Learning Priors

Maximum Entropy Principle

statistics \( \phi(x) \), observed statistics \( \psi \).

For example, the histograms of the derivative operators \( \|VI\| \) are very similar from image to image.

These statistics are fairly similar, but how to use them to get a probability distribution? \( \|VI\| \)

Maximum Entropy:

The entropy of a distribution is

\[
H[p] = -\sum_{x} p(x) \log p(x)
\]

The distribution with maximum entropy is constant. \( p(x) = c \) (provided this is normalized.)
Maximum Entropy consistent with observed constraints.

\[ \sum_x p(x, x, y) = -\sum_x p(x) \log p(x) + \sum_x \left( \sum_y p(x, y) \phi(x, y) - \psi \right) + \lambda \left( \sum_x p(x) \phi(x) - \psi \right) \geq 0 \]

Extremize \( \mathcal{S} \) w.r.t. \( p, \lambda, \lambda \).

\[ \frac{\partial \mathcal{S}}{\partial p} = -\frac{\lambda}{p(x)} - \frac{1}{\lambda}, \quad \frac{\partial \mathcal{S}}{\partial \lambda} = 0 \Rightarrow \frac{\lambda}{p(x)} = 1, \quad \frac{\partial \mathcal{S}}{\partial \lambda} = 0 \Rightarrow \sum_x p(x) \phi(x) = \psi. \]

\[ \frac{\partial \mathcal{S}}{\partial \phi} = -\log p(x) - 1 + \sum_x \lambda \phi(x) \]

\[ \Rightarrow p(x) = e^{\sum_x \lambda \phi(x)} \]

\[ \text{where } \sum_x \psi \phi(x) = \log \sum_x e^{\sum_x \lambda \phi(x)} \]

\( \lambda \) is chosen s.t. \( \sum_x p(x) \phi(x) = \psi \)

\[ \mathcal{S}[p] = \log \sum_x e^{\sum_x \lambda \phi(x)} - \sum_x \lambda \phi(x) \geq 0. \]
The value of $\hat{\mu}$ can be solved by minimizing

$$G(\hat{\mu}) = \log Z(\hat{\mu}) - \mu \cdot \psi.$$ 

This is a convex function of $\mu$ (check that $\frac{\partial}{\partial \mu} \log Z(\mu)$ is positive definite hence $\log Z(\mu)$ is convex).

There are algorithms which are guaranteed to decrease $G(\mu)$ and converge to $\hat{\mu}$. For example, Generalized Iterative Scaling (GIS)

$$\mu^{t+1} = \mu^t + \log \psi - \log \psi^t$$

where $\psi^t = \frac{1}{Z} \sum x \cdot e^{\mu^t \cdot \phi(x)}$.

Converge when $\psi = \psi^t$.

Problem - computing $\psi^t$ is difficult. It may require stochastic sampling methods.
In practice, we do not know what statistics $\phi(x)$ to use? An approach known as feature selection is needed (Della Pietra et al.) (Zhu, Wu, Humford.) See papers for more details.

What does this mean for vision?

In 1-dimension, consider the histogram of the filter $f_i(x) = x_{i+1} - x_i$.

Histogram: $H_x(z) = \frac{1}{N} \sum_{i=1}^{N} \delta(f_i(x), z)$

Allow $z$ to take $m$ distinct values.

Use the histogram as the statistic for maximum entropy

$$p(x) = \frac{1}{Z[\mu]} e^{-\frac{1}{2} \sum \mu(z) H_x(z) + H_x(z)}$$

Lagrange multiplied
\[ p(x) = \frac{1}{Z} e^{-\frac{1}{2} \mu(z) + H_x(z)} \]

Substitute:
\[ t_x(z) = \frac{1}{N} \sum_{i=1}^{N} \delta_{t_i}(x), z \]

\[ \frac{1}{Z} e^{-\frac{1}{2} \mu(z) + H_x(z)} = \frac{1}{N} \sum_{i=1}^{N} \mu(z) \delta_{t_i}(x), z \]

\[ = \frac{1}{N} \sum_{i=1}^{N} \mu(t_i(x)) = \frac{1}{N} \sum_{i=1}^{N} \mu(x_i - x) \]

Hence:
\[ p(x) = \frac{1}{Z} e^{-\frac{1}{2} \sum_{i=1}^{N} \mu(x_i - x)} \]

This is like the Bernoulli-Gamma model nearest neighbor interaction.

The form of \( \mu(\cdot) \) is determined by the empirical statistics (i.e. solution \( \frac{1}{N} \sum_{i=1}^{N} p(x|\mu) \phi(x_i) = \phi(x) \))

This has form \( \mu(x) \)

By contrast, Gamma-Gamma or Mixel-Shah are of the quadratic form.

\[ \mu(x) \]

Redi-Osher-Fatemi

\[ \uparrow \]

\[ \uparrow - \mu(\Delta x) \]

\[ \text{edge} \]