The Maximum Entropy on the Mean Method for Linear Inverse Problems (And Beyond)

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

International Conference on Continuous Optimization, Lehigh

July 26th, 2022

Motivation: Linear Inverse Problems

Canonical Example: $Cx \sim b$

$$\min_{x \in \mathbb{R}^d} \left\{ R(x) + \frac{\alpha}{2} F(Cx, b) \right\}$$

- R is a regularizer imposing constraints on the optimizers
- $F(Cx, b)$ is a fidelity term estimating the difference between Cx and b
- Many of these "norms" for regularization and fidelity have interpretations from statistical estimation

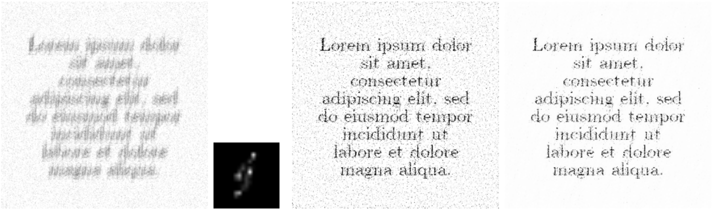


Figure: Image deblurring problem of the form $Cx \sim b$, Rioux et al. (2021)

An Information-Theoretic Approach

$$\min_{x\in\mathbb{R}^d} \left\{ R(x) + \frac{\alpha}{2} F(Cx, b) \right\}$$

- How can we choose R and F in a meaningful way?
- Idea: Work at the higher level of the probability distribution of the ground truth x.
- Given a prior distribution, P, we want to understand the distribution of the ground truth, Q.
- The Kullback-Leibler (KL) divergence [Kullback, Leibler (1951)] between σ-finite P and $Q \in \mathcal{P}(\Omega)$ is defined by

$$D_{\mathrm{KL}}(Q||P) := \begin{cases} \int_\Omega \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right) \mathrm{d}Q, & Q \ll P \\ +\infty, & \text{otherwise} \end{cases}$$

MEM Paradigm

- Maximum Entropy on the Mean: The state best describing a system is the mean of a distribution maximizing some measure of entropy (à la Principle of Maximum Entropy [Jaynes, 1957])

Definition (MEM Function)

The Maximum Entropy on the Mean (MEM) Function $\kappa_P \colon \mathbb{R}^d \to (-\infty, \infty]$ is defined by [Rietsch, 1977]:

$$\kappa_P(y) := \inf \{ D_{\mathrm{KL}}(Q || P) \mid Q \ll P \text{ with } \mathbb{E}_Q = y \}$$

- Information-driven approach: Measure compliance of y with P via $\kappa_P(y)$
- Applications: crystallography [Navaza (1985)], seismic tomography [Fermín et al. (2006)], medical imaging [Amblard et al. (2004), Deslauriers-Gauthier et al. (2017), Cai et al. (2022)], image processing [Rioux et al. (2021)]

Alternate Formulation and the MEM Function

The MEM reformulation of our original inverse problem in the least squares setting:

$$\overline{x} = \mathbb{E}_{\overline{Q}}[X], \quad \overline{Q} = \operatorname{argmin}_{Q \in \mathcal{P}(\Omega)} \left\{ KL(Q \| P) + \frac{\alpha}{2} \| b - C\mathbb{E}_Q[X] \|_2^2 \right\}.$$

One can equivalently formulate as:

$$\overline{x} \; = \; \operatorname{argmin}_{y \in \mathbb{R}^d} \; \left\{ \frac{\alpha}{2} \| Cy - b \|^2 \; + \; \kappa_P(y) \right\}$$

where $\kappa_P(y) = \inf \{ D_{\text{KL}}(Q || P) \mid Q \ll P \text{ with } \mathbb{E}_Q = y \}$ is the MEM function.

Cramér's Function

The MEM function is defined by a seemingly intractable problem. How can we use it?

$$\kappa_P(y) := \inf \{ D_{\mathrm{KL}}(Q||P) \mid Q \ll P \text{ with } \mathbb{E}_Q = y \}$$

Under some conditions:

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

where $\psi_P(\theta) := \log \int_\Omega \exp(\langle y, \theta \rangle) dP(y)$ is the log-normalizer of P. The map ψ_P^* is known as Cramér's function (c.f. large deviations theory).

Legendre Type

Definition (Legendre Type)

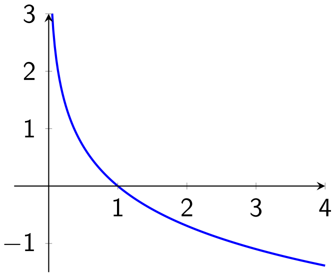
A function $\psi \in \Gamma_0$ is essentially smooth if it satisfies the following conditions:

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 $\{x^k\} \subseteq \text{int}(\text{dom}\,\psi)$ such that $x^k \to \bar{x} \in \partial(\text{dom}\,\psi)$

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

Ex: The function $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$

$$f(x) = \begin{cases} -\log x, & x > 0, \\ +\infty, & x \leq 0, \end{cases}$$



Legendre Functions and Conjugacy

Theorem (Rockafellar, 1970)

If $\psi \in \Gamma_0$ is of Legendre type then

1. *The convex conjugate ψ^* is of Legendre type*
2. $\nabla\psi$ *is a bijection from* $\mathrm{int}(\mathrm{dom}\,\psi)$ *to* $\mathrm{int}(\mathrm{dom}\,\psi^*)$ *with inverse* $(\nabla\psi)^{-1} = \nabla\psi^*$

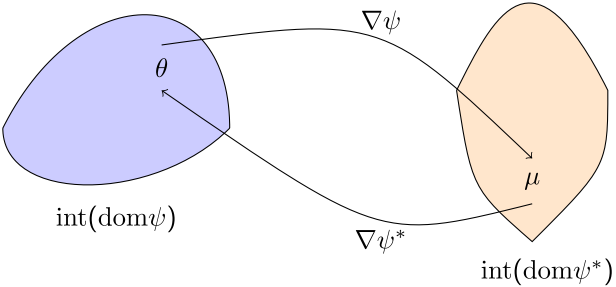


Figure: Illustration of the above theorem

Probability Theory

Let ρ be a σ-finite measure on measurable $\Omega \subseteq \mathbb{R}^d$. Some definitions:

- $\Omega_\rho = \mathsf{supp}(\rho)$ (support of ρ)
- $\Omega_\rho^{cc} = \mathsf{cl}(\mathsf{conv}\,\Omega_\rho)$ (convex support of ρ)

We consider two cases:

1. $(\Omega = \mathbb{R}^d, \nu = \text{Lebesgue})$
2. $(\Omega \subseteq \mathbb{R}^d, \nu = \text{Counting})$

Define $\mathcal{P}(\Omega) := \{P \text{ probability measure on } \Omega \mid P \ll \nu\}$.

Each such P has Radon-Nikodym Derivative $f_P := \frac{\mathrm{d}P}{\mathrm{d}\nu}$, expected value $\mathbb{E}_P$, and moment-generating function[1] M_P:

$$\mathbb{E}_P := \int_\Omega y dP(y) \in \mathbb{R}^d$$

$$M_P(\theta) := \int_\Omega \exp(\langle y, \theta \rangle) dP(y)$$

[1] We assume $\mathrm{int}(\mathsf{dom}\, M_P) \neq \emptyset$

Exponential Families

Let P be σ-finite, $P \ll \nu$. The natural parameter space for P is defined by

$$\Theta_P := \left\{ \theta \in \mathbb{R}^d \mid M_P(\theta) = \int_\Omega \exp(\langle y, \theta \rangle) dP(y) < \infty \right\}$$

Definition (Log-Normalizer)

The function $\psi_P \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ by

$$\psi_P(\theta) = \begin{cases} \log M_P(\theta), & \theta \in \Theta_P \\ +\infty, & \theta \notin \Theta_P \end{cases}$$

is called the log-normalizer.

Definition (Exponential Family)

The standard exponential family generated by P is

$$\mathcal{F}_P := \{ f_{P_\theta}(y) := \exp(\langle y, \theta \rangle - \psi_P(\theta)) \mid \theta \in \Theta_P \}$$

Exponential Family Properties

We assume $\operatorname{int}\Theta_P \neq \emptyset, \operatorname{int}\Omega_P^{cc} \neq \emptyset$ (an exponential family satisfying this is called minimal)

Theorem (Regularity of ψ_P, Brown 1986)

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

1. *The log-normalizer ψ_P is strictly convex on the convex set Θ_P*
2. $\psi_P \in C^\infty(\operatorname{int}\Theta_P)$, $\nabla\psi_P(\theta) = \mathbb{E}_{P_\theta}$

If ψ_P is essentially smooth we say $\mathcal{F}_P$ is steep.
Conclusion: If $\mathcal{F}_P$ is minimal and steep then ψ_P is of Legendre type.

Corollary (Mean Value Parametrization)

The natural parameter θ can be expressed as

$$\theta = \nabla\psi_P^*(\mu)$$

where $\mu = \mathbb{E}_{P_\theta} = \nabla\psi_P(\theta)$.

Domain of Cramér Function vs. MEM Function

Theorem (Domain of ψ_P^*, Barndorff-Nielsen 1978)

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

$$\operatorname{int} \Omega_P^{cc} \subseteq \operatorname{dom} \psi_P^* \subseteq \Omega_P^{cc}$$

Moreover, the following hold:

1. *If Ω_P is finite then* $\operatorname{dom} \psi_P^* = \Omega_P^{cc}$
2. *If Ω_P is countable then* $\operatorname{dom} \psi_P^* \supseteq \operatorname{conv} \Omega_P$
3. *If Ω_P is uncountable then* $\operatorname{dom} \psi_P^* = \operatorname{int} \Omega_P^{cc}$

Theorem (Domain of κ_P, Vaisbourd et al.)

Suppose P satisfies the same assumptions above. Then:

- *If Ω_P is countable then* $\operatorname{dom} \kappa_P = \operatorname{conv} \Omega_P$
- *If Ω_P is uncountable then* $\operatorname{dom} \kappa_P = \operatorname{int} \Omega_P^{cc}$

Key Inequality

Given ψ of Legendre type, its Bregman divergence is:

$$D_\psi(y, x) := \psi(y) - \psi(x) - \langle \nabla\psi(x), y - x\rangle$$

Lemma (MEM Upper Bound, Vaisbourd et al.)

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

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

for any $y \in \operatorname{dom}\kappa_P, Q \ll P$ with $\mathbb{E}_Q = y$, and $\theta \in \operatorname{int}\Theta_P$. Recall P_θ is defined by density $f_{P_\theta} = \exp(\langle\cdot, \theta\rangle - \psi_P(\theta)) \in \mathcal{F}_P$.

Proof of equality: If $y \in \operatorname{int}\Omega_P^{cc}$ then $\exists\ \theta \in \operatorname{int}\Theta_P$ s.t. $y = \nabla\psi_P(\theta) = \mathbb{E}_{P_\theta}$. Now take $Q = P_\theta$ above. □

What if y is on the boundary?

Equivalence of Cramér and MEM

Theorem (Equality Conditions, Vaisbourd et al.)

Suppose $P \in \mathcal{P}(\Omega)$ generates a minimal and steep exponential family. Moreover, suppose one of the following holds:

- *Ω_P is uncountable*
- *Ω_P is countable and $\operatorname{conv}\Omega_P$ is closed*

Then $\kappa_P = \psi_P^*$. In particular, κ_P is closed, proper, and convex.

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

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

Upshot: Suffices to compute scalar Cramér functions.

Examples

Reference Distribution (P)	Cramér Rate Function ($\psi_P^*(y)$)	dom ψ_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 $\alpha, \beta, \delta \in \mathbb{R}\colon \alpha \geq \lvert\beta\rvert$, $\delta > 0, \gamma := \sqrt{\alpha^2 - \beta^2}$	$\alpha\sqrt{\delta^2 + (y-\mu)^2} - \beta(y-\mu) - \delta\gamma$	$\mathbb{R}$
Multinomial ($p \in \Delta_d, n \in \mathbb{N}$)	$\sum_{i=1}^d y_i \log\left(\frac{y_i}{np_i}\right)$	$n\Delta_d \cap I(p)^2$

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).

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

The MEM Estimator

Maximum likelihood (ML) is a popular principle of statistical estimation

$$\theta_{ML} = \theta_{ML}(\hat{y}, F_\Theta, S) := \text{argmax}_{\theta \in S \cap \Theta} \{\log f_{P_\theta}(\hat{y})\}$$

where:

- $S \subseteq \mathbb{R}^d$ are admissible parameters
- F_Θ parameterized family of distributions $P_\theta, \theta \in \Theta \subseteq \mathbb{R}^d$ with densities f_{P_θ}
- $\hat{y} \in \mathbb{R}^d$ is a sample of observed data

Definition/Theorem (Vaisbourd et al.)

The MEM estimator $y_{MEM} \in \mathbb{R}^d$ is defined by:

$$y_{MEM} = y_{MEM}(\hat{y}, F_\Theta, S^*) := \text{argmin}_{y \in S^*} \left\{\psi^*_{P_{\hat{\theta}}}(y)\right\}$$

where $P_{\hat{\theta}} \in F_\Theta$ is such that $\hat{y} = \mathbb{E}_{P_{\hat{\theta}}}$. The existence and uniqueness of y_{MEM} is guaranteed under some mild assumptions.

Linear Models

- Bioinformatics, Image Processing, Machine Learning, ...
- $C \in \mathcal{C} \subseteq \mathbb{R}^{m \times d}$ (dictated by the problem)
- $F_\Theta = \{P_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^m\} \subseteq \mathcal{P}(\Omega)$

Reference distribution $P_{\hat{\theta}}$ is specified via $\hat{y} = \mathbb{E}_{P_{\hat{\theta}}}$ where $\hat{y}$ is our observation vector. Thus the MEM estimator of the linear model is:

$$\text{argmin}_{x \in X} \left\{ \psi^*_{P_{\hat{\theta}}}(Cx) \right\}, \quad (C \in \mathcal{C}, \hat{\theta} \in \Theta \colon \mathbb{E}_{P_{\hat{\theta}}} = \hat{y})$$

Reference Family	Objective Function ($\psi^*_{P_{\hat{\theta}}} \circ C$)
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 Model

Regularize to create well-posed problem:

$$\min_{x \in X} \left\{ \psi^*_{P_{\hat{\theta}}}(Ax) + \varphi(x) \right\}, \quad (A \in \mathcal{C}, \hat{\theta} \in \Theta\colon \mathbb{E}_{P_{\hat{\theta}}} = \hat{y})$$

- Here $\varphi\colon \mathbb{R}^d \to (-\infty, \infty]$ is closed, proper, convex.
- We can use Cramér's function to regularize
- Take $R \in \mathcal{P}(\Omega)$ as a prior distribution encoding info about the desired solution

$$\min_{x \in X} \left\{ \psi^*_{P_{\hat{\theta}}}(Cx) + \psi^*_R(x) \right\}$$

Application to Image Deblurring

QR code image deblurring:

$$\min_{x \in \mathbb{R}^d} \left\{ \frac{1}{2} \|Ax - \hat{y}\|_2^2 + \kappa_R(x) \right\}$$

- $\hat{y}$ - blurred and noisy image
- A - blurring matrix
- R - reference distribution (Bernoulli)

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

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

Figure: From Rioux et al. (2021)

Solving the Problem

Regularized model falls into the additive composite framework:

$$\min_{x\in\mathbb{R}^d} \{f(x) + g(x)\}$$

The Bregman proximal gradient algorithm is specified by a kernel function h that [Bauschke et al. (2017)]:

- is smooth adaptable w.r.t. f ($Lh - f$ is convex for some $L > 0$)
- induces a computationally tractable Bregman proximal operator with respect to g

Definition (Bregman Proximal Operator)

Let $g, h\colon \mathbb{R}^d \to (-\infty, +\infty]$ such that g is proper and closed, and h is Legendre type. Then for $\bar{x} \in \operatorname{int}(\operatorname{dom} h)$ we define the Bregman proximal operator to be

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

Bregman Proximal Gradient Algorithm

Algorithm 1: Bregman Proximal Gradient (BPG) Method

Input: Set $t \in (0, 1/L]$ and $x^0 \in \operatorname{int}(\operatorname{dom} h)$.
for $k = 0, 1, 2, \ldots$ **do**
    $x^{k+1} = \operatorname{prox}_{tg}^{h}(\nabla h^*(\nabla h(x^k) - t\nabla f(x^k)))$;
end

- $h = \frac{1}{2}\|\cdot\|_2^2$ - proximal gradient method
- $h = \frac{1}{2}\|\cdot\|_2^2,\ g = \delta_S$ - gradient projection method
- $h = \frac{1}{2}\|\cdot\|_2^2,\ g = 0$ - gradient descent method

Other variants and methods (acceleration, decomposition) rely on the same operators we derive in this work.

Bregman Proximal Operators

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[3]	$-\sum \log x_i$
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) for each each choice of h shown.

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

Summary

- MEM is a very useful tool for the incorporation of prior information into models for inverse problems.
- While much of the theory appears in the literature and was historically applied to a few inverse problems, it seems to have been forgotten.
- Revisit the theory and experiment with solving regularized MEM linear models.
- 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.