# Mathematics and numerics for data assimilation and state estimation – Lecture 11

Summer semester 2020

# Overview

# Summary of lecture 10

- Weak convergence of distributions $\mathbb{P}_k \Rightarrow \mathbb{P}$.

- Bayesian inversion in the linear-Gaussian setting

$$Y = AU + \eta, \qquad \pi_U, \pi_\eta \quad \text{Gaussian pdfs.}$$

- Consistency of posterior $\pi(u|y)$ in small noise limit when $\eta$ "disappears", when $A$ is overdetermined, determined and underdetermined.

# Overview

1. Bayesian inversion and optimization
2. Entropy and Kullback-Leibler divergence

# Problem setting

$$Y = G(U) + \eta \tag{1}$$

with $G : \mathbb{R}^d \to \mathbb{R}^k$, $\eta \sim \pi_\eta$, $U \sim \pi_U$ and $\eta \perp U$.

For an observation $Y = y$, we obtained

$$\pi(u|y) \propto \pi_\eta(y - Au)\pi_U(u)$$

And in the linear-Gaussian setting

$$\pi(u|y) \propto \exp\left( -\frac{1}{2}|y - G(u)|_\Gamma^2 - \frac{1}{2}|u - \hat{m}|_{\hat{C}}^2 \right) = \exp(-\mathrm{J}(u))$$

where, decomposing into loss and regularization terms,

$$\begin{aligned} \mathrm{L}(u) &:= -\log(\pi_\eta(y - G(u))) \quad \text{and} \quad \mathrm{R}(u) := -\log(\pi_U(u)) \\ \text{and} \quad \underbrace{\mathrm{J}(u)}_{\text{Objective fcn}} &:= \mathrm{L}(u) + \mathrm{R}(u) \end{aligned} \tag{2}$$

Assuming $\pi_\eta, \pi_U > 0$, we extend the notation (2) to general settings:

$$\pi(u|y) \propto \pi_\eta(y - Au)\pi_U(u) = \exp(-\mathrm{J}(u)) = \exp(-\mathrm{L}(u) - \mathrm{R}(u)).$$

# MAP estimators and Tikhonov regularization

Maximizing the posterior is equivalent to minimizing the objective function:

$$u_{MAP}[\pi(\cdot|y)] = \arg\max_{u\in\mathbb{R}^d} \pi(u|y) = \qquad = \arg\min_{u\in\mathbb{R}^d} J(u)$$

- In Gaussian setting, with $U|Y = y \sim N(m, C)$ and $U \sim N(0, \lambda^{-1} I)$,

$$u_{MAP} = m = \arg\min_{u\in\mathbb{R}^d} \frac{1}{2}|y - G(u)|_\Gamma^2 + \frac{\lambda}{2}|u|^2.$$

- This corresponds to Tikhonov regularization. Unique, closed form solution in linear setting $G(u) = Au$.

# Laplace-distributed prior and LASSO regression

- Alternatively, consider the prior with iid Laplace-distributed components

$$\pi_U(u) = \prod_{i=1}^{d} \pi_{U_i}(u_i) \propto \prod_{i=1}^{d} e^{-\lambda|u_i|} = e^{-\lambda|u|_1}$$

  where

$$|u|_p := \Big( \sum_{j=1}^{d} |u_j|^p \Big)^{1/p}, \qquad p > 0.$$

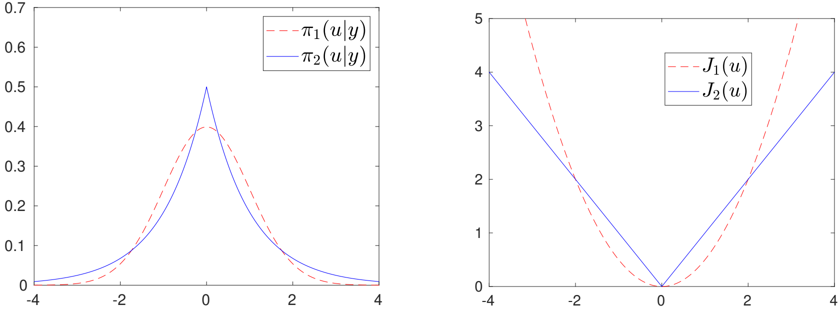
- This yields

$$\mathrm{R}(u) \propto \lambda|u|_1 \quad \text{and} \quad u_{MAP} = \arg\min_{u\in\mathbb{R}^d} \frac{1}{2}|y - G(u)|_\Gamma^2 + \lambda|u|_1$$

  which corresponds to lasso (least absolute shrinkage and slection operator) regression.
- Generally, lasso has no closed-form solution, but a solution is typically attainable. It tends to produce more sparse solutions than Tikhonov.

Posterior setting with R $\gg$ L and regularizers so that approximately

$$\pi_1(u|y) \propto \exp(-|u|^2/2) \quad \text{and} \quad \pi_2(u|y) \propto \exp(-|u|_1).$$



# Attainability of $u_{MAP}$

**Theorem 1**

*Assume that the objective fcn $J : \mathbb{R}^d \to \mathbb{R}$ is bounded from below, continuous and that $J(u) \to \infty$ as $|u| \to \infty$. Then $J$ attains its infimum, which implies that*

$$u_{MAP}[\pi(\cdot|y)] \quad \textit{is attained for} \quad \pi(u|y) \propto \exp(-J(u)).$$

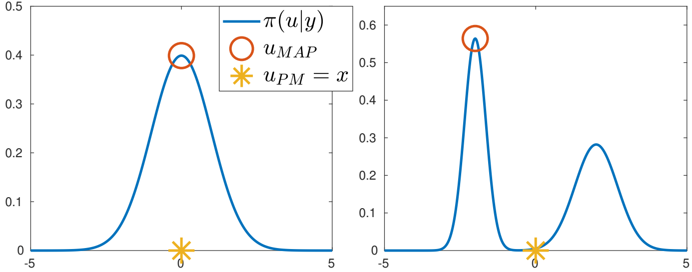
Sufficient conditions for attainable $u_{MAP}$:

- $G \in C(\mathbb{R}^d, \mathbb{R}^k)$ and $\eta \sim N(0, \Gamma)$,
- $\mathrm{R}(u) = \lambda |u|_p^p$ for any $\lambda, p > 0$
  (as this implies $\mathrm{J}(u) \to \infty$ as $|u| \to \infty$).

# Examples of the MAP performing poorly

- "All happy families are alike; each unhappy family is unhappy in its own way." Leo Tolstoy, in Anna Karenina
- Paraphrasing: "All unimodal densities are alike; each multimodal density is multimodal in its own way"

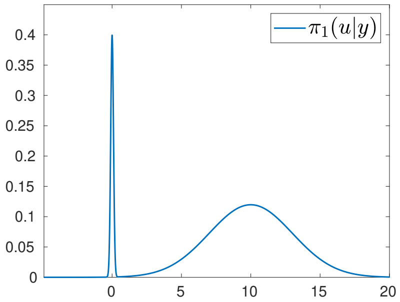
In Lecture 7 we already saw that $u_{MAP}$ can be of limited value for bimodal densities:

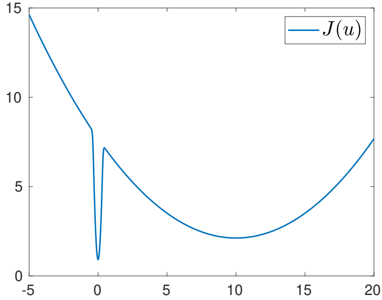


# Slab-spike figure

For

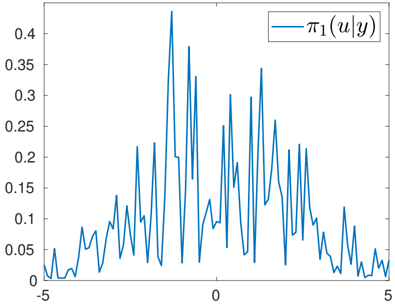
$$\pi(u|y) = \frac{\exp(-|u|^2/0.02) + 0.3\exp(-|u-10|^2/18)}{\sqrt{2\pi}}$$

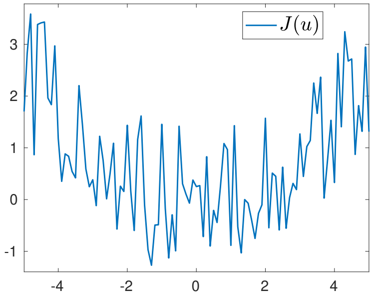


# Low-regularity objective function

```
normalF = @(x) (x).^2/10;
objective = normalF(x)+1.5*(1-2*rand(size(x)));
posterior = exp(-objective);
posterior = exp(-objective)/(trapz(posterior)*dx);
```



And low-regularity in higher dimensions . . .

Figure: Photo by Michel Royon / Wikimedia Commons

# Overview

# Low-rank approximations of posteriors

- We have seen that one-parameter/vector compression of a posterior, like MAP or posterior mean, may provide little information.
- Natural next step: Extend the compressed representations of posteriors to best fitting in a class of candidate densities:

$$p^* = \arg\inf_{p \in \mathcal{A}} d(p, \pi(\cdot|y))$$

  for some $d : \mathcal{M} \times \mathcal{M} \to [0, \infty)$

- Here we will restrict ourselves to

$$\mathcal{A} = \{p = PDF(N(\mu, C)) \mid \mu \in \mathbb{R}^d \text{ and } C \in \mathbb{R}^{d \times d} \text{ and pos definite}\}$$

  which can be viewed as a two-parameter (two-moment) compression of a posterior.

# Kullback-Leibler divergence

## Definition 2 (K-L divergence)

- For positive discrete measures: Let

$$\mathcal{P}_+ = \{\text{Probability measures on } A \mid \mathbb{P}(x), \mathbb{Q}(x) > 0 \text{ for all } x \in A\}.$$

For all $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_+$,

$$d_{KL}(\mathbb{P}||\mathbb{Q}) := \sum_{x \in A} \log\left(\frac{\mathbb{P}(x)}{\mathbb{Q}(x)}\right)\mathbb{P}(x).$$

- For positive pdfs on $\mathbb{R}^d$: Let

$$\mathcal{M}_+ := \{\pi \in \mathcal{M} \mid \pi(x) > 0 \quad \forall x \in \mathbb{R}^d\}.$$

For all $\pi, p \in \mathcal{M}_+$

$$d_{KL}(\pi||p) := \int_{\mathbb{R}^d} \log\left(\frac{\pi(x)}{p(x)}\right)\pi(x)\,dx = \mathbb{E}^{\pi}\left[\log\left(\frac{\pi}{p}\right)\right]$$

# Properties of the K-L divergence

For all $\pi, p \in \mathcal{M}_+$, it holds that $d_{KL}(\pi||p) \in [0, \infty]$ (similar result holds for prob measures).

Example of infinite K-L divergence:

$$p(x) \propto e^{-|x|}, \quad \pi \propto (1 + |x|)^{-2}, \quad x \in \mathbb{R}$$

Then

$$\begin{aligned}
d_{KL}(\pi||p) &= \int_{\mathbb{R}} \log \left(\frac{\pi(x)}{p(x)}\right) \pi(x)\, dx \\
&= C \int_{\mathbb{R}} \Big( \log(\pi(x)) - \log(p(x)) \Big) \pi(x)\, dx \\
&= C \int_{\mathbb{R}} \frac{-2\log((1 + |x|)) + |x|}{(1 + |x|)^2} \Big) \pi(x) dx \\
&= \infty.
\end{aligned}$$

# Properties of the K-L divergence

$d_{KL}$ is not a metric; neither does it saitisfy the triangle inequality nor is it symmetric in its arguments.

**Example:** Let $A = \{1, 2, 3\}$ and $\mathbb{P}(1) = \mathbb{P}(2) = \mathbb{P}(3) = 1/3$ and $\mathbb{Q}(1) = 1/2$, $\mathbb{Q}(2) = 1/3$, $\mathbb{Q}(3) = 1/6$. Then

$$\begin{aligned} d_{KL}(\mathbb{P}||\mathbb{Q}) &= \sum_{x_i \in A} \log\left(\frac{\mathbb{P}(x_i)}{\mathbb{Q}(x_i)}\right)\mathbb{P}(x_i) \\ &= \frac{\log(2/3) + \log(1) + \log(2)}{3} \approx 0.0959 \end{aligned}$$

while

$$d_{KL}(\mathbb{Q}||\mathbb{P}) = \frac{3\log(3/2) + 2\log(1) + \log(1/2)}{6} \approx 0.0872$$

# Properties of the K-L divergence

- K-L divergence has natural applications in information theory and thermodynamics.

- In Bayesian inference, for a prior $\pi_U$ and a posterior $\pi(\cdot|y)$, $d_{KL}(\pi(\cdot|y), \pi_U)$ is a measure of the information gain of replacing the prior by the posterior.

- The logarithm base in the definition of K-L divergence is flexible; use what is most suitable for the application (here, log denotes the natural logarithm).

## Lemma 3 (Lower bounds for K-L divergence, (SST 4.2))

*For any pi, $p \in \mathcal{M}_+$ it holds that*

$$d_H(\pi, p)^2 \leq \frac{1}{2} d_{KL}(\pi || p) \quad \textit{and} \quad d_{TV}(\pi, p)^2 \leq d_{KL}(\pi || p).$$

**Proof of first inequality:**

$$\begin{aligned}
d_H(\pi, p)^2 &= \frac{1}{2} \int_{\mathbb{R}^d} (\sqrt{\pi} - \sqrt{p})^2 \, dx \\
&= \\
&= \\
&= \int_{\mathbb{R}^d} \left(1 - \sqrt{\frac{p}{\pi}}\right) \pi \, dx \leq -\frac{1}{2} \int_{\mathbb{R}^d} \log\left(\frac{p}{\pi}\right) \pi \, dx = \frac{1}{2} d_{KL}(\pi || p).
\end{aligned}$$

where we used that

$$1 - \sqrt{x} \leq -\frac{1}{2} \log(x) \quad \forall x \in [0, \infty].$$

# Comments

- Second inequality follows from $d_{TV}(\pi, p) \leq \sqrt{2} d_H(\pi, p)$.

- The lemma implies that K-L divergence is point/density separating: For all $\pi, p \in \mathcal{M}_+$,
$$d_{KL}(\pi || p) \geq 0$$
and
$$d_{KL}(\pi || p) = 0 \iff p = \pi.$$
(Similar for measures.)

# Entropy in information theory

Suppose you want to transmit a very long text encoded in some alphabet, e.g., $A = \{a, b, c, d, e\}$,

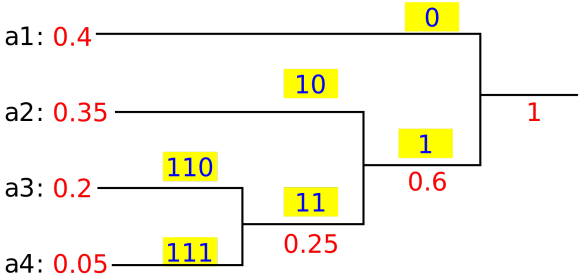
```
TEXT=  "abbedeeeedcaababecbddaeedeccabe..."
```

and that

- the data-transmission problem can to good approximation be viewed as transmitting a sequence iid characters drawn with relative frequencies $\mathbb{P}(a)$, $\mathbb{P}(b)$ etc.
- you want to send the text over a digital communication channel with alphabet $\{0, 1\}$. Hence, each letter in your original alphabet must be replaced with a codeword, e.g. $a = 101$, $b = 111$, and your want the digitally encoded text to be as short as possible.
- Core idea: assign shortest codeword to most frequent letter in the text, second shortest codeword to . . . (then there is a subtle issue with uniqueness/reversibility of encoding).

# Huffman encoding

Input alphabet: $A = \{a1, a2, a3, a4\}$.




Letter frequency: $\mathbb{P}(a1) = 0.4$, $\mathbb{P}(a2) = 0.35$ etc

Digital codewords: $a1 = 0$, $a2 = 10$, etc

NB! A shorter encoding is possible: $a1 = 0$, $a2 = 1$, $a3 = 10$ and $a4 = 11$ but this encoding is, unlike Huffman's, not uniquely reversible, since it is not injective when applied to strings:

$$a4 \mapsto 11 \quad a2a2 \mapsto 11$$

## Shannon's approach

Shannon relates the text-frequency of a letter to the information content:

**Definition 4 (Information content of a character)**

For an event/character $a$ which occurs with probability $\mathbb{P}(a)$ we define its information content by

$$I(a) := -\log_2(\mathbb{P}(a))$$

**Idealized motivation:** if there are $1/\mathbb{P}(a)$ many independent events, each occurring with probability $\mathbb{P}(a)$, how many bits do I need to distinguish all these events when encoded in $\{0, 1\}$?

**Example** Alphabet $A = \{a, b, c, d, e\}$ with uniform letter probability $1/5$. Then at least $-\lceil \log_2(1/5) \rceil = 3$ bits are needed to distinguish the letters/events.

# Shannon entropy

**Generalization:** Information content straightforwardly generalizes from a character to any text string $B$

$$I(B) := -\log_2(\mathbb{P}(B))$$

**where we recall that** letter sequences, e.g., $B = abeba$, are assumed to consist of iid characters,

$$\mathbb{P}(abeba) = \mathbb{P}(a)\mathbb{P}(b)\mathbb{P}(e)\mathbb{P}(b)\mathbb{P}(a)$$

**Lemma 5 (Information content of independent events)**

*Let $B$ and $C$ denote two independent events (i.e., text strings), then the information content of $B$ and $C$ is additive*

$$I(BC) := I(B) + I(C)$$

**Verification for two-character sequence:** Consider basic events $B = a$ and $C = b$. Then

$$I(ab) = -\log_2(\mathbb{P}(ab)) = -\log_2(\mathbb{P}(a)\mathbb{P}(b)) = I(a) + I(b)$$

# Shannon entropy

**Question:** Given a text encoded in the alphabet $A = \{a_1, \dots, a_n\}$ with relative frequencies $\{\mathbb{P}(a_k)\}_k$ , and a digital encoding representing the letter $a_k$ by $I(a_k)$ bits (we allow fractional-bit encoding in this thought experiment) then if the original text consists of $N \gg 1$ characters, how long does the digitally encoded text become?

**Answer:**

$$N \times \text{mean num of bits for single } A\text{-character} = N \sum_{k=1}^{n} I(a_k)\mathbb{P}(a_k)$$

Introducing the information content rv

$$I_{\mathbb{P}}(a) := -\log_2(\mathbb{P}(a)), \quad (I_{\mathbb{P}} : A \to [0, \infty], \text{ and } \mathbb{P}_{I_{\mathbb{P}}}(I_{\mathbb{P}}(a)) = \mathbb{P}(a)),$$

we may associate the above with the expected information content/Shannon entropy

$$\mathbb{E}^{\mathbb{P}}[I_{\mathbb{P}}] = \sum_{k=1}^{n} I_{\mathbb{P}}(a_k)\mathbb{P}(a_k) = -\sum_{k=1}^{n} \log_2(\mathbb{P}(a_k))\mathbb{P}(a_k)$$

# Comparison of encoding methods

Assume that a text encoded in $A = \{a_1, \dots, a_n\}$ has true relative frequencies $\{\mathbb{P}(a_k)\}$, but that

- you only have an approximation of the relative frequencies $\{\mathbb{Q}(a_k)\}$
- and that given $\mathbb{Q}$, your encoding in $\{0, 1\}$ is optimal, meaning it uses $l_{\mathbb{Q}}(a_k) = -\log_2(\mathbb{Q}(a_k))$ bits to encode the letter $a_k$.

**K-L divergence is a comparison of efficiency $\mathbb{Q}$- vs $\mathbb{P}$-encoding:**

$$[\text{mean } \mathbb{Q}\text{-bits in encoded } A\text{-char}] \quad - \quad [\text{mean } \mathbb{P}\text{-bits in encoded } A\text{-char}]$$

$$= \sum_{k=1}^{n} (l_{\mathbb{Q}}(a_k) - l_{\mathbb{P}}(a_k))\mathbb{P}(a_k)$$

$$= \sum_{k=1}^{n} (\log_2(\mathbb{P}(a_k)) - \log_2(\mathbb{Q}(a_k))\mathbb{P}(a_k)$$

$$= \sum_{k=1}^{n} \log_2\left(\frac{\mathbb{P}(a_k)}{\mathbb{Q}(a_k)}\right)\mathbb{P}(a_k) = d_{KL}(\mathbb{P}||\mathbb{Q})$$

# Best encoding in a set

Given a collection of encodings, a natural task is to find the most efficient one:

$$\mathbb{Q}^* = \arg\min_{\mathbb{Q}\in\mathcal{A}} d_{KL}(\mathbb{P}||\mathbb{Q}).$$

**Example:** Let $A = \{a, b, c, d, e\}$ and $\mathbb{P}(a) = \mathbb{P}(b) = \ldots = \mathbb{P}(d) = 1/5$, and $\mathcal{A} = \{\mathbb{Q}_1, \mathbb{Q}_2\}$ with

$$\mathbb{Q}_1(a) = \mathbb{Q}_1(b) = \mathbb{Q}_1(c) = \mathbb{Q}_1(d) = 2^{-4}, \quad \mathbb{Q}_1(e) = 3/4$$

and

$$\mathbb{Q}_2(a) = \mathbb{Q}_2(b) = \mathbb{Q}_2(c) = \mathbb{Q}_2(d) = 2^{-5}, \quad \mathbb{Q}_2(e) = 7/8.$$

**Result:** $\mathbb{Q}^* = \mathbb{Q}_1$ as

$$d_{KL}(\mathbb{P}||\mathbb{Q}_1) = \frac{4\log_2(16/5) + \log_2(4/15)}{5} \approx 0.9611$$

and

$$d_{KL}(\mathbb{P}||\mathbb{Q}_2) = \frac{4\log_2(32/5) + \log_2(8/35)}{5} \approx 1.7166$$

# Connecting information theory and random variables

For discrete distributions $\mathbb{P}$ and $\mathbb{Q}$ on $A$ we defined the information content rv

$$I_{\mathbb{P}}(a) = -\log(\mathbb{P}(a)), \quad I_{\mathbb{Q}}(a) = -\log(\mathbb{Q}(a))$$

and the K-L divergence from $\mathbb{Q}$ to $\mathbb{P}$ takes the form

$$d_{KL}(\mathbb{P}||\mathbb{Q}) = \mathbb{E}^{\mathbb{P}}[I_{\mathbb{Q}} - I_{\mathbb{P}}] = \sum_{a\in A} \log\left(\frac{\mathbb{P}(a)}{\mathbb{Q}(a)}\right)\mathbb{P}(a)$$

For continuous rv $X, Y$ with densities $\pi_X, \pi_Y \in \mathcal{M}_+$, we define the information content as

$$I_{\pi_X}(x) = -\log(\pi_X(x)), \quad I_{\pi_Y}(x) = -\log(\pi_Y(x))$$

and

$$d_{KL}(\pi_X||\pi_Y) = \mathbb{E}^{\pi_X}[I_{\pi_Y} - I_{\pi_X}] = \int_{\mathbb{R}^d} \log\left(\frac{\pi_X(x)}{\pi_Y(x)}\right)\pi_X(x)\, dx$$

# Expected information gain Bayesian inversion

For the additive Gaussian inverse problem

$$Y = G(U) + \eta \tag{3}$$

with $\pi_\eta, \pi_U \in \mathcal{M}_+$ and $U \perp \eta$, the posterior is also a strictly positive pdf

$$\pi(u|y) = \frac{\exp(-\mathrm{L}(u))\pi_U(u)}{Z}. \tag{4}$$

Then

$$d_{KL}(\pi(\cdot|y)||\pi_U) = \mathbb{E}^{\pi(\cdot|y)}[I_{\pi_U} - I_{\pi(\cdot|y)}]$$

is a measure of the information gained by revising the prior $\pi_U$ into the posterior $d_{KL}(\pi(\cdot|y)||\pi_U)$

**Interpretation:** wrt $\pi(\cdot|y)$, $I_{\pi(\cdot|y)}$ yields the minimum expected information content, so, as we already know,

$$\mathbb{E}^{\pi(\cdot|y)}[I_{\pi_U} - I_{\pi(\cdot|y)}] \geq 0.$$

# Variational formulation of Bayes theorem

**Theorem 6 (SST Thm 4.9)**

*For the inverse problem* (3) *it holds that*

$$\pi(\cdot|y) = \arg\min_{p \in \mathcal{M}_+} d_{KL}(p||\pi_U) + \mathbb{E}^p[L(u)]$$

**Verification:** Recalling that $\pi(\cdot|y) = \frac{\exp(-\mathrm{L}(u))\pi_U(u)}{Z}$,

$$\begin{aligned}
d_{KL}(p||\pi(\cdot|y)) &= \int_{\mathbb{R}^d} \log\left(\frac{p\,\pi_U}{\pi(x|y)\,\pi_U}\right) p(x)\,dx \\
&= \int_{\mathbb{R}^d} \log\left(\frac{pZ\exp(\mathrm{L}(u))}{\pi_U}\right) p(x)\,dx \\
&= \int_{\mathbb{R}^d} \log\left(\frac{p}{\pi_U}) + \mathrm{L}(u)\right) p(x)\,dx + \log(Z) \\
&= d_{KL}(p||\pi_U) + \mathbb{E}^p[\mathrm{L}] + \log(Z)
\end{aligned}$$

and

$$\pi(\cdot|y) = \arg\min_{p \in \mathcal{M}_+} d_{KL}(p||\pi(\cdot|y)).$$

# Best Gaussian fit and K-L divergence

Consider again the posterior obtained from the inverse problem (3),

$$\pi(u|y) = \frac{\exp(-\mathrm{L}(u))\pi_U(u)}{Z}. \tag{5}$$

**Theorem 7**

*Assume that L is non-negative, continuous, and globally bounded from above and that $U \sim N(0, \lambda^{-1} I)$ for some $\gamma > 0$. Then there exists at least one pdf p in*

$$\mathcal{A} := \{\rho = PDF(N(\mu, C)) \mid \mu \in \mathbb{R}^d \text{ and } C \in \mathbb{R}^{d\times d} \text{ and pos definite}\}. \tag{6}$$

*which satisfies the best-Gaussian-fit-of-posterior condition*

$$d_{KL}(p||\pi(\cdot|y)) = \inf_{\rho\in\mathcal{A}} d_{KL}(\rho||\pi(\cdot|y))$$

Essential fitting idea:

$$\text{make} \quad \log\left(\frac{p(x)}{\pi(x|y)}\right) \quad \text{small} \quad \text{i.e.,} \quad \frac{p}{\pi(\cdot|y)} \lessapprox 1.$$

# Ideas in proof

For $p_{\mu,C} = PDF(N(\mu, C))$ it is possible to show that for

$$I(\mu, C) := d_{KL}(p_{\mu,C} || \pi(\cdot|y))$$

it holds that

$$I(0, I) < \infty, \quad \lim_{|\mu| \to \infty} I(\mu, C) = \infty$$

and

$$\lim_{trace(C) \to 0} I(\mu, C) = \lim_{trace(C) \to \infty} I(\mu, C) = \infty.$$

Consequently, there exists $R > r > 0$ s.t.

$$\arg \inf_{p \in \mathcal{A}} d_{KL}(p||\pi) \in \tilde{\mathcal{A}}_{r,R}$$

where

$$\tilde{\mathcal{A}}_{r,R} = \{p_{\mu,C} \in \mathcal{A} \mid |\mu| < R, \quad \text{and} \quad r < trace(C) < R\}.$$

# Best Gaussian fit by moment matching

One may also fit $p$ to $\pi$ by minimizing $d_{KL}(\pi(\cdot|y)||p)$

**Theorem 8 (SST Thm 4.5)**

*Let $\pi(\cdot|y)$ denote the posterior density of the inverse problem (3). If $\bar{\mu} = \mathbb{E}^{\pi(\cdot|y)}[u]$ is finite and $\bar{C} = \mathbb{E}^{\pi(\cdot|y)}[(u-\bar{\mu})(u-\bar{\mu})^T]$ is finite and positive definite then*

$$p_{\bar{\mu},\bar{C}} = \arg\inf_{p\in\mathcal{A}} d_{KL}(\pi||p),$$

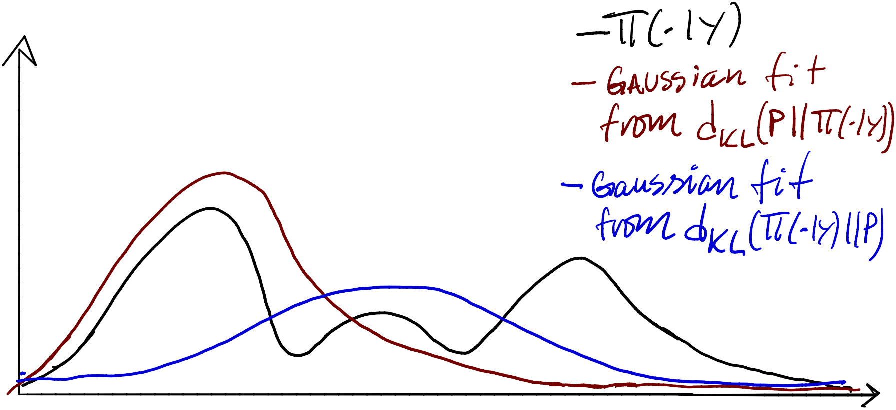
*and the minimizer $p_{\bar{\mu},\bar{C}}$ is unique.*

Essential fitting idea:

$$\text{make} \quad \log(\frac{\pi(x|y)}{p(x)}) \quad \text{small, i.e.,} \quad \frac{\pi(\cdot|y)}{p} \lessapprox 1.$$

## Comparison of the fitting approaches

- For $d_{KL}(p||\pi(\cdot|y))$: make $\frac{p}{\pi(\cdot|y)} \lessapprox 1$
- For $d_{KL}(\pi(\cdot|y)||p)$: make $\frac{\pi(\cdot|y)}{p} \lessapprox 1$



# Next time

- discrete time continuous state-space Markov chains
- Markov chain Monte Carlo methods
- introduction to smoothing and filtering in continuous state-space