Let $r_i = \underbrace{f_i(x)}_{\text{tuning curve}} + \underbrace{z_i}_{\text{noise model}}$

$\boxed{\Rightarrow \text{Given } r_i \text{, how much information do we have about } x\text{?}}$

Translates to: how much $\underline{\text{Fisher info}}$ is in the population?

We can think about this in terms of a $\underline{\text{local}}$ decoder w:

$$\hat{x} = x_0 + w^T(r - f(x_0))$$

We want this estimate to be

- unbiased:

$$\langle \hat{x} \rangle = x$$

$$\Leftrightarrow x_0 + w^T(\langle r \rangle - f(x_0)) = x$$

$$x_0 + w^T(f(x) - f(x_0))$$

$$\approx x_0 + w^T(\cancel{f(x_0)} + f'(x_0)(x - x_0) - \cancel{f(x_0)})$$

$$= x_0 + w' f(x_0)(x - x_0)$$

$$\Leftrightarrow w^T f'(x_0)(x - x_0) = x - x_0$$

$$\Leftrightarrow \underline{\underline{w^T f'(x_0) = 1}}$$

- minimum variance:

$$\mathrm{Var}\{\hat{x}\} = w^T \mathrm{Var}[r - f(x_0)] w$$

$$= w^T \Big[\big\langle (r - f(x_0))(r - f(x_0))^T \big\rangle$$

$$- \langle r - f(x_0) \rangle \langle r - f(x_0) \rangle^T \Big] w$$

$$= w^T \left[\langle r r^T \rangle - \langle r \rangle \langle r \rangle^T \right] w$$

$$= w^T \left[\left\langle (f(x) + \xi)(f(x) + \xi)^T \right\rangle - f(x) f(x)^T \right] w$$

$$= w^T \underbrace{\langle \xi \xi^T \rangle}\, w$$

$$= \underline{\underline{w^T R w}}$$

So, now find w that

minimizes $\mathrm{Var}[\hat{x}] = w^T R w$

subject to $w^T f'(x_0) = 1$

$$\Rightarrow \frac{\partial}{\partial w} \left[w^T R w + \lambda \left(w^T f'(x_0) - 1 \right) \right] = 0$$

$\iff \alpha R w = f'(x_0)$

$$\iff \underline{w = \alpha R^{-1} f'(x_0)}$$

$$\implies w^T f'(x_0) = \alpha f'(x_0)^T R^{-1} f'(x_0) = 1$$

$$\iff \alpha = \frac{1}{f'(x_0)^T R^{-1} f'(x_0)}$$

$$\implies \boxed{w = \frac{R^{-1} f'(x_0)}{f'(x_0)^T R^{-1} f'(x_0)}}$$

Giving us:

$$\mathrm{Var}[\hat{x}] = w^T R w = \frac{1}{f'(x_0)^T R^{-1} f'(x_0)}$$

We then define the Linear Fisher Information as

$$I = \frac{1}{\mathrm{Var}[\hat{x}]} = f'(x_0)^T R^{-1} f'(x_0)$$

~~In fact, this is dependent on the angle~~

Geometrically, the decoder is computing the gradient at x_0 and adjusting it by the directions of variance in the distribution over r (see slides).

Intuitively, then, the amount of information in the code should depend on those directions of highest variance, ~~e.g. variance along $f'(x_0)$ shouldn't be averaged away so should decrease this!~~ which come in at R^{-1}.

Also picture it in your head: variance perpendicular to $f'(x_0)$ shouldn't hurt $\hat{x}$.

~~...~~ Indeed, it turns out that we can rewrite the linear Fisher Info in terms of angles b/w directions of maximal variance (e-vectors of R) and $f'(x_0)$:

$$\Rightarrow I = f'(x_0)^T f'(x_0) \times \left(\frac{\cos^2\theta_1}{\sigma_1^2} + \frac{\cos^2\theta_2}{\sigma_2^2} \right)$$

Pf. Recall that any symmetric matrix can be rewritten as:

$$A v_k = \lambda_k v_k, \quad v_k^T v_j = \delta_{jk} \rightarrow \left(\sum_k \lambda_k v_k v_k^T \right) v_j = \lambda_j v_j$$

$$\Rightarrow A = \sum_k \lambda_k v_k v_k^T$$

where λ_k, v_k are the e-values, vectors of A

~~Thus, we can write~~

~~$R = \sum_k \sigma_k^2 v_k v_k^T$~~

~~$\Rightarrow R^{-1} = \sum_k \frac{v_k v_k^T}{\sigma_k^2}$~~

Then,

$$\left(\sum_k \lambda_k^{-1} v_k v_k^T \right) A$$

$$= \sum_{j,k} \frac{1}{\cancel{\lambda_k}} v_k v_k^T \cancel{\lambda_j} v_j v_j^T$$

$$= \sum_k v_k v_k^T = I$$

(since $\sum_k v_k v_k^T = \begin{bmatrix} v_1 \cdots v_k \end{bmatrix} \begin{bmatrix} -v_1- \\ \vdots \\ -v_k- \end{bmatrix} = QQ^T = I$ since Q is orthogonal)

$$\Rightarrow A^{-1} = \sum_k \frac{1}{\lambda_k} v_k v_k^T$$

We thus have

$$R^{-1} = \sum_k \frac{1}{\sigma_k^2} v_k v_k^T$$

$$\Rightarrow I = \sum_k \frac{\left(f'(x_0)^T v_k\right)^2}{\sigma_k^2}$$

$$= \sum_k \frac{|f'(x_0)|^2 \underbrace{|v_k|^2}_{=1} \cos^2\theta_k}{\sigma_k^2}$$

angle b/w $f'(x_0)$ and v_k (k^{th} e-vector of R)

$$= \left(\sum_k \frac{\cos^2\theta_k}{\sigma_k^2}\right) \underbrace{f'(x_0)^T f'(x_0)}_{= |f'(x_0)|^2}$$

We now note some properties:

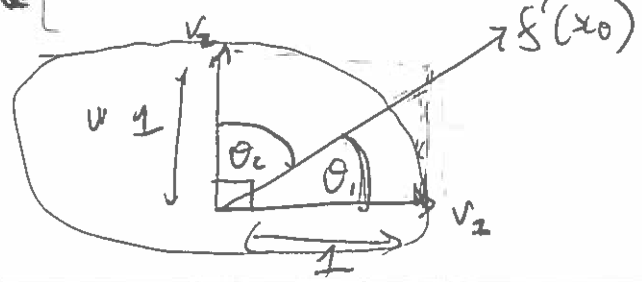
- $f'(x_0)^T f'(x_0) \propto N \sim$ big
- $\sum_k \cos^2\theta_k = 1$

↳ think about 2-D case:

since $v_1 \perp v_2$,

$$\cos\theta_2 = \cos\left(\frac{\pi}{2} - \theta_1\right) = \sin\theta_1$$

$$\Rightarrow \cos^2\theta_1 + \cos^2\theta_2 = \cos^2\theta_1 + \sin^2\theta_1 = \underline{\underline{1}}$$



$\cdot \sum_k \sigma_k^2 \propto N$

assuming all σ_k^2 are of similar size (i.e. spherical noise)

$$\Rightarrow \sum_k \frac{\cos^2\theta_k}{\sigma_k^2} = \text{convex combination of } \sigma_k^2\text{'s} \sim \mathcal{O}(1)$$

(approximately independent of N)

$\Rightarrow I \propto N$

assuming σ_k^2 are independent (spherical noise)

or weakly correlated such that σ_k^2's still similar size

What about when noise is not independent?

$\rightarrow$ some of the σ_k^2 then become proportional to N

Then, if $f'(x_0)$ is parallel to v_k with $\sigma_k^2 \propto N$, $\cos^2\theta_j = \delta_{jk}$ and thus

$I \propto \mathcal{O}(1)$ will be constant w.r.t. N (noise can't be averaged away)

But does this ever happen!

It seems unlikely that the v_k's with largest covariance (σ_k^2) would line up exactly with $f'(x_0)$ ...

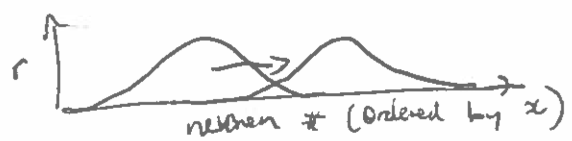
⇛ Shamir & Sompolinsky 2006

Ecker, Berens, Tolias & Bethge 2011

<u>NO</u> (more specifically, whenever tuning curves are not exactly the same (e.g. amplitude, width, etc.) across neurons, <u>which is the case in the brain!</u>)

So, under realistic conditions (i.e. heterogeneous tuning curves, lots of neurons) the directions of highest covariance do not exactly line up with the derivative of the tuning curve and thus <u>noise can be averaged away with large N.</u>

In other words, the only correlations that make I independent of N are those that push the responses of all the neurons in the same <u>direction</u> : we call these $\boxed{\text{differential correlations}}$



$\cancel{(mr_1 + mr_2)\left((1+c)^2 m^2 n^2 + 8\right)}$

$\cancel{16(1+c)\, m n r_1 r_2}$

Correlations Story

Shadlen et al 1994:

correlations limit SNR
in large N limit
→ homogeneous tuning curves

Dayan & Abbott '98:

correlations are good
for homogeneous tuning with
equally spaced preferences
→ except when correlations
a function of pref. stim.

Shamir & Sompolinsky 2006
Ecker et al. 2011:

For heterogeneous tuning curves,
correlations don't matter

Moreno-Bote et al 2014:

The only correlations that matters are differential correlations

→ Also, information cannot be infinite!

(finite input info, suboptimal computation)

→ Therefore, there must be differential correlations

(But hard to see, generally masked by other, 1 noise correlations)