CS/DS 541: Class 16

Jacob Whitehill

More probability theory

Expectation

- Recall that the expected value of a function *f* w.r.t. a probability distribution *P*(**z**) is defined as:

$$\mathbb{E}_P[f(\mathbf{z})] = \int_{\mathbf{z}} f(\mathbf{z}) P(\mathbf{z}) d\mathbf{z}$$

Expectation

- Note that *f* itself might be a probability distribution, e.g.:

$$\mathbb{E}_P[Q(\mathbf{z})] = \int_{\mathbf{z}} Q(\mathbf{z})P(\mathbf{z})d\mathbf{z}$$

Expectation

- In this case, we can interpret the same quantity as either:
  - The expected value of *Q* w.r.t. probability distribution *P*.
  - The expected value of *P* w.r.t. probability distribution *Q*.

$$
\begin{aligned}
\mathbb{E}_P[Q(\mathbf{z})] &= \int_{\mathbf{z}} Q(\mathbf{z})P(\mathbf{z})d\mathbf{z} \\
&= \int_{\mathbf{z}} P(\mathbf{z})Q(\mathbf{z})d\mathbf{z} \\
&= \mathbb{E}_Q[P(\mathbf{z})]
\end{aligned}
$$

Expectation

- Here are a few other examples:

$$\int_{\mathbf{z}} Q(\mathbf{z}) \log P(\mathbf{z}) d\mathbf{z} = \mathbb{E}_Q[\log P(\mathbf{z})]$$

Expectation

- Here are a few other examples:

$$\int_{\mathbf{z}} Q(\mathbf{z}) \log P(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} = \mathbb{E}_Q[\log P(\mathbf{z} \mid \mathbf{x})]$$

Here, x is another random variable that is independent of the integration variable z.

Estimating expectations by sampling

- We can estimated the expected value of *f* w.r.t. probability distribution *P* by sampling, e.g.:

$$\mathbb{E}_P[f(\mathbf{z})] \approx \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}^{(i)})$$

$$\text{where } \mathbf{z}^{(i)} \sim P(\mathbf{z})$$

Sampling: example

- Suppose that $Z \epsilon \{ 1, 2, 3, 4 \}$ and

  $P(Z=1) = 0.2$
  $P(Z=2) = 0.3$
  $P(Z=3) = 0.4$
  $P(Z=4) = 0.1$

- What is $E_P[f(z)]$ for $f(z)=z^2$?

Sampling: example

- Suppose that $Z \epsilon$ { 1, 2, 3, 4 } and

  P(Z=1) = 0.2
  P(Z=2) = 0.3
  P(Z=3) = 0.4
  P(Z=4) = 0.1

- What is $E_P[f(z)]$ for $f(z)=z^2$?

- We can compute this analytically:

$$\mathbb{E}_P[f(\mathbf{z})] = \sum_{z=1}^{4} P(Z=z) f(z) = 0.2 \times 1^2 + 0.3 \times 2^2 + 0.4 \times 3^2 + 0.1 \times 4^2 = 6.6$$

Sampling: example

- Suppose that $Z \epsilon$ { 1, 2, 3, 4 } and
  P(Z=1) = 0.2
  P(Z=2) = 0.3
  P(Z=3) = 0.4
  P(Z=4) = 0.1
- What is $E_P[f(z)]$ for $f(z)=z^2$?
- We can compute this analytically:

$$\mathbb{E}_P[f(\mathbf{z})] = \sum_{z=1}^{4} P(Z=z)f(z) = 0.2 \times 1^2 + 0.3 \times 2^2 + 0.4 \times 3^2 + 0.1 \times 4^2 = 6.6$$

- But we can also estimate it by sampling from P, e.g., for n=1000:
  - 1/1000 * (f(2) + f(1) + f(3) + f(2) + f(3) + …) =
    1/1000 * (2^2 + 1^2 + 3^2 + 2^2 + 3^2 + …) = 6.75

Sampling from a Gaussian

- Suppose $z \sim P(z) = \mathcal{N}(z; \mu, \sigma^2)$
- To sample *z*, we can **either**:
  - Sample from *P*(*z*) directly (Python: `scipy.random.normal(loc=mu, scale=sigma)`).

Sampling from a Gaussian

- Suppose $z \sim P(z) = \mathcal{N}(z; \mu, \sigma^2)$
- To sample *z*, we can **either**:
  - Sample from *P*(*z*) directly (Python: `scipy.random.normal(loc=mu, scale=sigma)`).
  - Sample from a standard normal, multiply by σ, and add μ

$$z' \sim \mathcal{N}(z'; 0, 1)$$

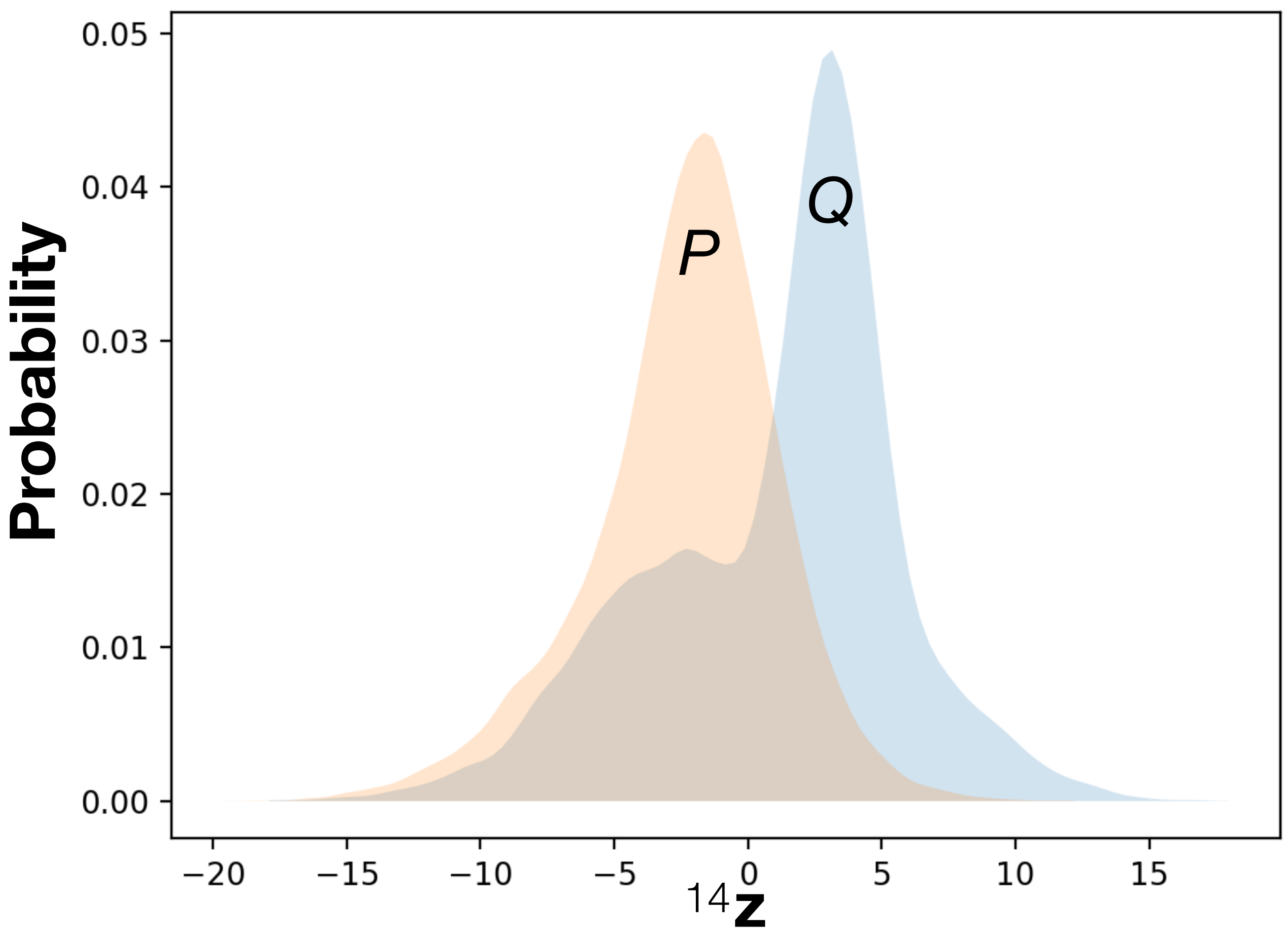
$$z = \sigma z' + \mu$$

(Python: `sigma*scipy.random.normal(0,1) + mu`).

Kullback-Leibler Divergence

- Consider two probability distributions $P(\mathbf{z})$, $Q(\mathbf{z})$.

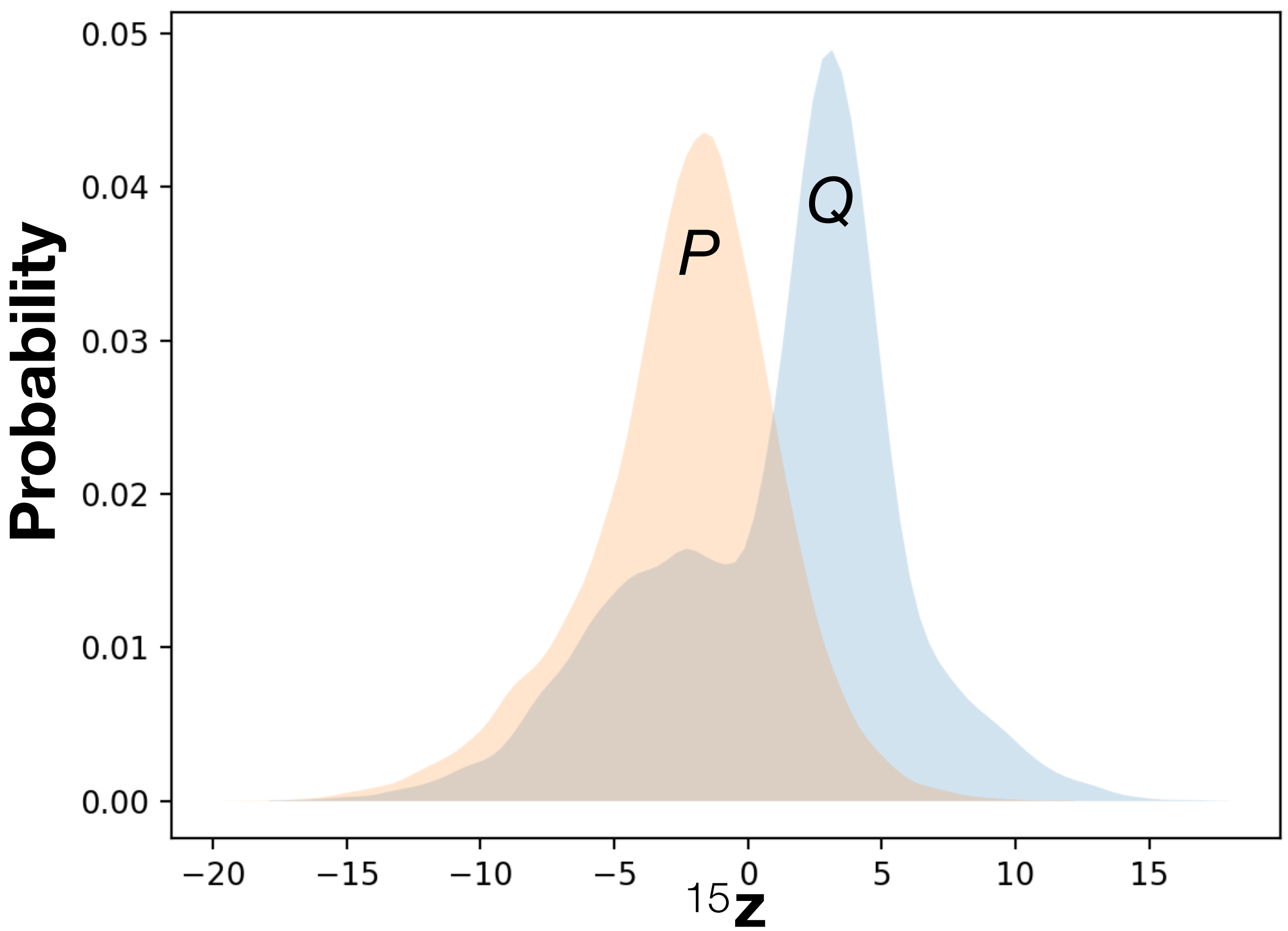
- How can we quantify the distance between them?



Kullback-Leibler Divergence

- The Kullback-Leibler (KL) divergence quantifies the distance of *Q* from *P* as the log difference in probabilities at each **z** weighted by the probability of **z** according to *P*.

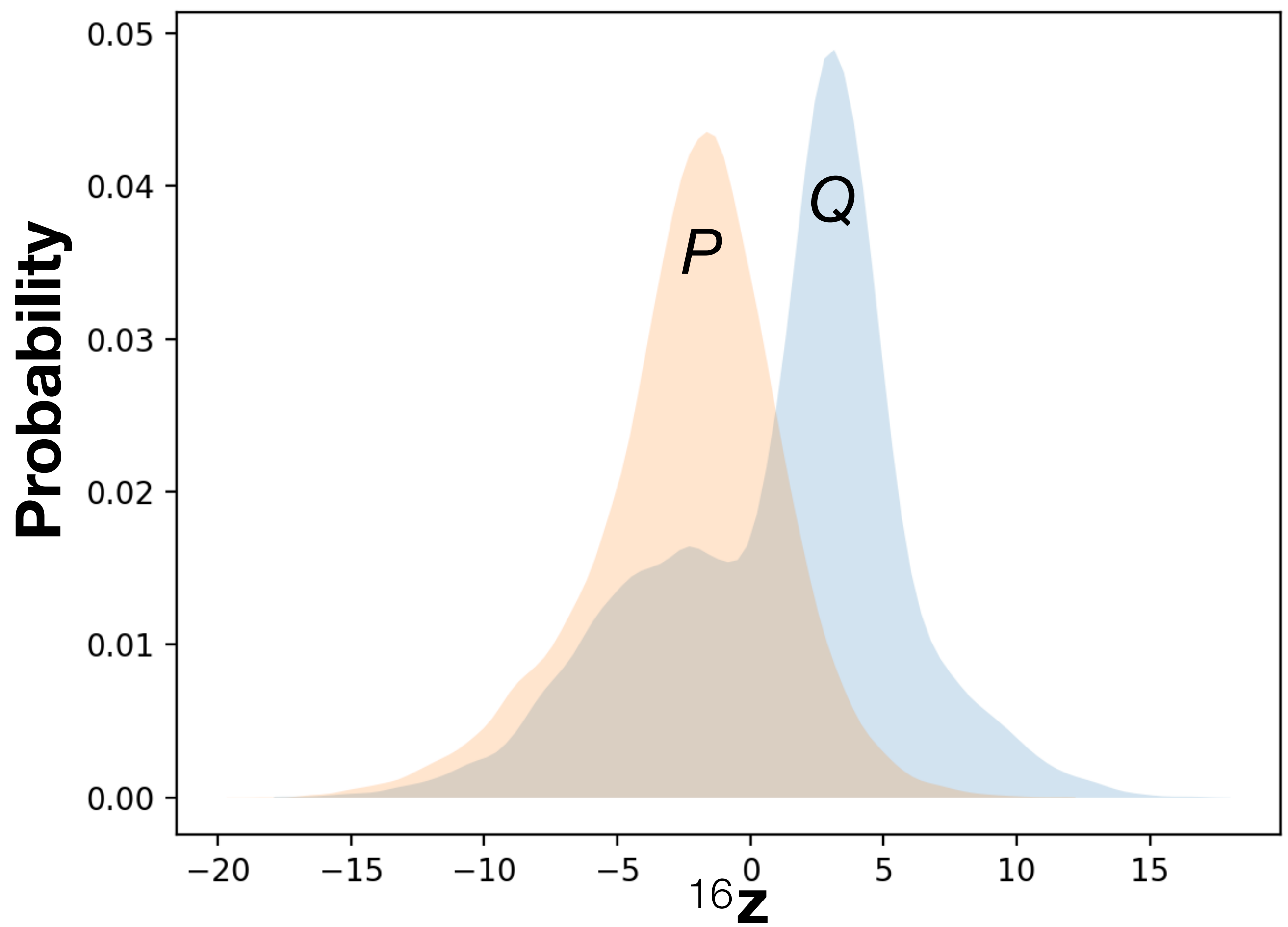
$$D_{\mathrm{KL}}(P(\mathbf{z}) \parallel Q(\mathbf{z})) = \int_{\mathbf{z}} P(\mathbf{z}) \boxed{\log \frac{P(\mathbf{z})}{Q(\mathbf{z})}} d\mathbf{z}$$



Kullback-Leibler Divergence

- The Kullback-Leibler (KL) divergence quantifies the distance of *Q* from *P* as the log difference in probabilities at each **z** weighted by the probability of **z** according to *P*.

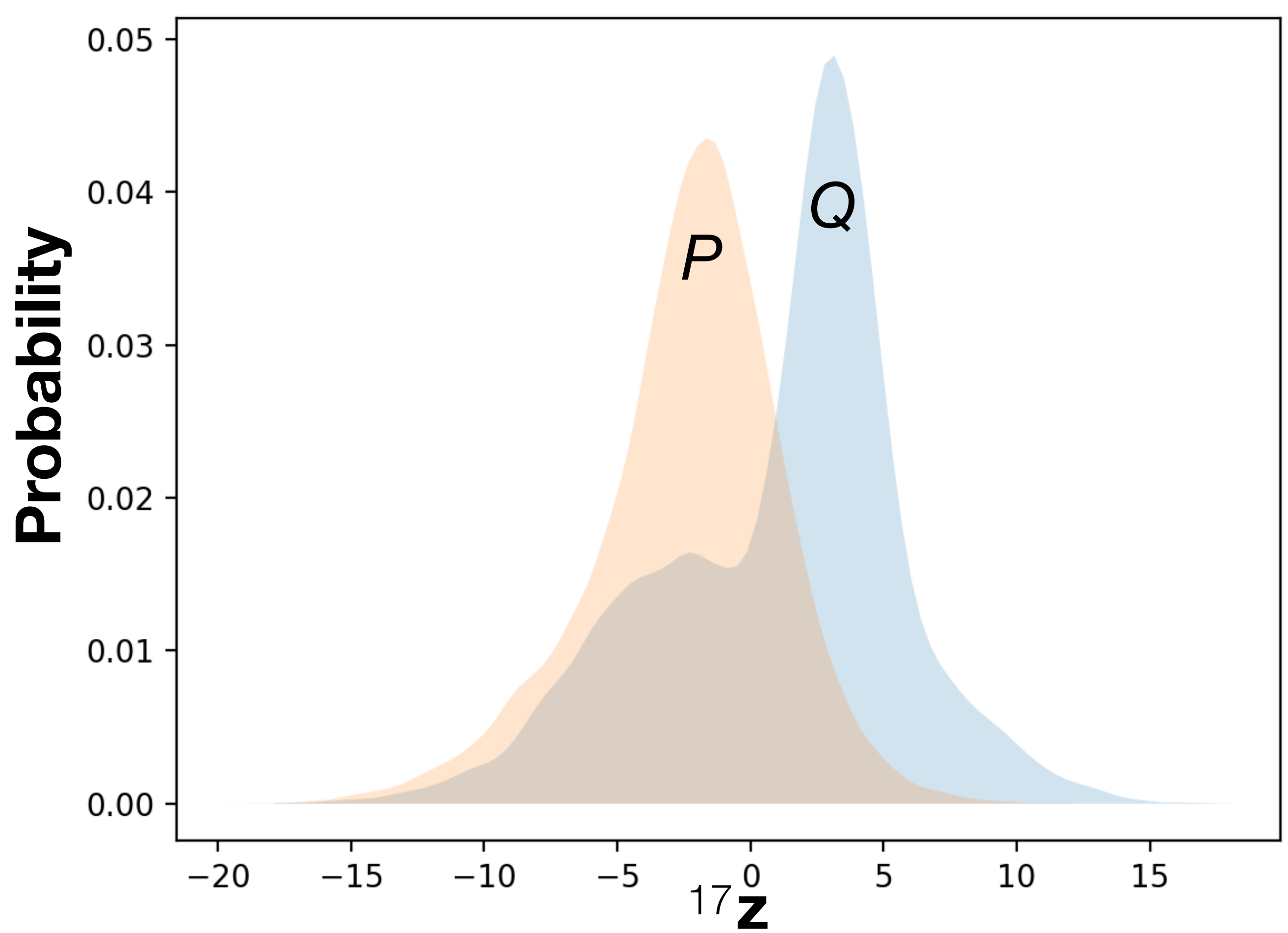
$$D_{\mathrm{KL}}(P(\mathbf{z}) \parallel Q(\mathbf{z})) = \int_{\mathbf{z}} \boxed{P(\mathbf{z})} \log \frac{P(\mathbf{z})}{Q(\mathbf{z})} d\mathbf{z}$$



Kullback-Leibler Divergence

- Note that the KL divergence is always non-negative.

$$D_{\mathrm{KL}}(P(\mathbf{z}) \parallel Q(\mathbf{z})) = \int_{\mathbf{z}} P(\mathbf{z}) \log \frac{P(\mathbf{z})}{Q(\mathbf{z})} d\mathbf{z} \quad \geqq 0$$



Kullback-Leibler Divergence

- We can also write the KL divergence as:

$$
\begin{aligned}
D_{\mathrm{KL}}(P(\mathbf{z}) \parallel Q(\mathbf{z})) &= \int_{\mathbf{z}} P(\mathbf{z}) \log \frac{P(\mathbf{z})}{Q(\mathbf{z})} d\mathbf{z} \\
&= -\int_{\mathbf{z}} P(\mathbf{z}) \log \frac{Q(\mathbf{z})}{P(\mathbf{z})} d\mathbf{z}
\end{aligned}
$$

Kullback-Leibler Divergence

- Note that the KL divergence is **not** symmetric:

$$D_{\mathrm{KL}}(P(\mathbf{z}) \parallel Q(\mathbf{z})) = \int_{\mathbf{z}} P(\mathbf{z}) \log \frac{P(\mathbf{z})}{Q(\mathbf{z})} d\mathbf{z}$$

$$D_{\mathrm{KL}}(Q(\mathbf{z}) \parallel P(\mathbf{z})) = \int_{\mathbf{z}} Q(\mathbf{z}) \log \frac{Q(\mathbf{z})}{P(\mathbf{z})} d\mathbf{z}$$

KL-divergence for Gaussian distributions

- For the special case of two Gaussian distributions

  $$P(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}) \text{ and } Q(\mathbf{z}) = \mathcal{N}\left(\mathbf{z}; \mu, \text{diag}[\sigma_1^2, \ldots, \sigma_m^2]\right)$$

  there is a closed formula for the KL-divergence:

  $$D_{\text{KL}}(Q(\mathbf{z}) \parallel P(\mathbf{z})) = -\frac{1}{2}\sum_{j=1}^{m}(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2)$$

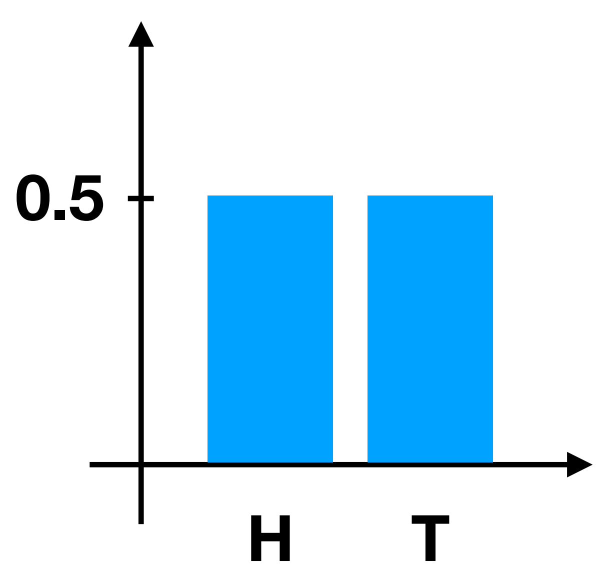
- Importantly, this function is differentiable in *μ* and *σ* (this will become useful later).

Entropy

- The entropy of a probability distribution $P(\mathbf{z})$ quantifies the amount of uncertainty in random variable **z**.

- Distributions that heavily favor some values of **z** over others are less “uncertain” than those that are more uniform, e.g.:

  - This distribution (over {H,T}) has higher entropy…

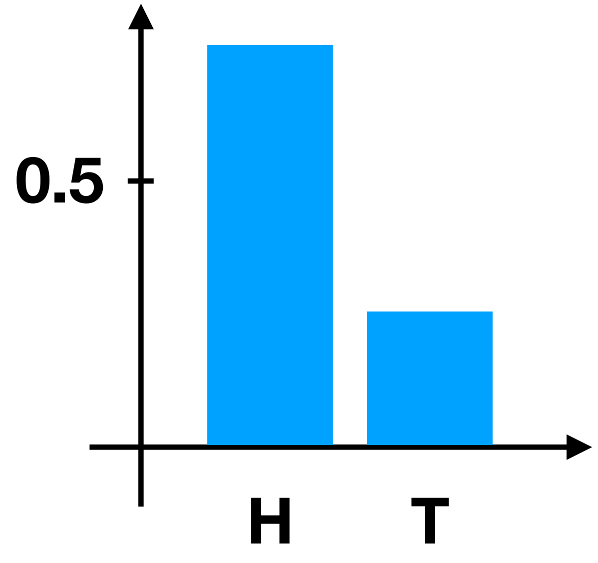


Entropy

- The entropy of a probability distribution $P(\mathbf{z})$ quantifies the amount of uncertainty in random variable $\mathbf{z}$.

- Distributions that heavily favor some values of $\mathbf{z}$ over others are less "uncertain" than those that are more uniform, e.g.:

  - …than this one.

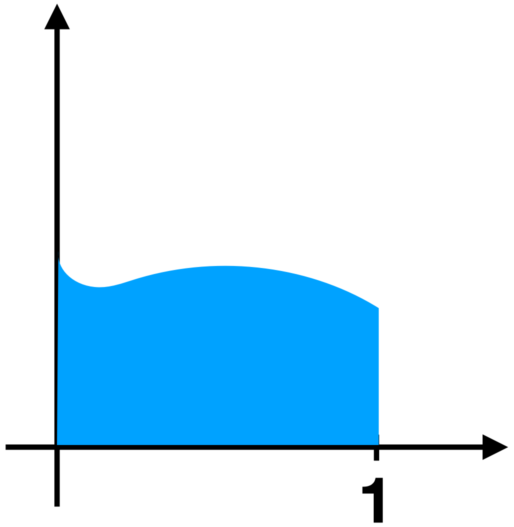


Entropy

- The entropy of a probability distribution $P(\mathbf{z})$ quantifies the amount of uncertainty in random variable $\mathbf{z}$.

- Distributions that heavily favor some values of $\mathbf{z}$ over others are less "uncertain" than those that are more uniform, e.g.:

  - This distribution (over [0,1]) has higher entropy…

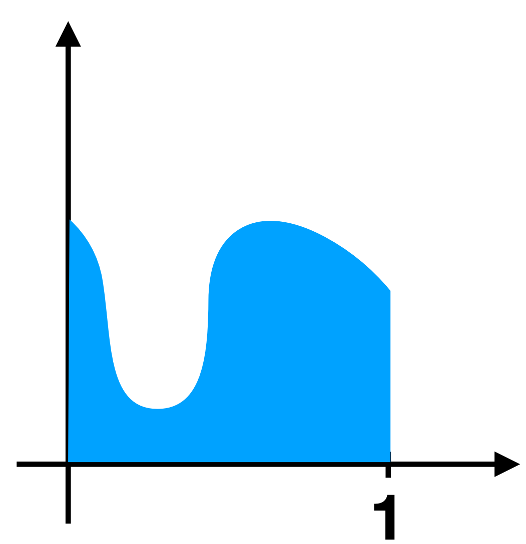


Entropy

- The entropy of a probability distribution $P(\mathbf{z})$ quantifies the amount of uncertainty in random variable $\mathbf{z}$.

- Distributions that heavily favor some values of $\mathbf{z}$ over others are less "uncertain" than those that are more uniform, e.g.:

  - …than this one.



Entropy

- For continuous RVs, we quantify entropy as the negative expected log probability over all values **z**:

$$H[\mathbf{z}] = -\int_{\mathbf{z}} P(\mathbf{z}) \log P(\mathbf{z}) d\mathbf{z}$$

- This equals the number of bits required to transmit the observed value of **z**, given both communicators know *P*(**z**).

Calculus of variations

Calculus of variations

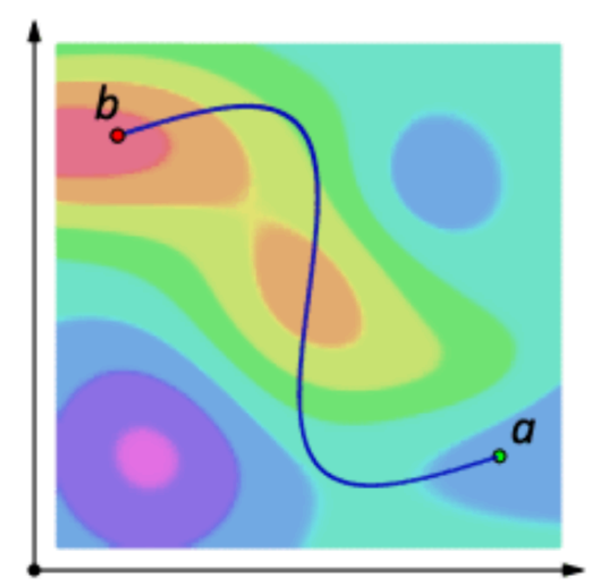
- Ordinary differential calculus is concerned with optimizing a function *f* w.r.t. scalar or vector-valued parameters **x**, e.g.:
  - Minimize w.r.t. *x* the function: $f(x) = 3x^2 + 2x - 1$

Calculus of variations

- Ordinary differential calculus is concerned with optimizing a function *f* w.r.t. scalar or vector-valued parameters **x**, e.g.:
  - Minimize w.r.t. *x* the function: $f(x) = 3x^2 + 2x - 1$
- In contrast, the calculus of variations can be used to optimize *f* w.r.t. a (infinite-dimensional) function ***r***, e.g.:
  - Minimize w.r.t. function **r** the function:

    $$f(\mathbf{r}) = \int_a^b g(\mathbf{r}(t))|\mathbf{r}'(t)|dt$$

    where *g* is a scalar field and **r** is a parametric curve from *a* to *b*.



https://en.wikipedia.org/wiki/Line_integral#Line_integral_of_a_scalar_field

Calculus of variations

- The calculus of variations can also identify the unique probability distribution over $\mathbb{R}$ that has maximum entropy:

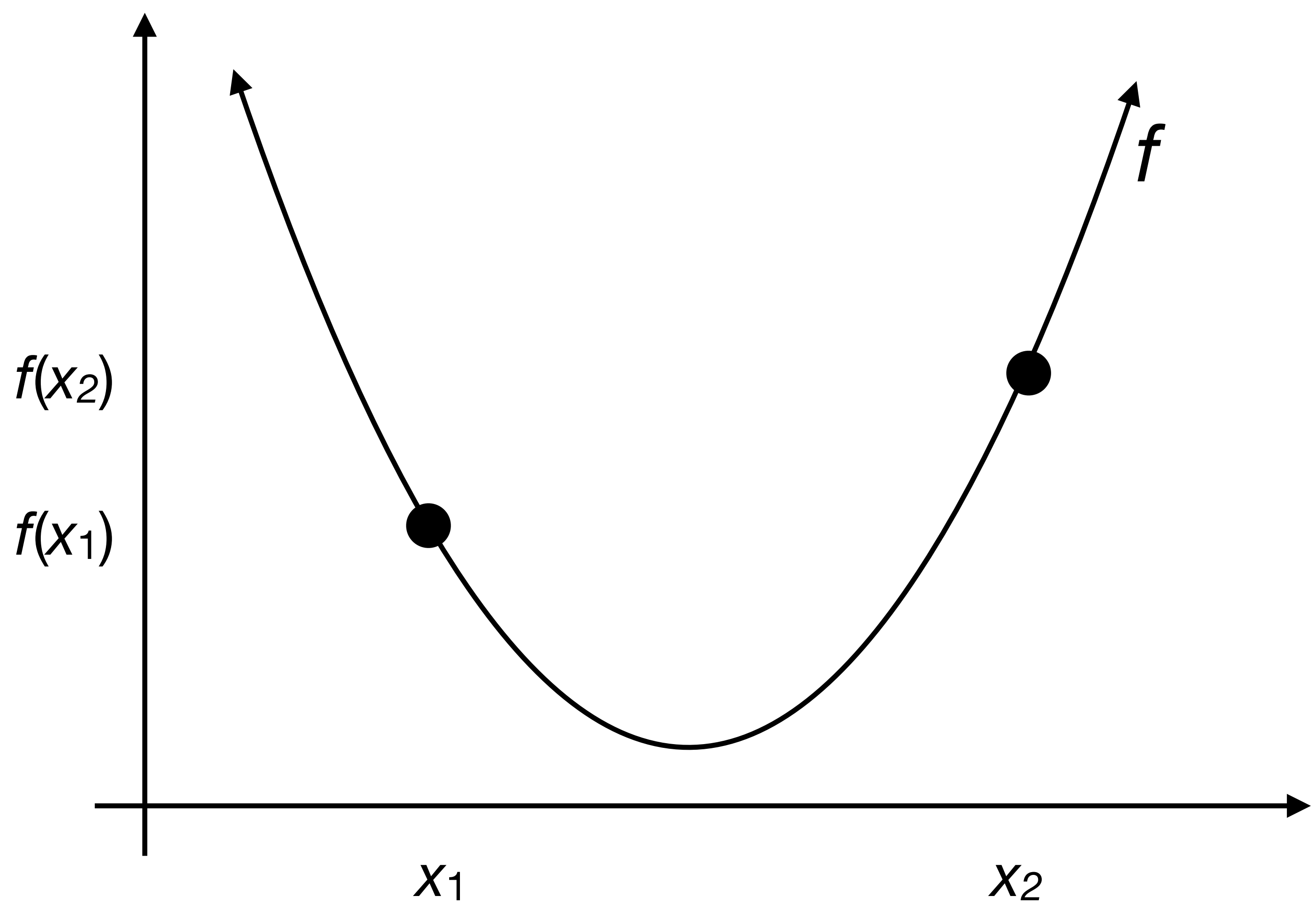
Calculus of variations

- The calculus of variations can also identify the unique probability distribution over $\mathbb{R}$ that has maximum entropy:
  - Gaussian.

Jensen’s inequality

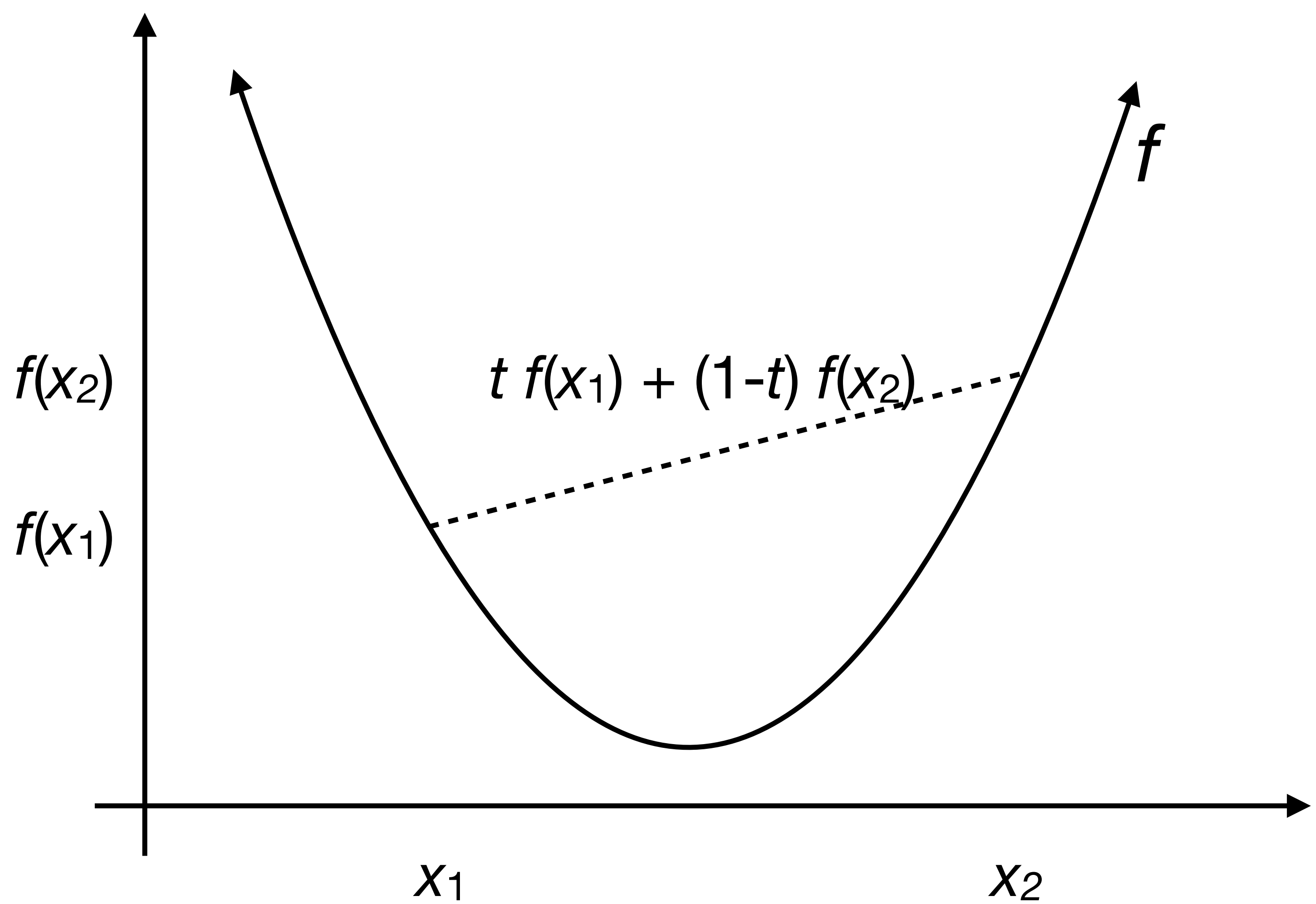
Jensen's inequality

- Consider any convex function f, and its value at any two points x_1, x_2 in its domain:



Jensen's inequality

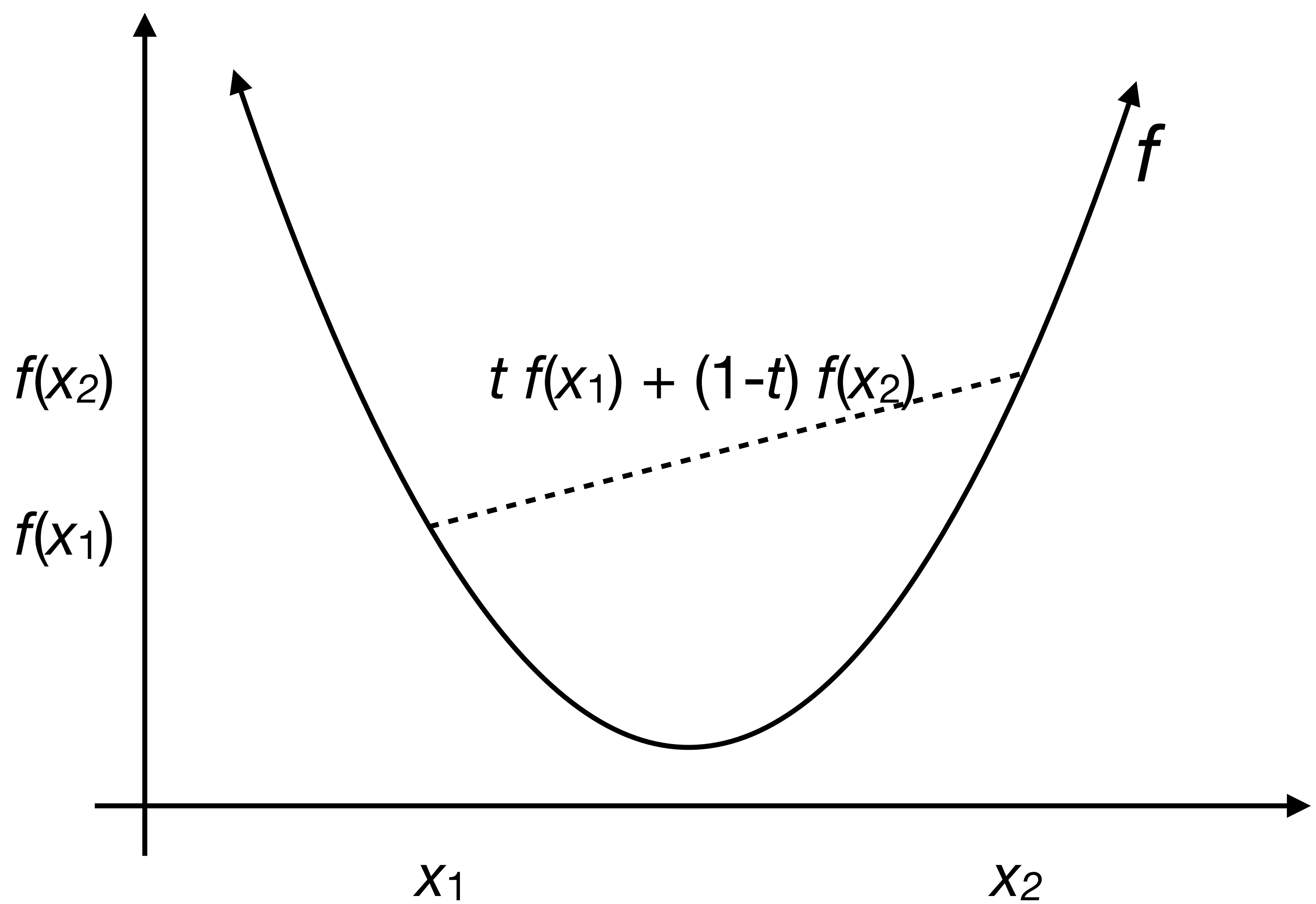
- We can form the secant line between x_1 and x_2:



Jensen's inequality

- At all points in [x_1, x_2], the secant is at least as large as *f*.

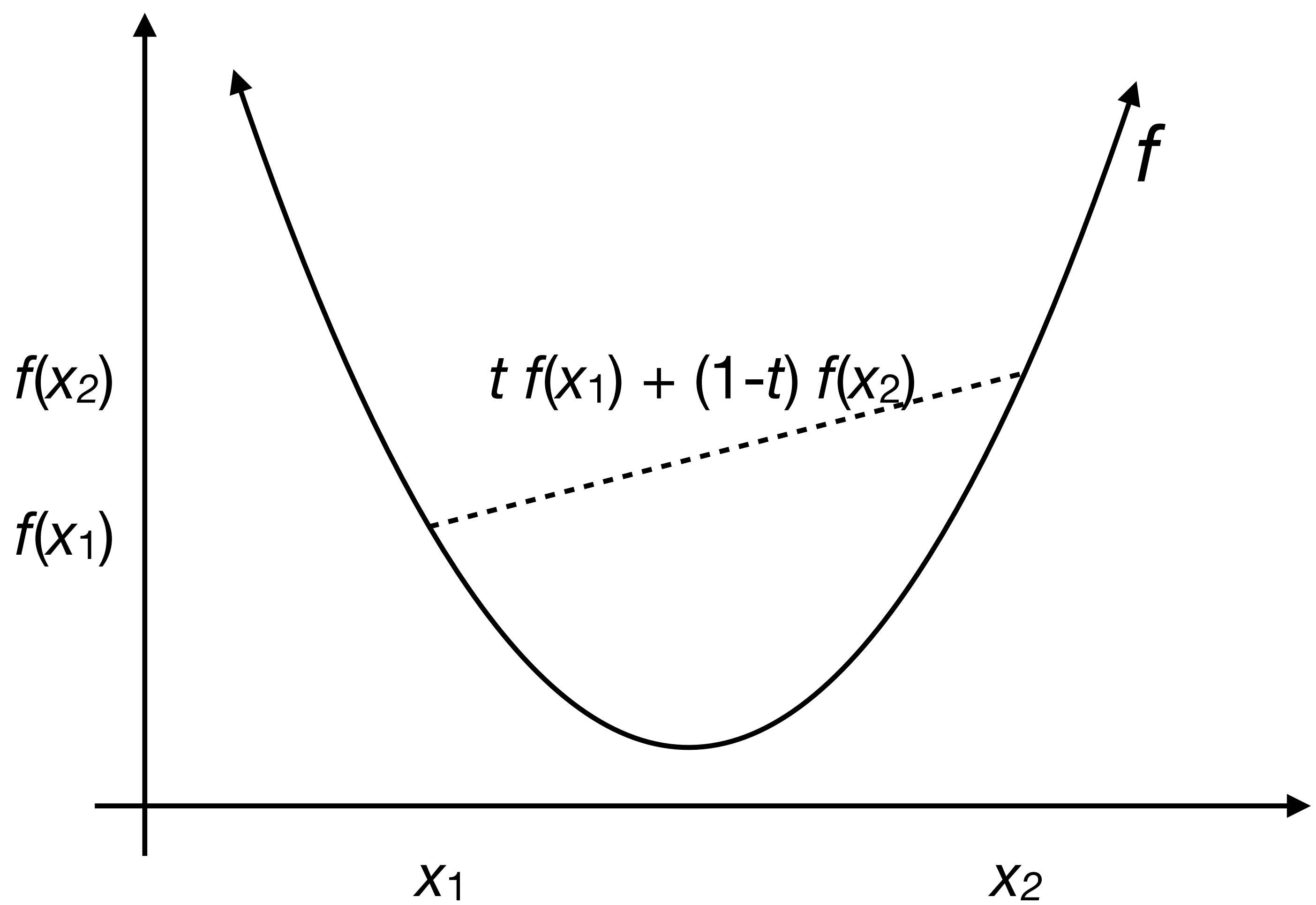
$$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2)$$



Jensen's inequality

- This is called **Jensen's inequality**.

$$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2)$$



Jensen’s inequality

- Jensen’s inequality can be generalized:

$$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2)$$

Note: this is not a derivation! It is just a list of generalizations of the inequality.

Jensen's inequality

- Jensen's inequality can be generalized:

$$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2)$$

$$\frac{\sum_{i=1}^{n} a_i f(x_i)}{\sum_{i=1}^{n} a_i} \geq f\left(\frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i}\right)$$

Note: this is not a derivation! It is just a list of generalizations of the inequality.

Jensen's inequality

- Jensen's inequality can be generalized:

$$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2)$$

$$\frac{\sum_{i=1}^{n} a_i f(x_i)}{\sum_{i=1}^{n} a_i} \geq f\left(\frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i}\right)$$

$$\sum_i \frac{1}{n} f(x_i) \geq f\left(\sum_i \frac{1}{n} x_i\right)$$

Note: this is not a derivation! It is just a list of generalizations of the inequality.

Jensen's inequality

- Jensen's inequality can be generalized:

$$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2)$$

$$\frac{\sum_{i=1}^{n} a_i f(x_i)}{\sum_{i=1}^{n} a_i} \geq f\left(\frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i}\right)$$

$$\sum_i \frac{1}{n} f(x_i) \geq f\left(\sum_i \frac{1}{n} x_i\right)$$

$$\frac{1}{n}\sum_i f(x_i) \geq f\left(\frac{1}{n}\sum_i x_i\right)$$

Note: this is not a derivation! It is just a list of generalizations of the inequality.

Jensen's inequality

- Jensen's inequality can be generalized:

$$t f(x_1) + (1-t) f(x_2) \geq f(t x_1 + (1-t) x_2)$$

$$\frac{\sum_{i=1}^{n} a_i f(x_i)}{\sum_{i=1}^{n} a_i} \geq f\left(\frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i}\right)$$

$$\sum_i \frac{1}{n} f(x_i) \geq f\left(\sum_i \frac{1}{n} x_i\right)$$

$$\frac{1}{n} \sum_i f(x_i) \geq f\left(\frac{1}{n} \sum_i x_i\right)$$

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

Note: this is not a derivation! It is just a list of generalizations of the inequality.

Jensen's inequality

- Jensen's inequality can be generalized:

$$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2)$$

$$\frac{\sum_{i=1}^n a_i f(x_i)}{\sum_{i=1}^n a_i} \geq f\left(\frac{\sum_{i=1}^n a_i x_i}{\sum_{i=1}^n a_i}\right)$$

$$\sum_i \frac{1}{n} f(x_i) \geq f\left(\sum_i \frac{1}{n} x_i\right)$$

$$\frac{1}{n}\sum_i f(x_i) \geq f\left(\frac{1}{n}\sum_i x_i\right)$$

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

$$\int_x f(x)P(x)dx \geq f\left(\int_x xP(x)dx\right)$$

Note: this is not a derivation! It is just a list of generalizations of the inequality.

Jensen's inequality

- Consider $f(x)=\log(x)$.
- f is concave (opposite of convex) because its second derivative is negative everywhere in its domain.

Jensen's inequality

- Consider *f*(*x*)=log(*x*).
- *f* is concave (opposite of convex) because its second derivative is negative everywhere in its domain.
- Therefore, when we apply Jensen's inequality we reverse the sign, i.e.:

$$\int_x f(x)P(x)dx \geq f\left(\int_x xP(x)dx\right) \implies$$

$$\log \int_x xP(x)dx \geq \int_x \log(x)P(x)dx$$

Here, we can "pull" the log into the integral.