Bayesian Methods

Paolo Favaro

Contents

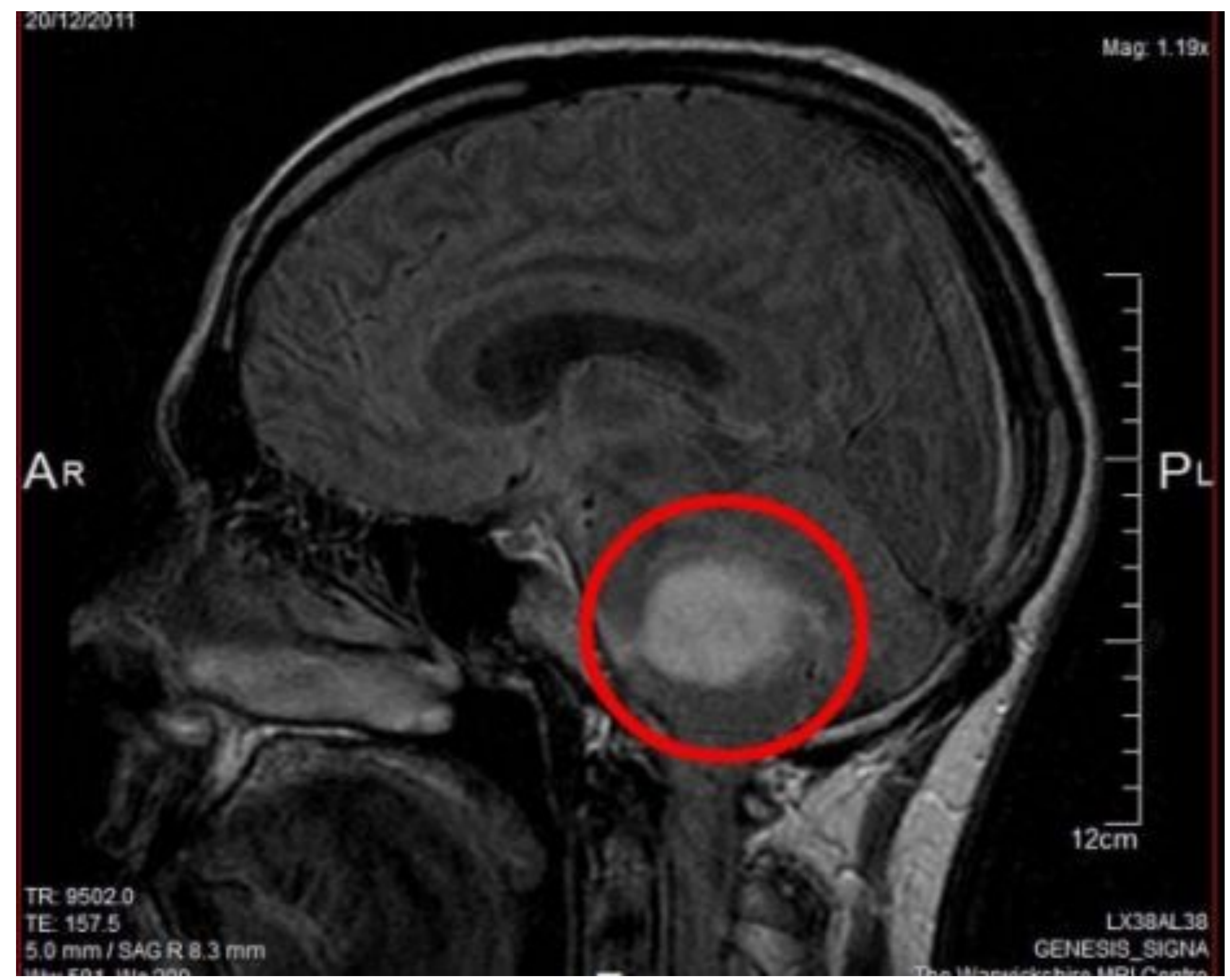
- Bayesian Decision Theory
- Majorization-Minimization
- Expectation-Maximization

Task

- Observe an X-ray image of a patient and decide whether the patient has a tumor or not

- Input/data = image

- Output/target = yes/no



Task

- Observe an X-ray image of a patient and decide whether the patient has a tumor or not

- Input/data = image

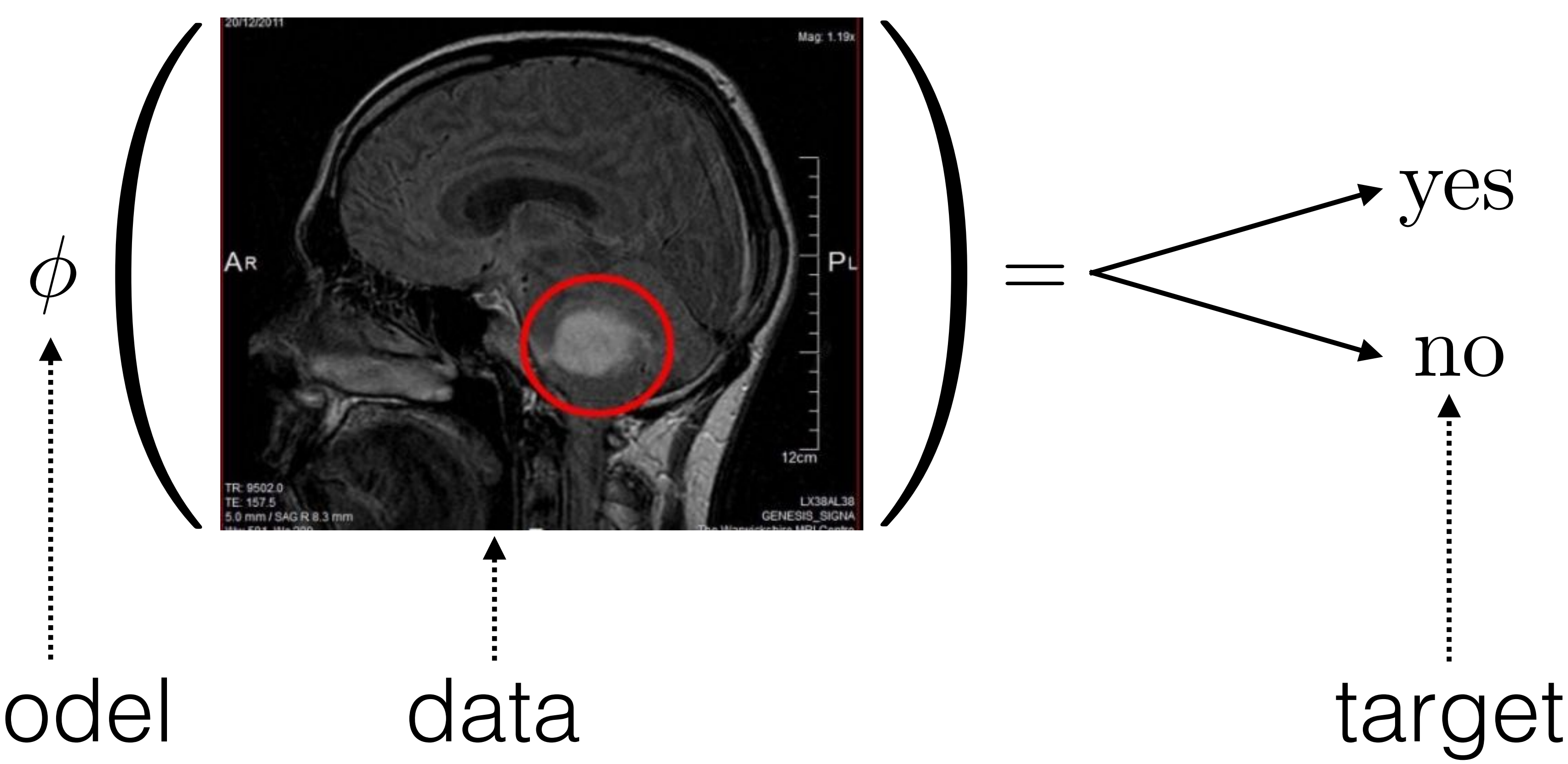
- Output/target = yes/no

- **How do we pose this as a numerical problem?**

Solving a task

- **Define a model to do the task**

- The model is a function that maps given inputs to desired outputs



Solving a task

- **Measure how well the model works on the task**

- Count the mistakes or how close we are to the desired output (**performance**)

	ground truth	error
$\phi\Big($  $\Big) =$ yes	yes	0
$\phi\Big($  $\Big) =$ no	yes	1
$\phi\Big($  $\Big) =$ no	no	0
$\phi\Big($  $\Big) =$ yes	no	1

performance

2

Basic notation

- Suppose that we have an observation vector $x \in \mathcal{X}$ together with a target vector $y \in \mathcal{Y}$
- Our goal is to predict y given x
- The space $\mathcal{Y}$ where y lives is continuous for a **regression** problem and discrete for a **classification** problem
- The joint probability $p(x, y)$ captures all the knowledge about x and y

Decision rule

- Given m observations $x_1, \ldots, x_m$
- Obtain an estimate ϕ for each $y_1, \ldots, y_m$ that best describes them
- ϕ is a **decision rule** (the model) and maps x to $\phi(x)$

Decision rule

- Examples of decision rules for **classification**

$$\phi(x) = \begin{cases} 1 & \text{if } w^\top x + b > 0 \\ 0 & \text{if } w^\top x + b \leq 0 \end{cases}$$

hyperplane

$$\phi(x) = \frac{1}{1 + e^{-(w^\top x + b)}}$$

logistic

$$\phi(x) = \begin{bmatrix} \frac{e^{a_1 x_1}}{\sum_{i=1}^{n} e^{a_i x_i}} \\ \frac{e^{a_2 x_2}}{\sum_{i=1}^{n} e^{a_i x_i}} \\ \cdots \\ \frac{e^{a_n x_n}}{\sum_{i=1}^{n} e^{a_i x_i}} \end{bmatrix}$$

softmax

Decision rule

- Examples of decision rules for **regression**

$$\phi(x) = w^\top x + b$$ hyperplane

$$\phi(x) = \sum_{i=0}^{n} w_i x^i$$ polynomial

$$\phi(x) = \sum_{i=1}^{n} w_i e^{-\frac{|w_i^\top x + b_i|^2}{\tau_i^2}}$$ radial basis function (RBF)

Loss function

- To choose the decision rule, we define a **loss function** L, which is a measure of how well ϕ describes the target variables

- L is a function of y and ϕ and defines their similarity

- Examples

  - $L(y, \phi, x) = |y - \phi(x)|^2$ **quadratic loss**

  - $L(y, \phi, x) = \mathbf{1}\{y \neq \phi(x)\}$ **0-1 loss**

Bayes risk

- **Bayes risk** is a measure of the **performance** across the whole distribution of observed and target variables of a decision rule given a certain loss function

$$E_{X,Y}[L(y,\phi,x)] = \int L(y,\phi,x)p(x,y)dxdy$$

Bayes risk

- **Bayes risk** is a measure of the **performance** across the whole distribution of observed and target variables of a decision rule given a certain loss function

$$
\begin{aligned}
E_{X,Y}[L(y,\phi,x)] &= \int L(y,\phi,x)p(x,y)dxdy \\
&= \int L(y,\phi,x)p(y|x)p(x)dxdy \\
&= E_X[E_{Y|X}[L(y,\phi,x)]]
\end{aligned}
$$

Bayes risk

- We define the optimal decision rule by solving

$$\hat{\phi} = \arg\min_{\phi} E_X[E_{Y|X}[L(y, \phi, x)]]$$

- Thus we can solve the problem element-wise via

$$\hat{\phi}(x) = \arg\min_{\phi(x)} E_{Y|X}[L(y, \phi(x), x)]$$

- The **posterior expected loss** is

$$E_{Y|X}[L(y, \phi, x)] = \int L(y, \phi(x), x) p(y|x) dy$$

Example #1

- Quadratic loss function

$$L(y, \phi, x) = |y - \phi(x)|^2$$

- Bayes risk minimization yields

$$\hat{\phi} = \arg\min_{\phi} \int |y - \phi(x)|^2 p(x, y) dx dy$$

Example #1

- Compute derivatives with respect to ϕ and set to 0

$$2\int(\phi(x)-y)p(x,y)dy=0$$

we separate the two terms

$$\phi(x)\int p(x,y)dy=\int yp(x,y)dy$$

and use marginalization

$$\phi(x)p(x)=\int yp(x,y)dy$$

Example #1

- We finally obtain the **conditional mean**

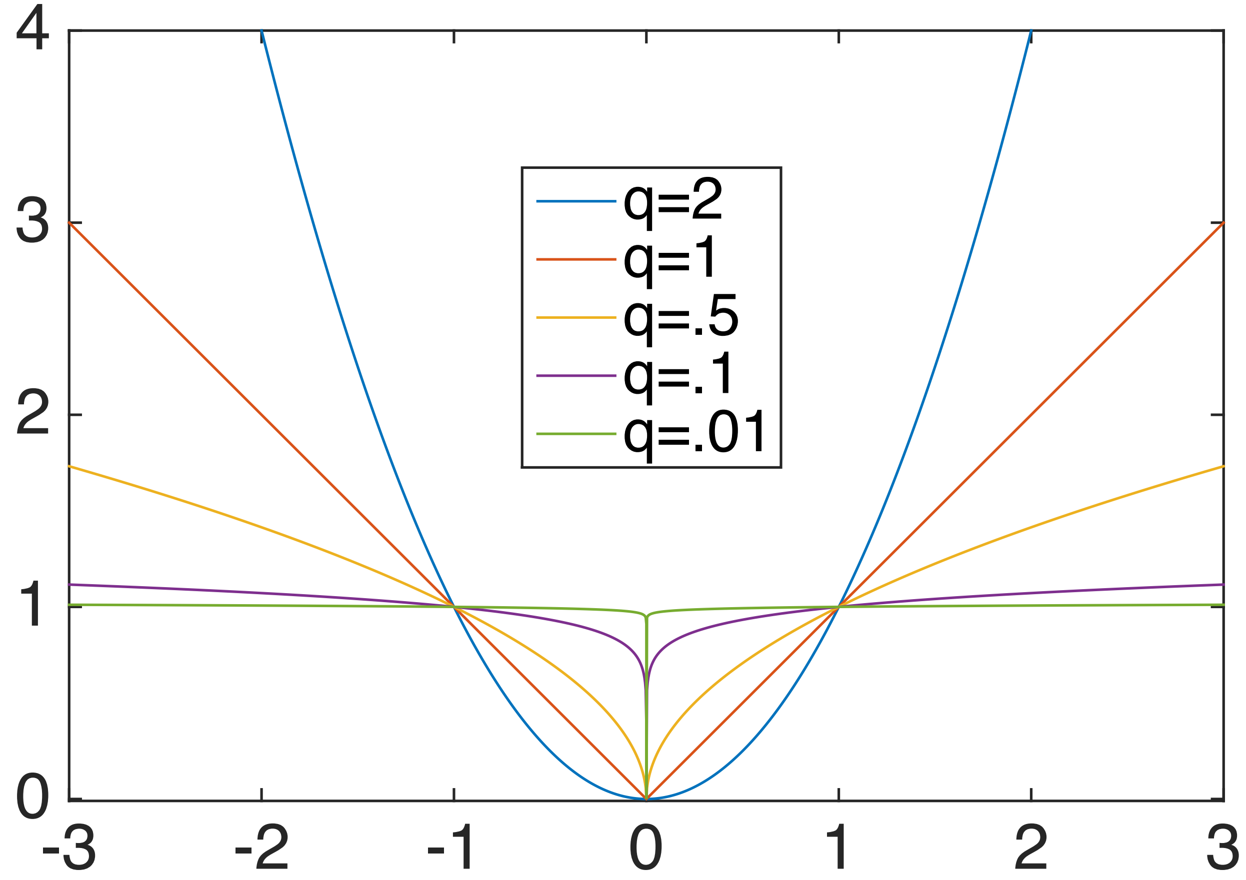
$$\phi(x) = \int y p(y|x) dy = E_{Y|X}[y]$$

and Bayes risk becomes

$$\begin{aligned} E_X[E_{Y|X}[|Y - \phi(X)|^2]] &= E_X[E_{Y|X}[|Y - E_{Y|X}[y]|^2]] \\ &= E_X[\mathrm{var}(Y|X)] \end{aligned}$$

Example #2

- Consider Minkowski's loss $L_q(y, \phi, x) = |y - \phi(x)|^q$



Example #2

- Consider Minkowski's loss

$$L_q(y, \phi, x) = |y - \phi(x)|^q$$

- Let $q=1$, then Bayes risk minimization gives

$$\hat{\phi} = \arg\min_{\phi} \int |y - \phi(x)| p(x, y) dx dy$$

Example #2

- Let us rewrite Bayes risk in a simpler form

$$
\begin{aligned}
E_{X,Y}[L_1(Y,\phi,X)] &= \int |y-\phi(x)|p(x,y)dxdy \\
&= \int \left(\int |y-\phi(x)|p(y|x)dy \right) p(x)dx \\
&= \int \left(\int\limits_{y|y\succ\phi(x)} (y-\phi(x))p(y|x)dy + \int\limits_{y|y\prec\phi(x)} (\phi(x)-y)p(y|x)dy \right) p(x)dx
\end{aligned}
$$

Example #2

- Take derivatives with respect to ϕ and set to 0

$$\frac{\delta E_{X,Y}[L_1(Y,\phi,X)]}{\delta\phi} = 0$$

Example #2

- Take derivatives with respect to ϕ and set to 0

$$\left(\int_{y|y \succ \phi(x)} p(y|x)dy - \int_{y|y \prec \phi(x)} p(y|x)dy \right) p(x) = 0$$

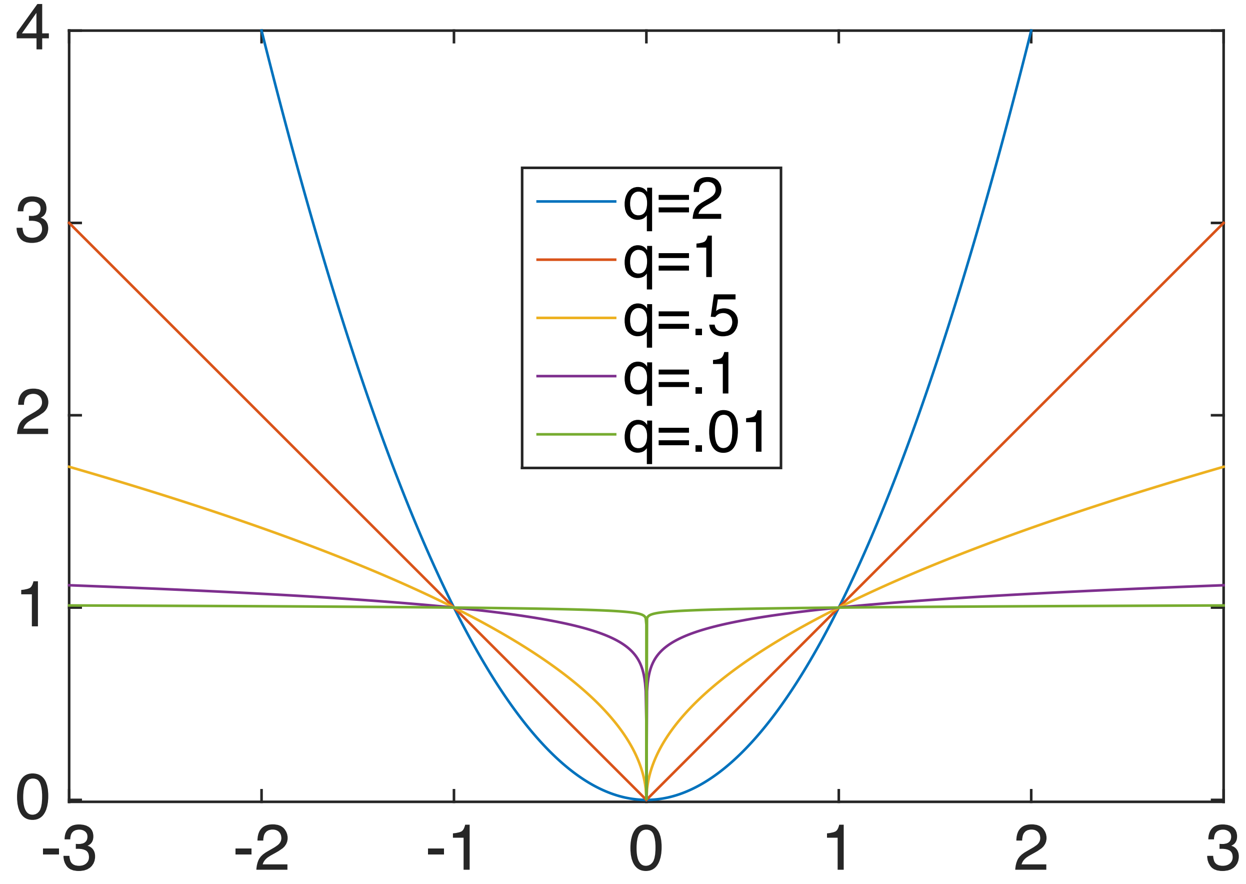
- That is, ϕ is the **conditional median**

$$\int_{y|y \succ \phi(x)} p(y|x)dy = \int_{y|y \prec \phi(x)} p(y|x)dy = \frac{1}{2}$$

Example #3

- Recall Minkowski's loss $\quad L_q(y, \phi, x) = |y - \phi(x)|^q$

$$q \to 0$$



Example #3

- Recall Minkowski's loss $L_q(y, \phi, x) = |y - \phi(x)|^q$
- When $q \to 0$ the loss converges to

$$\lim_{q \to 0} |y - \phi(x)|^q = \begin{cases} 1 & \text{if } y \neq \phi(x) \\ 0 & \text{if } y = \phi(x) \end{cases}$$

Example #3

- Recall Minkowski's loss $L_q(y, \phi, x) = |y - \phi(x)|^q$

- Let $q \to 0$, then Bayes risk minimization leads to

$$\begin{aligned}
\hat{\phi} &= \arg\min_{\phi} \int L_{q\to 0}(y, \phi, x) p(x, y) dx dy \\
&= \arg\min_{\phi} \int \left(\int L_{q\to 0}(y, \phi, x) p(y|x) dy \right) p(x) dx \\
&= \arg\min_{\phi} 1 - \int p(\phi(x)|x) p(x) dx
\end{aligned}$$

Maximum a Posteriori

- Recall Minkowski's loss $L_q(y, \phi, x) = |y - \phi(x)|^q$
- Let $q \to 0$, then Bayes risk minimization leads to **Maximum a Posteriori**

$$\hat{\phi}(x) = \arg\max_{\phi(x)} p(\phi(x)|x)$$

Maximum a Posteriori

- Can be rewritten as

$$
\begin{aligned}
\hat{\phi}(x) &= \arg\max_{\phi(x)} p(\phi(x)|x) \\
&= \arg\max_{\phi(x)} \frac{p(x, \phi(x))}{p(x)} \\
&= \arg\max_{\phi(x)} \frac{p(x|\phi(x))p_Y(\phi(x))}{p(x)} \\
&= \arg\max_{\phi(x)} p(x|\phi(x))p_Y(\phi(x)) \\
&= \arg\max_{\phi(x)} \underbrace{\log p(x|\phi(x))}_{\text{data}} + \underbrace{\log p_Y(\phi(x))}_{\text{prior}}
\end{aligned}
$$

Example #4

- Denoising problem
- We consider the following data model

$$x = y + n \qquad \text{with} \qquad n \sim \mathcal{N}(0, \sigma^2 I)$$

and the prior

$$y \sim \mathcal{N}(0, \sigma_Y^2 I)$$

Example #4

- From the Maximum a Posteriori formulation

$$\hat{\phi}(x) = \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$

we choose the data model

$$p(x|\phi(x)) \propto e^{-\frac{|x-\phi(x)|^2}{2\sigma^2}}$$

and the prior

$$p_Y(\phi(x)) \propto e^{-\frac{|\phi(x)|^2}{2\sigma_Y^2}}$$

Example #4

- We obtain

$$\hat{\phi}(x) = \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$
$$= \arg\min_{\phi(x)} \frac{|x-\phi(x)|^2}{2\sigma^2} + \frac{|\phi(x)|^2}{2\sigma_Y^2}$$

which gives the closed-form solution

$$\hat{\phi}(x) = \frac{\sigma_Y^2}{\sigma^2 + \sigma_Y^2} x$$

Example #5

- Denoising linear system
- We consider the following data model

$$x = Ay + n \qquad \text{with} \qquad n \sim \mathcal{N}(0, \sigma^2 I)$$

and the prior (e.g., to smooth the gradients)

$$\Delta y \sim \mathcal{N}(0, I)$$

Example #5

- From the Maximum a Posteriori formulation

$$\hat{\phi}(x) = \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$

we choose the data model

$$p(x|\phi(x)) \propto e^{-\frac{1}{2}(x - A\phi(x))^\top \Sigma^{-1}(x - A\phi(x))}$$

and the prior

$$p_Y(\phi(x)) \propto e^{-\frac{1}{2}|\Delta\phi(x)|^2}$$

Example #5

- We obtain

$$\hat{\phi}(x) = \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$
$$= \arg\min_{\phi(x)} \frac{1}{2}(x - A\phi(x))^\top \Sigma^{-1}(x - A\phi(x)) + \frac{1}{2}|\Delta\phi(x)|^2$$

which gives the closed-form solution

$$\hat{\phi}(x) = \left(A^\top \Sigma^{-1} A + \Delta^\top \Delta\right)^{-1} A^\top \Sigma^{-1} A x$$

Example #6

- From the Maximum a Posteriori formulation

$$\hat{\phi}(x) = \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x))$$

we choose the data model

$$p(x|\phi(x)) \propto e^{-\frac{|x - A\phi(x)|^2}{2\sigma^2}}$$

and the prior

$$p_Y(\phi(x)) \propto e^{-|\nabla\phi(x)|_{TV}}$$

Example #6

- We obtain

$$\begin{aligned}\hat{\phi}(x) &= \arg\max_{\phi(x)} \log p(x|\phi(x)) + \log p_Y(\phi(x)) \\ &= \arg\min_{\phi(x)} \frac{1}{2\sigma^2}|x - A\phi(x)|^2 + |\nabla\phi(x)|_{TV}\end{aligned}$$

which has no known closed-form solution

$$\hat{\phi}(x) = \ ?$$

Example #6

- How do we solve

$$\hat{\phi}(x) = \arg\min_{\phi(x)} \frac{1}{2\sigma^2} |x - A\phi(x)|^2 + |\nabla\phi(x)|_{TV}$$

- Recall the techniques in the previous lectures: Discretize the energy, compute the energy gradient, and solve the gradient equation $\nabla_\phi E = 0$ with gradient descent or linearization

Example #6

- If we use gradient descent we iterate

$$\phi^{t+1}(x) = \phi^t(x) - \epsilon \nabla_\phi E[\phi^t]$$

where

$$E[\phi] = \frac{1}{2\sigma^2}|x - A\phi(x)|^2 + |\nabla\phi(x)|_{TV}$$

and then let

$$\hat{\phi}(x) = \phi^\tau(x)$$

Example #6

- Issues with the original energy
  - Computation of the gradient $\nabla_\phi E[\phi]$ at each iteration might be computationally intensive (e.g., inversion of large matrices)
  - Gradient might be not defined (e.g., absolute value)
  - Difficult to incorporate additional constraints

Example #6

- An approach to minimize these energies is to use Majorization Minimization

- We describe this method in the next part