The maximum entropy on the mean method for linear inverse problems

Tim Hoheisel (McGill)

Joint work with Rustum Choksi (McGill), Ariel Goodwin (McGill), Carola-Bibiane Schönlieb (Cambridge), and Yakov Vaisbourd (McGill)

Based on earlier work with Gabriel Rioux (Cornell), Pierre Maréchal (Toulouse), and Christopher Scarvelis (MIT).

IAM Distinguished Colloquium
Vancouver, September 26, 2022

Higher level approach to linear inverse problems

The canonical linear inverse problem $Cx \approx b$ is usually solved via an optimization problem

$$\min_{x \in \mathbb{R}^d} \left\{ \frac{1}{2} \|Cx - b\|^2 + R(x) \right\}$$

- C: linear (forward) operator
- b: measurement vector
- R: (convex) regularizer

Higher level approach: Interpret the ground truth as a random vector with unknown distribution. Solve for a distribution Q that is close to a prior (guess) P and such that its expectation[1] E_Q satisfies $C \cdot E_Q \approx b$.

What is the information theoretic foundation for this?

Principle of Maximum Entropy: ”The probability distribution which is maximally non-committal with regard to missing information among all the distributions that agree with the present knowledge is the one with the maximum entropy.” (E.T. Jaynes, 1957)

[1] i.e. $E_Q = \int_\Omega y dQ(y)$

Measuring compliance: the KL divergence

Let P be a (prior) distribution, i.e. a probability measure on $\Omega \subset \mathbb{R}^n$.

The measure of compliance of another distribution Q with P is measured by the **Kullback-Leibler divergence** $\mathrm{KL}(\cdot \mid \cdot) : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega)^{2} \to \mathbb{R} \cup \{+\infty\}$,

$$\mathrm{KL}(Q \mid P) = \begin{cases} \int_\Omega \log\left(\frac{dQ}{dP}\right)\, dQ, & Q \ll P,^{3} \\ +\infty, & \text{otherwise,} \end{cases}$$

where $\frac{dQ}{dP}$ is the *Radon-Nikodym derivative*.

- $\mathrm{KL}(\cdot \mid \cdot)$ is convex, $\mathrm{KL}(\cdot \mid P)$ strictly convex for all $P \in \mathcal{P}(\Omega)$.
- $\mathrm{KL}(Q \mid P) \geq 0$; equality if and only if $Q = P$ a.e.

[2] $\mathcal{P}(\Omega)$: (convex) set of probability measures on Ω.
[3] $Q \ll P \;\; :\Longleftrightarrow \;\; P(A) = 0 \Rightarrow Q(A) = 0.$

KL divergence concretely

Let $P \in \mathcal{P}(\Omega)$ be our prior/reference distribution. We are mainly interested in two cases.

1. $\Omega = \mathbb{R}^n$ and P is absolutely continuous w.r.t. the Lebesgue measure μ, i.e. has a density $p = \frac{dP}{d\mu}$. In this case, if $Q \ll P$, Q has a density q, and
$$\mathrm{KL}(Q \mid P) = \int_{\mathbb{R}^n} \log\left(\frac{q(x)}{p(x)}\right) q(x) dx.$$
2. P is a discrete probability distribution, i.e. Ω is countable, and the probability mass function $p(x) = P(\{x\})$ has $\sum_{x \in \Omega} p(x) = 1$. Then $Q \ll P$ implies that Q has a probability mass function q and it holds that
$$\mathrm{KL}(Q \mid P) = \sum_{x \in \Omega} q(x) \log\left(\frac{q(x)}{p(x)}\right).$$

Example: Let P be the uniform distribution on $\Omega := \{1, \ldots, N\}$, i.e. $p(i) = 1/N$ for all $i = 1, \ldots, N$. Then for $Q \ll P$ with PMF q, we have
$$\mathrm{KL}(Q \mid P) = \log(N) + \sum_{i=1}^{N} \log(q(i))q(i).$$

The MEMM formulation and its dual

Given a prior $P \in \mathcal{P}(\Omega)$, the *maximum entropy on the mean method (MEMM)* for the linear inverse problem $Cx \approx b$ reads:

Determine $\bar{Q}$ as the solution of

$$\min_{Q \in \mathcal{P}(\Omega)} \left\{ \frac{\alpha}{2} \|C \cdot E_Q - b\|^2 + \mathsf{KL}(Q \mid P) \right\}, \tag{1}$$

and set $\bar{x} := E_{\bar{Q}}$ to be the estimate for the ground truth.

A dual approach for finding $\bar{x}$: Let $\psi_P : \mathbb{R}^d \to \mathbb{R}$ be given by the *cumulant generating function* of P, i.e.

$$\psi_P(y) = \log \int_\Omega \exp \langle y, \cdot \rangle \, dP = \log(M_P(y)).$$

Under suitable assumptions[4], the (Fenchel) dual of (1) reads (Rioux et al. '21):

$$\max_{\lambda \in \mathbb{R}^d} \left\{ \langle b, \lambda \rangle - \frac{1}{2\alpha} \|\lambda\|^2 - \psi_P(C^T \lambda) \right\}. \tag{2}$$

Given the maximizer $\bar{\lambda}$ of (2) one can recover $\bar{x}$ via $\bar{x} = \nabla \psi_P(C^T \bar{\lambda})$.

[4] E.g. Ω compact.

Applications

To solve the dual problem, one can use standard solvers like e.g. L-BFGS which was successfully done for (blind and non-blind) deblurring of

- *Barcodes/QR-codes*.
  Prior P: Bernoulli.
  Reference: G. Rioux et al.: *Blind Deblurring of Barcodes via Kullback-Leibler Divergence.* IEEE TPAMI 43(1), 2021, pp.77-88.
- *General images.*
  Prior P: Uniform on box.
  Reference: G. Rioux et al.: *The Maximum Entropy on the Mean Method for Image Deblurring. Inverse Problems 37, 2021*

Fig. 11. Out of focus image of a QR code.

Fig. 12. Result of applying our method to a processed version of Fig. 11.

[Rioux et al. (2021)]

The reformulated problem and the MEM functional

We observe that the (primal) MEMM problem can be reformulated as follows:

$$\inf_{Q \in \mathcal{P}(\Omega)} \left\{ \frac{\alpha}{2} \|C \cdot E_Q - b\|^2 + \mathsf{KL}(Q \mid P) \right\} \quad = \quad \inf_{y \in \mathbb{R}^d} \left\{ \frac{\alpha}{2} \|C \cdot y - b\|^2 + \underbrace{\inf_{\substack{Q \in \mathcal{P}(\Omega): \\ E_Q = y}} \mathsf{KL}(Q \mid P)}_{:= \kappa_P(y)} \right\}$$

We define the *MEM functional* $\kappa_P : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$,

$$\kappa_P(y) = \inf_{Q \in \mathcal{P}(\Omega)} \{\mathsf{KL}(Q \mid P) + \delta_{\{0\}}(E_Q - y)\}.$$

Then we obtain the *reformulated problem*

$$\min_{y \in \mathbb{R}^d} \frac{\alpha}{2} \|C \cdot y - b\|^2 + \kappa_P(y).$$

Since $\kappa_P \geq 0$, and $\kappa_P(y) = 0$ iff $y = E_P$, κ_P can be interpreted as a regularizer promoting proximity to the prior distribution.

Q: *Is this reformulation useful at all?*

Interlude: convex analysis[6] basics - the epigraphical perspective

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$.

- $\operatorname{dom} f := \left\{ x \in \mathbb{R}^d \mid f(x) < +\infty \right\}$ (domain);
- $\operatorname{epi} f := \left\{ (x, \alpha) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \leq \alpha \right\}$ (epigraph).

We call f

- *convex* if $\operatorname{epi} f$ is convex;
- *closed* (or *lower semicontinuous*) if $\operatorname{epi} f$ is closed;
- *proper* if $\operatorname{dom} f \neq \emptyset$.
- $\Gamma_0 := \left\{ f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\} \mid f \text{ closed, proper, convex} \right\}$.

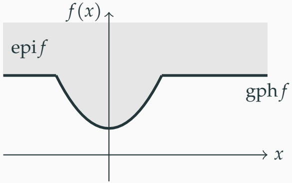


Figure 1: Epigraph of $f : \mathbb{R} \to \mathbb{R}$

Affine minorization principle: Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ convex and proper, and $\bar{x} \in \operatorname{ri}(\operatorname{dom} f)$[5]. Then there exists $v \in \mathbb{R}^n$ such that

$$f(x) \geq f(\bar{x}) + \langle v, x - \bar{x} \rangle \quad \forall x \in \mathbb{R}^n.$$

[5] The relative interior of a convex set is its interior in the relative topology w.r.t. its affine hull.

[6] *'What's dead may never die!'*

Interlude: convex analysis basics - the Fenchel conjugate

Let $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be the function whose epigraph encodes the affine minorants of $\operatorname{epi} f$ in that

$$\operatorname{epi} f^* \stackrel{!}{=} \{(v, \beta) \mid \langle v, x\rangle - \beta \le f(x) \quad \forall x \in \mathbb{R}^n\}.$$

Thus

$$f^*(v) \le \beta \quad \iff \quad \sup_{x \in \mathbb{R}^n} \{\langle v, x\rangle - f(x)\} \le \beta \quad \forall (v, \beta) \in \mathbb{R}^n \times \mathbb{R}.$$

Therefore

$$f^*(v) = \sup_{x \in \mathbb{R}^n} \{\langle v, x\rangle - f(x)\} \quad \forall v \in \mathbb{R}^n,$$

which is called the *(Fenchel) conjugate* of f. We set $f^{**} := (f^*)^*$.

- f^* closed and convex - proper if f has an affine minorant
- If f is convex and proper, then f^* is proper (closed, convex), and

$$f^{**}(x) = (\operatorname{cl} f)(x)^7.$$

- $f = f^{**} \quad \iff \quad f \in \Gamma_0 \quad$ (Fenchel-Moreau)

[7] $(\operatorname{cl} f) : x \in \mathbb{R}^n \mapsto \liminf_{z \to x} f(z)$, the closure of f, is the largest lsc minorant of f.

Cramér's function

Recall the *cumulant generating function* $\psi_P : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ of $P \in \mathcal{P}(\Omega)$, given by

$$\psi_P(\theta) := \log \int_\Omega \exp(\langle \theta, \cdot \rangle) dP = \log(M_P(\theta)).$$

The conjugate $\psi_P^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$,

$$\psi_P^*(y) := \sup_{\theta \in \mathbb{R}^d} \{\langle y, \theta \rangle - \psi_P(\theta)\}$$

is called *Cramér's function*[8] (fundamental in *large deviations theory*).

The key to computational tractability of the reformulated MEMM problem is to establish conditions (on P) under which Cramér's function equals the MEM functional, i.e.

$$\kappa_P = \psi_P^*.$$

Key ingredient: Exponential families and Legendre-type functions.

[8]Named after Swedish mathematician and statistician Harald Cramér who is considered as *'one of the giants of statistical theory'*.

Legendre-type functions

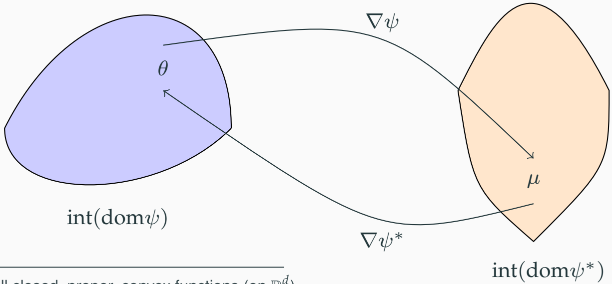
A function $\psi \in \Gamma_0$ [9] is *essentially smooth* if it satisfies the following conditions:

1. $\text{int}\,(\text{dom}\,\psi) \neq \emptyset$
2. ψ is differentiable on $\text{int}\,(\text{dom}\,\psi)$
3. $\|\nabla\psi(x^k)\| \to \infty$ for any $\{x^k \in \text{int}\,(\text{dom}\,\psi)\} \to \bar{x} \in \text{bd}\,(\text{dom}\,\psi)$

If, in addition, ψ is strictly convex on $\text{int}\,(\text{dom}\,\psi)$ then ψ is called of *Legendre type.*

Rockafellar (1970): For $\psi \in \Gamma_0$ of Legendre type, we have:

- ψ^* is of Legendre type.
- $\nabla\psi : \text{int}\,(\text{dom}\,\psi) \to \text{int}\,(\text{dom}\,\psi^*)$ is a bijection (with $(\nabla\psi)^{-1} = \nabla\psi^*$).



[9] Γ_0 : set of all closed, proper, convex functions (on $\mathbb{R}^d$)

Exponential families

Let $(\Omega, \mathcal{A}, P)$ be a probability space[10] with $P \ll \nu$[11]. The *natural parameter space* for P is defined by

$$\Theta_P := \left\{ \theta \in \mathbb{R}^d \mid \int_\Omega \exp(\langle \theta, \cdot \rangle) dP < +\infty \right\} (= \operatorname{dom} \psi_P) \, .$$

The *standard exponential family* generated by P is given by

$$\mathcal{F}_P := \left\{ f_{P_\theta} \;\middle|\; f_{P_\theta}(y) := \exp(\langle y, \theta \rangle - \psi_P(\theta)), \quad \theta \in \Theta_P \right\} .$$

Properties and connections

- $\int_\Omega f_{P_\theta} dP = 1$, thus $P_\theta := P \circ f_{P_\theta}^{-1}$ is a probability measure with $\frac{dP_\theta}{dP} = f_\theta$ $(\theta \in \Theta_P)$.
- Under suitable assumptions: $\underset{Q : E_Q = y}{\text{argmin}} \{ \text{KL}(Q \mid P) \} \in \mathcal{F}_P$
- $\theta_1 \in \Theta_P, \theta_2 \in \operatorname{int}(\Theta_P)$: $\text{KL}(P_{\theta_2} \mid P_{\theta_1}) = D_{\psi_P}(\theta_1, \theta_2)$ (Bregman distance).

[10] $(\Omega, \mathcal{A})$ measurable and P σ-finite works, too.
[11] ν : Lebesgue measure ($\Omega = \mathbb{R}^d$) or counting measure (Ω countable).

Regularity of standard exponential family

The (standard) exponential family $\mathcal{F}_P$ is called

- *minimal* [12] if $\operatorname{int}\Theta_P \neq \emptyset$ and $\operatorname{int}(\operatorname{conv} S_P) \neq \emptyset$ [13];
- *steep* if ψ_P is essentially smooth (automatically satisfied if Θ_P open).

Theorem (Regularity of ψ_P, Brown 1986)

Let $\mathcal{F}_P$ be a minimal exponential family. Then:

(a) *The log-cumulant generating function ψ_P is strictly convex on (the convex set) Θ_P.*

(b) $\psi_P \in C^\infty(\operatorname{int}\Theta_P)$*, and then* $\nabla\psi_P(\theta) = \mathbb{E}_{P_\theta}$.

Corollary

Let the exponential family $\mathcal{F}_P$ be minimal and steep. Then:

(a) ψ_P *(and hence ψ_P^*) is of Legendre type.*

(b) $\theta = \nabla\psi_P^*(\mathbb{E}_{P_\theta})$.

---

[12] This can essentially be assumed w.l.o.g.

[13] S_P: support of P, i.e. the smallest closed set $A \subset \Omega$ s.t. $P(\Omega \setminus A) = 0$.

Domain correspondences and the key inequality

Given ψ of Legendre type, its *Bregman distance* is:

$$D_\psi(y, x) := \psi(y) - \psi(x) - \langle \nabla\psi(x), y - x\rangle \quad \forall (x, y) \in \operatorname{int}(\operatorname{dom}\psi) \times \operatorname{dom}\psi.$$

- $D_\psi \geq 0$ and $D_\psi(x, y) = 0 \iff x = y$;
- D_ψ *not* symmetric in general, but $D_{\frac{1}{2}\|\cdot\|^2} = \frac{1}{2}\|x - y\|^2$;
-

Lemma (Vaisbourd et al.)
Suppose $P \in \mathcal{P}(\Omega)$ generates a minimal and steep exponential family. Then:

(a) *(Domain relations)*

(i) *If S_P is countable, then $\operatorname{dom}\kappa_P = \operatorname{conv} S_P \subset \operatorname{dom}\psi_P^*$;*

(ii) *If S_P is uncountable, then $\operatorname{dom}\kappa_P = \operatorname{int}(\operatorname{conv} S_P) = \operatorname{dom}\psi_P^*$.*

(b) *For all $y \in \operatorname{dom}\kappa_P, Q \ll P$ s.t. $\mathbb{E}_Q = y$ and for all $\theta \in \operatorname{int}\Theta_P$ we have*

$$\psi_P^*(y) \leq \kappa_P(y) \leq \psi_P^*(y) + KL(Q \mid P_\theta) - D_{\psi_P^*}(y, \nabla\psi_P(\theta)). \tag{3}$$

Equivalence of MEM functional and Cramér's function

$\psi_P^* = \kappa_P$?

- $y \in \operatorname{int}(\operatorname{conv} S_P)$:

$$\operatorname{int}(\operatorname{conv} S_P) \subset \operatorname{int}(\operatorname{dom}\psi^*),\ \psi^* \text{ Legendre-type}$$
$$\Longrightarrow \quad \exists \theta \in \operatorname{int}(\operatorname{dom}\psi) = \operatorname{int}\Theta_P :\ y = \nabla\psi_P(\theta) = E_{P_\theta}$$
$$\overset{(3)}{\Longrightarrow} \quad \psi_P^*(y) \le \kappa_P(y) \le \psi^*(y) + \underbrace{\mathrm{KL}(P_\theta \mid P_\theta)}_{=0} - \underbrace{D_{\psi_P^*}(\nabla\psi_P(\theta), \nabla\psi_P(\theta))}_{=0}$$

- $y \in \operatorname{bd}(\operatorname{conv} S_P)$: Can only occur when S_P is countable.

Theorem ($\psi_P^* = \kappa_P$, Vaisbourd et al.)
Suppose $P \in \mathcal{P}(\Omega)$ generates a minimal and steep exponential family. Moreover, suppose one of the following holds:

- *S_P is uncountable*
- *S_P is countable and $\operatorname{conv} S_P$ is closed (which is always the case if S_P is finite).*

Then $\boxed{\kappa_P = \psi_P^*.}$ In this case $0 \le \kappa_P \in \Gamma_0$ is of Legendre type and coercive.

How is $\kappa_P = \psi_P^*$ useful?

If $P \in \mathcal{P}(\Omega)$ is separable (i.e. $P = P_1 \times P_2 \times \cdots \times P_d$), then $M_P(\theta) = \prod_{i=1}^d M_{P_i}(\theta_i)$. Hence

$$\begin{aligned}\psi_P^*(y) &= \sup_{\theta \in \mathbb{R}^d} \{\langle y, \theta \rangle - \log M_P(\theta)\} \\ &= \sum_{i=1}^d \sup_{\theta_i \in \mathbb{R}} \left\{y_i \theta_i - \log M_{P_i}(\theta_i)\right\}.\end{aligned}$$

In many cases this yields analytic formulas for ψ_P^*, i.e. κ_P (even without separability!).

Example: If P is the multivariate normal distribution $N(\mu, \Sigma)$ for $\Sigma \succ 0$, i.e. $M_P(\theta) = \exp\left(\langle \mu, \theta \rangle + \frac{1}{2}\theta^T \Sigma \theta\right)$, then

$$\begin{aligned}\psi_P^*(y) &= \sup_{\theta \in \mathbb{R}^n} \{\langle y, \theta \rangle - \log M_P(\theta)\} \\ &= \sup_{\theta \in \mathbb{R}^n} \left\{\langle y - \mu, \theta \rangle - \frac{1}{2}\theta^T \Sigma \theta\right\} \\ &= \frac{1}{2}(y - \mu)^T \Sigma^{-1} (y - \mu).\end{aligned}$$

Examples of Cramér's function

Reference Distribution (P)	Cramér Rate Function ($\psi_P^*(y)$)	$\operatorname{dom} \psi_P^*$
Multivariate Normal $\mu \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d, \Sigma \succ 0$	$\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)$	$\mathbb{R}^d$
Poisson ($\lambda \in \mathbb{R}_{++}$)	$y \log(y/\lambda) - y + \lambda$	$\mathbb{R}_+$
Gamma ($\alpha, \beta \in \mathbb{R}_{++}$)	$\beta y - \alpha + \alpha \log\left(\frac{\alpha}{\beta y}\right)$	$\mathbb{R}_{++}$
Normal-inverse Gaussian $\alpha, \beta, \delta \in \mathbb{R} : \alpha \geq \lvert\beta\rvert,$ $\delta > 0, \gamma := \sqrt{\alpha^2 - \beta^2}$	$\alpha\sqrt{\delta^2 + (y-\mu)^2} - \beta(y-\mu) - \delta\gamma$	$\mathbb{R}$
Multinomial ($p \in \Delta_d, n \in \mathbb{N}$)	$\sum_{i=1}^d y_i \log\left(\frac{y_i}{np_i}\right)$	$n\Delta_d \cap I(p)$[14]

In addition: Laplace, (Negative) Multinomial, Continuous/Discrete Uniform, Logistic, Exponential/Chi-Squared/Erlang (via Gamma), Binomial/Bernoulli/Categorical (via Multinomial), Negative Binomial & Shifted Geometric (via Negative Multinomial).

[14] $I(p) := \left\{x \in \mathbb{R}^d \mid x_i = 0 \text{ if } p_i = 0\right\}$

The MEM estimator

Let the following be given:

- $\hat{y} \in \mathbb{R}^d$: observed data;
- $S^* \subset \mathbb{R}^d$: admissible parameters;
- $F_\Lambda := \{P_\lambda \mid \lambda \in \Lambda \subset \mathbb{R}^d\}$: parameterized family of distributions[15];
- $P_{\hat{\lambda}} \in F_\Lambda$: reference distribution such that $\hat{y} = E_{P_{\hat{\lambda}}}$.

We define the *MEM estimator* $y_{MEM} \in \mathbb{R}^d$ by

$$y_{MEM}(\hat{y}, F_\Lambda, S^*) := \operatorname*{argmin}_{y \in S^*} \psi^*_{P_{\hat{\lambda}}}(y).$$

Under suitable assumptions on $P_{\hat{\lambda}}$, the function $\psi^*_{P_{\hat{\lambda}}}$ is coercive and strictly convex, which guarantees well-definedness of the MEM estimator.

[15] not necessarily exponential

MEM vs. ML estimation

Let the following be given:

- $\hat{y} \in \mathbb{R}^d$: observation;
- $S \subset \mathbb{R}^m$: set of admissible parameters;
- $F_\Lambda := \{P_\lambda \mid \lambda \in \Lambda \subset \mathbb{R}^m\}$: parameterized family of distributions with densities f_λ;

The ubiquitous *maximum likelihood estimator is given by*

$$\lambda_{ML}(\hat{y}, F_\Lambda, S) := \operatorname{argmax}_{\lambda \in S \cap \Lambda} \log f_\lambda(\hat{y}).$$

It induces a distribution that is most likely to produce the given observation.

When F_Λ is an exponential family induced by P, and $\hat{\lambda} := \nabla \psi_P^*(\hat{y})$ then (under some technical assumptions) we have

$$y_{MEM} = \psi_P^*(\lambda_{MEM})$$

for

$$\lambda_{MEM} = \underset{\lambda \in S}{\text{argmin}}\ \text{KL}(P_\lambda \mid P_{\hat{\lambda}}),$$

whereas

$$\lambda_{ML} = \underset{\lambda \in S}{\text{argmin}}\ \text{KL}(P_{\hat{\lambda}} \mid P_\lambda).$$

Linear model based on MEM

Consider the linear inverse problem $Cx \approx \hat{y}$ for some

- $\hat{y} \in D \subset \mathbb{R}^m$: measurement vector;
- $C \in \mathcal{C} \subset \mathbb{R}^{m \times d}$: measurement matrix (dictated by the problem).

Now consider:

- $F_\Lambda = \{P_\lambda \mid \lambda \in \Lambda \subset \mathbb{R}^m\} \subset \mathcal{P}(\Omega)$: reference family;
- $\hat{P} := P_{\hat{\lambda}}$: reference distribution with $E_{\hat{P}} = \hat{y}$;
- $S^* := \{Cx \mid x \in X\}$: set of admissible parameters.

The linear model based on the MEM functional reads

$$\min_{x \in X} \psi^*_{\hat{P}}(Cx).$$

Reference Family	Objective Function ($\psi^*_{\hat{P}} \circ C$)
Normal	$\frac{1}{2}\|Cx - \hat{y}\|^2$
Poisson	$\sum_{i=1}^m [\langle c_i, x\rangle \log(\langle c_i, x\rangle / \hat{y}_i) - \langle c_i, x\rangle + \hat{y}_i]$
Gamma $(\beta = 1)$	$\sum_{i=1}^m [\langle c_i, x\rangle - \hat{y}_i \log(\langle c_i, x\rangle) - (\hat{y}_i - \hat{y}_i \log \hat{y}_i)]$

Regularized linear model

In case of ill-posedness or to incorporate prior information we consider the

MEM regularized linear model:

$$\min \left\{ \kappa_{P_{\hat{\theta}}}(Ax) + \varphi(x) : x \in X \right\},$$

where

$$\varphi(x) = \begin{cases} \kappa_R(x) & \\ \kappa_R(Lx) & (L \in \mathbb{R}^{r \times d}) \\ \sum_{i=1}^{d} \kappa_R(L_i x) & (L_i \in \mathbb{R}^{r \times d}, i = 1, 2, \ldots, d), \end{cases}$$

with $R \in \mathcal{P}(\Omega)$ reference distribution.

Q: How can we efficiently solve this problem?

Bregman proximal gradient method for MEM linear model

The regularized model falls into the additive composite framework

$$\min_{x \in \mathbb{R}^d} \{f(x) + g(x)\} \quad (g \in \Gamma_0, f \in C^1(\cap \Gamma_0)).$$

The *Bregman proximal gradient* algorithm

Initialization. Pick $t \in (0, 1/L]$ and $x^0 \in \operatorname{int}(\operatorname{dom} h)$.

Procedure. For $k = 0, 1, 2, \ldots$:

$$x^{k+1} = \quad \operatorname{prox}_{tg}^{h}\left(\nabla h^* \left(\nabla h(x^k) - t\nabla f(x^k)\right)\right)$$

is specified by a *kernel* $h \in \Gamma_0 \cap C^1$ that [Bauschke et al., 2017]:

- is *smooth adaptable* w.r.t. f i.e. $Lh - f$ is convex for some $L > 0$.
- has computationally tractable *Bregman proximal operator* with respect to g:

$$\operatorname{prox}_g^h(\bar{x}) := \underset{x \in \mathbb{R}^d}{\text{argmin}} \{g(x) + D_h(x, \bar{x})\}.$$

Bregman Proximal Operators

The h-Bregman proximal operator of ψ_R^* is always well defined under mild assumptions (on R and h), and can be efficiently evaluated, often has closed form:

Reference Distribution	Proximal Operator	Kernel ($h(x)$)
Multivariate Normal $\mu \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d, \Sigma \succ 0$	$x^+ = (tI + \Sigma)^{-1}(\Sigma\bar{x} + t\mu)$	$(1/2)\|x\|_2^2$
Gamma $(\alpha, \beta \in \mathbb{R}_{++})$	$x^+ = \left(\bar{x} - t\beta + \sqrt{(\bar{x} - t\beta)^2 + 4t\alpha}\right)/2$	$(1/2)\|x\|_2^2$
Laplace $(\mu \in \mathbb{R},\ b \in \mathbb{R}_{++})$	$x^+ = \begin{cases} \mu, & \mu = \bar{x}, \\ \mu + b\rho, & \mu \neq \bar{x}, \end{cases}$ where ρ is the unique real root of a cubic[16]	$-\sum \log x_i$
Poisson $(\lambda \in \mathbb{R}_{++})$	$x^+ = (\bar{x}\lambda^t)^{\frac{1}{t+1}}$	$\sum x_i \log x_i$
Multinomial $(p \in \Delta_d, n \in \mathbb{N})$	$x^+ = \left(\frac{n(np_i)^{\frac{t}{t+1}} \bar{x}_i^{\frac{1}{t+1}}}{\sum_{i=1}^d (np_i)^{\frac{t}{t+1}} \bar{x}_i^{\frac{1}{t+1}}} \right)_{i=1}^d$	$\sum x_i \log x_i$

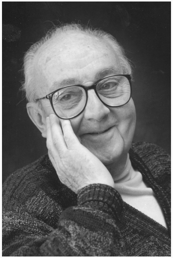
In addition: Normal-inverse Gaussian, Negative Multinomial, Continuous/Discrete Uniform, Logistic, Exponential/Chi-Squared/Erlang (via Gamma), Binomial/Bernoulli/Categorical (via Multinomial), Negative Binomial & Shifted Geometric (via Negative Multinomial).

[16] With closed-form coefficients dependent on b, μ, $\bar{x}$, t

Summary

All models are wrong, but some are useful.

George E. P. Box



- MEM is a useful tool for incorporating prior information into models for inverse problems.
- The application of MEM to inverse problems is scarce in the literature.
- We unify and extend much of the theory that appears in the literature, while providing an algorithmic framework.
- arXiv preprint and *computational toolbox* of Cramér functions, prox operators, and algorithms, to appear online shortly.
- Ongoing work: Obtain the Cramér function (or log-MGF) via (deep) learning.

References

1. J.M. Borwein and A.S. Lewis: *Partially finite convex programming, Part I: Quasi relative interiors and duality theory.* Mathematical Programming 57(1)1992, pp. 15–48.
2. *L.D. Brown:Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, 1986.
3. H.H. Bauschke, J. Bolte, M. Teboulle:*A descent Lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications.* Math. Oper. Res. 42(2), 2016, pp. 330–348.
4. E.T. Jaynes:*Information theory and statistical mechanics.* Phys. Rev. 106(4), 1957, pp. 620–630.
5. G. Le Besnerais, J. Bercher, and G. Demoment: *A new look at entropy for solving linear inverse problems.* IEEE Transactions on Information Theory 45(5), 1999, pp. 1565–1578.
6. P. Maréchal: *On the principle of maximum entropy on the mean as a methodology for the regularization of inverse problems.* In B. Grigelionis et al. (Eds.), Probability Theory and Mathematical Statistics 1999, pp. 481–492.