

that will allow us to easily apply it to other estimation problems in which there are also latent variables, and which will allow us to give a convergence guarantee.

11.2 Jensen's inequality

We begin our discussion with a very useful result called **Jensen's inequality**

Let f be a function whose domain is the set of real numbers. Recall that f is a convex function if $f''(x) \geq 0$ (for all $x \in \mathbb{R}$). In the case of f taking vector-valued inputs, this is generalized to the condition that its hessian H is positive semi-definite ($H \geq 0$). If $f''(x) > 0$ for all x , then we say f is **strictly** convex (in the vector-valued case, the corresponding statement is that H must be positive definite, written $H > 0$). Jensen's inequality can then be stated as follows:

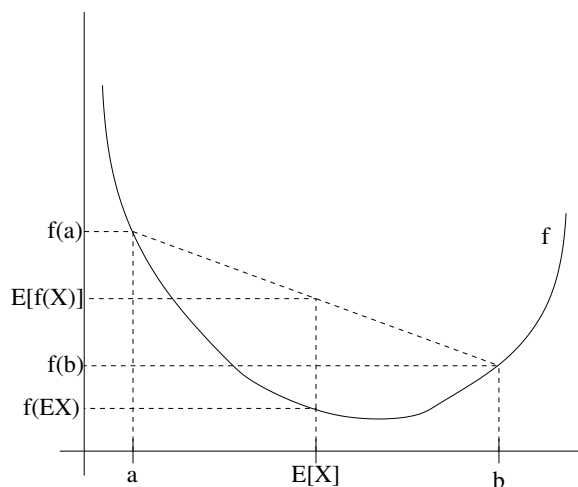
Theorem. Let f be a convex function, and let X be a random variable. Then:

$$E[f(X)] \geq f(EX).$$

Moreover, if f is strictly convex, then $E[f(X)] = f(EX)$ holds true if and only if $X = E[X]$ with probability 1 (i.e., if X is a constant).

Recall our convention of occasionally dropping the parentheses when writing expectations, so in the theorem above, $f(EX) = f(E[X])$.

For an interpretation of the theorem, consider the figure below.



Here, f is a convex function shown by the solid line. Also, X is a random variable that has a 0.5 chance of taking the value a , and a 0.5 chance of

taking the value b (indicated on the x -axis). Thus, the expected value of X is given by the midpoint between a and b .

We also see the values $f(a)$, $f(b)$ and $f(E[X])$ indicated on the y -axis. Moreover, the value $E[f(X)]$ is now the midpoint on the y -axis between $f(a)$ and $f(b)$. From our example, we see that because f is convex, it must be the case that $E[f(X)] \geq f(E[X])$.

Incidentally, quite a lot of people have trouble remembering which way the inequality goes, and remembering a picture like this is a good way to quickly figure out the answer.

Remark. Recall that f is [strictly] concave if and only if $-f$ is [strictly] convex (i.e., $f''(x) \leq 0$ or $H \leq 0$). Jensen's inequality also holds for concave functions f , but with the direction of all the inequalities reversed ($E[f(X)] \leq f(E[X])$, etc.).

11.3 General EM algorithms

Suppose we have an estimation problem in which we have a training set $\{x^{(1)}, \dots, x^{(n)}\}$ consisting of n independent examples. We have a latent variable model $p(x, z; \theta)$ with z being the latent variable (which for simplicity is assumed to take finite number of values). The density for x can be obtained by marginalized over the latent variable z :

$$p(x; \theta) = \sum_z p(x, z; \theta) \quad (11.1)$$

We wish to fit the parameters θ by maximizing the log-likelihood of the data, defined by

$$\ell(\theta) = \sum_{i=1}^n \log p(x^{(i)}; \theta) \quad (11.2)$$

We can rewrite the objective in terms of the joint density $p(x, z; \theta)$ by

$$\ell(\theta) = \sum_{i=1}^n \log p(x^{(i)}; \theta) \quad (11.3)$$

$$= \sum_{i=1}^n \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta). \quad (11.4)$$

But, explicitly finding the maximum likelihood estimates of the parameters θ may be hard since it will result in difficult non-convex optimization prob-