## 1 Statistics

### 1.1 Definitions

- **Data:** \( \{d_i\} = \vec{d} = d_0, d_1, ..., d_{M-1} \in \mathbb{R}^M \)
- **Parameters:** \( \{\theta\} = \vec{\theta} = \theta_0, \theta_1, ..., \theta_{N-1} \in \mathbb{R}^N \)
- **Nonlinear Model:** \( \{y_i(\vec{\theta}) = \vec{C}_i(\tau_i, \theta_\alpha)\} : \mathbb{R}^N \rightarrow \mathbb{R}^M \)

Curved surface in data space = Model manifold, with coordinates \( \vec{\theta} \)

\( \tau_i \): Value of continuous experimental/control variables (ex: temperature, time, etc.)

Later, smooth variations in \( \tau \) ⇒ hyperribbon

Components of \( \vec{y} \) are discrete measured variables (ex: chemical species)

- **Residuals:** \( \{r_i\} \)

\[
 r_1 = \frac{y_i(\vec{\theta}) - d_i}{\sigma_i} 
\]

Best fit \( \leftrightarrow \) Finding \( \vec{\theta} \) which minimizes the residual vector

- **Likelihood function:** \( P(\vec{d} | \vec{\theta}) \)

Probability density that data \( \vec{d} \) would have been generated by model with parameters \( \vec{\theta} \) and errors \( \vec{\sigma} \)

\( \text{ex: } \) "Normal" error distribution

\[
 P(\vec{d} | \vec{\theta}) \propto e^{-\sum r_i^2 / 2} = e^{-\sum \frac{(y_i(\vec{\theta}) - d_i)^2}{2\sigma_i^2}} 
\]

- **Cost Function:** \( C(\vec{\theta}) = -\log(P(\vec{d} | \vec{\theta})) = \chi^2 / 2 = -l(\vec{d} | \vec{\theta}) \)

\[
 \chi^2 = 2 \sum r_i^2 
\]

\( \Rightarrow C(\vec{\theta}) = \chi^2 / 2 = \sum r_i^2 = \sum \frac{(y_i(\vec{\theta}) - d_i)^2}{2\sigma_i^2} \)

- **Jacobian:** \( J_{\alpha\beta} = \frac{\partial y_i}{\partial \theta_\beta} \)

- **Approximate Hessian:** \( H_{\alpha\beta} = (J^T J)_{\alpha\beta} = g_{\alpha\beta} \)

\[
 H_{\alpha\beta} = \frac{\partial^2 C}{\partial \theta_\alpha \partial \theta_\beta} = \frac{1}{2} \frac{\partial}{\partial \theta_\alpha} \left( \frac{\partial}{\partial \theta_\beta} \sum_i \frac{(y_i(\vec{\theta}) - d_i)^2}{2\sigma_i^2} \right) = \frac{\partial}{\partial \theta_\alpha} \left( \sum_i \frac{(y_i(\vec{\theta}) - d_i) \partial y_i(\vec{\theta})}{\sigma_i^2} \right) = \sum_i \frac{1}{\sigma_i^2} \left( \frac{\partial y_i(\vec{\theta})}{\partial \theta_\alpha} \frac{\partial y_i(\vec{\theta})}{\partial \theta_\beta} + (y_i(\vec{\theta}) - d_i) \frac{\partial^2 y_i(\vec{\theta})}{\partial \theta_\alpha \partial \theta_\beta} \right) 
\]
\[ \approx \sum_i \frac{1}{\sigma_i^2} \left( \frac{\partial y_i(\vec{\theta})}{\partial \theta_\alpha} \frac{\partial y_i(\vec{\theta})}{\partial \theta_\beta} \right) = (J^T J)_{\alpha \beta} \]  

(3)

We can therefore express the likelihood function as:

\[ P(\vec{d} | \vec{\theta}) \propto e^{-\frac{1}{2} \vec{d}^T H(\vec{\theta}) \vec{d}} \]  

(4)

- **Covariance Matrix:** Inverse of the Hessian at best fit
  
  Let \( \vec{\theta}^* \) be \( \vec{\theta} \) at best fit.

  \( (H^{-1})_{\alpha \beta} = \langle (\theta_\alpha - \theta^*_\alpha)(\theta_\beta - \theta^*_\beta) \rangle \)  

(5)

- **Fisher Information Matrix:** \( F_{\alpha \beta} \)

  \[ F_{\alpha \beta} = -\frac{\partial^2}{\partial \theta_\alpha \partial \theta_\beta} \log P(\vec{d} | \vec{\theta}) \bigg|_{\vec{\theta}} \]  

(6)

- **Metric Tensor:** \( g_{\alpha \beta} \)

  \[ dy_i = \sum_\alpha \frac{\partial y_i}{\partial \theta_\alpha} d\theta_\alpha \Rightarrow dy_i^2 = \sum_{\alpha \beta} \left( \frac{\partial y_i}{\partial \theta_\alpha} d\theta_\alpha \right) \left( \frac{\partial y_i}{\partial \theta_\beta} d\theta_\beta \right) = \sum_{\alpha \beta} g_{\alpha \beta} d\theta_\alpha d\theta_\beta \]  

(7)

1.2 Bayes & Priors

Baye's Theorem:

\[ P(\vec{\theta} | \vec{d}) = \frac{\text{Prior} \ P(\vec{d} | \vec{\theta}) P(\vec{\theta})}{\text{Boring} \ P(\vec{d})} \]  

(8)

Proof)

\[ \text{prob of } \vec{\theta} \& \vec{d} = \text{prob of } \vec{d} \text{ prob of } \vec{\theta} \]

\[ P(\vec{\theta}, \vec{d}) = P(\vec{d}) P(\vec{\theta} | \vec{d}) = P(\vec{\theta}) P(\vec{d} | \vec{\theta}) \]  

(9)

A prior \( P(\vec{\theta}) \):

- Previous knowledge
- "Flat": essentially equal likelihood to all parameters
• "Non-informative"/objective: if it has a minimal impact on $P(\vec{\theta} | \vec{d})$ (if it is flat)

Jeffrey’s Prior: Weights $\theta$-volume by data volume = $|J|$. 

$$|J| = \det(J) = \sqrt{\det(J^T J)} = \sqrt{|g_{\mu\nu}|}$$ (10)

$$\rho(\vec{\theta}) = \sqrt{|g_{\mu\nu}|} = \prod_i \sqrt{\lambda_i}$$ (11)

Where $\lambda_i$ are the eigenvalues of $J$

Monte Carlo Random Sample $\rho(\theta)$

$$\theta \rightarrow \theta + \Delta \theta$$ (12)

With rate $\Gamma_{\theta \rightarrow \theta'}$

Detailed Balance: $\rho(\theta) \Gamma_{\theta \rightarrow \theta'} = \rho(\theta') \Gamma_{\theta' \rightarrow \theta}$

Try $\Gamma_{\theta \rightarrow \theta'} = \Gamma_{\Delta \theta}$ (random step)

If $\rho(\theta') > \rho(\theta)$ accept always

If $\rho(\theta') < \rho(\theta)$ accept with probability $\frac{\rho(\theta')}{\rho(\theta)} = \sqrt{\frac{|g_{\mu\nu}(\theta')|}{|g_{\mu\nu}(\theta)|}}$

2 Geodesics

Want to find the shortest path $y(\theta(t))$ such that $y(\theta(a)) = y_0$ and $y(\theta(b)) = y_1$

The length of a path is given by $^1$:

$$L = \int_a^b \sqrt{\sum_i \left(\frac{dy_i}{dt}\right)^2} dt$$

$$= \int_a^b \sqrt{\sum_i \left(\frac{\partial y_i}{\partial \theta^\alpha} \dot{\theta}^\alpha \right) \left( \frac{\partial y_i}{\partial \theta^\beta} \dot{\theta}^\beta \right)} dt$$

$$= \int_a^b \sqrt{\sum_i J_{\alpha i} \dot{\theta}^\alpha \dot{\theta}^\beta dt}$$

$$= \int_a^b \sqrt{g_{\alpha \beta} \dot{\theta}^\alpha \dot{\theta}^\beta} dt$$ (13)

It is difficult to work with the square root, however we can use the Cauchy-Schwartz inequality, which, for integrals of functions, states that $^2$

$$\left| \int f(t)g(t) dt \right|^2 \leq \int |f(t)|^2 dt \int |g(t)|^2 dt$$ (14)

$^1$For simplicity, Einstein summation notation will be used

$^2$Note: $(x \cdot y)^2 = x^2 \cdot y^2$ if $x \propto y \iff \sqrt{g_{\alpha \beta} \dot{\theta}^\alpha \dot{\theta}^\beta} = \sqrt{\sum \left(\frac{dy_i}{dt}\right)^2} \iff$ constant speed
If we let $f(t) = \sqrt{g_{\alpha\beta} \dot{\theta}^\alpha \dot{\theta}^\beta}$ and $g(t) = 1$, then we obtain:

$$\left| \int_a^b f(t) \, dt \right|^2 \leq \int_a^b dt \int_a^b f(t)^2 \, dt = (b-a) \int_a^b f(t)^2 \, dt$$  \hspace{1cm} (15)

And so, to minimize $L$, it is sufficient to minimize

$$E = \int_a^b g_{\alpha\beta} \dot{\theta}^\alpha \dot{\theta}^\beta \, dt$$  \hspace{1cm} (16)

This is accomplished through calculus of variation. We find the condition which sets $\delta E = 0$.

$$\delta E = E(\theta + \delta \theta) - E(\theta)$$

$$= \int_a^b g_{\alpha\beta} (\theta + \delta \theta)(\dot{\theta}^\alpha + \delta \dot{\theta}^\alpha)(\dot{\theta}^\beta + \delta \dot{\theta}^\beta) \, dt - \int_a^b g_{\alpha\beta} \dot{\theta}^\alpha \dot{\theta}^\beta \, dt$$

$$= \int_a^b \left( \partial_\gamma g_{\alpha\beta} \delta \theta^\gamma \dot{\theta}^\alpha \dot{\theta}^\beta + g_{\alpha\beta} \dot{\theta}^\alpha \delta \dot{\theta}^\beta + g_{\alpha\beta} \dot{\theta}^\beta \delta \dot{\theta}^\alpha + O(\delta \theta^2) \right) \, dt$$  \hspace{1cm} (17)

Integrating the second (or third) terms by parts with respect to $t$, we obtain:

$$\int_a^b g_{\alpha\beta} \dot{\theta}^\alpha \delta \dot{\theta}^\beta \, dt = g_{\alpha\beta} \dot{\theta}^\beta \bigg|_a^b - \int_a^b \frac{d(g_{\alpha\beta} \dot{\theta}^\gamma)}{dt} \delta \dot{\theta}^\beta \, dt$$

$$= -\int_a^b \left( \partial_\gamma g_{\alpha\beta} \dot{\theta}^\gamma \dot{\theta}^\alpha + g_{\alpha\beta} \ddot{\theta}^\alpha \right) \delta \dot{\theta}^\beta \, dt$$  \hspace{1cm} (18)

And so, we get that

$$\delta E = \int_a^b \left( \partial_\gamma g_{\alpha\beta} \delta \theta^\gamma \dot{\theta}^\alpha \dot{\theta}^\beta - \partial_\gamma g_{\alpha\beta} \dot{\theta}^\gamma \delta \theta^\alpha \ddot{\theta}^\beta - g_{\alpha\beta} \delta \theta^\alpha \ddot{\theta}^\beta - \partial_\gamma g_{\alpha\beta} \ddot{\theta}^\gamma \delta \theta^\alpha - g_{\alpha\beta} \ddot{\theta}^\beta \delta \theta^\alpha \right) \, dt$$  \hspace{1cm} (19)

By rearranging indices and grouping terms, we obtain:

$$\delta E = \int_a^b \left[ (\partial_\gamma g_{\alpha\beta} - \partial_\beta g_{\alpha\gamma} + \partial_\alpha g_{\gamma\beta}) \dot{\theta}^\alpha \dot{\theta}^\beta + 2g_{\alpha\beta} \ddot{\theta}^\gamma \right] \delta \theta^\gamma \, dt$$  \hspace{1cm} (20)

this term must = 0
And so, we obtain the following condition:

\[ g_{\alpha\beta} \ddot{\theta}^\gamma + \frac{1}{2} (\partial_\alpha g_{\gamma\beta} + \partial_\beta g_{\alpha\gamma} - \partial_\gamma g_{\alpha\beta}) \dot{\theta}^\alpha \dot{\theta}^\beta = 0 \]  

(21)

Christoffel symbols: \( \Gamma_{\gamma\alpha\beta} \)

To raise an index by \( g^{\mu\nu} = (g^{-1})_{\mu\nu} \), and so we can re-write the Christoffel symbols in the form \( \Gamma^\gamma_{\alpha\beta} = g^{\gamma\delta} \Gamma_{\delta\alpha\beta} \) and obtain the traditional geodesic equation:

\[ \ddot{\theta}^\gamma = \Gamma^\gamma_{\alpha\beta} \dot{\theta}^\alpha \dot{\theta}^\beta \]  

(22)

Given an initial \( \theta^\alpha \) and \( \frac{\partial \theta^\alpha}{\partial t} = v^\alpha \), we obtain a straight path. We can use this to find the width of a model manifold (see fig. 1). Let:

- \( \overrightarrow{\theta} (a) \) be the best fit
- \( \frac{\partial \overrightarrow{\theta}}{\partial t} = \overrightarrow{v} = \text{nth Eigen parameter} \)
- \( \overrightarrow{\theta} (b) \) be the edge of the manifold
- Then, let \( \frac{\partial \overrightarrow{\theta}}{\partial t} = -\overrightarrow{v} \), and find the other edge

Trick: Singular value decomposition on the Jacobian, an \( M \times N \) matrix,

\[ J_{i\alpha} = U_{i\beta} \Sigma_{\beta\gamma} V_{\gamma\alpha}^T \]  

(23)

Where \( U \) and \( V^T \) both have orthonormal columns.

The projection matrix, \( P \), which projects tangent vectors, can be written as:

\[ P = U U^T \]  

(24)

\( \overrightarrow{v}_{\alpha\beta} = \hat{v}_{(\beta)} \alpha \) forms an eigenbasis of \( \theta \) space.

\[ J^T J = V \Sigma U^T U \Sigma V^T = V \Sigma^2 V^T \]  

(25)

5
3 Sloppy Ising

3.1 Diffusion

\[ P_{t+1}(j) = \sum_{i=-N}^{N} \theta_{j-i} P_t(i) \]  

(26)

Start with \( P_0(i) - \delta_{i0} \) (all at zero).

"One step" data \( d_i = P_t(j) = \theta_j \)

\[ \sum \theta_i = R = 1 \Rightarrow \text{probability is conserved.} \]

\[ \frac{\partial d_i}{\partial \theta_\alpha} = \delta_{i\alpha} = 1 \]  

(27)

\( \Rightarrow \) NOT sloppy

3.2 Continuum Hydrodynamic Theory

"Many Steps" data \( \Rightarrow \) Diffusion equation

\[ \frac{\partial \rho}{\partial t} = D \frac{\partial^2 \rho}{\partial x^2} \]

\[ \rho_0(x) = \delta(x) \]

\[ \rho_t(x) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{x^2}{4Dt}} \]  

(28)

More generally,

\[ \frac{\partial \rho}{\partial t} = -v \frac{\partial \rho}{\partial x} + D \frac{\partial^2 \rho}{\partial x^2} \]  

(29)

\( V = \sum \mu \theta^\mu \) average out

\( D = \sum \mu^2 \theta^\mu u - v^2 \) = diffusion

3.3 Sloppy Equations

At large \( t \) eigenvalues

\[ \lambda_n \sim t^2 \left( \frac{Dt}{N^2} \right)^{-n-1/2} \]

\[ n = 0 \rightarrow R \]

\[ n = 1 \rightarrow V \]

\[ n = 2 \rightarrow D \]  

(30)

Sloppy spectrum: Only a few microscopic combinations matter
4 Additional Notes

Data points: \( \{x_1, x_2, ..., x_N\} \)

Probability distributions \( P_1 \) and \( P_2 \)

Probability of generating \( \mathcal{D} = \prod_i P_1(x_i) = e^{\sum_i \log P_1(x_i)} \)

For Large \( N \), we have:

\[
\prod_x P_1(x_i)^{N P_1(x)} = e^{N \sum_x P_1(x) \log P_1(x)} = e^{-NS}
\] (31)

Where \( S_1 \) is the entropy (in nats).

How likely is \( P_2 \) to generate a typical ensemble generated by \( P_1 \)? This is given by the following equation:

\[
\prod_x P_2(x)^{N P_1(x)} = e^{N \sum_x P_1(x) \log P_2(x)}
\] (32)

How much MORE likely is a typical ensemble from \( P_1 \) came from \( P_1 \), rather than \( P_2 \)? This is given by the following equation:

\[
\prod_x \left( \frac{P_1(x)}{P_2(x)} \right)^2 = e^{-N D_{KL}(P_1||P_2)}
\] (33)

Where \( D_{KL} \) is the Kullback-Liebler Divergence. It is a statistical measure of how distinguishable \( P_1 \) is from \( P_2 \) from its data. However, it is NOT a proper distance, since

\[
D_{KL}(P_1||P_2) \neq D_{KL}(P_2||P_1)
\] (34)

However, for ”close-by” models, \( D_{KL} \) does become symmetric

\[
D_{KL}(P_\theta, P_{\theta+\Delta \theta}) = g_{\mu \nu} \Delta \theta^\mu \Delta \theta^\nu + O(\Delta \theta^3)
\] (35)

Where \( g_{\mu \nu} \) is the Fisher information matrix (FIM), given by:

\[
g_{\mu \nu}(P_\theta) = -\sum_x P_\theta(x) \frac{\partial^2}{\partial \theta^\mu \partial \theta^\nu} \log P_\theta(x)
\] (36)

Use the FIM as a metric on parameter space, defining a Riemannian manifold, where each point specifies a probability distribution. \( g_{\mu \nu} \) is a metric, since:

- It is symmetric (derivatives commute)
- It is positive semi-definite (no model fits any model better than itself)
- It undergoes the correct transformations under reparametrization

Distance on the manifold is (locally) a measure of how distinguishable 2 models are from their data, in units of dimensionless standard deviations. This is an important difference between informational geometry and general relativity.