CS-E4891 Deep Generative Models
Lecture 1: Statistical methods

Harri Lähdesmäki

Department of Computer Science
Aalto University

March 17, 2025

Outline

- Divergence measures (Sections 2.7, 5.1 from (Murphy, 2023))
- Monte Carlo (Sec. 11.2)
- Importance sampling (Sec. 11.5)
- Neural network likelihood models (Sec. 16.3)

- Reading: parts of Sections 2.7, 5.1, 11.2, 11.5, and 16.3 from (Murphy, 2023)

Commonly used training objectives in probabilistic modeling

- Likelihood, maximum likelihood: e.g.
  - Bernoulli distribution for binary classification (cross-entropy loss)
  - Categorical distribution for multi-class classification (cross-entropy loss)
  - Gaussian distribution for regression (a special case: MSE loss)
  - Generalized models with link function, typically for exponential family
- Bayesian inference:
  - Likelihood, prior, Bayes theorem, posterior distribution
- Divergence measures between probability distributions, e.g.:
  - f-divergences: e.g. Kullback-Leibler
  - Integral probability metrics: e.g. maximum mean discrepancy (MMD)

Divergence measures between probability distributions

- Goal: compare two probability distributions, P and Q, defined on the same space $\mathcal{X}$
- Two settings: compute a divergence measure $D(P, Q)$ assuming distributions are defined by
  1. Probability densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}$
  2. Samples from the distributions

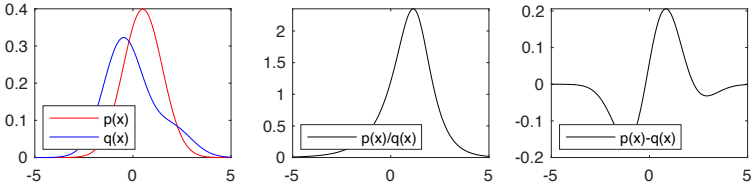
$$X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}, \text{ where } \boldsymbol{x}_i \sim P$$
$$X' = \{\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_M\}, \text{ where } \boldsymbol{x}'_j \sim Q$$

Divergence measures between probability distributions

- Goal: compare two probability distributions, P and Q, defined on the same space $\mathcal{X}$
- Two settings: compute a divergence measure $D(P, Q)$ assuming distributions are defined by
  1. Probability densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}$
  2. Samples from the distributions

$$X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}, \text{ where } \boldsymbol{x}_i \sim P$$
$$X' = \{\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_M\}, \text{ where } \boldsymbol{x}'_j \sim Q$$

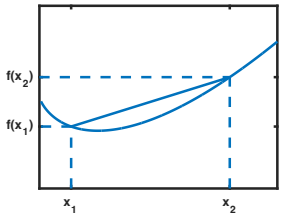
- Two main ways to compute the divergence between a pair of distributions:
  1. In terms of their ratio: P/Q
  2. In terms of their difference: $P - Q$



Convex and concave functions

A real-valued function $f : \mathcal{X} \to \mathbb{R}$ is[a] convex if for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and for all $0 \leq t \leq 1$:

$$f(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) \leq tf(\boldsymbol{x}_1) + (1-t)f(\boldsymbol{x}_2)$$



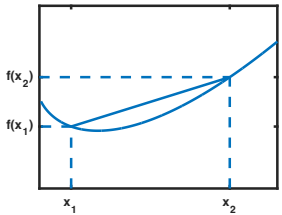
$f(x) = x \log(x)$

[a] Assume $\mathcal{X}$ is a convex (sub)set, e.g. $\mathcal{X} = \mathbb{R}$, $\mathcal{X}$ is an interval in $\mathbb{R}$, $\mathcal{X} = \mathbb{R}^2$, etc.

Convex and concave functions

A real-valued function $f : \mathcal{X} \to \mathbb{R}$ is[a] convex if for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and for all $0 \leq t \leq 1$:

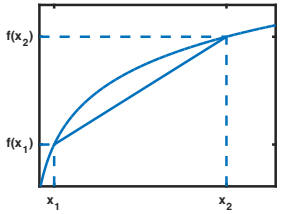
$$f(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) \leq tf(\boldsymbol{x}_1) + (1-t)f(\boldsymbol{x}_2)$$



$f(x) = x\log(x)$

A real-valued function $f : \mathcal{X} \to \mathbb{R}$ is concave if for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and for all $0 \leq t \leq 1$:

$$f(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) \geq tf(\boldsymbol{x}_1) + (1-t)f(\boldsymbol{x}_2)$$



$f(x) = \log(x)$

If f is convex, then $-f$ is concave

[a] Assume $\mathcal{X}$ is a convex (sub)set, e.g. $\mathcal{X} = \mathbb{R}$, $\mathcal{X}$ is an interval in $\mathbb{R}$, $\mathcal{X} = \mathbb{R}^2$, etc.

f-divergence

f-divergence between two probability distributions P and Q is defined via the density ratio $r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$

$$D_f(p \,||\, q) = \int q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x},$$

where $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex function, and $f(1) = 0$

Special attention needed such that the ratio $\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ and $f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)$ are defined

f-divergence

f-divergence between two probability distributions P and Q is defined via the density ratio $r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$

$$D_f(p \,||\, q) = \int q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x},$$

where $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex function, and $f(1) = 0$

Special attention needed such that the ratio $\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ and $f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)$ are defined

Important special cases of the f-divergence

- Kullback-Leibler divergence: $f(r) = r\log(r)$
- Alpha divergence: $f(r) = \frac{4}{1-\alpha^2}(1 - r^{(1+\alpha)/2})$, $\alpha \neq \pm 1$
- Hellinger distance: $f(r) = (\sqrt{r} - 1)^2$
- Chi-squared distance: $f(r) = (r-1)^2$

Jensen's inequalities

Jensen's inequality for concave functions

If $\boldsymbol{x}$ is a random variable with probability distribution P and f is a concave function, then

$$f(\mathbb{E}_{p(\boldsymbol{x})}[\boldsymbol{x}]) \geq \mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x})],$$

where $\mathbb{E}_{p(\boldsymbol{x})}[\cdot]$ is the expectation w.r.t. probability density $p(\boldsymbol{x})$

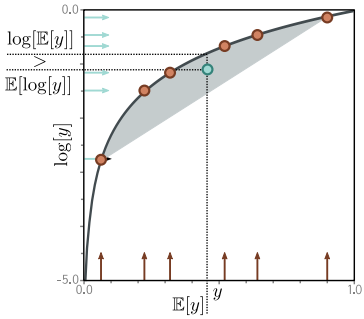


Figure 17.4 from (Prince, 2023)
Note: y has the role of $\boldsymbol{x}$

Jensen's inequalities

Jensen's inequality for concave functions

If $\boldsymbol{x}$ is a random variable with probability distribution P and f is a concave function, then

$$f(\mathbb{E}_{p(\boldsymbol{x})}[\boldsymbol{x}]) \geq \mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x})],$$

where $\mathbb{E}_{p(\boldsymbol{x})}[\cdot]$ is the expectation w.r.t. probability density $p(\boldsymbol{x})$

Jensen's inequality for convex functions

If $\boldsymbol{x}$ is a random variable with probability distribution P and f is a convex function, then

$$f(\mathbb{E}_{p(\boldsymbol{x})}[\boldsymbol{x}]) \leq \mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x})]$$

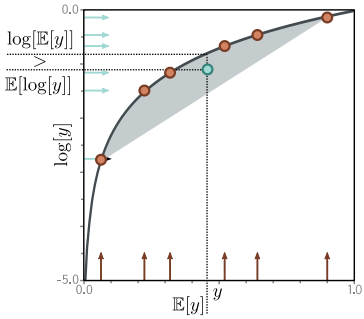


Figure 17.4 from (Prince, 2023)
Note: y has the role of $\boldsymbol{x}$

f-divergence: properties

- f-divergence can be written as

$$D_f(p \,||\, q) = \int q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x} = \mathbb{E}_{q(\boldsymbol{x})}\left[f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\right]$$

f-divergence: properties

- f-divergence can be written as

$$D_f(p \,||\, q) = \int q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x} = \mathbb{E}_{q(\boldsymbol{x})}\left[f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\right]$$

- Jensen's inequality for convex functions gives

$$\mathbb{E}_{q(\boldsymbol{x})}\left[f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\right] \geq f\left(\mathbb{E}_{q(\boldsymbol{x})}\left[\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right]\right) = f\left(\int q(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x}\right) = f\left(\int p(\boldsymbol{x})d\boldsymbol{x}\right) = f(1) = 0$$

f-divergence: properties

- f-divergence can be written as

$$D_f(p \,||\, q) = \int q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x} = \mathbb{E}_{q(\boldsymbol{x})}\left[f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\right]$$

- Jensen's inequality for convex functions gives

$$\mathbb{E}_{q(\boldsymbol{x})}\left[f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\right] \geq f\left(\mathbb{E}_{q(\boldsymbol{x})}\left[\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right]\right) = f\left(\int q(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x}\right) = f\left(\int p(\boldsymbol{x})d\boldsymbol{x}\right) = f(1) = 0$$

- Thus, f-divergence satisfies
  - $D_f(p \,||\, q) \geq 0$
  - $D_f(p \,||\, p) = 0$
- f-divergence is not a valid distance measure because
  - it is not generally symmetric
  - it does not satisfy triangle inequality

Kullback-Leibler divergence

The Kullback-Leibler divergence between two probability distributions P and Q (divergence of P from Q) is defined as

$$D_{\mathbb{KL}}(p \,||\, q) = \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x}$$
$$= \mathbb{E}_{p(\boldsymbol{x})}\left[\log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\right].$$

In other words, the KL divergence is the f-divergence with $f(r) = r\log(r)$.

- KL divergence satisfies
  1. $D_{\mathbb{KL}}(p \,||\, q) \geq 0$ (since KL is an f-divergence)
  2. $D_{\mathbb{KL}}(p \,||\, q) = 0$ iff $p = q$ for all $\boldsymbol{x}$
- KL divergence is not symmetric, i.e., generally $D_{\mathbb{KL}}(p \,||\, q) \neq D_{\mathbb{KL}}(q \,||\, p)$

KL divergence for independent random variables

For probability distributions of D-dimensional random variables where dimensions are independent,

$$q(\boldsymbol{x}) = \prod_{i=1}^{D} q_i(x_i) \quad \text{and} \quad p(\boldsymbol{x}) = \prod_{i=1}^{D} p_i(x_i),$$

the KL factorizes across dimensions

$$\begin{aligned}
D_{\mathbb{KL}}(p \,||\, q) &= \mathbb{E}_{p(\boldsymbol{x})}\left[\log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\right] = \mathbb{E}_{p(\boldsymbol{x})}\left[\log\left(\frac{\prod_{i=1}^{D} p(x_i)}{\prod_{i=1}^{D} q(x_i)}\right)\right] = \mathbb{E}_{p(\boldsymbol{x})}\left[\log\left(\prod_{i=1}^{D} \frac{p(x_i)}{q(x_i)}\right)\right] \\
&= \mathbb{E}_{p(\boldsymbol{x})}\left[\sum_{i=1}^{D} \log\left(\frac{p(x_i)}{q(x_i)}\right)\right] = \sum_{i=1}^{D} \mathbb{E}_{p(\boldsymbol{x})}\left[\log\left(\frac{p(x_i)}{q(x_i)}\right)\right] \\
&= \sum_{i=1}^{D} \mathbb{E}_{p(x_i)}\left[\log\left(\frac{p(x_i)}{q(x_i)}\right)\right] \\
&= \sum_{i=1}^{D} D_{\mathbb{KL}}(p_i \,||\, q_i),
\end{aligned}$$

KL divergence: chain rule

Chain rule for probability distributions: $p(\boldsymbol{x}) = p(x_1)p(x_2 \mid x_1) \dots p(x_D \mid x_1, \dots, x_{D-1})$

KL divergence satisfies a natural chain rule

$$
\begin{aligned}
D_{\mathbb{KL}}(p(x_1, x_2) \mid\mid q(x_1, x_2)) &= \iint p(x_1, x_2) \log \left(\frac{p(x_1, x_2)}{q(x_1, x_2)} \right) dx_1 dx_2 \\
&= \iint p(x_1) p(x_2 \mid x_1) \left[\log \left(\frac{p(x_1)}{q(x_1)} \right) + \log \left(\frac{p(x_2 \mid x_1)}{q(x_2 \mid x_1)} \right) \right] dx_1 dx_2 \\
&= D_{\mathbb{KL}}(p(x_1) \mid\mid q(x_1)) + \mathbb{E}_{p(x_1)} \left[D_{\mathbb{KL}}(p(x_2 \mid x_1) \mid\mid q(x_2 \mid x_1)) \right]
\end{aligned}
$$

KL divergence: two Gaussian distributions

The KL divergence between two univariate Gaussian distributions

$$D_{\mathbb{KL}}\left(\mathcal{N}(x \mid \mu_1, \sigma_1^2) \,||\, \mathcal{N}(x \mid \mu_2, \sigma_2^2)\right) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

KL divergence: two Gaussian distributions

The KL divergence between two univariate Gaussian distributions

$$D_{\mathbb{KL}}\left(\mathcal{N}(x \mid \mu_1, \sigma_1^2) \,||\, \mathcal{N}(x \mid \mu_2, \sigma_2^2)\right) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

The KL divergence between two multivariate Gaussian distributions

$$D_{\mathbb{KL}}\left(\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \,||\, \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)\right) = \frac{1}{2}\Big[\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - D + \log\left(\frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)}\right)\Big]$$

KL divergence: two Gaussian distributions

The KL divergence between two univariate Gaussian distributions

$$D_{\mathbb{KL}}\left(\mathcal{N}(x \mid \mu_1, \sigma_1^2) \,||\, \mathcal{N}(x \mid \mu_2, \sigma_2^2)\right) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

The KL divergence between two multivariate Gaussian distributions

$$\begin{aligned} D_{\mathbb{KL}}\left(\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \,||\, \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)\right) = \frac{1}{2}\Big[&\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &- D + \log\left(\frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)}\right)\Big] \end{aligned}$$

If $\boldsymbol{\Sigma}_1 = \mathrm{diag}(\sigma_{11}^2, \ldots \sigma_{1D}^2)$ and $\boldsymbol{\Sigma}_2 = \mathrm{diag}(\sigma_{21}^2, \ldots \sigma_{2D}^2)$, then due to the independence property

$$D_{\mathbb{KL}}\left(\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \,||\, \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)\right) = \sum_{i=1}^{D} D_{\mathbb{KL}}\left(\mathcal{N}(x_i \mid \mu_{1i}, \sigma_{1i}^2) \,||\, \mathcal{N}(x_i \mid \mu_{2i}, \sigma_{2i}^2)\right)$$

KL divergence: two Gaussian distributions

The KL divergence between two univariate Gaussian distributions

$$D_{\mathbb{KL}}\left(\mathcal{N}(x \mid \mu_1, \sigma_1^2) \,||\, \mathcal{N}(x \mid \mu_2, \sigma_2^2)\right) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

The KL divergence between two multivariate Gaussian distributions

$$\begin{aligned} D_{\mathbb{KL}}\left(\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \,||\, \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)\right) = \frac{1}{2}\Big[&\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &- D + \log\left(\frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)}\right)\Big] \end{aligned}$$

If $\boldsymbol{\Sigma}_1 = \mathrm{diag}(\sigma_{11}^2, \ldots \sigma_{1D}^2)$ and $\boldsymbol{\Sigma}_2 = \mathrm{diag}(\sigma_{21}^2, \ldots \sigma_{2D}^2)$, then due to the independence property

$$D_{\mathbb{KL}}\left(\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \,||\, \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)\right) = \sum_{i=1}^{D} D_{\mathbb{KL}}\left(\mathcal{N}(x_i \mid \mu_{1i}, \sigma_{1i}^2) \,||\, \mathcal{N}(x_i \mid \mu_{2i}, \sigma_{2i}^2)\right)$$

These are useful results for developing deep generative models because

- Distributions are often assumed to be Gaussians
- Often $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_D^2)$

KL divergence has closed-form solution for any two exponential family distributions from the same family

KL divergence and maximum likelihood

Suppose $p_{\mathcal{D}}$ is the empirical data distribution of data set $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$p_{\mathcal{D}}(\boldsymbol{x}) = \frac{1}{N} \sum_{n=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_n),$$

where $\delta(\boldsymbol{x})$ is the delta distribution

Suppose we want to find distribution q that is as close as possible to $p_{\mathcal{D}}$

$$q^* = \arg\min_q D_{\mathbb{KL}}(p_{\mathcal{D}} \,||\, q)$$

KL divergence and maximum likelihood

Suppose $p_{\mathcal{D}}$ is the empirical data distribution of data set $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$p_{\mathcal{D}}(\boldsymbol{x}) = \frac{1}{N} \sum_{n=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_n),$$

where $\delta(\boldsymbol{x})$ is the delta distribution

Suppose we want to find distribution q that is as close as possible to $p_{\mathcal{D}}$

$$q^* = \arg\min_q D_{\mathbb{KL}}(p_{\mathcal{D}} \,||\, q)$$

The objective function can be written as

$$\begin{aligned} D_{\mathbb{KL}}(p_{\mathcal{D}} \,||\, q) &= \underbrace{\int p_{\mathcal{D}}(\boldsymbol{x}) \log p_{\mathcal{D}}(\boldsymbol{x}) d\boldsymbol{x}}_{\text{constant } C} - \int p_{\mathcal{D}}(\boldsymbol{x}) \log q(\boldsymbol{x}) d\boldsymbol{x} \\ &= C - \int \left[\frac{1}{N} \sum_{n=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_n) \right] \log q(\boldsymbol{x}) d\boldsymbol{x} \\ &= C - \frac{1}{N} \sum_{n=1}^{N} \int \delta(\boldsymbol{x} - \boldsymbol{x}_n) \log q(\boldsymbol{x}) d\boldsymbol{x} \\ &= C - \frac{1}{N} \sum_{n=1}^{N} \log q(\boldsymbol{x}_n) \\ &= C - \frac{1}{N} \log \prod_{n=1}^{N} q(\boldsymbol{x}_n) \end{aligned}$$

Minimizing $D_{\mathbb{KL}}(p_{\mathcal{D}} \,||\, q)$ maximizes the likelihood

KL divergence: minimization

Given a true distribution P, minimize the KL divergence w.r.t. an approximate distribution Q

Two optimization approaches due to the asymmetry of the KL divergence

1. Forwards KL: $q = \arg\min_q D_{\mathbb{KL}}(p \,||\, q)$
   - Results in mode covering, i.e., a wide (under-confident) distribution q
2. Reverse KL: $q = \arg\min_q D_{\mathbb{KL}}(q \,||\, p)$
   - Results in mode seeking, i.e., a narrow (over-confident) distribution q

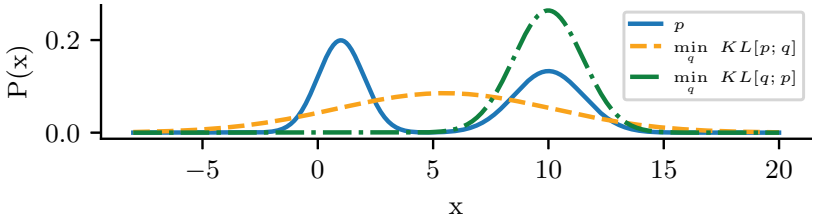


Figure 5.1 from (Murphy, 2023)

Forwards KL minimization for two Gaussians

Assume the true distribution is a 2-D correlated Gaussian $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$

Assume that $q(\boldsymbol{x})$ is a mean-field Gaussian $q(\boldsymbol{x} \mid \boldsymbol{m}, \mathbf{V}) = \mathcal{N}(x_1 \mid m_1, v_1)\mathcal{N}(x_2 \mid m_2, v_2)$

Forwards KL minimization for two Gaussians

Assume the true distribution is a 2-D correlated Gaussian $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$

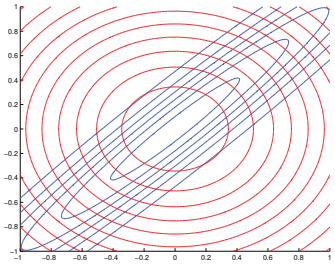
Assume that $q(\boldsymbol{x})$ is a mean-field Gaussian $q(\boldsymbol{x} \mid \boldsymbol{m}, \mathbf{V}) = \mathcal{N}(x_1 \mid m_1, v_1)\mathcal{N}(x_2 \mid m_2, v_2)$

It can be show that

- Optimizing the forwards KL $D_{\mathbb{KL}}(p \,||\, q)$ w.r.t. q results in moment matching, where

$$q(\boldsymbol{x}) = N(x_1 \mid \mu_1, \Sigma_{11})N(x_2 \mid \mu_2, \Sigma_{22})$$

- This holds more generally for exponential family
- Forwards KL is called moment projection



(a)

Figure 5.2 a) from (Murphy, 2023)

Reverse KL minimization for two Gaussians

Assume the true distribution is a 2-D correlated Gaussian $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$

Assume that $q(\boldsymbol{x})$ is a mean-field Gaussian $q(\boldsymbol{x} \mid \boldsymbol{m}, \mathbf{V}) = \mathcal{N}(x_1 \mid m_1, v_1)\mathcal{N}(x_2 \mid m_2, v_2)$

Reverse KL minimization for two Gaussians

Assume the true distribution is a 2-D correlated Gaussian $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$

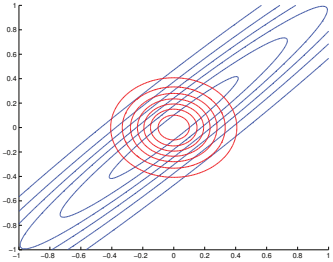
Assume that $q(\boldsymbol{x})$ is a mean-field Gaussian $q(\boldsymbol{x} \mid \boldsymbol{m}, \mathbf{V}) = \mathcal{N}(x_1 \mid m_1, v_1)\mathcal{N}(x_2 \mid m_2, v_2)$

It can be show that

- Optimizing the reverse KL $D_{\mathbb{KL}}(q \,||\, p)$ w.r.t. q results in

$$q(\boldsymbol{x}) = N(x_1 \mid \mu_1, \Lambda_{11})N(x_2 \mid \mu_2, \Lambda_{22})$$

- Reverse KL is called information projections
- Reverse KL is typically easier to optimize because expectations are w.r.t. q which we can choose



(b)

Figure 5.2 b) from (Murphy, 2023)

Jensen-Shannon divergence

- Jensen-Shannon divergence is a symmetric variant of the KL

$$\mathrm{JSD}(p, q) = \frac{1}{2} D_{\mathbb{KL}}(p \mid\mid m) + \frac{1}{2} D_{\mathbb{KL}}(q \mid\mid m) = \mathrm{JSD}(q, p),$$

  where $m(\boldsymbol{x}) = \frac{1}{2} p(\boldsymbol{x}) + \frac{1}{2} q(\boldsymbol{x})$
- This divergence is important when training implicit generative models, such as generative adversarial networks (GANs)

Integral probability metric (IPM)

- f-divergences are defined in terms of density ratio P/Q
- Many methods use the KL divergence as their objective function
- Divergence measures can also be defined in terms of $P - Q$
- Integral probability metric is defined as

$$D_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x})] - \mathbb{E}_{q(\boldsymbol{x}')}[f(\boldsymbol{x}')]|,$$

  where $\mathcal{F}$ is a class of smooth functions
- Depending on the choice of $\mathcal{F}$ we get different divergence measures
- This class of divergences is important when training implicit generative models, such as generative adversarial networks (GANs)

Total variation distance

- The total variation (TV) distance between two probability distributions P and Q for a continuous random variable $\boldsymbol{x}$

$$D_{\mathrm{TV}} = \frac{1}{2} \int |p(\boldsymbol{x}) - q(\boldsymbol{x})| d\boldsymbol{x}$$

- TV satisfies the properties of a distance: symmetric, positive, definite, triangle inequality
- TV is the (only) divergence that is both f-divergence (with $f(r) = \frac{1}{2}|r-1|$) and IPM

Monte Carlo integration

Monte Carlo methods are a stochastic approach to solve numerical integration problems

- A random variable $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^n$ with probability density $p(\boldsymbol{x})$, and a function $f : \mathbb{R}^n \to \mathbb{R}^m$
- Interested in computing the expected value of function $f(\boldsymbol{x})$

$$\mathbb{E}\left[f(\boldsymbol{x})\right] = \int f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

Monte Carlo integration

Monte Carlo methods are a stochastic approach to solve numerical integration problems

- A random variable $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^n$ with probability density $p(\boldsymbol{x})$, and a function $f : \mathbb{R}^n \to \mathbb{R}^m$
- Interested in computing the expected value of function $f(\boldsymbol{x})$

$$\mathbb{E}\left[f(\boldsymbol{x})\right] = \int f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

Monte Carlo integration

Monte Carlo integration estimate of $\mathbb{E}\left[f(\boldsymbol{x})\right]$ is

$$\mathbb{E}\left[f(\boldsymbol{x})\right] \approx \bar{f}_N = \frac{1}{N} \sum_{s=1}^{N} f(\boldsymbol{x}_s), \quad \text{where } \boldsymbol{x}_s \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$$

and N is the number of samples

- By design, function $f(\boldsymbol{x})$ is only evaluated at points $\boldsymbol{x}$ that have non-neglible probability

Monte Carlo integration: example

- The area of a circle with radius r is $A = \pi r^2$
- Alternatively, A equals

$$A = \int_{-r}^{r} \int_{-r}^{r} \mathbb{I}(x^2 + y^2 \leq r^2) dx dy$$

where $\mathbb{I}(\cdot)$ is the indicator function (value is 1 inside the circle, 0 outside)

Monte Carlo integration: example

- The area of a circle with radius r is $A = \pi r^2$
- Alternatively, A equals

$$A = \int_{-r}^{r} \int_{-r}^{r} \mathbb{I}(x^2 + y^2 \leq r^2) dx dy$$

  where $\mathbb{I}(\cdot)$ is the indicator function (value is 1 inside the circle, 0 outside)
- Define e.g. $p(x, y) = p(x)p(y)$, where $p(x)$ and $p(y)$ are uniform over $[-r, r]$
- Monte Carlo estimate of A

$$\begin{aligned} A &= (2r)(2r) \int \int \mathbb{I}(x^2 + y^2 \leq r^2) p(x) p(y) dx dy \\ &\approx 4r^2 \frac{1}{N} \sum_{s=1}^{N} \mathbb{I}(x_s^2 + y_s^2 \leq r^2), \qquad x_s, y_s \sim p(x, y) \end{aligned}$$

Monte Carlo integration: example

- The area of a circle with radius r is $A = \pi r^2$
- Alternatively, A equals

$$A = \int_{-r}^{r} \int_{-r}^{r} \mathbb{I}(x^2 + y^2 \leq r^2) dx dy$$

  where $\mathbb{I}(\cdot)$ is the indicator function (value is 1 inside the circle, 0 outside)
- Define e.g. $p(x, y) = p(x)p(y)$, where $p(x)$ and $p(y)$ are uniform over $[-r, r]$
- Monte Carlo estimate of A

$$A = (2r)(2r) \int \int \mathbb{I}(x^2 + y^2 \leq r^2) p(x) p(y) dx dy$$

$$\approx 4r^2 \frac{1}{N} \sum_{s=1}^{N} \mathbb{I}(x_s^2 + y_s^2 \leq r^2), \qquad x_s, y_s \sim p(x, y)$$

- An estimate of π is then $\hat{\pi} = \hat{A}/(r^2)$

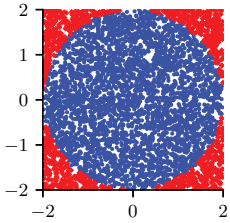


Figure 11.1 from (Murphy, 2023)

Monte Carlo integration: accuracy

Monte Carlo estimate $\bar{f}_N$ of $\mathbb{E}[f(\boldsymbol{x})]$

$$\bar{f}_N = \frac{1}{N}\sum_{s=1}^{N} f(\boldsymbol{x}_s), \quad \text{where } \boldsymbol{x}_s \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$$

Assuming the true mean $\mu = \mathbb{E}[f(\boldsymbol{x})]$ exists, by the law of large numbers

$$\bar{f}_N \to \mu \quad \text{when} \quad N \to \infty$$

Monte Carlo integration: accuracy

Monte Carlo estimate $\bar{f}_N$ of $\mathbb{E}[f(\boldsymbol{x})]$

$$\bar{f}_N = \frac{1}{N}\sum_{s=1}^{N} f(\boldsymbol{x}_s), \quad \text{where } \boldsymbol{x}_s \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$$

Assuming the true mean $\mu = \mathbb{E}[f(\boldsymbol{x})]$ exists, by the law of large numbers

$$\bar{f}_N \to \mu \quad \text{when} \quad N \to \infty$$

Assuming a scalar-valued function f and finite variance $\sigma^2 = \mathbb{V}[f(\boldsymbol{x})] = \mathbb{E}\left[(f(\boldsymbol{x}) - \mu)^2\right] < \infty$, by the central limit theorem

$$(\bar{f}_N - \mu) \to \mathcal{N}\left(0, \frac{\sigma^2}{N}\right)$$

Monte Carlo integration: accuracy

Monte Carlo estimate $\bar{f}_N$ of $\mathbb{E}[f(\boldsymbol{x})]$

$$\bar{f}_N = \frac{1}{N}\sum_{s=1}^{N} f(\boldsymbol{x}_s), \quad \text{where } \boldsymbol{x}_s \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$$

Assuming the true mean $\mu = \mathbb{E}[f(\boldsymbol{x})]$ exists, by the law of large numbers

$$\bar{f}_N \to \mu \quad \text{when} \quad N \to \infty$$

Assuming a scalar-valued function f and finite variance $\sigma^2 = \mathbb{V}[f(\boldsymbol{x})] = \mathbb{E}\left[(f(\boldsymbol{x}) - \mu)^2\right] < \infty$, by the central limit theorem

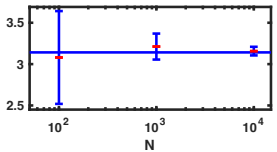
$$(\bar{f}_N - \mu) \to \mathcal{N}\left(0, \frac{\sigma^2}{N}\right)$$

Variance σ^2 can be estimated

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{s=1}^{N} \left(f(\boldsymbol{x}_s) - \bar{f}_N\right)^2$$

Thus, for large N

$$P\left(\bar{f}_N - 1.96\frac{\hat{\sigma}^2}{\sqrt{N}} \leq \mu \leq \bar{f}_N + 1.96\frac{\hat{\sigma}^2}{\sqrt{N}}\right) \approx 0.95$$



An example of π estimates and empirical standard errors for varying N

Monte Carlo integration: accuracy

To summarize, accuracy of Monte Carlo integration

- is independent of dimensionality of $\boldsymbol{x}$
- increases with increasing sample size N

Importance sampling

- Consider again using Monte Carlo method to approximate integrals of the form

$$\mathbb{E}\left[f(\boldsymbol{x})\right] = \int f(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x},$$

  where $f(\boldsymbol{x})$ is the target function and $\pi(\boldsymbol{x})$ is now the target distribution
- If $\pi(\boldsymbol{x})$ is difficult to sample, we can instead sample from a proposal distribution $q(\boldsymbol{x})$
- We need to adjust the Monte Carlo method for this inaccuracy by associating weights with each sample

$$\mathbb{E}[f(\boldsymbol{x})] \approx \frac{1}{N}\sum_{s=1}^{N} W_n f(\boldsymbol{x}_n)$$

- Two different importance sampling variants:
  - Direct importance sampling: assume that, albeit difficult to sample, we can still evaluate the normalized target distribution $\pi(\boldsymbol{x})$
  - Self-normalized importance sampling: we can only evaluate unnormalized target distribution $\tilde{\pi}(\boldsymbol{x})$

Direct importance sampling

- Write the target integral as

$$\mathbb{E}\left[f(\boldsymbol{x})\right] = \int f(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x} = \int f(\boldsymbol{x})\frac{\pi(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x}$$

  assuming $q(\boldsymbol{x}) > 0$ whenever $\pi(\boldsymbol{x}) > 0$, i.e., $\sup \pi \subseteq \sup q$

Direct importance sampling

- Write the target integral as

$$\mathbb{E}\left[f(\boldsymbol{x})\right] = \int f(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x} = \int f(\boldsymbol{x})\frac{\pi(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x}$$

  assuming $q(\boldsymbol{x}) > 0$ whenever $\pi(\boldsymbol{x}) > 0$, i.e., $\sup \pi \subseteq \sup q$

Direct importance sampling

Direct importance sampling estimate of $\mathbb{E}[f(\boldsymbol{x})]$ is

$$\mathbb{E}[f(\boldsymbol{x})] \approx \frac{1}{N}\sum_{s=1}^{N}\frac{\pi(\boldsymbol{x}_s)}{q(\boldsymbol{x}_s)}f(\boldsymbol{x}) = \frac{1}{N}\sum_{s=1}^{N}\tilde{w}_s f(\boldsymbol{x}_s), \quad \text{where } \boldsymbol{x}_s \overset{\text{i.i.d.}}{\sim} q(\boldsymbol{x})$$

and $\tilde{w}_s = \frac{\pi(\boldsymbol{x}_s)}{q(\boldsymbol{x}_s)}$ are the importance weights

Direct importance sampling: proposal distribution

Performance of the importance sampling depends crucially on the quality of $q(\boldsymbol{x})$

Ideally, the proposal distribution should

- cover the target distribution, $\pi(\boldsymbol{x}) > 0 \Rightarrow q(\boldsymbol{x}) > 0$
- be similar with $\pi(\boldsymbol{x})$
- take into account properties of f

One good strategy is to learn a proposal

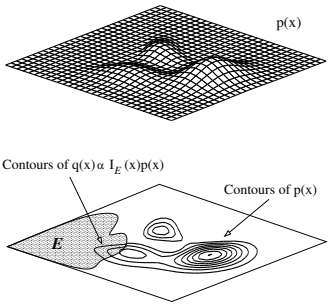


Figure 11.6 from Murphy (2023) illustrating the proposal $q(\boldsymbol{x})$

Self-normalized importance sampling

- It is often much easier to evaluate the unnormalized distribution $\tilde{\pi}(\boldsymbol{x}) = Z\pi(\boldsymbol{x})$, where $Z = \int \tilde{\pi}(\boldsymbol{x})d\boldsymbol{x}$

Self-normalized importance sampling

- It is often much easier to evaluate the unnormalized distribution $\tilde{\pi}(\boldsymbol{x}) = Z\pi(\boldsymbol{x})$, where $Z = \int \tilde{\pi}(\boldsymbol{x})d\boldsymbol{x}$

Self-normalized importance sampling

Self-normalized importance sampling estimate of $\mathbb{E}[f(\boldsymbol{x})]$ is

$$\mathbb{E}[f(\boldsymbol{x})] = \int f(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x} = \int f(\boldsymbol{x})\frac{1}{Z}\tilde{\pi}(\boldsymbol{x})d\boldsymbol{x} = \frac{\int f(\boldsymbol{x})\tilde{\pi}(\boldsymbol{x})d\boldsymbol{x}}{\int \tilde{\pi}(\boldsymbol{x})d\boldsymbol{x}} = \frac{\int \left[\frac{\tilde{\pi}(\boldsymbol{x})}{q(\boldsymbol{x})}f(\boldsymbol{x})\right] q(\boldsymbol{x})d\boldsymbol{x}}{\int \left[\frac{\tilde{\pi}(\boldsymbol{x})}{q(\boldsymbol{x})}\right] q(\boldsymbol{x})d\boldsymbol{x}}$$

$$\approx \frac{\frac{1}{N}\sum_{s=1}^{N} \tilde{w}_s f(\boldsymbol{x}_s)}{\frac{1}{N}\sum_{s=1}^{N} \tilde{w}_s}, \quad \text{where } \boldsymbol{x}_s \overset{\text{i.i.d.}}{\sim} q(\boldsymbol{x})$$

$\tilde{w}_s = \frac{\tilde{\pi}(\boldsymbol{x}_s)}{q(\boldsymbol{x}_s)}$ are the unnormalized importance weights, and assuming $\sup \tilde{\pi} \subseteq \sup q$

Multilayer perceptron (MLP)

- Multilayer perceptron (MLP) $\boldsymbol{f} : \mathbb{R}^D \to \mathbb{R}^C$ consists of L layers of linear computations and element-wise non-linearities φ (here $L = 2$)

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}) \\ &= \boldsymbol{\varphi}(\mathbf{W}_2 \boldsymbol{\varphi}(\mathbf{W}_1 \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2)\end{aligned}$$

where $\boldsymbol{\theta} = (\mathbf{W}_1, \ldots, \mathbf{W}_L, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_L)$ and

$$\mathbf{W}_1 \in \mathbb{R}^{D \times K_1}$$
$$\mathbf{W}_l \in \mathbb{R}^{K_{l-1} \times K_l}$$
$$\mathbf{W}_L \in \mathbb{R}^{K_{L-1} \times C}$$

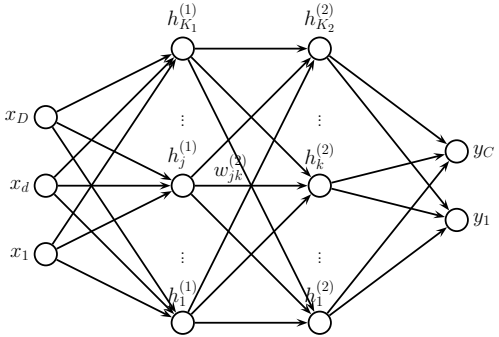


Figure 16.10 from (Murphy, 2023)

MLPs as likelihood models

Throughout this course we will use neural nets to parameterize probabilistic models

Example: regression with homoscedastic Gaussian noise

- Univariate response variable $y \in \mathbb{R}$ with MLP $f : \mathbb{R}^D \to \mathbb{R}$

$$p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) = \mathcal{N}(y \mid f(\boldsymbol{x}; \boldsymbol{\theta}), \sigma^2_{\text{noise}}),$$

where $f(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{E}(y \mid \boldsymbol{x})$

MLPs as likelihood models

Throughout this course we will use neural nets to parameterize probabilistic models

Example: regression with homoscedastic Gaussian noise

- Univariate response variable $y \in \mathbb{R}$ with MLP $f : \mathbb{R}^D \to \mathbb{R}$

$$p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) = \mathcal{N}(y \mid f(\boldsymbol{x}; \boldsymbol{\theta}), \sigma^2_{\text{noise}}),$$

  where $f(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{E}(y \mid \boldsymbol{x})$

- Multivariate response variable $\boldsymbol{y} \in \mathbb{R}^D$ with MLP $f : \mathbb{R}^D \to \mathbb{R}^C$

$$p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{\Sigma}),$$

  where $\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})$

MLPs as likelihood models

Throughout this course we will use neural nets to parameterize probabilistic models

Example: regression with homoscedastic Gaussian noise

- Univariate response variable $y \in \mathbb{R}$ with MLP $f : \mathbb{R}^D \to \mathbb{R}$

$$p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) = \mathcal{N}(y \mid f(\boldsymbol{x}; \boldsymbol{\theta}), \sigma^2_{\text{noise}}),$$

  where $f(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{E}(y \mid \boldsymbol{x})$
- Multivariate response variable $\boldsymbol{y} \in \mathbb{R}^D$ with MLP $f : \mathbb{R}^D \to \mathbb{R}^C$

$$p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{\Sigma}),$$

  where $\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})$

Example: multiclass classification $y \in \{1, \dots, C\}$

$$p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) = \text{Cat}\big(y \mid \text{softmax}(\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}))\big),$$

where

$$\text{softmax}(\boldsymbol{a}) = \left(\frac{e^{a_1}}{\sum_{c=1}^{C} e^{a_c}}, \dots, \frac{e^{a_C}}{\sum_{c=1}^{C} e^{a_c}} \right)$$

MLPs as likelihood models

Throughout this course we will use neural nets to parameterize probabilistic models

Example: regression with homoscedastic Gaussian noise

- Univariate response variable $y \in \mathbb{R}$ with MLP $f : \mathbb{R}^D \to \mathbb{R}$

$$p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) = \mathcal{N}(y \mid f(\boldsymbol{x}; \boldsymbol{\theta}), \sigma^2_{\text{noise}}),$$

  where $f(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{E}(y \mid \boldsymbol{x})$
- Multivariate response variable $\boldsymbol{y} \in \mathbb{R}^D$ with MLP $f : \mathbb{R}^D \to \mathbb{R}^C$

$$p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{\Sigma}),$$

  where $\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})$

Example: multiclass classification $y \in \{1, \dots, C\}$

$$p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) = \text{Cat}\big(y \mid \text{softmax}(\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}))\big),$$

where

$$\text{softmax}(\boldsymbol{a}) = \left(\frac{e^{a_1}}{\sum_{c=1}^{C} e^{a_c}}, \dots, \frac{e^{a_C}}{\sum_{c=1}^{C} e^{a_c}} \right)$$

We use interchangeably $p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) = p(y \mid \boldsymbol{x}, \boldsymbol{\theta})$

All neural network architectures can be converted to likelihood models with appropriate link functions

If neural network parameters $\boldsymbol{\theta}$ are random variables, then these are called Bayesian neural networks

MLP for heteroscedastic regression

- Heteroscedastic nonlinear regression = nonlinear regression with input-dependent noise variance
- Assume univariate response variable $y \in \mathbb{R}$ with Gaussian noise
- MLP has two outputs ($C = 2$) which approximate $f_\mu(\boldsymbol{x}) \approx \mathbb{E}[y|\boldsymbol{x}, \boldsymbol{\theta}]$ and $f_\sigma(\boldsymbol{x}) \approx \mathbb{V}[y|\boldsymbol{x}, \boldsymbol{\theta}]$ such that

$$p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}) = \mathcal{N}(y \mid \boldsymbol{w}_\mu^\top f(\boldsymbol{x}; \boldsymbol{w}_{\text{shared}}), \sigma_+(\boldsymbol{w}_\sigma^\top f(\boldsymbol{x}; \boldsymbol{w}_{\text{shared}}))),$$

where e.g. $\sigma_+(a) = \log(1 + e^a)$

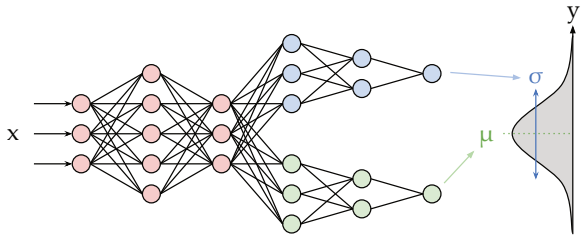


Figure 16.11 from (Murphy, 2023)

References

- Murphy K, Probabilistic Machine Learning: Advanced Topics, 2023.
- Prince SJD, Understanding Deep Learning, The MIT Press, 2023.