ICASSP 2016

Low-rank Matrix Recovery via Entropy Function

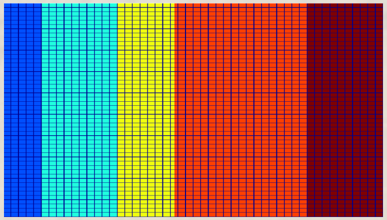
Dung N. Tran[1], Shuai Huang[1], Sang Chin[2] and Trac D. Tran[1]

[1]The Johns Hopkins University
[2]Draper Laboratory and Boston University

Low-rank matrix recovery

- ***Main assumption***: input matrix is low-rank

$$\text{rank}(\boldsymbol{X}) = r \ll \min\{n_1, n_2\}$$



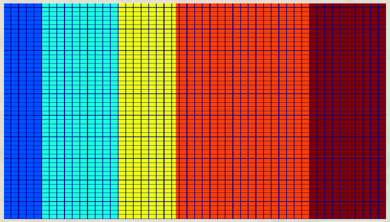
$\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$

- ***Measurements:***

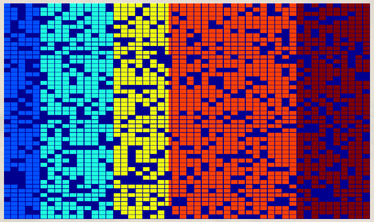
$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}) + \boldsymbol{\epsilon}$$

  - $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^m$: linear sampling operator
  - $\boldsymbol{y} \in \mathbb{R}^m$: measurement vector
  - $\boldsymbol{\epsilon} \in \mathbb{R}^m$: noise vector
  - $m < n_1 n_2$

- ***Reconstruction:*** given $(\boldsymbol{y}, \boldsymbol{A})$, recover $\boldsymbol{X}$.



$\boldsymbol{X}$

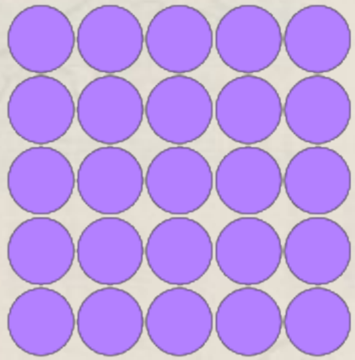


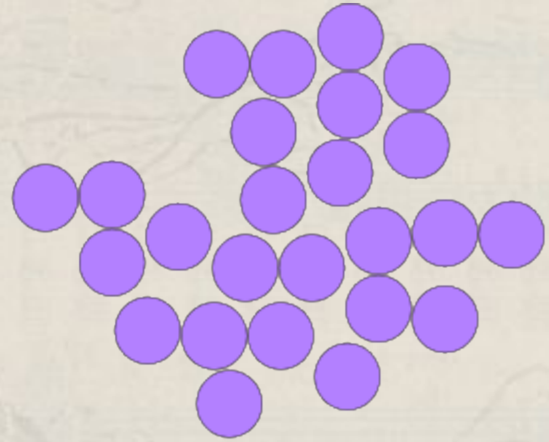
$\boldsymbol{y}$

$\boldsymbol{X}$?

Images courtesy of the research group of Prof. Yi Ma at UIUC.

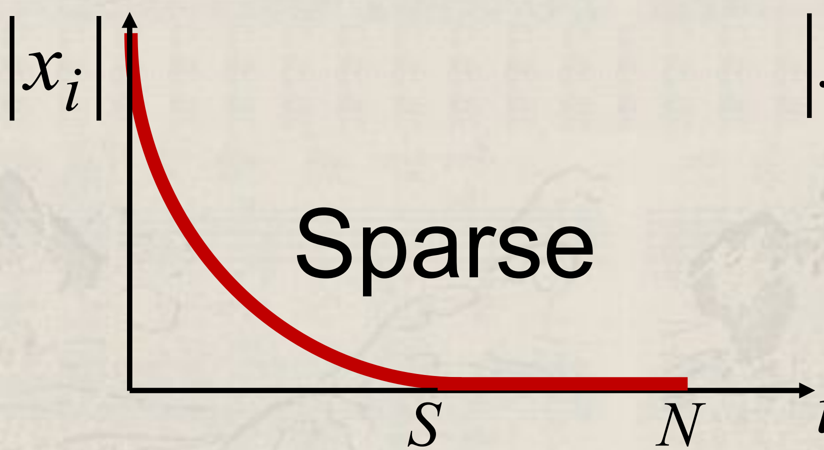
Entropy

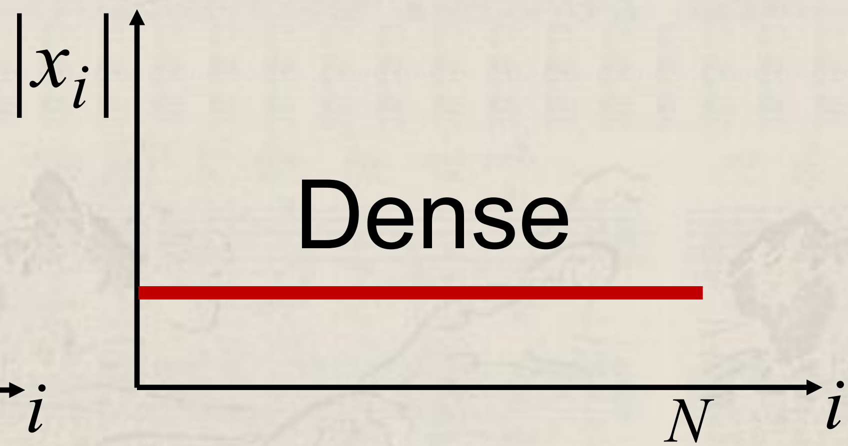




Low Entropy | High Entropy

Low Rank | High Rank





Entropy Minimization

- ***Entropy Minimization for low-rank matrix recovery***

$$\min_{\boldsymbol{X}} \ h(\boldsymbol{\sigma}(\boldsymbol{X})) \quad \text{s.t.} \quad \mathcal{A}(\boldsymbol{X}) = \boldsymbol{y},$$

where

$h(\boldsymbol{x}) = -\sum_i \frac{|x_i|}{\|\boldsymbol{x}\|_1} \log \frac{|x_i|}{\|\boldsymbol{x}\|_1}$ is the entropy function,

$\boldsymbol{\sigma}(\boldsymbol{X}) = (\sigma_1(\boldsymbol{X}), \ldots, \sigma_n(\boldsymbol{X}))$ is the *vector of singular values* of $\boldsymbol{X}$.

- ***Questions of interest:***
  - Why entropy minimization? ➔ *Sparsity inducing property*
  - How to solve it? ➔ *ENM algorithm*
  - What do we gain? ➔ *Faster sampling rate (show empirically)*
  - Why does it work theoretically? ➔ *(Future work)*

Entropy Function Induces Sparsity

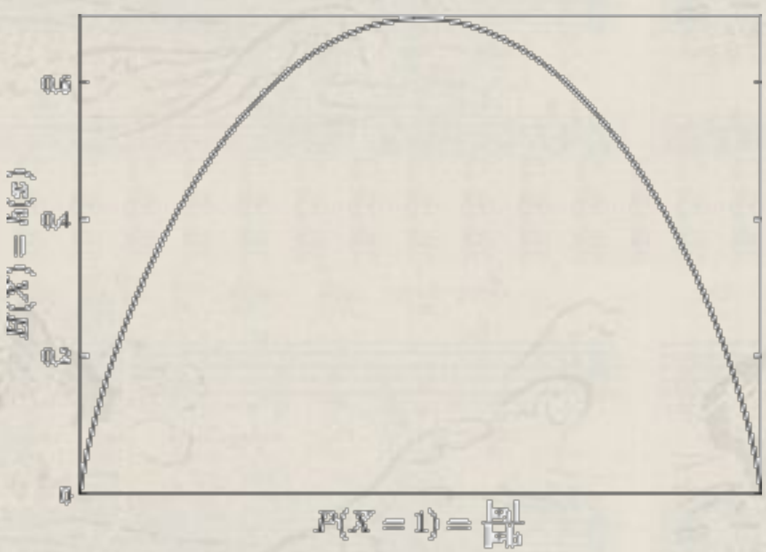
Recall $h(\boldsymbol{x}) = -\sum_i \frac{|x_i|}{\|\boldsymbol{x}\|_1} \log \frac{|x_i|}{\|\boldsymbol{x}\|_1}$ for $\boldsymbol{x} \in \mathbb{R}^n$

Let X : be a discrete random variable with possible values $\{1, \ldots, n\}$:

$$P(X = i) = \frac{|x_i|}{\|\boldsymbol{x}\|_1}$$

$\Longrightarrow \{\frac{|x_1|}{\|\boldsymbol{x}\|_1}, \ldots, \frac{|x_n|}{\|\boldsymbol{x}\|_1}\}$ is the distribution of X and $H(X) = h(\boldsymbol{x})$

Here, $H(\boldsymbol{X})$ is the Shannon entropy of $\boldsymbol{X}$.



Example: $\boldsymbol{x} \in \mathbb{R}^2$

$h(\boldsymbol{x}) = H(X)$ attains its maximum when $x_1 = x_2$ whereas its minima occur when $\boldsymbol{x}$ is 1 sparse

Entropy Function Induces Sparsity

Consider a nonnegative diagonal matrix $\boldsymbol{X}$

Let $\boldsymbol{x} = \mathrm{diag}(\boldsymbol{X})$, then $\boldsymbol{x} = \boldsymbol{\sigma}(\boldsymbol{X})$. Assume $\boldsymbol{x}$ is sparse.

$$\Rightarrow \quad \min_{\boldsymbol{x}} \; h(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}.$$

Lemma. *If there exists two solutions* $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ *to* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$*, with* $\boldsymbol{b} \neq \boldsymbol{0}$*, in the same d-dimensional orthant* ($d \leq n$)*, then there is at least one solutions* $\boldsymbol{x}'$ *in some* d'*-dimensional orthant such that* $d' < d$ *and* $h(\boldsymbol{x}') < \min\{h(\boldsymbol{x}_1), h(\boldsymbol{x}_2)\}$.

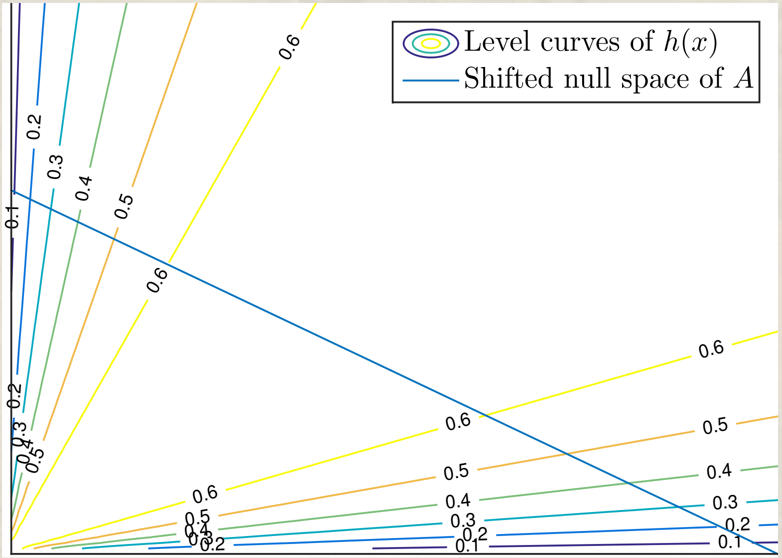


Fig. Illustration of Lemma 1: minimum entropy occurs at 1-sparse solution.

ENM Algorithm

- ***Robust variant:***

$$\min_{\boldsymbol{X}} \lambda h(\boldsymbol{\sigma}(\boldsymbol{X})) + f(\boldsymbol{X}; \mathcal{A}, \boldsymbol{y}),$$

for some loss function $f : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}_+$ with Lipchitz continuous gradient.

- ***Technique: linearization***

At current estimate $\boldsymbol{\sigma}^t$:

$$h(\boldsymbol{\sigma}) \approx h(\boldsymbol{\sigma}^t) + \nabla h(\boldsymbol{\sigma}^t)^T(\boldsymbol{\sigma} - \boldsymbol{\sigma}^t)$$

$$f(\boldsymbol{X}) \approx f(\boldsymbol{X}^t) + \nabla f(\boldsymbol{X}^t)^T(\boldsymbol{X} - \boldsymbol{X}^t) + \frac{\rho}{2}\|\boldsymbol{X} - \boldsymbol{X}^t\|_F^2,$$

$$\begin{aligned}\Longrightarrow \boldsymbol{X}^{t+1} &= \operatorname*{argmin}_{\boldsymbol{X}} \lambda \nabla h(\boldsymbol{\sigma}^t)^T \boldsymbol{\sigma} + f(\boldsymbol{X}^t) \\ &\quad + \nabla f(\boldsymbol{X}^t)^T(\boldsymbol{X} - \boldsymbol{X}^t) + \frac{\rho}{2}\|\boldsymbol{X} - \boldsymbol{X}^t\|_F^2 \\ &= \operatorname*{argmin}_{\boldsymbol{X}} \lambda \nabla h(\boldsymbol{\sigma}^t)^T \boldsymbol{\sigma} + \frac{\rho}{2}\left\|\boldsymbol{X} - \left(\boldsymbol{X}^t - \frac{1}{\rho}\nabla f(\boldsymbol{X}^t)\right)\right\|_F^2.\end{aligned}$$

ENM Algorithm

$$\boldsymbol{X}^{t+1} = \operatorname*{argmin}_{\boldsymbol{X}} \lambda \nabla h(\boldsymbol{\sigma}^t)^T \boldsymbol{\sigma} + \frac{\rho}{2} \left\| \boldsymbol{X} - \left(\boldsymbol{X}^t - \frac{1}{\rho} \nabla f(\boldsymbol{X}^t) \right) \right\|_F^2 .$$

Lemma. *Let h be the entropy function, and let $\boldsymbol{\sigma}$ be a positive vector, then*

$$\frac{\partial h(\boldsymbol{\sigma})}{\partial \sigma_i} = -\frac{\log \sigma_i}{\|\boldsymbol{\sigma}\|_1} + \frac{\sum_j \sigma_j \log \sigma_j}{\|\boldsymbol{\sigma}\|_1^2}.$$

Lemma. *If $\sigma_1^t \geq \sigma_2^t \geq ... \geq \sigma_n^t \geq 0$, then $0 \leq \frac{\partial h(\boldsymbol{\sigma}^t)}{\partial \sigma_1} \leq \frac{\partial h(\boldsymbol{\sigma}^t)}{\partial \sigma_2} \leq ... \leq \frac{\partial h(\boldsymbol{\sigma}^t)}{\partial \sigma_n}$.*

Lemma. *Let $\lambda > 0$, $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, and $0 \leq w_1 \leq w_2 \leq ... \leq w_n$, where $n = \min\{n_1, n_2\}$. Let $\boldsymbol{X}^*$ be the optimal solution of the minimization problem*

$$\min_{\boldsymbol{X}} \lambda \sum_i w_i \sigma_i(\boldsymbol{X}) + \frac{1}{2} \|\boldsymbol{X} - \boldsymbol{Z}\|_F^2, \tag{1}$$

then

$$\boldsymbol{X}^* = \boldsymbol{U} \mathcal{D}_{\lambda \boldsymbol{w}}(\boldsymbol{\Sigma}) \boldsymbol{V}^T, \tag{2}$$

where $\boldsymbol{Z} = \boldsymbol{U \Sigma V}^T$ and $\mathcal{D}_{\lambda \boldsymbol{w}}(\boldsymbol{\Sigma}) = diag\{(\sigma_i - \lambda w_i)_+\}$ is the singular value shrinkage operator.

ENM Algorithm

Algorithm 1 ENtropy-Minimization (ENM)

input: measurements $(\mathcal{A}, \boldsymbol{y})$, $\lambda > 0$, and $\rho > L_f$.
initialization: $\boldsymbol{X}^0$.
while not converged **do**

Update the weights:

$$w_i^t = -\frac{\log \sigma_i^t}{\|\boldsymbol{\sigma}^t\|_1} + \frac{\sum_j \sigma_j^t \log \sigma_j^t}{\|\boldsymbol{\sigma}^t\|_1^2}, \; i = 1, ..., n \tag{1}$$

Update the estimate:

$$\boldsymbol{X}^{t+1} = \boldsymbol{U}\mathcal{D}_{(\lambda/\rho)\boldsymbol{w}}(\boldsymbol{\Sigma})\boldsymbol{V}^T, \tag{2}$$

where $\boldsymbol{X}^t - \frac{1}{\rho}\nabla f(\boldsymbol{X}^t) = \boldsymbol{U\Sigma V}^T$.

end while
output: Estimated solution $\hat{\boldsymbol{X}}$.

Experiment Results

- ***Random subsampling (matrix completion):***

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_* \quad \text{s.t.} \quad X_{ij} = M_{ij}, \quad (i,j) \in \Omega$$

- ***Synthetic data***

$\boldsymbol{X} \in \mathbb{R}^{100\times 100}$

$m = 0.5 n_1 n_2 = 5000$ samples

$r = \text{rank}(\boldsymbol{X})$ varies

$\frac{\|\hat{\boldsymbol{X}} - \boldsymbol{M}\|_F^2}{\|\boldsymbol{M}\|_F^2} < 10^{-3} \rightarrow$ successful

100 trials for each r

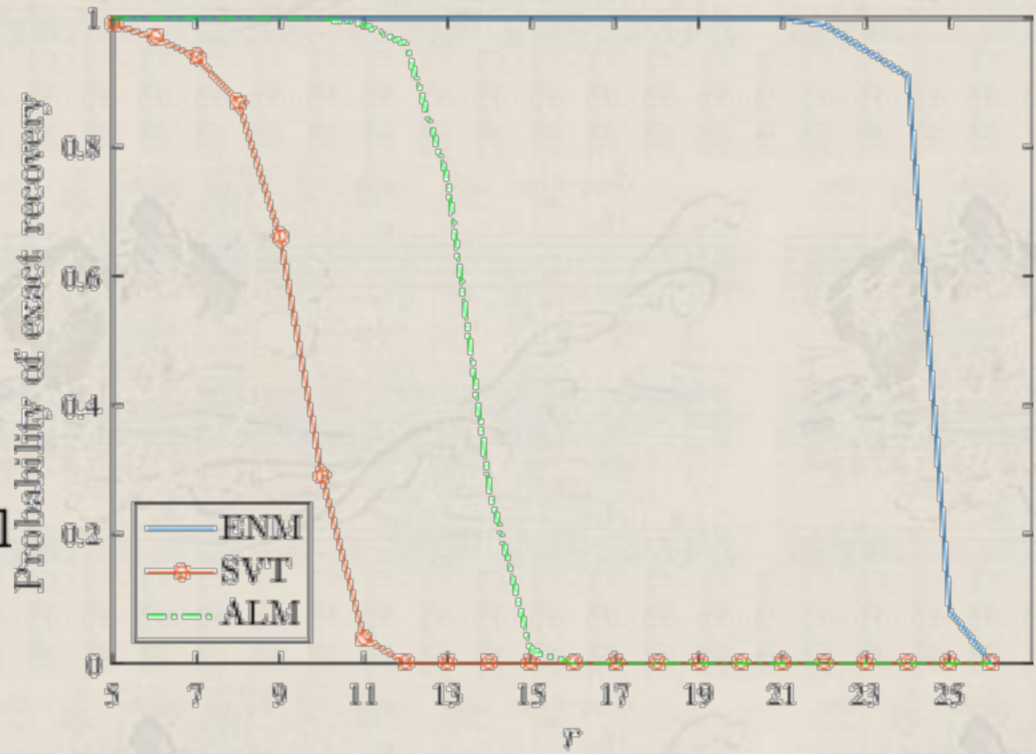


Fig. Probability of exact recovery on synthetic data.

Algorithms:

- ENtropy Minimization (ENM)
- Singular Value Thresholding (SVT) [Cai *et al.*]
- Augmented Lagrange Multiplier (ALM) [Lin *et al.*]

Experiment Results

- ***Random subsampling (matrix completion):***

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_* \quad \text{s.t.} \quad X_{ij} = M_{ij}, \quad (i,j) \in \Omega$$

- ***Real data: MIT Logo***

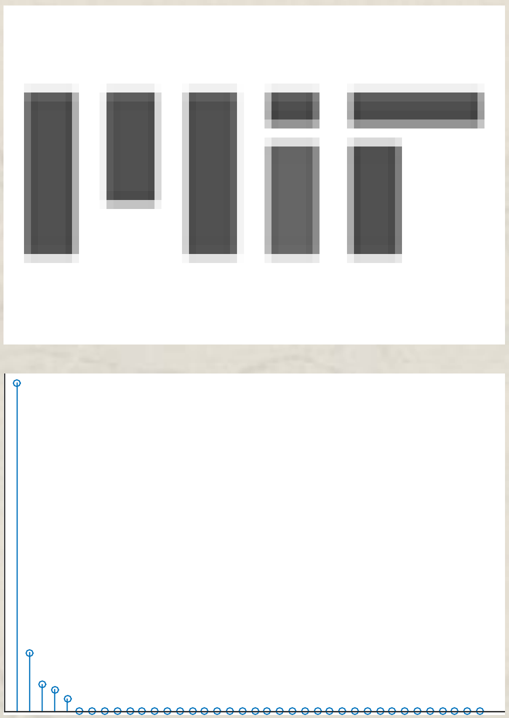


Fig. MIT logo and its singular values

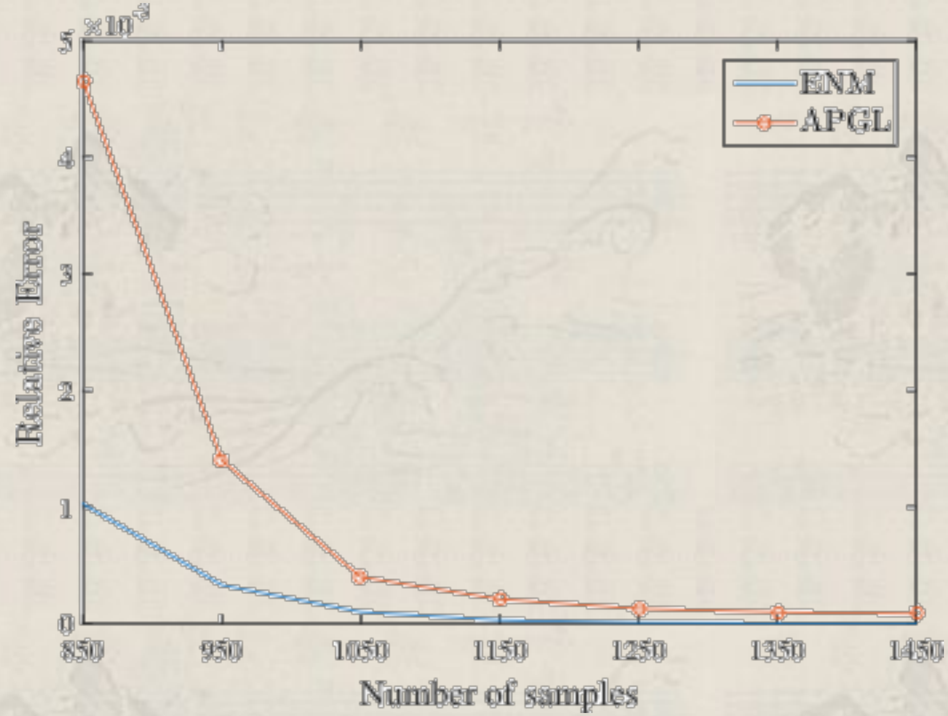


Fig. Low-rank matrix completion on MIT logo.

Algorithms:

- ENM vs. Accelerated Proximal Gradient with Line Search (APGL) [Toh & Yun]

What About Re-weighted ℓ_1?

- ENM

$$\min_{\mathbf{x}} h(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{y}$$

Not Separable in Its Parameters

- Reweighted ℓ_1

$$\min_{\mathbf{x}} \| \log(\mathbf{x}) \|_1 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{y}$$

- Common optimization problem

$$\min_{\mathbf{w},\mathbf{x}} \| \mathbf{w} \circ \mathbf{x} \|_1 + \frac{\mu}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$$

Key Difference

Discussion: ENM Weights

- Closed-form thresholds

$$w_i^t = -\frac{\log \sigma_i^t}{\left\|\boldsymbol{\sigma}^t\right\|_1} + \frac{\sum_j \sigma_j^t \log \sigma_j^t}{\left\|\boldsymbol{\sigma}^t\right\|_1^2}$$

- Threshold proven to be larger when associated singular value gets smaller ➔ lower-rank solution from shrinkage operation!
- ENM encourages singular values to have a Laplacian distribution

Final Discussion

- Both ENM and Re-weighted ℓ_1 can be solved with the same strategy
- ENM seems to empirically offer a better weighting scheme than Re-weighted ℓ_1
- Critical Questions:
  - How can we theoretically justify that?
  - Any other interesting applications of information theory tools/concepts to sparse problems?

Entropy Function Induces Sparsity

Lemma. *If there exists two solutions* $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ *to* $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, *with* $\boldsymbol{b} \neq \boldsymbol{0}$, *in the same d-dimensional orthant* ($d \leq n$), *then there is at least one solutions* $\boldsymbol{x}'$ *in some* d'*-dimensional orthant such that* $d' < d$ *and* $h(\boldsymbol{x}') < \min\{h(\boldsymbol{x}_1), h(\boldsymbol{x}_2)\}$.

Proof.(sketch) For any $\boldsymbol{x} \in \mathbb{R}^n$, define $\hat{\boldsymbol{x}} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1}$, then $h(\boldsymbol{x}) = h(\hat{\boldsymbol{x}})$.

Let $\boldsymbol{x}' = (1-\lambda)\boldsymbol{x}_1 + \lambda\boldsymbol{x}_2$, then there is an one-to one mapping between $\boldsymbol{x}'$ and $\hat{\boldsymbol{x}}'$.

Furthermore, $\hat{\boldsymbol{x}}' = (1-\hat{\lambda})\hat{\boldsymbol{x}}_1 + \hat{\lambda}\hat{\boldsymbol{x}}_2$, for some constant $\hat{\lambda}$.

As $h(\hat{\boldsymbol{x}}')$ is concave, its minima are archived at the boundaries of the current orthant.

This also true for $h(\boldsymbol{x}')$ by the one-to-one mapping between $\boldsymbol{x}'$ and $\hat{\boldsymbol{x}}'$.

Therefore, we can tune λ so that $\boldsymbol{x}'$ lies at a boundary of the current orthant and minimum entropy is archived.

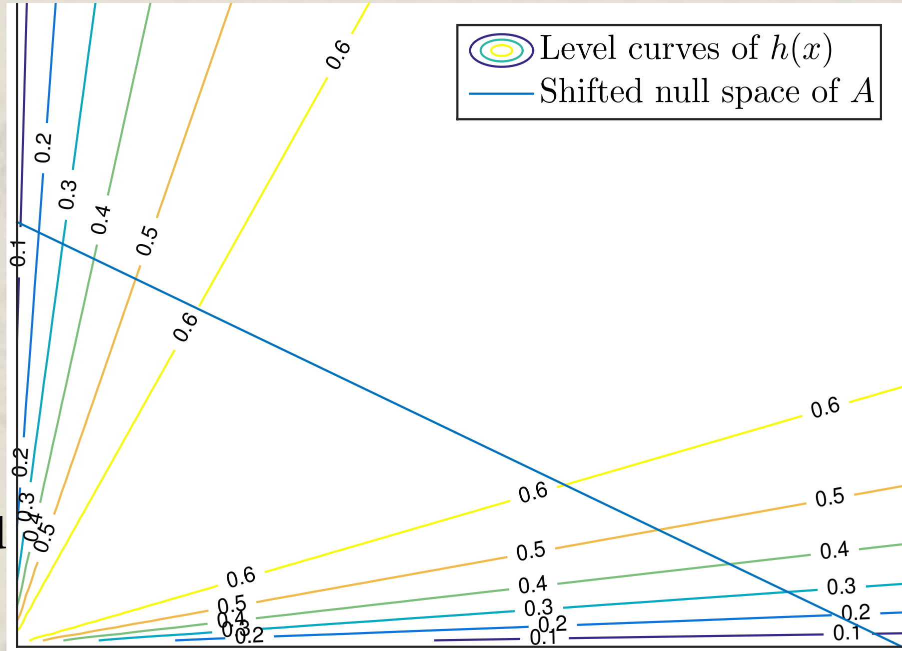


Fig. Illustration of Lemma 1: minimum entropy occurs at 1-sparse solution.