Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Thirteenth lecture, July 10, 2023

The setting

We work in the inverse problem setting of finding $x \in \mathbb{R}^d$ from $y \in \mathbb{R}^k$ given by

$$y = F(x) + \eta$$

with noise $\eta \sim \nu$ and prior $x \sim \pi$ such that $\eta \perp x$. The posterior density π^y of $x|y$ is given by Bayes' theorem

$$\pi^y(x) = \frac{1}{Z}\nu(y - F(x))\pi(x).$$

We have the negative log-likelihood:

$$\mathrm{L}(x) = -\log \nu\big(y - F(x)\big),$$

and a regularizer

$$\mathrm{R}(x) = -\log \pi(x).$$

So far we have mainly discussed point estimators: the MAP estimate

$$\hat{x}_{\mathsf{MAP}} = \arg\max_{x \in \mathbb{R}^d} \pi^y(x) = \arg\min_{x \in \mathbb{R}^d} (\mathsf{L}(x) + \mathsf{R}(x))$$

requires solving an optimization problem, and the CM estimate

$$\hat{x}_{\mathsf{CM}} = \int_{\mathbb{R}^d} x \pi^y(x) \, \mathrm{d}x$$

requires solving a high-dimensional integral. Recall that the latter can be achieved, e.g, by using MCMC to draw a sufficiently large sample from the posterior and computing the sample average. If we have a sample drawn from the posterior, we can use the sample to estimate other statistics such as the variance or credibility regions as well. Some alternatives to MCMC include importance sampling, high-dimensional cubature rules, etc.

Using point estimators reduces the complexity of Bayesian inference from determination of an entire distribution to determination of a single point. However, the approach has some limitations, in particular for noisy, multi-peaked or high-dimensional posterior distributions, where a point estimator may not capture enough information about the density.

Unimodal distributions

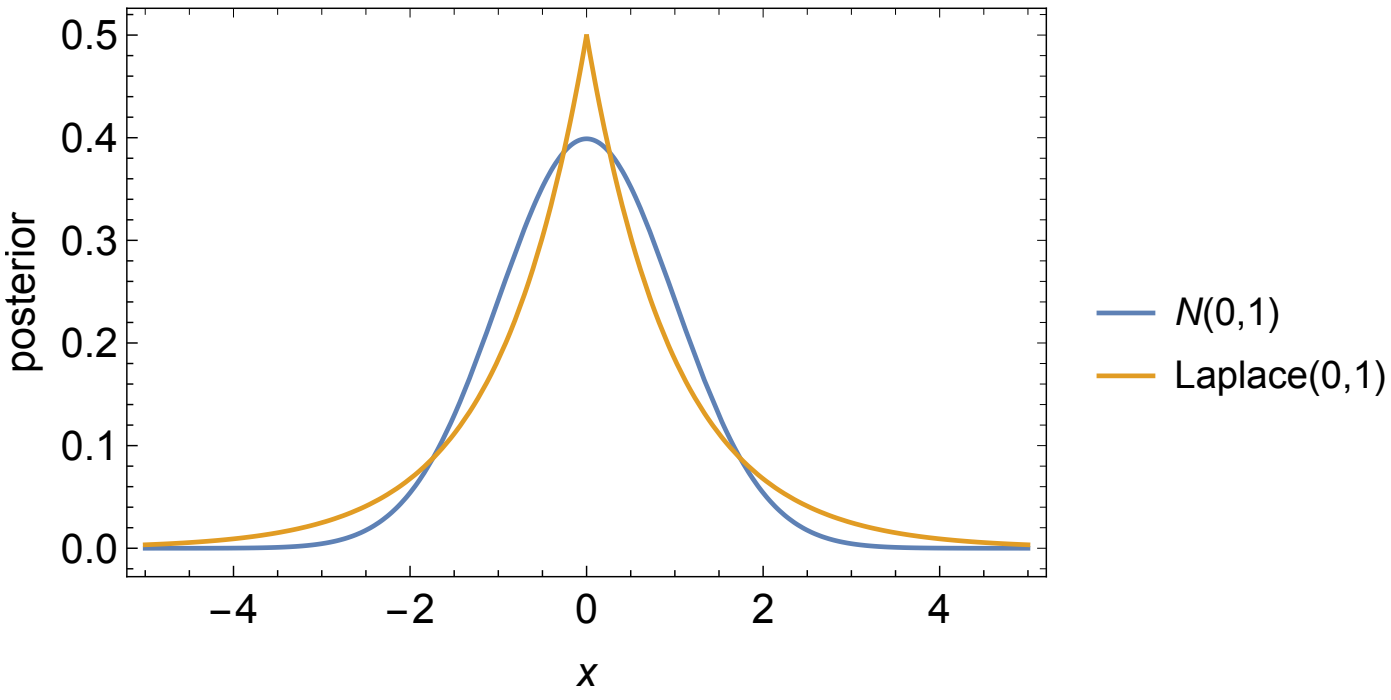


Figure: If the posterior is single-peaked, the MAP estimator reasonably summarizes the most likely value of the unknown parameter.

Problems with uneven distributions

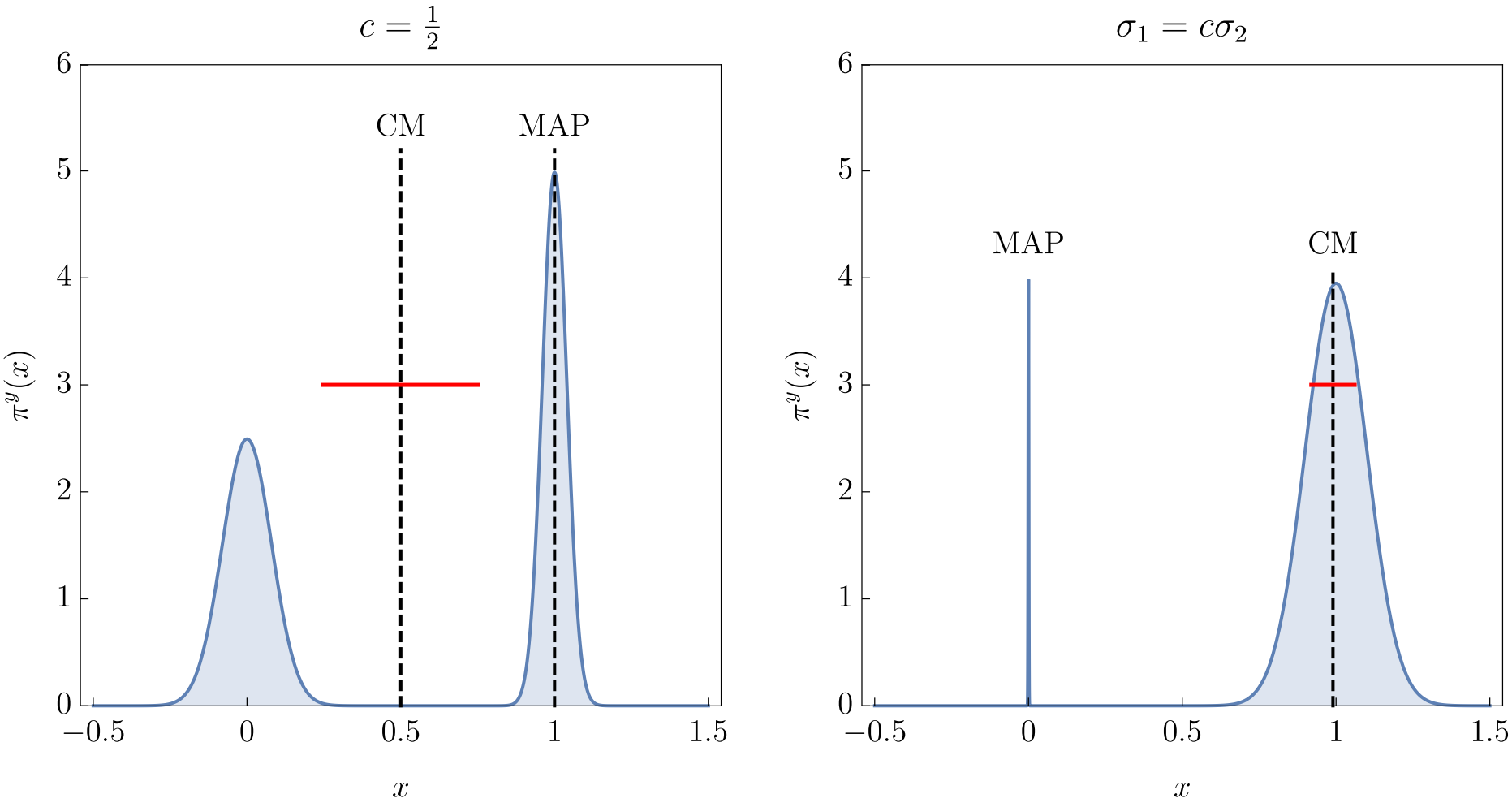


Figure: If the posterior is unevenly distributed, then it is less clear that the MAP or CM estimators usefully summarize the posterior.

Problems with rough distributions

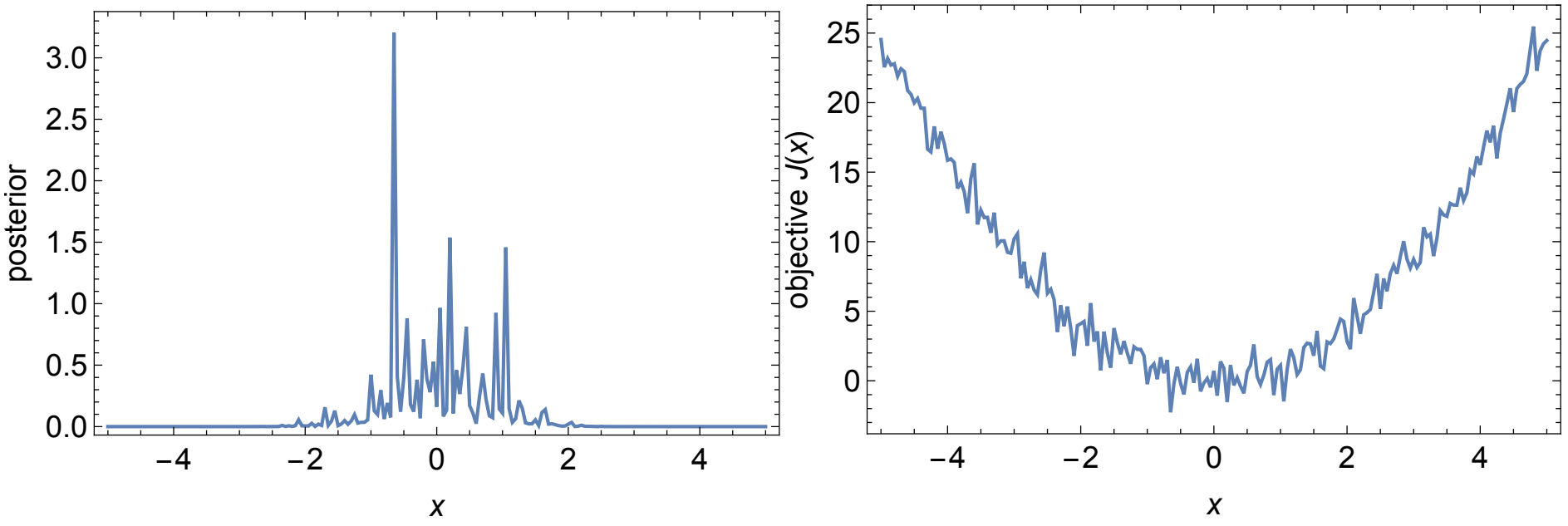


Figure: If the objective function J(x) is very rough (here it is a quadratic function contaminated with white noise), then the resulting posterior density is very rough.

The objective function has small-scale roughness, but it has a larger pattern. The MAP estimator cannot capture this larger pattern as it is found by minimizing the objective function. Arguably, $x = 0$ might be a better point estimate.

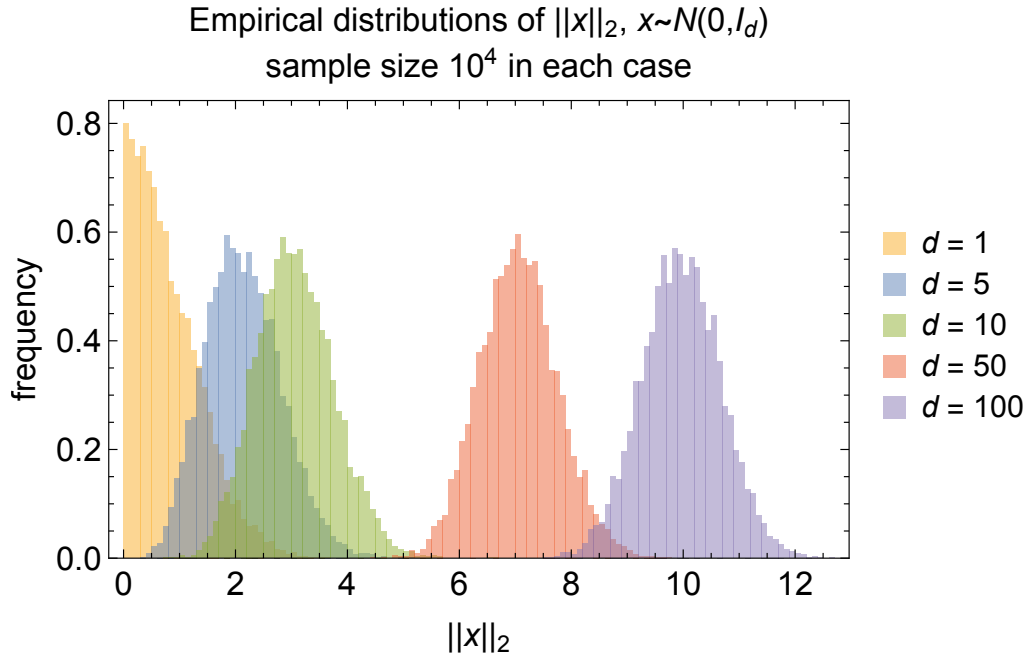
Problems with high dimension

Gaussian Annulus Theorem: Nearly all the probability of a d-dimensional spherical Gaussian distribution with unit variance is concentrated in a thin annulus of width $\mathcal{O}(1)$ at radius $\sqrt{d}$.

For example, if $x \sim \mathcal{N}(0, I_d)$, then

d	$\mathbb{P}(\|x\| < 5)$
10	0.99465
50	0.00119
100	1.135e-15

A point estimator may not capture enough information about the density.



Gaussian approximation

Instead of seeking a point estimator, we can try seeking a Gaussian distribution $p = \mathcal{N}(\mu, \Sigma)$ that minimizes the Kullback–Leibler divergence from the posterior $\pi^y(x)$. Since the Kullback–Leibler divergence is not symmetric this leads to two distinct problems, which we will consider separately.

The Kullback–Leibler divergence

Definition

Let $\pi, \pi' > 0$ be two probability distributions on $\mathbb{R}^d$. The *Kullback–Leibler (KL) divergence*, or *relative entropy*, of π with respect to π' is defined by

$$\begin{aligned} d_{\mathrm{KL}}(\pi \| \pi') &:= \int_{\mathbb{R}^d} \log\left(\frac{\pi(x)}{\pi'(x)}\right) \pi(x)\,\mathrm{d}x \\ &= \mathbb{E}^{\pi}\left[\log\left(\frac{\pi}{\pi'}\right)\right] \\ &= \mathbb{E}^{\pi'}\left[\log\left(\frac{\pi}{\pi'}\right)\frac{\pi}{\pi'}\right]. \end{aligned}$$

Kullback–Leibler is a divergence in that $d_{\mathrm{KL}}(\pi \| \pi') \geq 0$, with equality if and only if $\pi = \pi'$ a.e. However, unlike Hellinger and total variation, it is not a distance. In particular, the KL divergence is not symmetric: in general

$$d_{\mathrm{KL}}(\pi \| \pi') \neq d_{\mathrm{KL}}(\pi' \| \pi).$$

The KL divergence is useful for at least the following reasons:

- it provides an upper bound for many distances;
- its logarithmic structure allows explicit computations that are difficult using actual distances;
- it satisfies many convenient analytical properties such as being convex in both arguments and lower-semicontinuous in the topology of weak convergence;
- it has an information-theoretic and physical interpretation.

Lemma

The KL divergence provides the following upper bounds for Hellinger and total variation distance:

$$d_{\mathrm{H}}(\pi,\pi')^2 \leq \frac{1}{2} d_{\mathrm{KL}}(\pi\|\pi'), \quad d_{\mathrm{TV}}(\pi,\pi')^2 \leq d_{\mathrm{KL}}(\pi\|\pi').$$

Proof. Recall from Week 9 that $\frac{1}{\sqrt{2}} d_{\mathrm{TV}}(\pi,\pi') \leq d_{\mathrm{H}}(\pi,\pi')$ $\Leftrightarrow d_{\mathrm{TV}}(\pi,\pi')^2 \leq 2 d_{\mathrm{H}}(\pi,\pi')^2$. Thus the second inequality follows from the first one. We prove only the first inequality.

Consider the function $\phi\colon \mathbb{R}_+ \to \mathbb{R}$ defined by

$$\phi(x) = x - 1 - \log x.$$

Note that

$$\begin{aligned}
\phi'(x) &= 1 - \frac{1}{x}, \\
\phi''(x) &= \frac{1}{x^2}, \\
\lim_{x\to+\infty} \phi(x) &= \infty = \lim_{x\to 0+} \phi(x).
\end{aligned}$$

Thus the function is convex on its domain. As the minimum of ϕ is attained at $x = 1$, and as $\phi(1) = 0$, we deduce that $\phi(x) \geq 0$ for all $x \in (0, \infty)$. Hence,

$$\begin{aligned}
x - 1 &\geq \log x \qquad \text{for all } x > 0, \\
\sqrt{x} - 1 &\geq \frac{1}{2} \log x \qquad \text{for all } x > 0.
\end{aligned}$$

We can use this last inequality to bound the Hellinger distance:

$$
\begin{aligned}
d_{\mathsf{H}}(\pi, \pi')^2 &= \frac{1}{2} \int_{\mathbb{R}^d} \left(1 - \sqrt{\frac{\pi'}{\pi}}\right)^2 \pi \, \mathrm{d}x \\
&= \frac{1}{2} \int_{\mathbb{R}^d} \left(1 + \frac{\pi'}{\pi} - 2\sqrt{\frac{\pi'}{\pi}}\right) \pi \, \mathrm{d}x \\
&= 1 - \int_{\mathbb{R}^d} \sqrt{\frac{\pi'}{\pi}} \pi \, \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \left(1 - \sqrt{\frac{\pi'}{\pi}}\right) \pi \, \mathrm{d}x \\
&\leq -\frac{1}{2} \int_{\mathbb{R}^d} \log\left(\frac{\pi'}{\pi}\right) \pi \, \mathrm{d}x = \frac{1}{2} \int_{\mathbb{R}^d} \log\left(\frac{\pi}{\pi'}\right) \pi \, \mathrm{d}x = \frac{1}{2} d_{\mathsf{KL}}(\pi \| \pi').
\end{aligned}
$$

Lemma

$d_{\mathrm{KL}}(\pi\|\pi') = 0$ if and only if $\pi = \pi'$ a.e.

Proof. The sufficient direction is trivial. For the necessary direction, suppose that $d_{\mathrm{KL}}(\pi\|\pi') = 0$. From the previous lemma, we deduce that

$$0 \leq d_{\mathrm{TV}}(\pi, \pi')^2 \leq d_{\mathrm{KL}}(\pi\|\pi') = 0$$

and therefore

$$d_{\mathrm{TV}}(\pi, \pi') = \frac{1}{2} \int_{\mathbb{R}^d} |\pi(x) - \pi'(x)| \, \mathrm{d}x = 0,$$

which can only hold if $\pi = \pi'$ a.e. □

Best Gaussian approximation

Let π be the target distribution, e.g., the posterior. We consider two different minimization problems, both leading to a "best Gaussian":

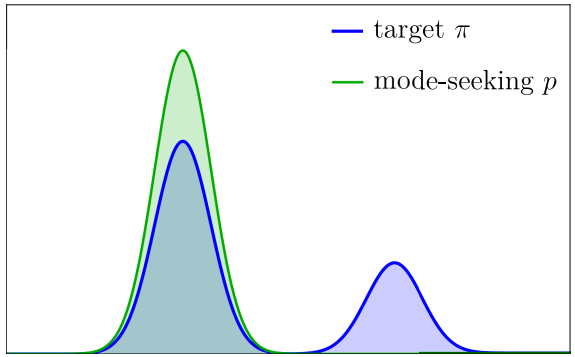
$$\inf_{p \in \mathcal{A}} d_{\mathrm{KL}}(p \| \pi) \qquad \text{("Mode-seeking Gaussian approximation")}$$

and

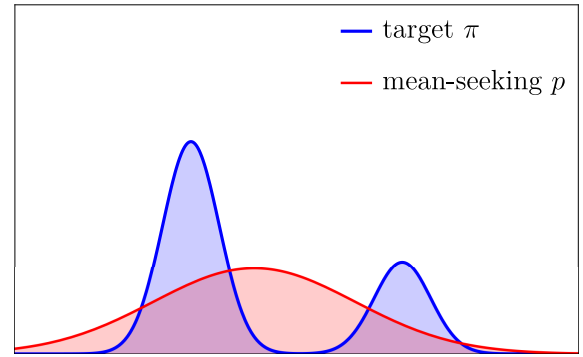
$$\inf_{p \in \mathcal{A}} d_{\mathrm{KL}}(\pi \| p), \qquad \text{("Mean-seeking Gaussian approximation")}$$

where the minimization is performed over the set of Gaussian distributions on $\mathbb{R}^d$ with positive definite covariance, i.e.,

$$\mathcal{A} := \{\mathcal{N}(\mu, \Sigma) \mid \mu \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d \times d} \text{ positive definite}\}.$$



(a) Minimizing $d_{\mathrm{KL}}(p\|\pi)$



(b) Minimizing $d_{\mathrm{KL}}(\pi\|p)$

- Fig. (a): Minimizing $d_{\mathrm{KL}}(p\|\pi)$ may miss out components of π – we want $\log\left(\frac{p}{\pi}\right)p$ to be small, which can happen when $p \approx \pi$ or $p \ll \pi$. Minimizing $d_{\mathrm{KL}}(p\|\pi)$ over Gaussians p can only give a single mode approximation which is achieved by matching one of the modes; we may think of this as "mode-seeking".
- Fig. (b): Minimizing $d_{\mathrm{KL}}(\pi\|p)$ over Gaussians p we want $\log\frac{\pi}{p}$ to be small where p appears as the denominator. Wherever π has some mass we must let p also have some mass there in order to keep $\frac{\pi}{p}$ as close as possible to one. The mass of p is allocated in a way such that on average the divergence between p and π attains its minimum; hence, it can be thought of as "mean-seeking".

Different applications will favor different choices between the mean and mode seeking approaches to Gaussian approximation.

Best Gaussian fit by minimizing $d_{\mathrm{KL}}(p\|\pi)$ ("mode-seeking")

Theorem (Best Gaussian approximation / "mode-seeking")

Suppose that the loss function $\mathsf{L}(x) := -\log \nu\big(y - F(x)\big)$ *is non-negative and bounded above and that the prior* $\pi \sim \mathcal{N}(0, \lambda^{-1} I)$. *Then there exists at least one probability distribution* $p \in \mathcal{A}$ *at which the infimum*

$$\inf_{p \in \mathcal{A}} d_{\mathrm{KL}}(p\|\pi^y)$$

is attained.

Proof. Let $p(x) = \frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \mathrm{e}^{-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}}}$, $\pi^y(x) = \frac{1}{Z}\mathrm{e}^{-\mathsf{L}(x)-\frac{\lambda}{2}\|x\|^2}$. Then

$$d_{\mathrm{KL}}(p\|\pi^y) = \mathbb{E}^p\left[\log\left(\frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \mathrm{e}^{-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}}}\right) - \log\left(\frac{1}{Z}\mathrm{e}^{-\mathsf{L}(x)-\frac{\lambda}{2}\|x\|^2}\right)\right]$$

$$= -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma + \log Z + \mathbb{E}^p\left[-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}} + \mathsf{L}(x) + \frac{\lambda}{2}\|x\|^2\right]$$

$$d_{\mathsf{KL}}(p\|\pi^y) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma + \log Z + \mathbb{E}^p\bigg[-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}} + \mathsf{L}(x) + \frac{\lambda}{2}\|x\|^2\bigg].$$

Note that Z is the normalization constant for π and is independent of p and hence of μ and Σ. We can represent a given random variable $x \sim p$ by writing $x = \mu + \Sigma^{1/2}\xi$, where $\xi \sim \mathcal{N}(0, I)$, and hence

$$\|x-\mu\|^2_{\Sigma^{-1}} = \|\Sigma^{1/2}\xi\|^2_{\Sigma^{-1}} = \|\xi\|^2 \quad \Rightarrow \quad \mathbb{E}^p\bigg[-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}}\bigg] = -\frac{d}{2}.$$

Moreover,

$$\begin{aligned}
\mathbb{E}^p[\|x\|^2] &= \int_{\mathbb{R}^d} \|x-\mu+\mu\|^2 p(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} \|x-\mu\|^2 p(x)\,\mathrm{d}x + 2\langle\mu, \int_{\mathbb{R}^d} x p(x)\,\mathrm{d}x\rangle - 2\langle\mu, \int_{\mathbb{R}^d} \mu p(x)\,\mathrm{d}x\rangle + \int_{\mathbb{R}^d} \|\mu\|^2 p(x)\,\mathrm{d}x \\
&= \mathsf{tr}(\Sigma) + 2\langle\mu,\mu\rangle - 2\langle\mu,\mu\rangle + \|\mu\|^2 = \mathsf{tr}(\Sigma) + \|\mu\|^2.
\end{aligned}$$

We obtain

$$d_{\mathrm{KL}}(p\|\pi^y) = -\frac{d}{2} - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma + \mathbb{E}^p \mathsf{L}(x) + \frac{\lambda}{2}\|\mu\|^2 + \frac{\lambda}{2}\mathrm{tr}(\Sigma) + \log Z.$$

Define $\mathcal{I}(\mu,\Sigma) = \mathbb{E}^p \mathsf{L}(x) + \frac{\lambda}{2}\|\mu\|^2 + \frac{\lambda}{2}\mathrm{tr}(\Sigma) - \frac{1}{2}\log\det\Sigma$. Note that there is a correspondence between minimizing $d_{\mathrm{KL}}(p\|\pi^y)$ over $p \in \mathcal{A}$ and minimizing $\mathcal{I}(\mu,\Sigma)$ over $\mu \in \mathbb{R}^d$ and positive definite Σ. Moreover:

- $\mathcal{I}(0, I) < \infty$.
- For any Σ, $\mathcal{I}(\mu,\Sigma) \to \infty$ as $\|\mu\| \to \infty$.
- For any μ, $\mathcal{I}(\mu,\Sigma) \to \infty$ as $\mathrm{tr}(\Sigma) \to 0$ or $\mathrm{tr}(\Sigma) \to \infty$.

Therefore, there are $M, r, R > 0$ such that the infimum of $\mathcal{I}(\mu,\Sigma)$ over $\mu \in \mathbb{R}^d$ and positive definite Σ is equal to the infimum of $\mathcal{I}(\mu,\Sigma)$ over

$$\tilde{\mathcal{A}} := \{(\mu,\Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d} \text{ positive-definite symmetric}, \|\mu\| \le M, r \le \mathrm{tr}(\Sigma) \le R\}.$$

Since $\mathcal{I}$ is continuous in $\tilde{\mathcal{A}}$ it achieves its infimum and the proof is complete. □

We remark that the theorem establishes the existence of a best Gaussian approximation. However, minimizers need not be unique.

Best Gaussian fit by minimizing $d_{\mathrm{KL}}(\pi\|p)$ ("mean-seeking")

The best Gaussian approximation in Kullback–Leibler with respect to its second argument is unique and given by moment matching.

Theorem (Best Gaussian by moment matching / "mean-seeking")

Assume that $\bar{\mu} := \mathbb{E}^{\pi}[x]$ *is finite and that* $\bar{\Sigma} := \mathbb{E}^{\pi}[(x-\bar{\mu})(x-\bar{\mu})^{\mathrm{T}}]$ *is positive definite. (Here,* π *denotes the target distribution, e.g., the posterior.) Then the infimum*

$$\inf_{p \in \mathcal{A}} d_{\mathrm{KL}}(\pi\|p)$$

is attained by $p = \mathcal{N}(\bar{\mu}, \bar{\Sigma})$.

Proof. Note that $d_{\mathrm{KL}}(\pi\|p) = -\mathbb{E}^{\pi}[\log p] + \overbrace{\mathbb{E}^{\pi}[\log \pi]}^{\text{independent of } p}$. Since we want a Gaussian minimizer, write $p(x) = ((2\pi)^d |\det \Sigma|)^{-1/2} \exp\big(-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}}\big)$

$$\begin{aligned}
\Rightarrow -\mathbb{E}^{\pi}[\log p] &= -\mathbb{E}^{\pi}[\log\big((2\pi)^{-d/2}(\det\Sigma)^{-1/2}\mathrm{e}^{-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}}}\big)] \\
&= \tfrac{1}{2}\mathbb{E}^{\pi}[\|x-\mu\|^2_{\Sigma^{-1}}] + \tfrac{1}{2}\log\det\Sigma + \tfrac{d}{2}\log(2\pi).
\end{aligned}$$

Note that the final term is irrelevant for the optimization problem.

Let $\Lambda := \Sigma^{-1}$. Our task is equivalent to finding the minimizer of

$$I(\mu, \Lambda) := \frac{1}{2}\mathbb{E}^{\pi}[(x-\mu)\Lambda(x-\mu)^{\mathrm{T}}] - \frac{1}{2}\log\det\Lambda.$$

Let $\Lambda = (\Lambda_{ij})_{i,j=1}^{d}$. We can view the above functional as the $d + d^2$ variate function $I(\mu_1, \ldots, \mu_d, \Lambda_{11}, \Lambda_{12}, \ldots, \Lambda_{dd})$. Thus, we only need to show that

$$\nabla I(\bar{\mu}, \bar{\Sigma}^{-1}) = 0 \quad \text{and} \quad \nabla^2 I(\mu, \Sigma^{-1}) > 0 \quad \text{for all } \mu, \Sigma.$$

($(\bar{\mu}, \bar{\Sigma}^{-1})$ is the critical point and the objective function is convex.)

By defining the notations $\partial_\mu f := \left(\frac{\partial f}{\partial \mu_i}\right)_{i=1}^{d}$ (gradient w.r.t. vector μ) and $\partial_\Lambda f := \left(\frac{\partial f}{\partial \Lambda_{ji}}\right)_{i,j=1}^{d}$ (gradient w.r.t. vector $(\Lambda_{11}, \Lambda_{12}, \ldots, \Lambda_{dd})$, reshaped into a $d \times d$ matrix), we easily see that $\nabla I = 0$ can be expressed as the pair

$$\begin{cases} 0 = \partial_\mu I & = -\mathbb{E}^{\pi}[\Lambda(x-\mu)] = 0 \\ 0 = \partial_\Lambda I & = \frac{1}{2}\partial_\Lambda(\mathbb{E}^{\pi}[(x-\mu)\Lambda(x-\mu)^{\mathrm{T}}]) - \frac{1}{2\det\Lambda}\partial_\Lambda \det\Lambda \\ & = \frac{1}{2}\mathbb{E}^{\pi}[(x-\mu)(x-\mu)^{\mathrm{T}}] - \frac{1}{2}\Lambda^{-1}, \end{cases}$$

where we used a special case of Jacobi's formula $\partial_\Lambda \det\Lambda = \det\Lambda \cdot \Lambda^{-1}$. Clearly, $(x, \Lambda) = (\bar{\mu}, \bar{\Sigma}^{-1})$ is the critical point satisfying the above condition.

Finally, we need to show that $\nabla^2 I(\mu, \Sigma^{-1})$ is positive definite. To this end, we note that

$$p(x) = \sqrt{\frac{\det \Lambda}{(2\pi)^d}} \mathrm{e}^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Lambda(x-\mu)} = \sqrt{\frac{\det \Lambda}{(2\pi)^d}} \mathrm{e}^{-\frac{1}{2}x^{\mathrm{T}}\Lambda x + \mu^{\mathrm{T}}\Lambda x - \frac{1}{2}\mu^{\mathrm{T}}\Lambda\mu}$$
$$= \sqrt{\frac{\det \Lambda}{(2\pi)^d}} \mathrm{e}^{-\frac{1}{2}\mu^{\mathrm{T}}\Lambda\mu} \mathrm{e}^{-\frac{1}{2}x^{\mathrm{T}}\Lambda x + \mu^{\mathrm{T}}\Lambda x} = \frac{\mathrm{e}^{-\frac{1}{2}x^{\mathrm{T}}\Lambda x + \mu^{\mathrm{T}}\Lambda x}}{\int_{\mathbb{R}^d} \mathrm{e}^{-\frac{1}{2}x^{\mathrm{T}}\Lambda x + \mu^{\mathrm{T}}\Lambda x}\,\mathrm{d}x}.$$

Noting that $x^{\mathrm{T}}\Lambda x = \sum_{i,j=1}^d \Lambda_{ij} x_i x_j = \sum_{i,j=1}^d \Lambda_{ij}(xx^{\mathrm{T}})_{ij}$, we can write $x^{\mathrm{T}}\Lambda x = \mathrm{vec}(\Lambda) \cdot \mathrm{vec}(xx^{\mathrm{T}})$, where we define

$$\mathrm{vec}(M) := (M_{11}, M_{12}, \ldots, M_{dd})^{\mathrm{T}} \quad \text{for } M \in \mathbb{R}^{d\times d}.$$

In particular,

$$-\frac{1}{2}x^{\mathrm{T}}\Lambda x + \mu^{\mathrm{T}}\Lambda x = \underbrace{\begin{bmatrix} \Lambda\mu \\ -\frac{1}{2}\mathrm{vec}(\Lambda) \end{bmatrix}}_{=:\theta}^{\mathrm{T}} \underbrace{\begin{bmatrix} x \\ \mathrm{vec}(xx^{\mathrm{T}}) \end{bmatrix}}_{=:T(x)}$$

and we can write $p_\theta(x) := p(x) = \frac{1}{Z(\theta)} \mathrm{e}^{\theta^{\mathrm{T}} T(x)}$, $Z(\theta) := \int_{\mathbb{R}^d} \mathrm{e}^{\theta^{\mathrm{T}} T(x)}\,\mathrm{d}x$.

The importance of the characterization

$$p_\theta(x) = \frac{1}{Z(\theta)} \mathrm{e}^{\theta^{\mathrm{T}} T(x)}, \quad Z(\theta) := \int_{\mathbb{R}^d} \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x,$$

lies in the fact that *every possible Gaussian PDF* can be parameterized by the vector $\theta = (\theta_1, \ldots, \theta_{d+d^2})^{\mathrm{T}}$. Thus, the KL divergence $d_{\mathsf{KL}}(\pi \| p_\theta)$ that we are interested in can be recast as

$$\begin{aligned} H(\theta) :=& d_{\mathsf{KL}}(\pi \| p_\theta) = -\mathbb{E}^\pi[\log p_\theta] + \mathbb{E}^\pi[\log \pi] \\ =& -\theta^{\mathrm{T}} \mathbb{E}^\pi[T(x)] + \log Z(\theta) + \mathbb{E}^\pi[\log \pi]. \end{aligned}$$

Noting that $\nabla_\theta^2(\theta^{\mathrm{T}} \mathbb{E}^\pi[T(x)]) = 0$ and $\frac{\partial \log Z(\theta)}{\partial \theta_i} = \frac{1}{Z(\theta)} \int_{\mathbb{R}^d} \frac{\partial}{\partial \theta_i} \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x$ $= \frac{1}{Z(\theta)} \int_{\mathbb{R}^d} T_i(x) \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x$, we compute

$$\begin{aligned} &[\nabla_\theta^2 H(\theta)]_{ij} = \frac{\partial^2 \log Z(\theta)}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_j} \left(\frac{1}{Z(\theta)} \int_{\mathbb{R}^d} T_i(x) \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x \right) \\ &= -\frac{1}{Z(\theta)^2} \left(\int_{\mathbb{R}^d} T_i(x) \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x \right) \left(\int_{\mathbb{R}^d} T_j(x) \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x \right) + \frac{1}{Z(\theta)} \int_{\mathbb{R}^d} T_i(x) T_j(x) \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x \\ &= \mathbb{E}^{p_\theta}[T_i T_j] - \mathbb{E}^{p_\theta}[T_i] \mathbb{E}^{p_\theta}[T_j] = [\mathrm{Cov}^{p_\theta}(T)]_{ij}, \end{aligned}$$

which is positive definite. □

Remark. Notice that the preceding proof of convexity holds for any distribution p that can be parameterized by the following more general expression:

$$p_\theta(x) = h(x)\exp\left(\theta^T T(x) - A(\theta)\right) \tag{1}$$

$$\text{with} \qquad A(\theta) = \log\left[\int_{\mathbb{R}^d} h(x)\exp\left(\theta^T T(x)\right) dx\right].$$

Since $h(x)$ is independent of θ, the conclusion of the previous theorem carries over to distributions with the form of (1). Such distributions belong to the *exponential family* in the statistics literature. Here, θ is called the natural parameter, $T(x)$ the sufficient statistic, $h(x)$ the base measure, and $A(\theta)$ the log-partition.

The Gaussian distribution is a special case in which $h(x)$ is constant with respect to x.

Variational formulation of Bayes' theorem

We have been concerned with finding the best Gaussian approximations to a measure with respect to KL divergences. Bayes' theorem itself can be formulated through a closely related minimization principle. Consider a posterior $\pi^y(x)$ in the following form:

$$\pi^y(x) = \frac{1}{Z}\exp\big(-\mathsf{L}(x)\big)\pi(x),$$

where $\pi(x)$ is the prior, $\mathsf{L}(x)$ is the negative log-likelihood, and Z the normalization constant. We assume here for exposition that all densities are positive. Let p be an *arbitrary* PDF. Then we can express $d_{\mathsf{KL}}(p\|\pi^y)$ as

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

If we define

$$\mathcal{J}(p) = d_{\mathsf{KL}}(p\|\pi) + \mathbb{E}^p[\mathsf{L}(x)]$$

then we have the following:

Theorem (Bayes' theorem as an optimization principle)

The posterior distribution π^y is given by the following minimization principle:

$$\pi^y = \arg\min_{p\in\mathscr{P}} \mathcal{J}(p),$$

where $\mathscr{P}$ contains all probability densities on $\mathbb{R}^d$.

Proof.
Since Z is the normalization constant for π^y and is independent of p, the minimizer of $d_{\mathsf{KL}}(p\|\pi^y)$ will also be the minimizer of $\mathcal{J}(p)$. Since the global minimizer of $d_{\mathsf{KL}}(p\|\pi^y)$ is attained at $p = \pi^y$ the result follows. □

Why is it useful to view the posterior as the minimizer of an energy?

- The variational formulation provides a natural way to approximate the posterior by restricting the minimization problem to distributions satisfying some computationally desirable property.
    - For instance, variational Bayes methods often restrict the minimization to densities with product structure and in this chapter we have studied restriction to the class of Gaussian distributions.
- Variational formulations provide natural paths, defined by a gradient flow, towards the posterior. Understanding these flows and their rates of convergence is helpful in the choice of sampling algorithms.

Appendix

The material on slides 28–33 was **not** considered during the 2023 course and it is **not** part of the course exam.

Consider still the problem of finding $x \in \mathbb{R}^d$ from $y \in \mathbb{R}^k$ given by

$$y = F(x) + \eta$$

with noise $\eta \sim \nu$ and prior $x \sim \pi$ such that $\eta \perp x$. The posterior density π^y of $x|y$ is given by Bayes' theorem

$$\pi^y(x) = \frac{1}{Z}\nu(y - F(x))\pi(x).$$

We have the negative log-likelihood:

$$\mathsf{L}(x) = -\log \nu\big(y - F(x)\big),$$

and a regularizer

$$\mathsf{R}(x) = -\log \pi(x).$$

When added together these two functions of x comprise an objective function of the form

$$\mathsf{J}(x) = \mathsf{L}(x) + \mathsf{R}(x).$$

Furthermore

$$\pi^y(x) = \frac{1}{Z}\nu\big(y - F(x)\big)\pi(x) \propto e^{-\mathsf{J}(x)}.$$

We see that minimizing the objective function J is equivalent to maximizing the posterior π^y. Therefore, the MAP estimator can be rewritten in terms of J as follows:

$$\hat{x}_{\mathsf{MAP}} = \arg\max_{x\in\mathbb{R}^d} \pi^y(x) = \arg\min_{x\in\mathbb{R}^d} \mathsf{J}(x).$$

Let us consider conditions under which the MAP estimator is attained, and characterize the MAP estimator in terms of small ball probabilities – this interpretation generalizes the definition of MAP estimators to measures that do not possess a Lebesgue density.

For any optimization problem for an objective function with a finite infimum, it is of interest to determine whether the infimum is attained.

Theorem (Attainable MAP estimator)

Assume that J *is non-negative, continuous and that* $\mathrm{J}(x) \to \infty$ *as* $|x| \to \infty$. *Then* J *attains its infimum. Therefore, the MAP estimator of* x *based on the posterior* $\pi^y(x) \propto \exp\big(-\mathrm{J}(x)\big)$ *is attained.*

Proof.
By the assumed growth and non-negativity of J, there is R such that $\inf_{x\in\mathbb{R}^d} \mathrm{J}(x) = \inf_{x\in\bar{B}(0,R)} \mathrm{J}(x)$ where $\bar{B}(0,R)$ denotes the closed ball of radius R around the origin. Since J is assumed to be continuous, its infimum over $\bar{B}(0,R)$ is attained and the proof is complete. □

Remark. The assumption that $\mathrm{J}(x) \to \infty$ is not restrictive: this condition needs to hold in order to be able to normalize $\pi^y(x) \propto \exp\big(-\mathrm{J}(x)\big)$ into a PDF, which is implicitly assumed in the second part of the theorem statement.

Example. Suppose that

1. $F : \mathbb{R}^d \to \mathbb{R}^k$ is continuous and $\eta \sim \mathcal{N}(0, \Gamma)$;
2. the objective function $\mathrm{J}(x) = \mathrm{L}(x) + \mathrm{R}(x)$ has Γ-weighted L^2 loss

$$\mathrm{L}(x) = \frac{1}{2}\|y - F(x)\|^2_{\Gamma^{-1}}$$

and L^p regularizer

$$\mathrm{R}(x) = \frac{\lambda}{p}\|u\|_p^p, \quad p \in (0, \infty).$$

Then the assumptions on J in the previous theorem are satisfied, and the infimum of J is attained at the MAP estimator of the corresponding Bayesian problem with posterior PDF proportional to $\exp\big(-\mathrm{J}(u)\big)$.

Intuitively the MAP estimator maximizes posterior probability. We make this precise in the following theorem which links the objective function J to small ball probabilities.

Theorem (Objective function and posterior probability)

Assume that J *is non-negative, continuous and that* $\mathsf{J}(x) \to \infty$ *as* $|x| \to \infty$. *Let*

$$\alpha(x, \delta) := \int_{B(x,\delta)} \pi^y(v)\,\mathrm{d}v = \mathbb{P}^{\pi^y}\big(B(x, \delta)\big),$$

be the posterior probability of a ball with radius δ *centered at* x. *Then, for all* $x_1, x_2 \in \mathbb{R}^d$, *we have*

$$\lim_{\delta \to 0} \frac{\alpha(x_1, \delta)}{\alpha(x_2, \delta)} = e^{\mathsf{J}(x_2) - \mathsf{J}(x_1)}.$$

Remark: For fixed x_2, the right-hand side is maximized at point x_1 that minimizes J. Independently of the choice of any fixed x_2, the above result shows that the probability of a small ball of radius δ centered at x_1 is, approximately, maximized by choosing the centre at a minimizer of J.

This result essentially characterizes the MAP estimate and, since it makes no reference to Lebesgue density, it can be generalized to infinite dimensions.

Proof. Let $x_1, x_2 \in \mathbb{R}^d$, $\varepsilon > 0$. By continuity of J, for all sufficiently small δ:

$$x \in \bar{B}(x_j, \delta) \quad \Rightarrow \quad |\mathsf{J}(x) - \mathsf{J}(x_j)| \le \varepsilon, \quad j \in \{1, 2\},$$

and therefore

$$\begin{aligned} e^{-\mathsf{J}(x_1)-\varepsilon} \le e^{-\mathsf{J}(v)} \le e^{-\mathsf{J}(x_1)+\varepsilon} \quad & \text{for all } v \in B(x_1, \delta), \\ e^{-\mathsf{J}(x_2)-\varepsilon} \le e^{-\mathsf{J}(v)} \le e^{-\mathsf{J}(x_2)+\varepsilon} \quad & \text{for all } v \in B(x_2, \delta). \end{aligned}$$

It follows, for all δ sufficiently small, that

$$\begin{aligned} B_\delta e^{-\mathsf{J}(x_1)-\varepsilon} &\le \int_{B(x_1,\delta)} e^{-\mathsf{J}(v)} \, \mathrm{d}v \le B_\delta e^{-\mathsf{J}(x_1)+\varepsilon}, \\ B_\delta e^{-\mathsf{J}(x_2)-\varepsilon} &\le \int_{B(x_2,\delta)} e^{-\mathsf{J}(v)} \, \mathrm{d}v \le B_\delta e^{-\mathsf{J}(x_2)+\varepsilon}, \end{aligned}$$

where B_δ is the Lebesgue measure of a ball with radius δ. Taking the ratio of α's and using the above bounds we obtain that, for all δ sufficiently small,

$$e^{\mathsf{J}(x_2)-\mathsf{J}(x_1)-2\varepsilon} \le \frac{\alpha(x_1, \delta)}{\alpha(x_2, \delta)} \le e^{\mathsf{J}(x_2)-\mathsf{J}(x_1)+2\varepsilon}.$$

Since ε was arbitrary, the desired result follows. □