a relative entropy cost function and a Dirichlet-β prior

$$P_0(\mathbf{p})\mathrm{d}\mathbf{p} \propto \prod_k p_k^{\beta-1}\mathrm{d}p_k. \tag{5}$$

Common examples of Dirichlet priors include the "flat" Lebesgue measure ($\beta = 1$), and Jeffreys' prior ($\beta = \frac{1}{2}$). Given any prior, we can minimize expected relative entropy by: (1) updating the prior to a posterior via Bayes' Rule; and (2) reporting its mean value. For the Dirichlet-β prior, this gives the "add β" rule.

The "add β" rule is *not* intrinsically Bayesian, however. A näive estimator following Eq. 2 can simulate it by adding β dummy observations of each letter k. This yields new frequencies $\{n_k + \beta\}$ and a total of $N + K\beta$ observations. To generalize to non-integer β, we observe that the likelihood function is $\mathcal{L}(\mathbf{p}) = \Pr(\{n_k\}|\mathbf{p}) = \prod_k p_k^{n_k}$, and adding β dummy observations of each letter yields a *hedged* likelihood function

$$\mathcal{L}'(\mathbf{p}) = \prod_k p_k^{n_k+\beta} = \left(\prod_k n_k^{\beta}\right)\mathcal{L}(\mathbf{p}), \tag{6}$$

whose maximum value is achieved by Eq. 4. When β is not an integer, the hedged likelihood (Eq. 6) remains well-defined, and the "add β" rule still maximizes it.

Quantum Hedging: The quantum analogue of a distribution $\mathbf{p}$ is a $d \times d$ density matrix ρ. It cannot be observed directly; observing a sample of ρ requires choosing a particular measurement $\mathcal{M}$. Experimentalists often divide the samples into groups and measure $\mathcal{M}_j$ on the N_j samples in group j, but $\mathcal{L}(\rho)$ depends only on observed events, *not* the unobserved alternatives, so we may pretend that all N samples were measured by $\mathcal{M} = \bigcup_j w_j \mathcal{M}_j$, where $w_j = \frac{N_j}{N}$. $\mathcal{M}$ corresponds to a *POVM*, a set of positive operators $\{E_i\}$ summing to $\mathbb{1}$, which determine the probability of outcome "i" as

$$\Pr(i) = \mathrm{Tr}[\rho E_i]. \tag{7}$$

The frequencies $\{n_i\}$ thus provide information about ρ. Interpreting this information is the central problem of quantum state estimation.

The oldest and simplest procedure, linear inversion tomography [11], is based on Eq. 2. Inverting Born's Rule (Eq. 7) yields an estimate $\hat{\rho}_{\mathrm{tomo}}$ satisfying

$$\mathrm{Tr}[\hat{\rho}_{\mathrm{tomo}} E_i] = \frac{n_i}{N} \text{ for } i = 1 \ldots m. \tag{8}$$

If these equations are overcomplete, $\hat{\rho}_{\mathrm{tomo}}$ is chosen by least-squares fitting. Frequently, some of $\hat{\rho}_{\mathrm{tomo}}$'s eigenvalues are negative – a serious problem, for they represent probabilities. This occurs because linear inversion is blind to the shape of the space of quantum states (which assign probabilities to *all* measurements). It tries to fit data from a *single* POVM $\mathcal{M}$, and happily assigns negative probabilities for measurements that weren't performed.

The usual fix for this problem is MLE [2]. A likelihood function is derived from the data,

$$\mathcal{L}(\rho) = \Pr(\{n_i\}|\rho) = \prod_i \mathrm{Tr}[\rho E_i]^{n_i}, \tag{9}$$

and we assign the $\hat{\rho}$ that maximizes it. Maximization over *all* trace-1 Hermitian matrices yields $\hat{\rho}_{\mathrm{tomo}}$ (just as in the classical case), but restricting to $\rho \geq 0$ yields a non-negative $\hat{\rho}_{\mathrm{MLE}}$.

However, $\hat{\rho}_{\mathrm{MLE}}$ can still assign zero probabilities – just like its classical counterpart (Eq. 2). If $\hat{\rho}_{\mathrm{tomo}}$ is not strictly positive, $\hat{\rho}_{\mathrm{MLE}}$ will have at least one zero eigenvalue [6], so this is rather common. Moreover, the zero probabilities in $\hat{\rho}_{\mathrm{MLE}}$ are less justified than those in $\mathbf{p}_{\mathrm{MLE}}$, because they generally correspond to a measurement outcome $|\psi\rangle\langle\psi|$ that is not an element of the measured POVM, and could never have appeared. In contrast, Eq. 2 assigns $p_k = 0$ only when "k" has been given N chances to appear and (so far) has not. So although $\hat{\rho}_{\mathrm{MLE}}$ may be the right estimator for *some* task, its zero eigenvalues represent a level of confidence that is implausible and (for predictive tasks like gambling and compression) catastrophic. *Prediction demands a hedged estimator.*

Bayesian mean estimators are hedged, and with suitable priors they have extremely good predictive behavior [6]. However, for quantum estimation there are no closed-form solutions, and numerical integration is hard. This is unfortunate, because Bayes estimators for classical probabilities work very well. They yield "add β" rules when applied to Dirichlet-β priors, and Dirichlet priors are well motivated. Jeffreys' prior ($\beta = \frac{1}{2}$) yields asymptotically minimax-optimal estimators for data compression [12], Krichevskiy showed that "add $0.50922\ldots$" outperforms all other rules for predicting the next letter [13], and Braess et al [14] pointed out that $\beta \approx 1$ generally works well because large-N behavior depends only weakly on β.

This suggests adapting "add β" to quantum state estimation (independent of Bayesian arguments). However, obvious methods like adding dummy counts don't work. Suppose we estimate a qubit source by measuring σ_x, σ_y, and σ_z ten times each, and – by unlikely chance – all the outcomes are $+1$. $\hat{\rho}_{\mathrm{tomo}}$ lies well outside the Bloch sphere, and $\hat{\rho}_{\mathrm{MLE}}$ is the projector onto its largest eigenvector. Now, if we add $\beta = 1$ dummy counts, $\hat{\rho}_{\mathrm{tomo}}$ is still outside the Bloch sphere, and $\hat{\rho}'_{\mathrm{MLE}}$ is unchanged!

The underlying problem is that MLE tries to fit the observed data, with no consideration of unobserved measurements – but the resulting quantum state makes predictions about those unobserved measurements. Adding dummy data works in the classical case because there are only K different events that can be observed *or* predicted, so by adding a dummy observation of each one, we rule out the possibility of assigning $p_k = 0$ to any event. A quantum state assigns probabilities to infinitely