Measures everywhere

Variation analysis on measures

Sergei Zuyev

University of Strathclyde, Glasgow, U.K.

Outline of the course

- Measures and constrained optimisation
- Optimal design of experiments
- General Poisson processes
- High intensity optimisation
- Steepest descent algorithms
- Other applications: FGM, Clustering, etc.

Measures everywhere!

- All the statistics is about: Estimation of an unknown underlying probability distribution: $\mathbf{P}$
- An estimate $\widehat{\mathbf{P}}$ minimises a given Goal functional $\psi(\mathbf{P})$ (–Likelihood, distance to the empirical distribution, etc.) usually under some constraints (e.g., within a given parametric class $\mathbf{P}_\theta,\ \theta \in \Theta$).
- Probability is a measure, so it is a particular case of optimisation in the class of non-negative measures $\mathbb{M}_+$ subject to a total mass fixed to 1 and possibly other constraints.

What is it?

We are given: a set X – phase space, a system of its subsets $\mathcal{B}$ closed under countable intersections and complements and containing empty set $\emptyset$ (σ-algebra).

Signed Measure (or Charge), is a function $\mu : \mathcal{B} \mapsto \mathbb{R}$ such that

1. $\mu(\emptyset) = 0$;
2. $\mu(A \cup B) = \mu(A) + \mu(B)$ whenever $A \cap B = \emptyset$;
3. $\mu(\cap_{n=1}^{\infty} B_n) = \lim_{n\to\infty} \mu(B_n)$ for any $B_1 \supseteq B_2 \supseteq \dots$

Positive measure (or just Measure) is a charge such that
$\mu(B) \geq 0 \, \forall B \in \mathcal{B}$.

❑ *Examples*

- Length in $X = \mathbb{R}$; Area in $X = \mathbb{R}^2$; Volume in $X = \mathbb{R}^d, \ d \geq 3$;
- Mass, Potential, Charge in physics;
- Probability is a positive measure such that $\mu(X) = 1$.

Banach space $\mathbb{M}$

- Measures can be added and multiplied by a number: $(\mu+\nu)(B) \stackrel{\text{def}}{=} \mu(B)+\nu(B)$; $(t\mu)(B) \stackrel{\text{def}}{=} t\mu(B)$.

- Jordan decomposition: of a signed measure $\mu = \mu^+ - \mu^-$, where $\mu^+,\ \mu^- \geq 0$ and orthogonal: $\mu^+(B) > 0 \Rightarrow \mu^-(B) = 0$; and $\mu^-(B) > 0 \Rightarrow \mu^+(B) = 0$.

- Total variation norm: $\|\mu\| = \mu^+(X) + \mu^-(X)$.

- The set $\mathbb{M}$ of all signed measures with finite norm thus forms a Banach space.

Cone $\mathbb{M}_+$

Positive measures with a finite norm form a cone $\mathbb{M}_+$ in $\mathbb{M}$:
if $\mu,\ \nu \in \mathbb{M}_+$, then $\mu + \nu \in \mathbb{M}_+$ and $t\mu \in \mathbb{M}_+$ for $t \geq 0$.

❑ Subtlety of $\mathbb{M}_+$ is that it does not contain inner points unless X is a finite set.

Lebesgue integral

For $\mu \in \mathbb{M}_+$, if $f(x) = \sum_i f_i \, \mathbb{I}_{B_i}(x)$ – a step-function then

$$\int f \, d\mu = \int f(x) \, \mu(dx) \stackrel{\text{def}}{=} \sum_i f_i \mu(B_i) \, .$$

For a general f,

$$\int f \, d\mu \stackrel{\text{def}}{=} \lim_n \int f_n \, d\mu$$

for any sequence of step-functions $f_n(x)$ uniformly converging to $f(x)$.

For $\mu \in \mathbb{M}$,

$$\int f \, d\mu \stackrel{\text{def}}{=} \int f \, d\mu^+ - \int f \, d\mu^- \, .$$

Differentiability on $\mathbb{M}$

A function $\psi : \mathbb{M} \mapsto \mathbb{R}$ is Fréchet (strongly) differentiable if

$$\psi(\nu + \eta) - \psi(\nu) = D\psi(\nu)[\eta] + o(\|\eta\|) \quad \text{as } \|\eta\| \to 0\,,$$

where $D\psi(\nu)[\eta]$ is a bounded linear continuous functional of η.

In this case for any $\eta \in \mathbb{M}$ there also exists Gateaux (directional) derivative:

$$\lim_{t \downarrow 0} t^{-1}(\psi(\nu + t\eta) - \psi(\nu)) = D\psi(\nu)[\eta]$$

Finite dimensional triviality

Let $X = \{1, \dots, n\}$. Finite measures on X are $\nu = (m_1, \dots, m_n)$, i. e. $\mathbb{M} = \mathbb{R}^n$ and $\mathbb{M}_+ = \mathbb{R}^n_+$.

$D\psi(\nu)$ is then a usual differential (linear mapping) at the point $\nu \in \mathbb{R}^n$, so that there is a vector $(d_1, \dots, d_n) = d(x, \nu),\ x \in X$ – gradient, such that for any increment $\eta(x) = (\eta_1, \dots, \eta_n)$ one has

$$D\psi(\nu)[\eta] = \sum_{x=1}^{n} d_x \eta_x = \int_X d(x, \nu)\, \eta(dx)\,.$$

Countable infinity

Let $X = \mathbb{N}$. Then finite measures on X are sequences $\nu = (\nu_1, \nu_2, \dots)$ such that $\|\nu\| = \sum_i |\nu_i| < \infty$, i. e. $\mathbb{M} = \ell_1$.

As the dual space $\ell_1^* = \ell_\infty$, the bounded linear functional $D\psi(\nu)$ can be represented as

$$D\psi(\nu)[\eta] = \sum_{x=1}^{n} d_x \eta_x = \int d(x,\nu)\,\eta(dx)\,,$$

where $d(x,\nu) = \{d_x\},\ x \in \mathbb{N}$ is a bounded sequence (gradient).

Gradient function

- ❑ Does a gradient (function, necessarily bounded) always exist for a general X, so that

$$D\psi(\nu)[\eta] = \int_X d(x,\nu)\eta(dx) \quad \forall \eta \in \mathbb{M} \; ?$$

If so, then

$$D\psi(\nu)[\delta_x] = \int_X d(y,\nu)\delta_x(dy) = d(x,\nu)\,,$$

i. e. the gradient $d(x,\nu)$ is the *directional derivative* of ψ at ν in 'direction' of δ_x (cf. finite-dimensional case)

- ❑ Answer is: **NO**, unless X is at most countable (as above).

Contre-example

Let $X = [0, 1]$ and μ_λ be the part of Lebesgue decomposition of μ which is absolutely continuous w.r.t. Lebesgue measure λ. Then the linear bounded functional $L : \mu \mapsto \mu_\lambda(X)$ cannot be represented as an integral w.r.t. μ.

Indeed, assume that $d(x)$ is such a gradient function. Then for any $y \in X$,

$$\int_0^1 d(x)\,dx = L(\lambda) = L(\lambda + \delta_y) = \int_0^1 d(x)\,dx + d(y)$$

so that $d(y) \equiv 0$, thus $L(\lambda) = 0$ – contradiction.

Nature is not that bad!

For most interesting differentiable functionals the gradient function **does exist**.

Example 1: $\psi(\nu) = \nu(X)$ – linear function of ν.

$$\psi(\nu + \eta) - \psi(\nu) = \eta(X) = \int_X 1 \, \eta(dx)$$

so that $d(x, \nu) \equiv 1$. Another way:

$$\nu(X) = \int 1 \, \nu(dx)$$

already an integral form of the linear functional, so that $d(x, \nu) \equiv 1$.

Example 2: μ - is a probability distribution on $\mathcal{B}(X),\ X \subseteq \mathbb{R}$,

$$\psi(\mu) = \mathbf{var}(\mu) = \int x^2 \mu(dx) - \Big[\int x\mu(dx)\Big]^2 .$$

By the Chain rule

$$d(x, \mu) = x^2 - 2\int x\mu(dx) \cdot x = x^2 - 2x\,\mathbf{E}(\mu) .$$

Note that $D\,\mathbf{var}(\mu)[\eta]$ does *not* exist for all $\eta \in \mathbb{M}$, and thus $\mathbf{var}(\mu)$ is not strongly differentiable, unless X is compact.

❑ From now on we consider only strongly differentiable functionals possessing a gradient function.

Variational analysis

Let ν provides min to ψ on $\mathbb{M}$. Then

$$D\psi(\nu)[\eta] \geq 0 \text{ for all } \eta\,.$$

If ψ possesses a gradient function, then

$$D\psi(\nu)[\eta] = \int d(x,\nu)\,\eta(dx) \geq 0\,.$$

- Taking $\eta = \delta_x$ implies $d(x,\nu) \geq 0$.
- Taking $\eta = -\delta_x$ implies $d(x,\nu) \leq 0$.

❑ Thus we have shown

Theorem 1. *If ν provides min to ψ on $\mathbb{M}$, then $d(x,\nu) = 0$ for all $x \in X$ (i. e. all directional derivatives are 0).*

Constrained optimisation.

Let ν provides min to ψ on $\mathbb{A} \subseteq \mathbb{M}$. Then

$$D\psi(\nu)[\eta] \geq 0 \text{ for all admissible } \eta\,,$$

i. e. for such η that $\nu + t\eta \in \mathbb{A}$ for all sufficiently small $t > 0$.

Closure of all admissible 'directions' at ν is called tangent cone

$$T_{\mathbb{A}}(\nu) = \liminf_{t \downarrow 0} \frac{\mathbb{A} - \nu}{t}$$

So we need to characterise $T_{\mathbb{A}}(\nu)$ for $\mathbb{A}$ of interest.

Tangent cone to $\mathbb{M}_+$

Take $\mu \in \mathbb{M}_+$ and $\eta \in \mathbb{M}$ such that $\eta^- \ll \mu$. Consider a sequence of measures $\eta_n(\bullet) = \int_\bullet \min\{h(x), n\}\, \mu(dx)$, where $h(x) = \frac{d\eta^-}{d\mu}(x)$. Then for any $B \in \mathcal{B}$,

$$(\mu + t\eta_n)(B) = \int_B (1 - t\min\{h(x), n\})\, \mu(dx) + t\eta_n^+(B)\,,$$

which is non-negative for all $t \le 1/n$. Thus $\eta_n \in T_{\mathbb{M}_+}(\mu)$ for all n. Next

$$\|\eta - \eta_n\| = \int h(x)\, \mathbb{1}_{h(x)>n}\, \mu(dx) \to 0$$

by dominated convergence as $\int h(x)\mu(dx) = \eta^-(X) < \infty$.
But $T_{\mathbb{M}_+}(\mu)$ is closed, so that $\lim \eta_n = \eta \in T_{\mathbb{M}_+}(\mu)$.

Consider now $\eta \in \mathbb{M}$ such that $\eta^- \not\ll \mu$, i.e. there is $B \in \mathcal{B}$ such that $\mu(B) = 0, \eta^+(B) = 0$, but $\eta^-(B) > 0$. Then $(\mu + t\eta)(B) = -t\eta^-(B) < 0$ for all $t > 0$ so that such $\eta \notin T_{\mathbb{M}_+}(\mu)$.

❑ Thus we have shown

Theorem 2. *For $\mu \in \mathbb{M}_+$ we have*

$$T_{\mathbb{M}_+}(\mu) = \{\eta \in \mathbb{M} : \ \eta^- \ll \mu\} .$$

Optimisation on $\mathbb{M}_+$

For μ providing minimum of ψ on $\mathbb{M}_+$ we should have

$$D\psi(\mu)[\eta] = \int d(x,\mu)\eta(dx) \geq 0 \quad \text{for all } \eta \in T_{\mathbb{M}_+}(\mu)\,.$$

- Take $\eta = \delta_x$. Then $d(x,\mu) \geq 0$.
- Take $\eta = -\mu(\bullet \cap B)$. Then $-\int_B d(s,\mu)\mu(dx) \geq 0$. Since this is true for all B, then $d(x,\mu) \leq 0$ μ-almost everywhere.

❑ Combining this,

Theorem 3. *If $\mu \in \mathbb{M}_+$ provides minimum of ψ over $\mathbb{M}_+$ then $d(x,\mu) \geq 0\, \forall x$[a] and $d(x,\mu) = 0$ μ-almost everywhere.*

[a] For maximisation, the inequality turns to the opposite

General constrained optimisation: regularity

Let Y be a Banach space and $\mathbb{A} \subseteq \mathbb{M}$, $C \subseteq Y$ be closed convex sets.

Consider

$$\psi(\nu) \to \inf \quad \text{subject to } \nu \in \mathbb{A},\ H(\nu) \in C\,, \qquad (1)$$

where $\psi:\ \mathbb{M} \mapsto \mathbb{R}$ and $H\ :\ \mathbb{M} \mapsto Y$ are strongly differentiable.

- ν is called *regular* for (1) if

$$\operatorname{cone}(H(\nu) + DH(\nu)[\mathbb{A} - \nu] - C) = Y\,,$$ [a]

where $\operatorname{cone}(B) = \{tb :\ b \in B,\ t \geq 0\}$.

---

[a] Equivalently, $0 \in \operatorname{core}(H(\nu) + DH(\nu)[\mathbb{A} - \nu] - C)$, where $\operatorname{core}(B)$ for $B \subseteq Y$ is $\{b \in B :\ \forall y \in Y\ \exists t_1$ such that $b + ty \in B\ \forall 0 < t \leq t_1\}$. For $Y = \mathbb{R}^d,\ \operatorname{core}(B) = \operatorname{int}(\mathrm{B})$.

1st-order necessary condition for $\inf$

Let Y^* denote the dual space to Y and $u \cdot y$ be the canonical bi-linear form for $y \in Y$ and $u \in Y^*$.

Theorem 4. *Let ν such that $H(\nu) \in C$ provide a local minimum point for Problem (1). Then*

$$D\psi(\nu)[\eta] \geq 0 \quad \text{for all } \eta \in T_{\mathbb{A}\cap H^{-1}(C)}(\nu)\,. \tag{2}$$

Moreover, if ν is regular, there exists Lagrange multiplier (or Kuhn-Tucker vector) $u \in Y^*$ such that $u \cdot y \geq 0$ for any $y \in T_C(H(\nu))$ and for the Lagrangian function $L(\nu) = \psi(\nu) - u \cdot H(\nu)$ one has

$$DL(\nu)[\eta] = D\psi(\nu)[\eta] - u \cdot DH(\nu)[\eta] \geq 0 \quad \text{for all } \eta \in T_{\mathbb{A}}(\nu)\,. \tag{3}$$

Finitely many constraints on $\mathbb{M}_+$

$$\psi(\mu) \to \inf\,, \quad \mu \in \mathbb{M}_+ \tag{4}$$

subject to

$$\begin{cases} H_i(\mu) = 0\,, & i = 1, \dots, l\,; \\ H_i(\mu) \le 0\,, & i = l+1, \dots, m\,. \end{cases} \tag{5}$$

where ψ and H_i are Fréchet differentiable functions with gradients $d(x, \mu)$ and $h_i(x, \mu)$, respectively.

Constraint qualification

For constraints (5) the regularity condition becomes:

- linear independence of the gradients $h_1, \dots, h_l$; and
- existence of $\eta \in \mathbb{M}$ such that

$$\begin{cases} \int h_i(x)\,\eta(dx) = 0 & \text{for all } i = 1, \dots, l, \\ \int h_i(x)\,\eta(dx) < 0 & \text{for all } i \in \{l+1, \dots, m\} \text{ verifying } H_i(\nu) = 0^{\text{a}}\,. \end{cases}$$

It can be shown that for a regular ν,

$$T_{\mathbb{A}\cap H^{-1}(C)}(\nu) = T_{\mathbb{A}}(\nu) \cap (DH(\nu))^{-1}[T_C(H(\nu))]\,.$$

---

[a] e.g., for the saturated inequality constraints

1st-order necessary condition on $\mathbb{M}_+$

Theorem 5. *Let* $\mu \in \mathbb{M}_+$ *be a regular local minimum of* ψ *subject to* (5)*. Then there exist Lagrange multipliers* $u_1, \dots, u_m$ *with* $u_j \le 0$ *if* $H_j(\mu) = 0$ *and* $u_j = 0$ *if* $H_j(\mu) < 0$ *for* $j \in \{l+1, \dots, m\}$*, such that*

$$\begin{cases} d(x,\mu) = \sum_{i=1}^{m} u_i h_i(x,\mu) & \mu - a.e.\,, \\ d(x,\mu) \ge \sum_{i=1}^{m} u_i h_i(x,\mu) & \forall x \in X\,. \end{cases} \tag{6}$$

Proof. Apply Theorem 3 to the Lagrangian function $L(\mu) = \psi(\mu) - \sum_{i=1}^{m} u_i H_i(\mu)$. □

Optimisation with a fixed total mass

Let μ be a local minimum of ψ subject to $\mu(X) = a$. Then there exists u such that

$$\begin{cases} d(x,\mu) = u & \mu - a.e.\,, \\ d(x,\mu) \geq u & \forall x \in X\,. \end{cases} \tag{7}$$

Optimisation with a limited cost

Let μ be a regular local minimum of ψ subject to $\mu(X) = a$ and $K(\mu) = \int \kappa(x)\mu(dx) \le C$. Then there exist u_1 and $u_2 < 0$ if $K(\mu) = C$ and $u_2 = 0$ otherwise, such that

$$\begin{cases} d(x,\mu) = u_1 + u_2\kappa(x) & \mu - a.e.\,, \\ d(x,\mu) \ge u_1 + u_2\kappa(x) & \forall x \in X\,. \end{cases} \tag{8}$$

Estimation of mixture distribution

$p_\theta(\cdot)$, $\theta \in \Theta\ (= X)$, is a parametric family of pdf's

$$p_\mu(y) = \int p_\theta(y)\mu(d\theta)$$

is the mixture density, μ is unknown mixing distribution

Aim: given a random sample $y_1, \ldots, y_n$, find μ that maximises the log-likelihood

$$\psi(\mu) = \sum_{i=1}^{n} \log p_\mu(y_i)\,.$$

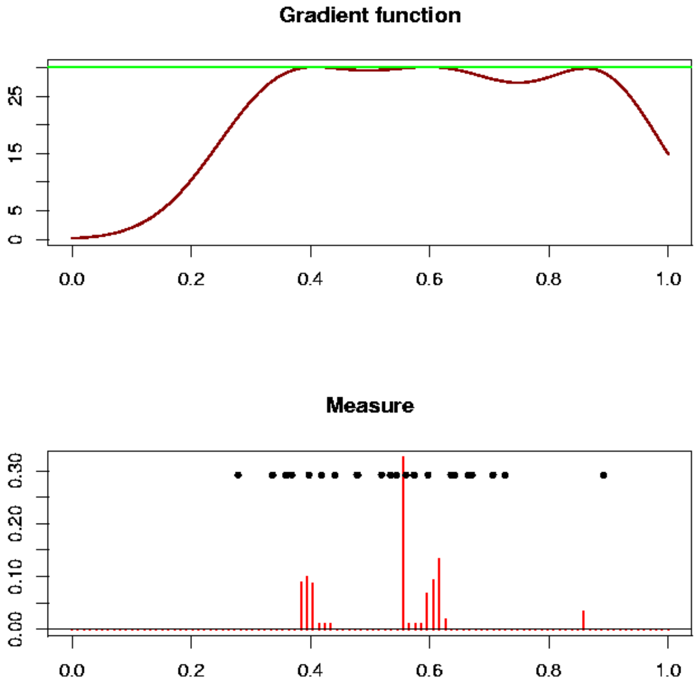
Note: ψ is concave w.r.t. μ so (7) becomes necessary *and sufficient*.

The gradient function (score function)

$$d(\theta, \mu) = \sum_{i=1}^{n} \frac{p_\theta(y_i)}{\int p_\theta(y)\mu(d\theta)}\,.$$

A synthetic example

$\Theta = [0, 1]$ discretised by 0.01, 30 observations, one third comes from $\mathcal{N}(0.4, 0.01)$ and two other thirds from $\mathcal{N}(0.6, 0.01)$. Looking to describe as mixture $\int \varphi_{(\theta,0.01)}(y)\mu(d\theta)$. Result:



- μ has 15 atoms
- The mass of μ in the neighbourhood of 0.4 is 0.3017 and in the neighbourhood of 0.6 is 0.666.
- Observe an artifact atom of mass 0.0323 at 0.859 due to an outlier observation point at 0.892.

References

- I. Molchanov and S. Zuyev. Tangent sets in the space of measures. *J. Math. Ann. Appl.*, **249**, 2000, 539–552.

❏ http://www.stams.strath.ac.uk/~sergei