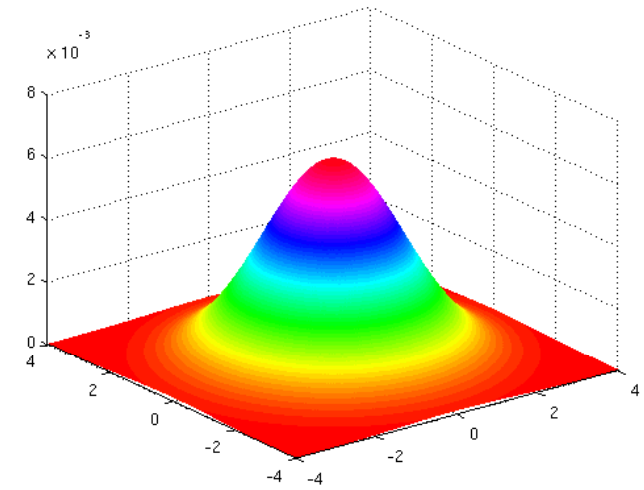
- subjectivity, different priors give rise to different posteriors
- difficult to express prior beliefs in prior probabilities $\rightarrow$ complex Bayesian models
- inference is computationally intensive, the data evidence

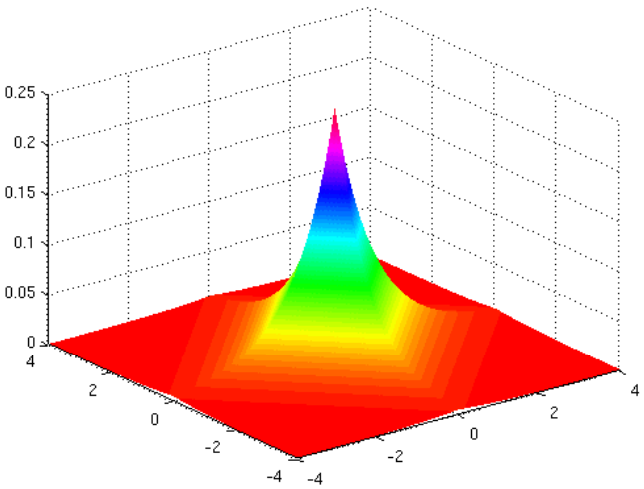
$$p(\mathbf{y}) = \int p(\mathbf{y}, \boldsymbol{\theta}) d\boldsymbol{\theta}$$

is intractable in most cases, hence, posterior approximations are largely based on sampling $\rightarrow$ MCMC

Bayesian modeling

deterministic vs probabilistic modeling





Deterministic scenario

Maximum likelihood estimation

- Tikhonov regularization:

$$\min_{\boldsymbol{\theta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \frac{\sigma_\eta^2}{2\sigma_{\boldsymbol{\theta}}^2} \|\boldsymbol{\theta}\|_2^2 \right\}$$

- sparse ℓ_1-norm regularization:

$$\min_{\boldsymbol{\theta}} \left\{ \frac{1}{2\sigma_\eta^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \right\}$$

Probabilistic scenario

Maximum a posteriori (MAP) estimation

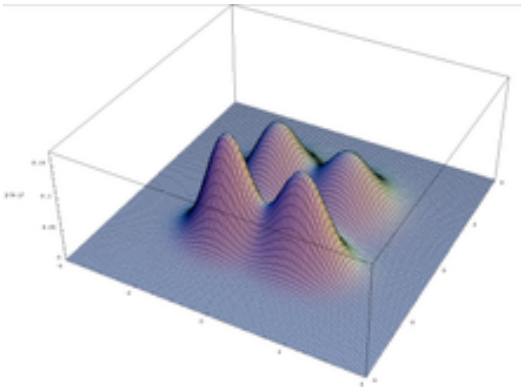
$$\min_{\boldsymbol{\theta}} \{-\log p(\mathbf{y}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\}$$

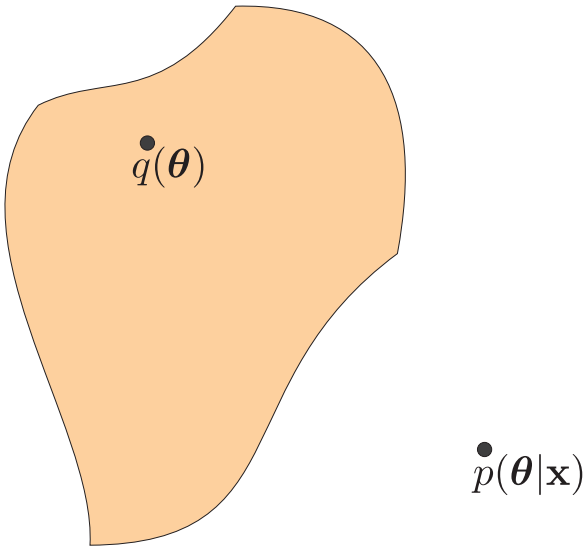
Gaussian prior $\rightarrow p(\boldsymbol{\theta}) \propto \exp\left[-\|\boldsymbol{\theta}\|_2^2/(2\sigma_\theta^2)\right]$

Laplace prior $\rightarrow p(\boldsymbol{\theta}) \propto \exp\left[-|\boldsymbol{\theta}|/b\right]$

Variational Bayes (VB)

...a fast alternative to MCMC





VB tranforms the statistical problem of computing $p(\boldsymbol{\theta}|\mathbf{y})$ to an optimization one

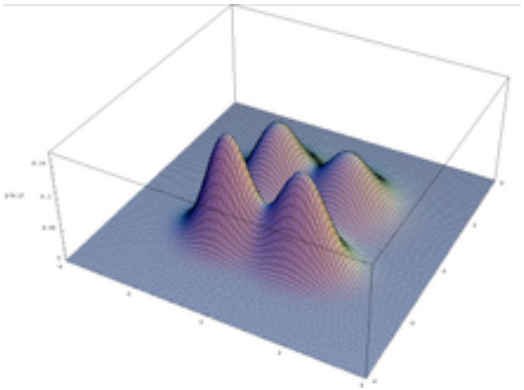
how?

- first, we assume an approximating pdf $q(\boldsymbol{\theta})$ for the true posterior $p(\boldsymbol{\theta}|\mathbf{y})$
- $q(\boldsymbol{\theta})$ is of specific form, e.g., it may belong to the exponential family
- in VB we minimize the Kullback-Leibler divergence $KL(q||p(\boldsymbol{\theta}|\mathbf{y}))$ among these two distributions

$$KL(q||p(\boldsymbol{\theta}|\mathbf{y})) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})} d\boldsymbol{\theta}$$

Variational Bayes (VB)

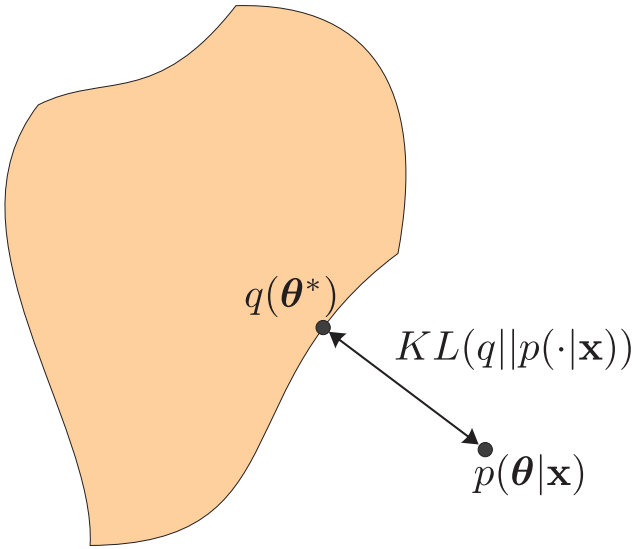
...a fast alternative to MCMC



...the task is to minimize

$$KL(q||p(\boldsymbol{\theta}|\mathbf{y})) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})} d\boldsymbol{\theta},$$

but is this possible? isn't $p(\boldsymbol{\theta}|\mathbf{y})$ unknown?



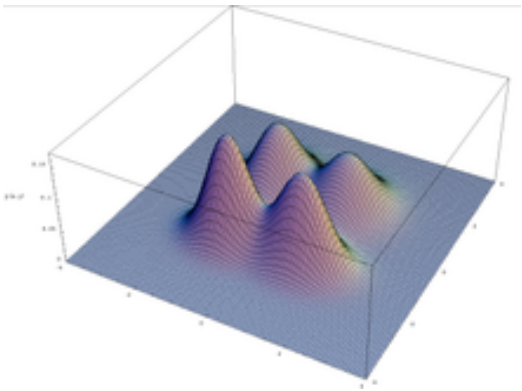
- $KL(q||p)$ can be expressed as

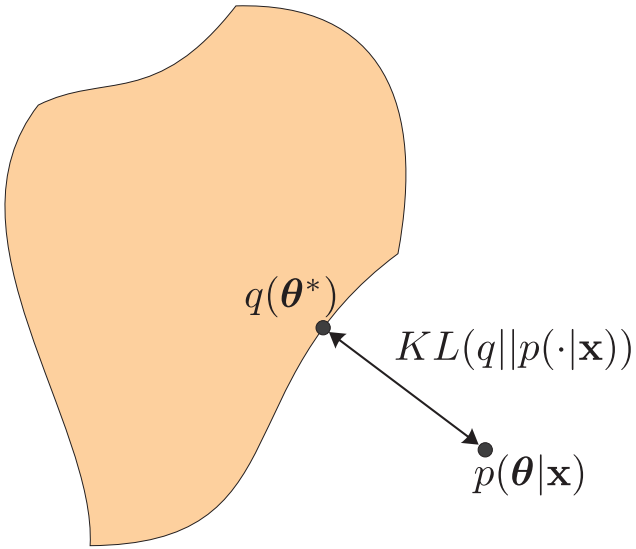
$$KL(q||p) = -(\mathbb{E}_q[p(\mathbf{y}, \boldsymbol{\theta})] - \mathbb{E}_q[q(\boldsymbol{\theta})]) + \log p(\mathbf{y})$$

- maximizing the evidence lower bound (ELBO) is equivalent to minimizing the $KL(q||p)$ divergence
- still a difficult problem to solve...

Variational Bayes (VB)

...a fast alternative to MCMC





...but if we assume independent approximating factors, based on the mean-field theory,

$$q(\boldsymbol{\theta}) = \prod_{n=1}^{N} q(\theta_n)$$

we can symplify further and compute a closed form solution for this problem, which is given by,

variational Bayes update

$$q(\theta_n) \propto \exp\left(\mathbb{E}_{q_{\neg n}}\left[\log p(\mathbf{y}, \boldsymbol{\theta})\right]\right)$$

where all factors $q(\theta_n), n = 1, 2, \ldots, N$, are updated in a sequential order

A simple linear regression example

Bayesian model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\eta}|\mathbf{0}, \beta^{-1}\mathbf{I})$$

Likelihood

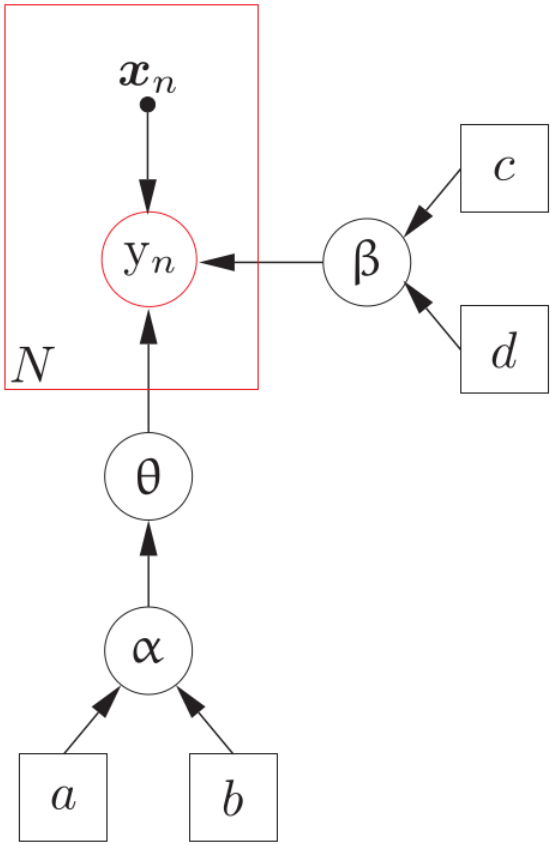
$$p(\mathbf{y}|\boldsymbol{\theta}, \beta) = \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\theta}, \beta^{-1}\mathbf{I})$$

Conjugate priors

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \prod_{k=1}^{K} \mathcal{N}(\theta_k|0, \alpha_k^{-1})$$

$$p(\beta) = \mathcal{G}(\beta|c, d)$$

$$p(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \mathcal{G}(\alpha_k|a, b)$$



Graphical illustration of dependencies among model parameters

A simple linear regression example

Variational Bayesian inference

Approximation

$$q(\boldsymbol{\theta}, \boldsymbol{\alpha}, \beta) = q(\boldsymbol{\theta})q(\boldsymbol{\alpha})q(\beta)$$

Approximate posteriors

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$$

$$q(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \mathcal{G}(\alpha_k|\tilde{a}, \tilde{b})$$

$$q(\beta) = \mathcal{G}(\beta|\tilde{c}, \tilde{d})$$

Batch variational Bayes algorithm

Initialize $\langle\boldsymbol{\alpha}\rangle, \langle\beta\rangle$

Set a, b, c, d to very small values

for $t = 1, 2, \ldots$

$$\boldsymbol{\Sigma}_\theta = (\mathbf{A} + \langle\beta\rangle \mathbf{X}^T\mathbf{X})^{-1}$$

$$\boldsymbol{\mu}_\theta = \langle\beta\rangle \boldsymbol{\Sigma}_\theta \mathbf{X}^T\mathbf{y}$$

$$\langle\beta\rangle = \frac{c + \frac{N}{2}}{d + \frac{1}{2}\langle\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2\rangle}$$

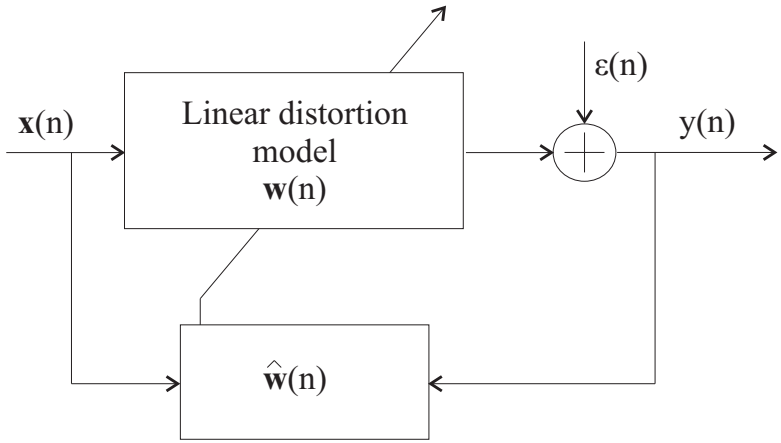
for $k = 1, 2, \ldots, K$

$$\langle\alpha_k\rangle = \frac{a + \frac{1}{2}}{b + \frac{1}{2}\langle\theta_k^2\rangle}$$

end for

end for

Online sparse regression



- input-output relation: $\boxed{y(n) = \mathbf{x}^T(n)\mathbf{w}(n) + \varepsilon(n)}$
- Goal: recursively estimate and track $\mathbf{w}(n)$ in time as new observations and input data pairs $\{y(n), \mathbf{x}(n)\}$ become available
- the recursive LS (RLS) algorithm solves the normal equations recursively in time in $O(N^2)$

Online sparse regression optimization function

$$\hat{\mathbf{w}}_{\ell_1}(n) = \arg\min_{\mathbf{w}(n)} \left[||\mathbf{\Lambda}^{\frac{1}{2}}(n)(\mathbf{y}(n) - \mathbf{X}(n)\mathbf{w}(n))||^2 + \tau||\mathbf{w}(n)||_1 \right]$$

Adaptive variational Bayes algorithm

Batch mode vs ...

- a set of observed data is available. Let $\boldsymbol{\theta} = [w_1, \ldots, w_N, \beta, \alpha_1, \ldots, \alpha_N, b_1, \ldots, b_N]^T$
- update each variational parameter θ_i sequentially until convergence
- each variational update minimizes the KL distance between $q(\theta_i)$ and $p(\theta_i|\mathbf{y}, \boldsymbol{\theta}_{\neg i})$ and increases monotonically the evidence lower bound

$$\log p(\mathbf{y}) \geq \mathbb{E}_q[\log p(\mathbf{y}, \boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{\theta})]$$

... adaptive mode

- a new data pair $\{\mathbf{x}(n), y(n)\}$ becomes available at time instant n
- perform a single update cycle on all variational parameters $\theta_i(n)$'s
- utilize the latest variational parameters, i.e.,

$$[\theta_1(n), \ldots, \theta_{i-1}(n), \theta_{i+1}(n-1), \ldots, \theta_{N_q}(n-1)]^T$$

- the evidence lower bound does not always increase (it gradually becomes accurate as learning proceeds)

Online sparse regression

Batch variational Bayes

Initialize $\langle \boldsymbol{\alpha} \rangle, \langle \beta \rangle$
Set $a, b, c, d \approx 0$
for $t = 1, 2, \ldots$

$$\boldsymbol{\Sigma}_\theta = (\mathbf{A} + \langle \beta \rangle \mathbf{X}^T \mathbf{X})^{-1}$$

$$\boldsymbol{\mu}_\theta = \langle \beta \rangle \boldsymbol{\Sigma}_\theta \mathbf{X}^T \mathbf{y}$$

$$\langle \beta \rangle = \frac{c + \frac{N}{2}}{d + \frac{1}{2} \langle \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 \rangle}$$

for $k = 1, 2, \ldots, K$

$$\langle \alpha_k \rangle = \frac{a + \frac{1}{2}}{b + \frac{1}{2} \langle \theta_k^2 \rangle}$$

end for
end for

Online variational Bayes

Initialize $\lambda, \hat{\mathbf{w}}(0), \mathbf{A}(-1), \mathbf{A}(0), \mathbf{R}(0), \mathbf{z}(0), d(0), \boldsymbol{\sigma}(0)$
Set $c, a, \rho, \delta, \kappa, \nu$ to very small values (10^{-6})
for $n = 1, 2, \ldots$

$$\mathbf{R}(n) = \lambda \mathbf{R}(n-1) + \mathbf{x}(n)\mathbf{x}^T(n) - \lambda \mathbf{A}(n-2) + \mathbf{A}(n-1)$$

$$\mathbf{z}(n) = \lambda \mathbf{z}(n-1) + \mathbf{x}(n) y(n)$$

$$d(n) = \lambda d(n-1) + y^2(n)$$

$$\beta(n) = \frac{N + (1-\lambda)^{-1} + 2\rho}{2\delta + d(n) - \mathbf{z}^T(n)\hat{\mathbf{w}}(n-1) + \mathbf{r}^T(n)\boldsymbol{\sigma}(n-1)}$$

for $i = 1, 2, \ldots, N$

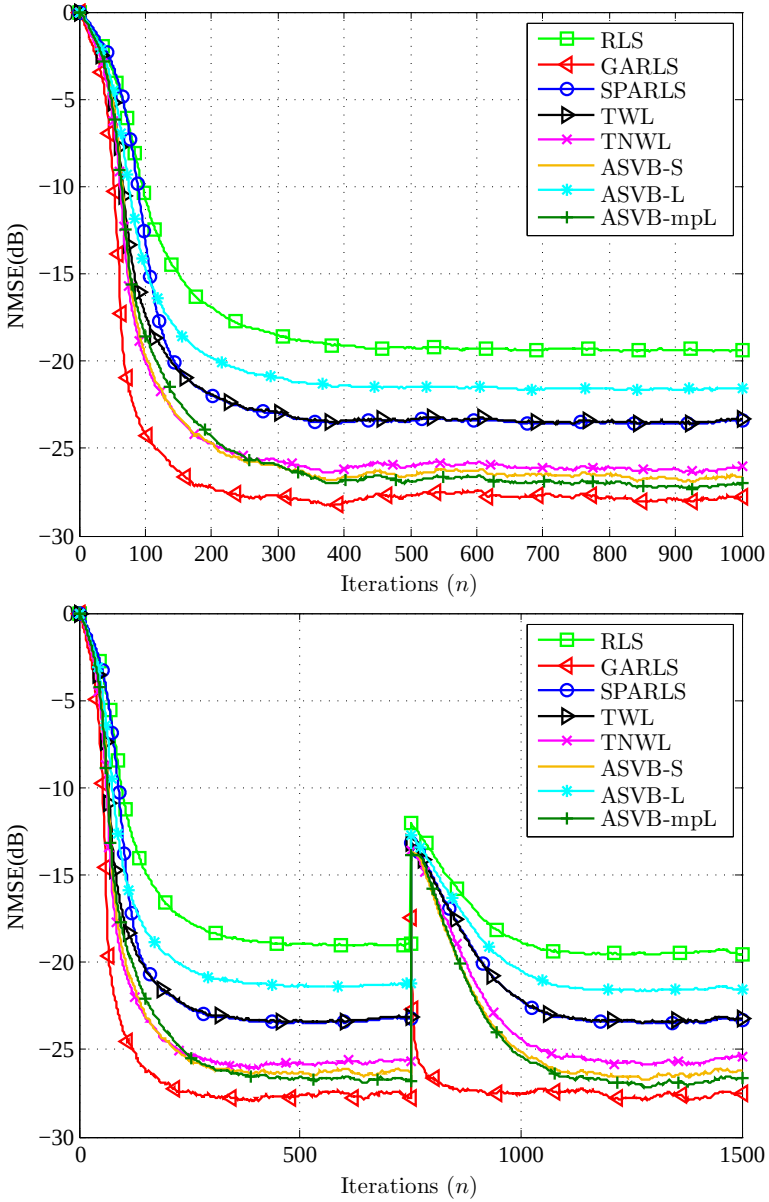
$$\sigma_i^2(n) = 1/(\beta(n) r_{ii}(n))$$

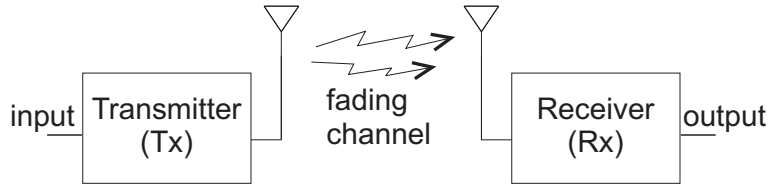
$$\hat{w}_i(n) = \frac{z_i(n) - \mathbf{r}_{\neg i}^T(n)\hat{\mathbf{w}}_{\neg i}(n)}{r_{ii}(n)}$$

$$\alpha_i(n) = \frac{2c + 1}{a + \beta(n)\hat{w}_i^2(n) + r_{ii}^{-1}(n)}$$

end for
end for

Simulations





Normalized mean square error (NMSE):

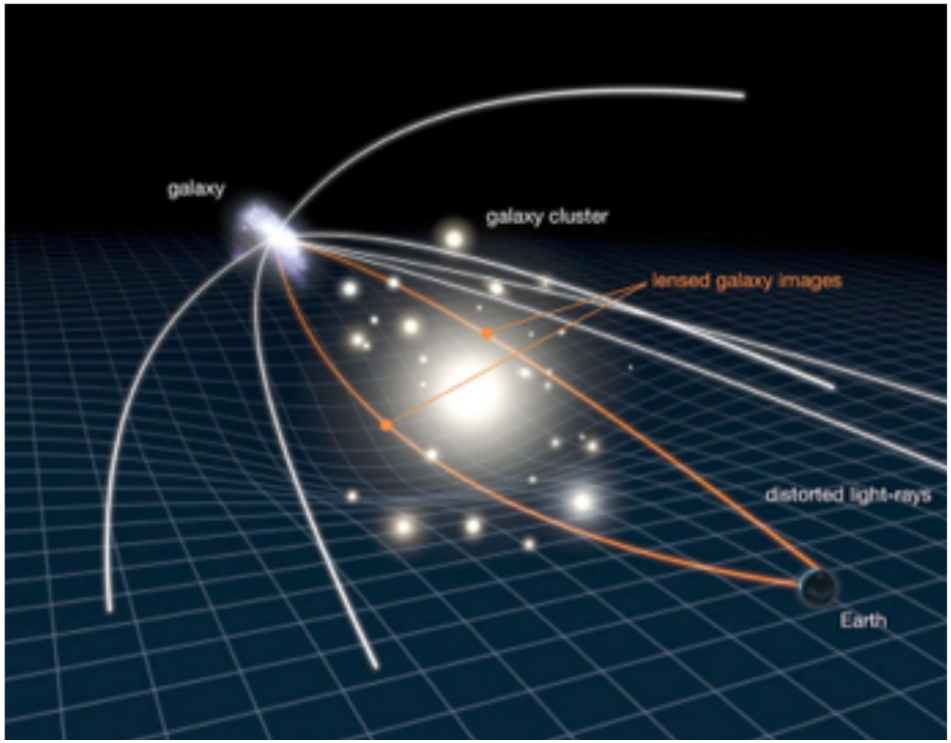
$$\mathrm{NMSE} = \frac{\left\langle \|\mathbf{w}(n) - \hat{\mathbf{w}}(n)\|^2 \right\rangle}{\left\langle \|\mathbf{w}(n)\|^2 \right\rangle}$$

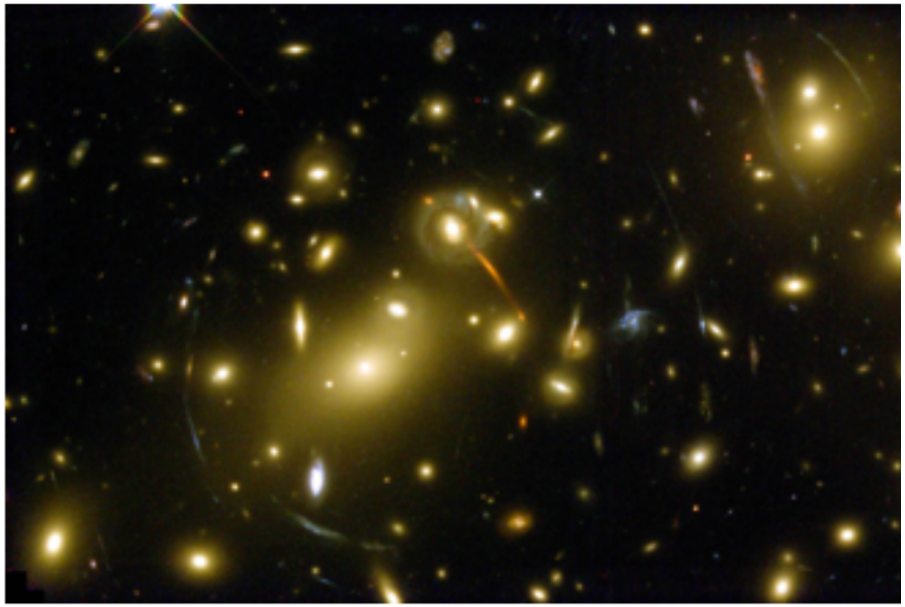
Adaptive filtering setup:

- 64-length time-varying channel.
- 8 nonzero coefficients, binary input.
- SNR is set to 15dB.
- A non-zero coefficient is added at the 750th time mark.

Weak gravitational lensing

Description





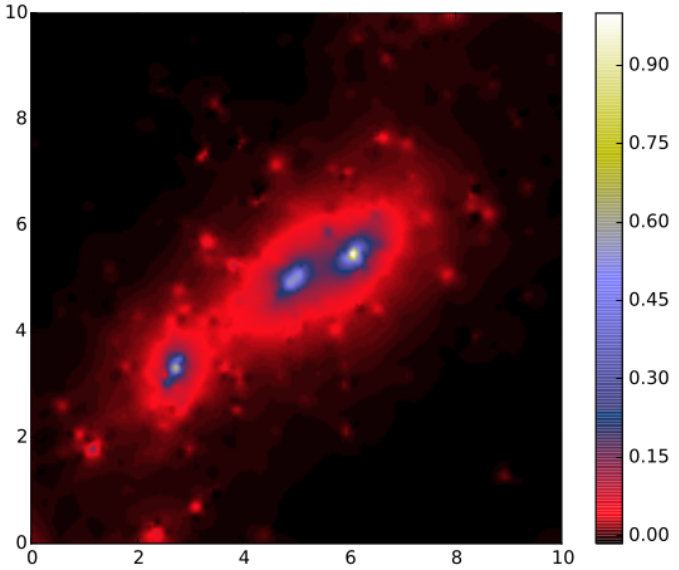
Credit: ESA/NASA

- light emitted by distant galaxies is curved, depending on the matter distribution in its path towards earth
- taking advantage of a plethora of small measured distortions, called 'shear', we can estimate the underlying matter distribution

Weak gravitational lensing

...towards a new version of GLIMPSE

- objective: estimate the convergence $\boldsymbol{\kappa}$, based on observed shear data $\boldsymbol{\gamma}$
- this is an ill-posed inverse problem



Credit: Lanusse et al, 2016

GLIMPSE algorithm

- provides a 2D or 3D reconstruction
- requires no binning of the galaxy shear
- imposes sparsity in a wavelet basis
- takes into account the reduced shear

$$\min_{\boldsymbol{\kappa}} \left\{ \frac{1}{2} \|\Sigma_{\gamma}^{-\frac{1}{2}} [\boldsymbol{\gamma} - \mathbf{TPF}^T \boldsymbol{\kappa}\|_2^2 + \lambda \|\mathbf{w} \circ \boldsymbol{\Phi}^T \boldsymbol{\kappa}\|_1 + i_{\mathcal{I}(\cdot)=0}(\boldsymbol{\kappa}) \right\}$$

$$\boldsymbol{\kappa} = \boldsymbol{\kappa}_{NL} + \boldsymbol{\kappa}_G$$

Thank you!