What is going on?

It is not at all obvious why, in reverse-mode auto-differentiation, you get a whole vector of information, the entire gradient of the function, from a single piece, setting gz = 1.

Here is an answer. First, some terminology.

The derivative of $f : \mathbb{R}^n \to \mathbb{R}^m$ at a point p is a linear transformation $df_p : \mathbb{R}^n \to \mathbb{R}^m$ which is a linear approximation to f at p; its matrix represenation has entries $(df_p)_{ij} = \frac{\partial df_i}{\partial x_j}$.

If $g : \mathbb{R}^m \to \mathbb{R}^k$ is another function, $d(g \circ f)_p = dg_{f(p)} \circ df_p$. That is, the deriviative of the composition is the composition of the derivatives. This is the chain rule.

Every linear transformation $A : \mathbb{R}^n \to \mathbb{R}^m$ has a transpose $A^t : \mathbb{R}^m \to \mathbb{R}^n$ and $(AB)^t = B^t A^t$. This is also true of derivative matrices.

If $f : \mathbb{R}^n \to \mathbb{R}$, then $df_p : \mathbb{R}^n \to \mathbb{R}$ is a row-vector of partial derivatives of f, the gradient of f at p. So $df_p^t : \mathbb{R} \to \mathbb{R}^n$ is the gradient as a column vector. These are same data packaged differently, but the difference is important.

To say what a linear transformation $\mathbb{R}^n \to \mathbb{R}$ is, you have to say how it acts on each of the n basis vectors; to say what a linear transformation $\mathbb{R} \to \mathbb{R}^n$ is, you just to say how it acts on the one and only basis vector, 1. In both forward-mode and backward-mode, we are feeding in a basis vector to get out what we want, but in forward-mode we are feeding a basis vector from $\mathbb{R}^n$ to df to obtain a single component of df whereas in reverse-mode we are feeding the one and only basis vector of $\mathbb{R}$ to df^t and getting the whole matrix df^t.

In this reverse-mode ad example, we have $z(a, b) = a + b$, $a(x, y) = xy$, $b(x, y) = \cos(x)$. Writing $\alpha(x, y) = (xy, \cos(x))$, we can summarize this as $z(x, y) = z(\alpha(x, y))$. So $\alpha : \mathbb{R}^2 \to \mathbb{R}^2$ and $z : \mathbb{R}^2 \to \mathbb{R}$ and $z \circ \alpha : \mathbb{R}^2 \to \mathbb{R}$. To compute the gradient we want, we are computing $d\alpha$, dz, taking their transposes, and composing them. "Setting $gz = 1$" is equivalent to feeding the basis vector 1 to this linear transformation $\mathbb{R} \to \mathbb{R}^2$.

$$
\begin{aligned}
d(z \circ \alpha)^t &= d\alpha^t dz^t \\
&= \begin{bmatrix} \frac{\partial a}{\partial x} & \frac{\partial b}{\partial x} \\ \frac{\partial a}{\partial y} & \frac{\partial b}{\partial y} \end{bmatrix} \begin{bmatrix} \frac{\partial z}{\partial a} \\ \frac{\partial z}{\partial b} \end{bmatrix} \\
&= \begin{bmatrix} y & \cos(x) \\ x & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} y + \cos(x) \\ x \end{bmatrix}
\end{aligned}
$$

It feels backward because we are literally going backward: we are starting in the tangent space to the target and mapping the lone basis vector back and back again until we land in the tangent space of the domain.

Of Derivatives and Differentials

Consider $\mathbb{R}^2$ with coordinates x, y. The tangent space to $\mathbb{R}^2$ at p, $T_p\mathbb{R}^2$, is the vector space with basis $\partial/\partial x, \partial/\partial y$. So, an element of the tangent space is a linear combination $a\partial/\partial x + b\partial/\partial y$. A tangent vector eats a functions $f : \mathbb{R}^2 \to \mathbb{R}$ and produces a number $a\partial f/\partial x + b\partial f/\partial y$, the directional derivative of f in the direction (a, b).

$T_p\mathbb{R}^2$ is clearly isomorphic to $\mathbb{R}^2$ and we routinely conflate the two via this isomorphism. Remembering that they are not the same will help us understand what is going on here.

The *co*tangent space is the dual to the tangent space. "Dual" means "function on" so "the dual foo" is "functions on foo". It is the vector space of linear functions $T_p\mathbb{R}^2 \to \mathbb{R}$. The cotangent space at p, $T_p^\star\mathbb{R}^2$, is the vector space with basis dx, dy.

This, too, is clearly isomorphic to both $\mathbb{R}^2$ and to $T_p\mathbb{R}^2$ and we often conflate it with those others too. But let's keep our vector spaces in a row.

So a cotangent vector, or 1-form (or "differential"), is a linear combination $adx + bdy$. 1-forms eat tangent vectors via the rule $dx_i(\partial/\partial x_j) = \delta_{ij}$, meaning 1 if $i = j$, 0 otherwise.

We can make 1-forms out of functions!: $df = \partial f/\partial x dx + \partial f/\partial y dy$. Looks familiar, right?

Here's another cool thing: given a vector $v = a\partial/\partial x + b\partial/\partial y$, evaluating v on a function f is the same as evaluating df on v!

$$v(f) = a\partial f/\partial x + b\partial f \partial y$$

and

$$\begin{aligned} df(v) &= \partial f/\partial x dx(v) + \partial f/\partial y dy(v) \\ &= a\partial f/\partial x + b\partial f/\partial y \end{aligned}$$

since dx and dy are linear.

So 1-forms eat tangent vectors and tangent vectors eat functions and we can use d to turn functions into 1-forms in which case they can then eat tangent vectors. Neat.

This clearly all holds for $\mathbb{R}^n$ in general with coordinates $x_1, ...x_n$, tangent basis ∂x_i and cotangent basis dx_i. The consumption rules are the same.

This notation might seem weird, but everyone knows this stuff: if you put a dot product on your vector space, you can turn vectors into functions on vectors by sending a vector v to the function "dot product with v". Vectors are kets and 1-forms are bras in physics-speak. The only thing that is new is defining tangent and cotangent spaces as "spaces of differential operators" and really strictly *not* conflating vector space we know are isomorphic.

This is where it gets good.

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$ induces a linear transformation $df_p : T_p\mathbb{R}^n \to T_p\mathbb{R}^m$ whose matrix representaiton, as we know, has entries the partial derivatives of f. We have already seen this, we just weren't distinguishing between $\mathbb{R}^n$ and its tangent space back then.