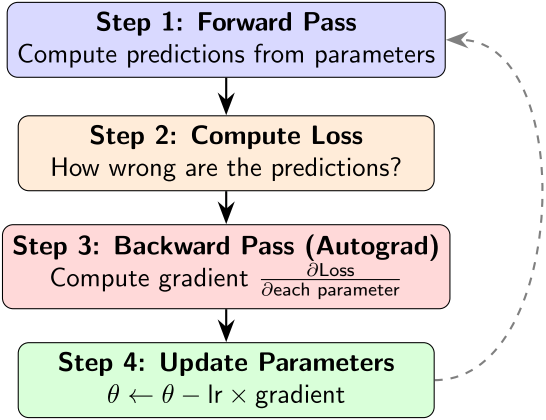
Centered vs. Non-Centered MAP: Same Model, Different Optimization

How ML Training Works (4 Steps, Every Epoch)



Step 1: Forward Pass
Compute predictions from parameters

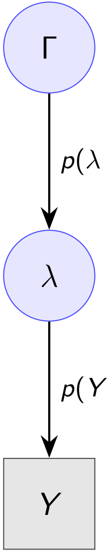
Step 2: Compute Loss
How wrong are the predictions?

Step 3: Backward Pass (Autograd)
Compute gradient $\frac{\partial \text{Loss}}{\partial \text{each parameter}}$

Step 4: Update Parameters
$\theta \leftarrow \theta - \text{lr} \times \text{gradient}$

Key rule: In Step 3, a parameter only gets a gradient from the loss if it was **used in Step 1** to compute the prediction. If a parameter isn't in the forward pass, autograd can't trace back to it.

The Generative Model (Same for Both)



The joint distribution (Giovanni's formulation):

$$\lambda - G\Gamma \sim \mathcal{N}(0, K_\theta) \iff p(\lambda \mid \Gamma) = \mathcal{N}(G\Gamma, K_\theta)$$

The Prior on Γ is Flat

Factorize the hierarchy:

$$p(Y, \lambda, \phi, \Gamma, \psi) = \underbrace{p(Y \mid \lambda, \phi, \psi)}_{\text{likelihood}} \cdot \underbrace{p(\lambda \mid \Gamma)}_{\text{GP on } \delta = \lambda - G\Gamma} \cdot \underbrace{p(\Gamma)}_{\propto 1}$$

- $p(\Gamma) \propto 1$ — flat. No prior on Γ.
- $p(\lambda \mid \Gamma) = \mathcal{N}(\lambda;\ G\Gamma,\ K_\theta)$ — this is the GP, and it's on δ, not on Γ.
- The GP penalty in the MAP loss is $W \cdot \delta^\top K_\theta^{-1} \delta$. It penalizes δ, not Γ.

Key point: The shrinkage of Γ we observe in practice is **not** from a prior on Γ. It comes from optimization dynamics: during gradient descent, Γ and δ compete to explain λ, and the GP penalty on δ shapes the landscape that Γ optimizes over.

The **centered parameterization starves Γ more** because the GP penalty has a direct gradient on Γ.

Centered Model: The Forward Pass

What the computer does each epoch:

1. Read λ from memory (it's a free `nn.Parameter`)

2. $\theta = \text{softmax}(\lambda)$ (mixing proportions)

3. $\pi = \theta \cdot \phi$ (disease probabilities)

4. $\text{Loss} = \text{NLL}(Y, \pi) + W \cdot (\lambda - G\Gamma)^\top K_\theta^{-1}(\lambda - G\Gamma)$

Notice: Γ appears **only in the prior term** (Step 4), never in the forward computation (Steps 1–3).

λ is read from memory $\rightarrow \theta \rightarrow \pi \rightarrow$ NLL.
The chain $Y \rightarrow \pi \rightarrow \theta \rightarrow \lambda$ does not pass through Γ.

Non-Centered Model: The Forward Pass

What the computer does each epoch:

1. Read δ and Γ from memory (both `nn.Parameters`)

2. $\lambda = G \cdot \Gamma + \delta$ (**Γ enters the forward pass!**)

3. $\theta = \mathrm{softmax}(\lambda)$ (mixing proportions)

4. $\pi = \theta \cdot \phi$ (disease probabilities)

5. $\mathrm{Loss} = \mathrm{NLL}(Y, \pi) + W \cdot \delta^\top K_\theta^{-1} \delta$

Now: Γ is **in the forward computation** (Step 2).
The chain is: $\Gamma \to \lambda \to \theta \to \pi \to \mathrm{NLL}$.

Autograd traces back through this chain, so Γ gets a gradient from the data.

The Backward Pass: Gradient Comparison

Centered model:

$$\frac{\partial \mathcal{L}}{\partial \Gamma_k} = \underbrace{\overset{=\mathbf{0}}{\frac{\partial \mathrm{NLL}}{\partial \Gamma_k}}}_{\text{Not in forward pass!}} + \underbrace{W \cdot (-2) \sum_i G_i^\top K_\theta^{-1} \delta_{ik}}_{\text{prior only } (W\text{=1e-4, weak})}$$

The Backward Pass: Gradient Comparison

Centered model:

$$\frac{\partial \mathcal{L}}{\partial \Gamma_k} = \underbrace{\overset{=\mathbf{0}}{\frac{\partial \mathrm{NLL}}{\partial \Gamma_k}}}_{\text{Not in forward pass!}} + \underbrace{W \cdot (-2) \sum_i G_i^\top K_\theta^{-1} \delta_{ik}}_{\text{prior only } (W{=}1\text{e-4, weak})}$$

Non-centered model:

$$\frac{\partial \mathcal{L}}{\partial \Gamma_k} = \underbrace{\frac{\partial \mathrm{NLL}}{\partial \pi} \cdot \frac{\partial \pi}{\partial \theta} \cdot \frac{\partial \theta}{\partial \lambda} \cdot G^\top}_{\text{chain rule through forward pass (strong!)}} + \underbrace{0}_{\delta \text{ indep. of } \Gamma}$$

Same loss, different gradient pathways:

- Centered: Γ gets a direct GP penalty gradient that **shrinks** it ($\propto W \cdot G^\top K^{-1} \delta$), but **no** NLL gradient — the data can't speak to Γ
- Non-centered: Γ gets **zero** GP penalty gradient, but a **strong** NLL gradient — the data drives Γ directly

Same Objective, Different Computation

Both models minimize the **same function**:

$$\mathcal{L} = -\log p(Y \mid \lambda) + W\|\lambda - G\Gamma\|^2_{K_\theta^{-1}}$$

Same Objective, Different Computation

Both models minimize the **same function**:

$$\mathcal{L} = -\log p(Y \mid \lambda) + W\|\lambda - G\Gamma\|^2_{K_\theta^{-1}}$$

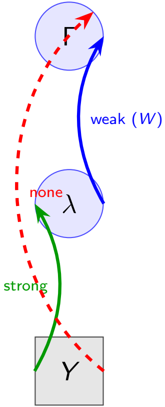
The difference is purely in **how the computer represents** λ:

	Centered	**Non-centered**
Free parameters	λ, Γ	δ, Γ
Forward pass	$\theta = \text{softmax}(\lambda)$	$\lambda = G\Gamma + \delta$ $\theta = \text{softmax}(\lambda)$
GP gradient on Γ	Nonzero (but $\propto W$)	Zero
NLL gradient on Γ	Zero	Strong (chain rule)
Net effect on Γ	Over-shrunk	Well-estimated

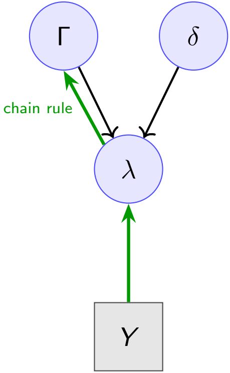
At any solution: $\delta^* = \lambda^* - G\Gamma^*$. If the problem were convex, both would find the same optimum. It is **non-convex**: different gradient flows → different local optima.

The Gradient Flow Picture

Centered:



Non-centered:



Both Γ and δ feed into λ in the forward pass → both get NLL gradients.
In centered, λ is a dead-end: NLL gradients stop there and never reach Γ.

Simulation: γ Recovery

Parameter recovery simulation (10k patients, 10 PRS, 5 signatures):

	Centered	**Reparam (**κ **free)**	**Nokappa (**$\kappa = 1$**)**
$r(\hat{\gamma}, \gamma_{\text{true}})$	0.796	0.953	0.954
Mean $\lvert\hat{\gamma}\rvert$	0.175	0.197	0.191
True mean $\lvert\gamma\rvert$	0.200		

- Non-centered recovers γ with $r = 0.954$ vs. $r = 0.796$ for centered
- Both show mild shrinkage (estimated $|\gamma|$ slightly below true)
- Residual shrinkage in non-centered is an **optimization phenomenon**: Γ and δ jointly determine λ, and the GP penalty on δ shapes the landscape Γ optimizes over
- κ free vs. fixed at 1: essentially no difference ($r = 0.953$ vs 0.954)

γ Stability Across 40 Training Batches

	Non-centered	**Centered**
Mean $\lvert\gamma\rvert$ (pooled)	0.065	0.005
Mean batch std	0.136	0.015
Median CV	2.50	5.91
Sign-flip rate	30%	39%
Batch-to-mean correlation	0.48	0.58

Key biological signals are stable (boxplot across 40 batches):

- T2D → Sig 15 (DM): consistently ~ 1.2, tight box, never crosses zero
- CAD → Sig 5 (Ischemic CVD): consistently ~ 0.25, all positive
- BMI → Sig 7 (Metabolic): all positive, tight range

The high median CV is driven by weak/noise entries. **Pooling across batches** regularizes γ by $\sqrt{B}$, acting as the post-hoc equivalent of the implicit prior.

ψ Stability: Much More Consistent

	γ **(genetics → sigs)**	ψ **(sigs → diseases)**
Median CV	2.50	0.14
Sign-flip rate	30%	2.4%
Batch-to-mean corr	0.48	0.96

Why is ψ more stable?

- ψ's tradeoff partner is ϵ (shared across all N patients)
- γ's tradeoff partner is δ (per-patient, $N \times K \times T$ free parameters)
- δ has more flexibility to absorb signal that γ would explain
- ϵ is constrained to work for all patients simultaneously

Psi switches from initialization: only 9/348 diseases changed max signature — all biologically reasonable reassignments.

Prediction AUC: Centered vs. Non-Centered

LOO evaluation: pool parameters from training batches, fit δ on held-out patients:

Metric	**Centered (nolr)**	**Non-centered**	Δ
Static 10-year	0.622	0.654	$+0.032$
Dynamic 10-year	0.624	0.629	$+0.005$
Dynamic 1-year	0.765	0.883	$+0.118$
Static 1-year	0.770	0.878	$+0.108$

Non-centered wins on all metrics, especially short-horizon:

- The genetically-informed prior mean $G\Gamma$ gives different starting points for high-risk vs. low-risk patients (**between-strata** discrimination)
- δ captures individual residual variation (**within-strata** discrimination)
- Centered model had to do both with a single λ

Why the Centered Model Starves Γ

Centered: λ is the free parameter, $\delta = \lambda - G\Gamma$ is derived.
The GP penalty gradient w.r.t. Γ_k:

$$\nabla_{\Gamma_k}\mathrm{GP} = -2W \sum_i G_i^\top K_\theta^{-1} \underbrace{(\lambda_{ik} - G_i\Gamma_k)}_{\delta_{ik}}$$

This is **nonzero** — it directly regularizes Γ toward values where $\delta \to 0$.

Non-centered: δ is the free parameter, $\lambda = G\Gamma + \delta$.
The GP penalty is $W \cdot \delta^\top K^{-1}\delta$. Its gradient w.r.t. Γ_k:

$$\nabla_{\Gamma_k}\mathrm{GP} = 0$$

Zero. δ doesn't depend on Γ. No penalty touches Γ.

Γ only sees the NLL gradient via the chain rule $Y \to \pi \to \theta \to \lambda \to \Gamma$. The data speaks directly to Γ without competition from a prior penalty.

A Simple Analogy

Minimize $f(x) = (x-3)^2$ with a preference that x be near μ:

Centered:

- Parameters: x and μ
- Loss $= (x-3)^2 + w \cdot (x-\mu)^2$
- μ only appears in the penalty. If w is small, μ barely moves.

Non-centered:

- Parameters: z and μ, where $x = \mu + z$
- Loss $= (\mu + z - 3)^2 + w \cdot z^2$
- μ appears in the **main loss**. It gets a strong gradient.

Same model, same answer ($x^* = 3$), but in the non-centered form μ learns where x wants to be. In the centered form, μ barely moves because w is tiny.

Summary

1. **Same model, flat prior on Γ:** $p(\Gamma) \propto 1$. The GP prior is on $\delta = \lambda - G\Gamma$, not on Γ.

2. **Centered starves Γ:** GP penalty has a direct gradient on Γ that shrinks it; the NLL has no gradient on Γ at all. Non-centered: GP gradient on Γ is zero; NLL gradient is strong.

3. **Simulation:** Non-centered recovers γ with $r = 0.954$ vs. 0.796 for centered. Residual shrinkage in non-centered is an optimization phenomenon, not a prior.

4. **Real data:** Key biological signals (T2D, CAD, BMI) are stable across 40 batches. ψ is very stable (CV $= 0.14$). Pooling regularizes γ.

5. **Prediction:** Non-centered wins on all AUC metrics ($+0.118$ on dynamic 1-year). Genetically-informed $G\Gamma$ provides between-strata discrimination; δ handles within-strata.

6. **Standard approach:** Non-centered parameterization is the