cairo.thws
Center for Artificial Intelligence
Technical University of
Applied Sciences
Würzburg-Schweinfurt

Magda Gregorová
magda.gregorova@thws.de

Meausure-theory view of probability
handwavy and informal

December 1, 2025

Outline

Outline

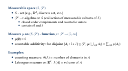
Measures and probability

Random variables and distributions

Three ways to describe a distribution

Transformations and push-forward

Change of variables at density level



1. σ-algebra $\mathscr{S}$ is collection of measurable sets (closed under complements and countable unions)
2. For practical purposes: think of $\mathscr{S}$ as "all reasonable subsets"
3. Lebesgue measure generalizes notions of length, area, volume
4. Probability measure is just a finite measure normalized to total mass 1
5. This abstraction allows us to talk about probability rigorously, but we'll soon move to more practical objects

Measures - assigning mass to sets

Measurable space $(S, \mathscr{S})$

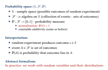
- S - set (e.g., $\mathbb{R}^d$, discrete set, etc.)
- $\mathscr{S}$ - σ-algebra on S (collection of measurable subsets of S)
    - closed under complements and countable unions
    - contains $\emptyset$ and S

Measure μ **on** $(S, \mathscr{S})$ **- function** $\mu : \mathscr{S} \to [0, \infty]$

- $\mu(\emptyset) = 0$
- countable additivity: for disjoint $\{A_i : i \in I\} \subseteq \mathscr{S}$, $\mu(\bigcup_{i \in I} A_i) = \sum_{i \in I} \mu(A_i)$

Examples:

- counting measure: $\#(A)$ = number of elements in A
- Lebesgue measure on $\mathbb{R}^d$: $\lambda(A)$ = volume of A



1. Probability measure is just a normalized measure - total mass equals 1
2. The abstract probability space $(S, \mathscr{S}, \mathbb{P})$ is often denoted $(\Omega, \mathscr{F}, \mathbb{P})$ in literature
3. Random experiment: think coin flip, dice roll, or any random process
4. Outcome $s \in S$: specific result of the experiment (e.g., "heads", or specific image)
5. Event $A \in \mathscr{S}$: set of outcomes we're interested in (e.g., "at least 2 heads in 3 flips")
6. This level of abstraction separates the experiment from what we measure
7. But as we'll see next, we usually work with random variables that map outcomes to values we care about

Probability measure

Probability space $(S, \mathscr{S}, \mathbb{P})$

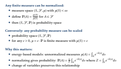
- S - sample space (possible outcomes of random experiment)
- $\mathscr{S}$ - σ-algebra on S (collection of events - sets of outcomes)
- $\mathbb{P} : \mathscr{S} \to [0,1]$ - probability measure
  - normalization: $\mathbb{P}(S) = 1$
  - countable additivity (same as before)

Interpretation:

- random experiment produces outcome $s \in S$
- event $A \in \mathscr{S}$ is set of outcomes
- $\mathbb{P}(A)$ is probability that outcome lies in A

Abstract formalism: $(S, \mathscr{S}, \mathbb{P})$ gives rigorous framework
In practice: we work with random variables and their distributions

1. Normalization is just dividing by total mass
2. This shows probability measures are special cases of finite measures
3. Connection to statistical physics and energy-based models
4. Partition function Z is the normalizing constant
5. When we transform measures, we transform probabilities - this is key for flows



Relationship: measures and probability

Any finite measure can be normalized:

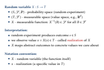
- measure space $(S, \mathscr{S}, \mu)$ with $\mu(S) < \infty$
- define $\mathbb{P}(A) = \frac{\mu(A)}{\mu(S)}$ for $A \in \mathscr{S}$
- then $(S, \mathscr{S}, \mathbb{P})$ is probability space

Conversely: any probability measure can be scaled

- probability space $(S, \mathscr{S}, \mathbb{P})$
- for any $c > 0$, $\mu = c \cdot \mathbb{P}$ is finite measure with $\mu(S) = c$

Why this matters:

- energy-based models: unnormalized measures $\mu(A) = \int_A e^{-E(s)} ds$
- normalizing gives probability: $\mathbb{P}(A) = \frac{1}{Z} \int_A e^{-E(s)} ds$ where $Z = \int_S e^{-E(s)} ds$
- change of variables preserves this relationship



1. Measurability is technical requirement - ensures we can measure probabilities after transformation
2. For continuous functions between nice spaces (e.g., $\mathbb{R}^d$ with Borel sets), measurability is automatic
3. The value space T is what we actually care about - numbers, vectors, images, etc.
4. Abstract sample space S is often just formal device
5. In deep learning: we care about generated image x, not the random seed that produced it
6. Capital letter = random variable (random), lowercase = realization (fixed value)

Random variable - moving to value space

Random variable $X : S \to T$

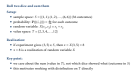
- $(S, \mathscr{S}, \mathbb{P})$ - probability space (random experiment)
- $(T, \mathscr{T})$ - measurable space (value space, e.g., $\mathbb{R}^d$)
- X - measurable function: $X^{-1}(B) \in \mathscr{S}$ for all $B \in \mathscr{T}$

Interpretation:

- random experiment produces outcome $s \in S$
- we observe value $x = X(s) \in T$ - called realization of X
- X maps abstract outcomes to concrete values we care about

Notation convention:

- X - random variable (the function itself)
- x - realization (a specific value in T)



1. Concrete example to ground the abstraction
2. Note: multiple outcomes in S can give same value in T (e.g., (2,6) and (3,5) both give sum 8)
3. This is why we need the distribution P_X on T - it aggregates probabilities
4. Example: $P_X(\{7\}) = \frac{6}{36}$ because 6 outcomes map to sum of 7

Example: dice sum

Roll two dice and sum them

Setup:

- sample space: $S = \{(1,1),(1,2),\dots,(6,6)\}$ (36 outcomes)
- probability: $\mathbb{P}(\{(i,j)\}) = \frac{1}{36}$ for each outcome
- random variable: $X(s_1,s_2) = s_1 + s_2$
- value space: $T = \{2,3,4,\dots,12\}$

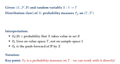
Realization:

- if experiment gives $(3,5) \in S$, then $x = X(3,5) = 8$
- $x = 8$ is a realization of random variable X

Key point:

- we care about the sum (value in T), not which dice showed what (outcome in S)
- this motivates working with distribution on T directly

1. The distribution P_X is completely determined by X and $\mathbb{P}$
2. Push-forward: probability "flows" from S to T through X
3. Once we have P_X, we can often forget about $(S, \mathscr{S}, \mathbb{P})$
4. In practice: we specify P_X directly without ever mentioning S
5. "X N(0,1)" directly specifies distribution on $\mathbb{R}$, no mention of underlying experiment



Distribution of random variable

Given: $(S, \mathscr{S}, \mathbb{P})$ **and random variable** $X : S \to T$

Distribution (law) of X**: probability measure** P_X **on** $(T, \mathscr{T})$

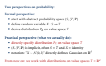
$$P_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(\{s \in S : X(s) \in B\}), \quad B \in \mathscr{T}$$

Interpretation:

- $P_X(B)$ = probability that X takes value in set B
- P_X lives on value space T, not on sample space S
- P_X is the push-forward of $\mathbb{P}$ by X

Notation: $X \sim P_X$ means "X has distribution P_X"

Key point: P_X is a probability measure on T - we can work with it directly!



1. This is the key conceptual shift for students
2. Abstract probability space is formal machinery, but not where we actually work
3. In generative modeling: we always work directly with distributions on $\mathbb{R}^d$
4. The triplet $(\Omega, \mathscr{F}, \mathbb{P})$ is rarely mentioned in ML papers
5. Examples: "$z \sim p_{\text{prior}}$", "$x \sim p_{\text{data}}$" - these directly specify distributions
6. We're working with push-forward measures, but we specify them directly
7. Next section: we'll see different ways to represent these distributions (measure, CDF, density)

Working directly with distributions

Two perspectives on probability:

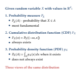
Formal perspective:

- start with abstract probability space $(S, \mathscr{S}, \mathbb{P})$
- define random variable $X : S \to T$
- derive distribution P_X on value space T

Practical perspective (what we actually do):

- directly specify distribution P_X on value space T
- $(S, \mathscr{S}, \mathbb{P})$ is implicit, often $S = T$ and $X =$ identity
- notation: "$X \sim \mathcal{N}(0, I)$" directly defines Gaussian on $\mathbb{R}^d$

From now on: we work with distributions on value spaces $T = \mathbb{R}^d$



Given random variable X with values in $\mathbb{R}^d$:

1. Probability measure P_X

- $P_X(A)$ - probability that $X \in A$
- most fundamental

2. Cumulative distribution function (CDF) F_X

- $F_X(x) = P_X((-\infty, x])$
- always exists

3. Probability density function (PDF) p_X

- $P_X(A) = \int_A p_X(x)\,dx$ when it exists
- does not always exist

Three views of the same distribution

1. This is a key conceptual point that students often miss
2. The measure, CDF, and density (when it exists) all describe the same distribution
3. Other representations exist: characteristic function $\phi_X(t) = E[e^{itX}]$, moment generating function $M_X(t) = E[e^{tX}]$, quantile function, etc.
4. We focus on these three because: (1) measure is fundamental, (2) CDF always exists, (3) density is what we compute with in ML
5. Measure is most general, CDF always exists, density only sometimes exists
6. In ML we mostly work with densities, but need to understand when they exist
7. Component-wise for CDF in $\mathbb{R}^d$: $F_X(x_1, \ldots, x_d) = P_X((-\infty, x_1] \times \cdots \times (-\infty, x_d])$

Three primary representations of a distribution

Given random variable X with values in $\mathbb{R}^d$:

1. Probability measure P_X

- $P_X(A)$ - probability that $X \in A$
- most fundamental

2. Cumulative distribution function (CDF) F_X

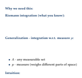
- $F_X(x) = P_X((-\infty, x])$
- always exists

3. Probability density function (PDF) p_X

- $P_X(A) = \int_A p_X(x)\,dx$ when it exists
- does not always exist

Three views of the same distribution

(other representations exist: characteristic function, MGF, etc.)



1. Riemann integration: only over intervals, uses length/area/volume
2. Measure integration: over any measurable set, uses any measure
3. The measure μ determines how we "weigh" different parts of the space
4. This is a generalization that will let us work with probability measures

Interlude: integration with respect to a measure

Why we need this: to connect measures with densities (functions)

Riemann integration (what you know):

$$\int_a^b f(x)\,dx = \text{area under curve}$$

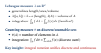
Generalization - integration w.r.t. measure μ:

$$\int_A f\,d\mu$$

- A - any measurable set
- μ - measure (weighs different parts of space)

Intuition: weighted sum of f over set A, weighted by measure μ

1. Lebesgue measure is the "natural" measure on Euclidean space
2. For nice functions on nice sets: Riemann = Lebesgue integration
3. But Lebesgue integration works for much wider class of functions and sets
4. The notation dx is shorthand for $d\lambda(x)$ (Lebesgue measure)
5. Counting measure: discrete sums become integrals!
6. Example with probability: $\int_{\mathbb{R}^d} f\, dP_X = E[f(X)]$ works for both discrete and continuous X
7. This framework unifies discrete and continuous cases - no need for separate notation



Interlude: Lebesgue measure and notation

cairo.thws
Center for Artificial Intelligence
Technical University of Applied Sciences Würzburg-Schweinfurt

Lebesgue measure λ on $\mathbb{R}^d$

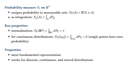
- generalizes length/area/volume
- $\lambda([a,b]) = b - a$ (length), $\lambda(A) =$ volume of A
- integration: $\int_A f\, d\lambda = \int_A f(x)\, dx$ (familiar!)

Counting measure # on discrete/countable sets

- $\#(A) =$ number of elements in A
- integration: $\int_A f\, d\# = \sum_{x\in A} f(x)$ (discrete sum!)

Key insight: integral notation unifies discrete and continuous

$$\int_A f\, d\mu \begin{cases} = \int_A f(x)\, dx & \text{continuous (Lebesgue)} \\ = \sum_{x\in A} f(x) & \text{discrete (counting)} \end{cases}$$



1. We work with measurable sets - for practical purposes, "all reasonable subsets"
2. Single points have zero probability for continuous distributions because they have zero volume (Lebesgue measure)
3. Intuition: $P_X(\{x_0\}) = \int_{\{x_0\}} dP_X = 0$ because the set has zero measure
4. This is why we can't talk about "probability at a point" for continuous distributions - only density
5. Integration w.r.t. probability measure uses the framework from the interlude
6. The normalization property connects to what we saw in Section 1: any finite measure can be normalized
7. Measure is the fundamental object, but we often use CDF or density for computation

Probability measure - the fundamental object

Probability measure P_X on $\mathbb{R}^d$

- assigns probability to measurable sets: $P_X(A) = \mathbb{P}(X \in A)$
- as integration: $P_X(A) = \int_A dP_X$

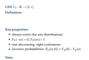
Key properties:

- normalization: $P_X(\mathbb{R}^d) = \int_{\mathbb{R}^d} dP_X = 1$
- for continuous distributions: $P_X(\{x_0\}) = \int_{\{x_0\}} dP_X = 0$ (single points have zero probability)

Properties:

- most fundamental representation
- works for discrete, continuous, and mixed distributions

1. CDF is unique representation - different distributions have different CDFs
2. Definition uses measure integration: integral of constant function 1 over the interval
3. For multidimensional case, CDF is defined component-wise
4. CDF can have jumps (discrete distributions) or be smooth (continuous distributions)
5. Derivative of CDF gives density (when density exists): $p_X(x) = \frac{dF_X}{dx}$
6. In practice: we rarely work with CDF directly in ML, but it's theoretically important



CDF $F_X : \mathbb{R} \to [0,1]$

Definition:

Key properties:

- always exists (for any distribution)
- $F_X(-\infty) = 0$, $F_X(\infty) = 1$
- non-decreasing, right-continuous
- recovers probabilities: $P_X((a,b]) = F_X(b) - F_X(a)$

Note:

Cumulative distribution function (CDF)

CDF $F_X : \mathbb{R} \to [0,1]$

Definition:

$$F_X(x) = P_X((-\infty, x]) = \int_{(-\infty,x]} dP_X$$

Key properties:

- always exists (for any distribution)
- $F_X(-\infty) = 0$, $F_X(\infty) = 1$
- non-decreasing, right-continuous
- recovers probabilities: $P_X((a,b]) = F_X(b) - F_X(a)$

Note: for $\mathbb{R}^d$: $F_X(x_1, \ldots, x_d) = P_X((-\infty, x_1] \times \cdots \times (-\infty, x_d])$

1. Radon-Nikodym derivative: formal way to define density as derivative of one measure w.r.t. another
2. The defining property says: if we integrate the density, we get the probability measure
3. This is analogous to: $df = f'(x)dx$ in calculus
4. Next slide: how to understand $p_X(x)$ at a point

PDF $p_X : \mathbb{R}^d \to [0, \infty)$

Definition via Radon-Nikodym derivative:

Note:

Defining property:

Notation:

Probability density function (PDF)

PDF $p_X : \mathbb{R}^d \to [0, \infty)$

Definition via Radon-Nikodym derivative:

$$p_X = \frac{dP_X}{d\lambda}$$

where λ is Lebesgue measure (volume)

Note: density does not always exist (requires absolute continuity)

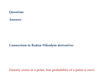
Defining property: for any set A,

$$P_X(A) = \int_A p_X \, d\lambda = \int_A p_X(x) \, dx$$

Notation: $dP_X = p_X \, d\lambda$ or $dP_X(x) = p_X(x) \, dx$

Lebesgue measure is "weighted" by p_X to give P_X

1. This limit IS the Radon-Nikodym derivative - now concrete!
2. The notation $\frac{dP_X}{d\lambda}$ means exactly this: derivative of one measure w.r.t. another
3. In 1D: $B_\epsilon(x) = [x-\epsilon, x+\epsilon]$, $\lambda(B_\epsilon(x)) = 2\epsilon$
4. So $p_X(x) = \lim_{\epsilon\to 0} \frac{P_X([x-\epsilon, x+\epsilon])}{2\epsilon}$ - this looks like the derivative from calculus!
5. In higher dimensions: ball has volume proportional to ϵ^d
6. This is why density can be > 1: it's a rate (derivative), not a probability
7. Common mistake: confusing $p_X(x)$ (density, can be > 1) with $P_X(\{x\})$ (probability, equals 0)
8. Interpretation: $p_X(x)$ measures local concentration of probability



Question:

Answer:

Connection to Radon-Nikodym derivative:

Density exists at a point, but probability of a point is zero!

Density at a point - the key subtlety

Question: We said $P_X(\{x\}) = 0$ for single points. So what is $p_X(x)$?

Answer: $p_X(x)$ is defined as a limit:

$$p_X(x) = \lim_{\epsilon\to 0} \frac{P_X(B_\epsilon(x))}{\lambda(B_\epsilon(x))}$$

where $B_\epsilon(x)$ is a small ball of radius ϵ around x

Connection to Radon-Nikodym derivative:

$$p_X(x) = \frac{dP_X}{d\lambda}(x) = \lim_{\epsilon\to 0} \frac{P_X(B_\epsilon(x))}{\lambda(B_\epsilon(x))}$$

This is literally the derivative of measure P_X w.r.t. measure λ at point x

Density exists at a point, but probability of a point is zero!

1. Measure is most fundamental but assigns probabilities to sets, not points
2. CDF always exists and can be evaluated at points
3. PDF only exists for absolutely continuous distributions
4. In ML: we work with PDFs (assume absolute continuity)
5. PDF is most convenient: can evaluate at points and optimize

Comparison of the three representations

How to convert between them

Relationship between the three

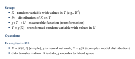
Comparison of the three representations:

	Measure P_X	**CDF** F_X	**PDF** p_X
Always exists?	yes	yes	no
Domain	sets	points	points
Range	$[0,1]$	$[0,1]$	$[0,\infty)$

How to convert between them:

From	**To**	**Formula**
P_X	F_X	$F_X(x) = P_X((-\infty, x])$
P_X	p_X	$p_X = \frac{dP_X}{d\lambda}$ (if abs. cont.)
p_X	P_X	$P_X(A) = \int_A p_X \, d\lambda$
p_X	F_X	$F_X(x) = \int_{-\infty}^{x} p_X(t)\, dt$
F_X	p_X	$p_X(x) = \frac{dF_X}{dx}$ (if smooth)

1. This is the fundamental question in generative modeling
2. We transform a simple distribution to get a complex one
3. The transformation g is typically a neural network
4. We work directly with distributions on value spaces (as established in Section 2)
5. No need to refer back to abstract probability space $(S, \mathscr{S}, \mathbb{P})$



Setup:
- X - random variable with values in T (e.g., $\mathbb{R}^d$)
- P_X - distribution of X on T
- $g : T \to U$ - measurable function (transformation)
- $Y = g(X)$ - transformed random variable with values in U

Question: If we know P_X, what is the distribution P_Y of $Y = g(X)$?

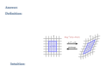
Examples in ML:
- $X \sim \mathcal{N}(0, I)$ (simple), g is neural network, $Y = g(X)$ (complex model distribution)
- data transformation: X is data, g encodes to latent space

Transformations of random variables

Setup:
- X - random variable with values in T (e.g., $\mathbb{R}^d$)
- P_X - distribution of X on T
- $g : T \to U$ - measurable function (transformation)
- $Y = g(X)$ - transformed random variable with values in U

Question: If we know P_X, what is the distribution P_Y of $Y = g(X)$?

Examples in ML:
- $X \sim \mathcal{N}(0, I)$ (simple), g is neural network, $Y = g(X)$ (complex model distribution)
- data transformation: X is data, g encodes to latent space



1. Push-forward is how distributions transform under functions
2. The pre-image $g^{-1}(C)$ is the set of points in T that map to C
3. Diagram shows: regular grid in T becomes warped in U, but probability is preserved
4. This is well-defined because g is measurable
5. In ML: g is neural network, P_X is simple distribution (e.g., Gaussian), P_Y is complex model distribution
6. We're working directly with distributions P_X and P_Y, not going back to abstract probability space

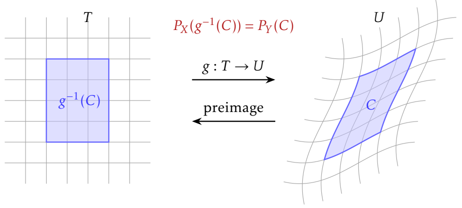
Push-forward measure

Answer: distribution of $Y = g(X)$ is the push-forward of P_X by g

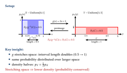
Definition: P_Y on U defined by

$$P_Y(C) = P_X(g^{-1}(C)) \quad \text{for } C \subseteq U$$

where $g^{-1}(C) = \{x \in T : g(x) \in C\}$ is the pre-image



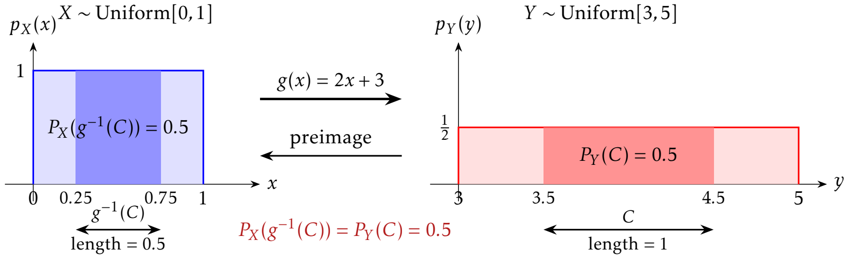
Intuition: probability of Y landing in C equals probability of X being in pre-image



1. This is a simple concrete example showing push-forward
2. The preimage calculation: solve $a \leq 2x + 3 \leq b$ for x
3. Uniform on $[0, 1]$: $P_X([a, b]) = b - a$ when $[a, b] \subseteq [0, 1]$
4. After transformation: still uniform, but on $[3, 5]$ instead of $[0, 1]$
5. The scaling factor 2 will become important when we look at densities (Jacobian)
6. This example shows: push-forward preserves the "shape" but changes location and scale

Example: uniform distribution and affine transformation

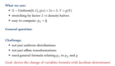
Setup: $X \sim \text{Uniform}[0,1]$, transformation $g(x) = 2x + 3$, find distribution of $Y = g(X)$



Key insight:

- g stretches space: interval length doubles ($0.5 \to 1$)
- same probability distributed over larger space
- density halves: $p_Y = \frac{1}{2} p_X$

Stretching space ⇒ lower density (probability conserved)



1. The uniform example was special - constant density, linear transformation
2. In ML: we use complex distributions (Gaussian, data distributions) and complex transformations (neural networks)
3. Need to understand how densities transform in general
4. This is the change of variables formula - fundamental for normalizing flows, diffusion models, etc.

From push-forward to change of variables

What we saw: uniform distribution + affine transformation

- $X \sim \text{Uniform}[0,1]$, $g(x) = 2x + 3$, $Y = g(X)$
- stretching by factor 2 $\Rightarrow$ density halves
- easy to compute: $p_Y = \frac{1}{2}$

General question:
Given arbitrary distribution p_X and transformation g, how do we compute p_Y?

Challenge:

- not just uniform distributions
- not just affine transformations
- need general formula relating p_Y to p_X and g

Goal: derive the change of variables formula with Jacobian determinant

1. This is the starting point - push-forward expressed with densities
2. Left side: probability in target space U (using p_Y)
3. Right side: probability in source space T (using p_X)
4. They're equal by push-forward definition
5. Next: use change of variables from calculus to transform the right integral

Recall:

Express using densities:

Key observation:

- left side: integrate p_Y over C in y-space
- right side: integrate p_X over $g^{-1}(C)$ in x-space
- same probability, different spaces

Idea:

Step 1: Start from push-forward

Recall: push-forward at measure level

$$P_Y(C) = P_X(g^{-1}(C))$$

Express using densities: (assuming densities exist)

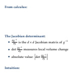
$$P_Y(C) = \int_C p_Y(y)\,dy = \int_{g^{-1}(C)} p_X(x)\,dx = P_X(g^{-1}(C))$$

Key observation:

- left side: integrate p_Y over C in y-space
- right side: integrate p_X over $g^{-1}(C)$ in x-space
- same probability, different spaces

Idea: change variables in right side from x to $y = g(x)$

1. This is the standard change of variables from multivariable calculus
2. The Jacobian matrix contains all partial derivatives
3. Its determinant tells us how volumes scale under the transformation
4. We need absolute value because we're integrating (measure must be positive)
5. For $d = 1$: this reduces to $|dx/dy|$ or $|1/g'(x)|$
6. We can also write in terms of g instead of g^{-1} (next slide)



Step 2: Change of variables in integrals

From calculus: change of variables formula

$$\int_{g^{-1}(C)} f(x)\,dx = \int_C f(g^{-1}(y)) \left|\det \frac{\partial g^{-1}}{\partial y}(y)\right| dy$$

Assuming $g : \mathbb{R}^d \to \mathbb{R}^d$ is invertible and differentiable

The Jacobian determinant:

- $\frac{\partial g^{-1}}{\partial y}$ is the $d \times d$ Jacobian matrix of g^{-1}
- $\det \frac{\partial g^{-1}}{\partial y}$ measures local volume change
- absolute value: $\left|\det \frac{\partial g^{-1}}{\partial y}\right|$

Intuition: $dx = \left|\det \frac{\partial g^{-1}}{\partial y}\right| dy$ (infinitesimal volume scaling)

1. This is the key step: combining push-forward with change of variables
2. We have two expressions for the same integral
3. Both integrate over C in y-space
4. The integrands must be equal
5. This gives us the formula for p_Y in terms of p_X

Step 3: Apply to our setting

Apply change of variables:

$$\int_{g^{-1}(C)} p_X(x)\,dx = \int_C p_X(g^{-1}(y)) \left|\det \frac{\partial g^{-1}}{\partial y}(y)\right| dy$$

But we also have:

$$\int_{g^{-1}(C)} p_X(x)\,dx = \int_C p_Y(y)\,dy$$

by push-forward

Therefore:

$$\int_C p_Y(y)\,dy = \int_C p_X(g^{-1}(y)) \left|\det \frac{\partial g^{-1}}{\partial y}(y)\right| dy$$

Since this holds for all C, the integrands must be equal!

1. This is the change of variables formula!
2. First form: uses Jacobian of g^{-1} (inverse transformation)
3. Second form: uses Jacobian of g (forward transformation) - usually more convenient
4. Inverse function theorem: $J_{g^{-1}}(y) = [J_g(g^{-1}(y))]^{-1}$
5. Determinants: $\det(A^{-1}) = 1/\det(A)$
6. The Jacobian determinant in denominator - this is key for normalizing flows

Step 4: Change of variables formula

Result:

$$p_Y(y) = p_X(g^{-1}(y)) \left|\det \frac{\partial g^{-1}}{\partial y}(y)\right|$$

Alternative form: using inverse function theorem

$$\det \frac{\partial g^{-1}}{\partial y}(y) = \frac{1}{\det \frac{\partial g}{\partial x}(g^{-1}(y))}$$

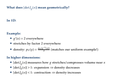
Therefore:

$$p_Y(y) = p_X(g^{-1}(y)) \left|\det \frac{\partial g}{\partial x}(g^{-1}(y))\right|^{-1}$$

Common notation: $J_g(x) = \frac{\partial g}{\partial x}(x)$ (Jacobian of g)

$$\boxed{p_Y(y) = \frac{p_X(g^{-1}(y))}{|\det J_g(g^{-1}(y))|}}$$

1. The Jacobian determinant has clear geometric meaning
2. It tells us how volumes change under the transformation
3. In 1D: just the derivative (slope)
4. Our uniform example: constant Jacobian = 2, density halves
5. In general: Jacobian varies with position x
6. This is why neural networks can create complex distributions - varying Jacobian
7. Key principle: stretching space → lower density, compressing space → higher density



Geometric interpretation of Jacobian

What does $|\det J_g(x)|$ mean geometrically?

Local volume scaling factor at point x

In 1D: $g : \mathbb{R} \to \mathbb{R}$

$$|\det J_g(x)| = |g'(x)| = \text{local stretching factor}$$

Example: $g(x) = 2x + 3$

- $g'(x) = 2$ everywhere
- stretches by factor 2 everywhere
- density: $p_Y(y) = \frac{p_X(g^{-1}(y))}{2}$ (matches our uniform example!)

In higher dimensions:

- $|\det J_g(x)|$ measures how g stretches/compresses volume near x
- $|\det J_g(x)| > 1$: expansion ⇒ density decreases
- $|\det J_g(x)| < 1$: contraction ⇒ density increases