Regularization as Prior Injection

Variational Quantum Algorithms · Module 5 — Regularization

Node 5.1

Position in Knowledge Graph

System: Variational Quantum Algorithms **Structural Domain:** Regularization **Node:** 5.1

This is the **standard machine-learning lens** on what regularization is. It is the lens you will encounter in any textbook on supervised learning, statistics, or inverse problems. We start here because Module 5 must first explain regularization in its conventional form before we can argue (in node 5.5) that the conventional form is incomplete in the VQA setting.

Formal Definition

Given an optimization problem

$$\min_{\boldsymbol{\theta}} \; C(\boldsymbol{\theta})$$

a **regularizer** is an auxiliary penalty function $R(\boldsymbol{\theta})$ added to the objective:

$$\min_{\boldsymbol{\theta}} \; C(\boldsymbol{\theta}) + \lambda \, R(\boldsymbol{\theta}),$$

where $\lambda \geq 0$ is the **regularization strength**.

The Bayesian interpretation is that this is a **maximum a posteriori (MAP)** estimate under a prior $p(\boldsymbol{\theta}) \propto \exp(-\lambda R(\boldsymbol{\theta}))$:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \; p(\text{data} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \; \big[-\log p(\text{data} \mid \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) \big].$$

The negative log-likelihood is the cost C. The negative log-prior is λR.

Intuition

A regularizer is a way of saying:

> Among all parameter values that fit the data equally well, prefer the ones that are "small," "smooth," "sparse," or otherwise structured.

The prior is your belief about what counts as a "reasonable" $\boldsymbol{\theta}$ before you see any data.

The data tugs the optimizer in one direction. The prior tugs it in another. The MAP estimate is the equilibrium.

In the limit $\lambda \to 0$, the prior is irrelevant and you recover the maximum-likelihood estimate. In the limit $\lambda \to \infty$, the data is irrelevant and you recover the prior mode. The art is choosing λ so that the two are in productive tension.

Structural Role

This node is the **anchor** for the standard regularization techniques in nodes 5.2 (geometric), 5.3 (spectral), and 5.4 (generalization bounds). All three are concrete instances of the prior-injection framework.

It is also the **foil** for node 5.5 (regularization as dynamical modification). The thesis claim is that prior-injection is the wrong framing for VQA, because in VQA the optimizer never reaches the MAP solution. It lives in a stochastic transient regime where dynamics, not steady states, determine outcomes. You cannot understand 5.5 without first having 5.1 in hand.

Relations to Other Concepts

- **L2 / weight decay** — Gaussian prior on parameters. $R(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$.
- **L1 / Lasso** — Laplace prior. $R(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$. Promotes sparsity.
- **Tikhonov regularization** — the inverse-problems name for L2.
- **Ridge regression** — linear regression with L2.
- **Dropout** — implicit regularization via stochastic pruning of computation paths.

Worked Derivations

L2 regularization is a Gaussian prior

Let $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \sigma^2 I)$. Then

$$-\log p(\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\|\boldsymbol{\theta}\|_2^2 + \text{const.}$$

Setting $\lambda = 1/(2\sigma^2)$ gives the standard L2 penalty. Smaller prior variance is equivalent to stronger regularization.

MAP versus full posterior

The MAP estimate is a single point, not a distribution. It throws away all information about the posterior's spread. Bayesian methods that integrate over the full posterior (variational inference, MCMC) recover this information at higher computational cost. For VQA, even MAP is hard — see node 5.5 for why we don't actually reach MAP in practice.

Visualization

Pending. Diagram: 2D parameter space with the data-likelihood contour ellipse, the prior contour circle (centered at origin), and the MAP point at the intersection of the two pulls.

Exercises

1. Show that L2 regularization with strength λ on a linear regression problem shrinks each least-squares coefficient by a factor of $\sigma_i^2/(\sigma_i^2 + \lambda)$, where σ_i is the i-th singular value of the design matrix.
2. Derive the MAP estimate for a Bernoulli likelihood with a Beta prior. What does the regularization strength correspond to in terms of "effective prior counts"?
3. (VQA) Consider a parameterized circuit with parameters $\boldsymbol{\theta} \in [-\pi, \pi]^d$. Why does an L2 penalty have a less natural interpretation here than in unconstrained Euclidean parameter spaces?

Research Extensions

- **Empirical Bayes** — learning the prior hyperparameters from the data itself.
- **Hierarchical priors** — putting priors on the regularization strength λ.
- **Implicit regularization** — gradient descent itself acts as a regularizer in overparameterized models, even without an explicit penalty term. This is the bridge to node 5.5.

Cross-links to Other Courses

- **Statistics Module 6** — Bayesian Inference (priors, posteriors, MAP).
- **Statistics Module 10** — Model Selection (penalized likelihood, AIC/BIC, cross-validation as implicit regularization).
- **Math-Intuition** — Lagrangian formulation of constrained optimization as regularization.

Variational Quantum Algorithms · Module 5 — Regularization · Node 5.1
Concepts: regularization, cost-function, prior-posterior, bayesian-updating
Prerequisites: 3.1
Enables: 5.2, 5.3, 5.4, 5.5

hbar.university — own.your.cognition