When Does Preconditioning Help or Hurt Generalization?

A Bias- Variation Perspective

Mitchell Scott

Department of Mathematics, Emory University

February 17, 2026

Motivation

Second-order methods (natural gradient, Newton, etc.):

- Often accelerate optimization
- Modify geometry of parameter updates

But:

- Gradient descent (GD) has implicit bias
- In overparameterized models, this affects generalization

Central Question:

When does preconditioning help or hurt generalization?

EMORY UNIVERSITY

Setting: Overparameterized Linear Regression

Student–teacher model:

$$y_i = x_i^\top \theta^* + \varepsilon_i$$

Assumptions:

- $x_i = \Sigma_X^{1/2} z_i$, isotropic z_i
- $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
- $d > n$, with $d/n \to \gamma > 1$

We study gradient flow with fixed preconditioner P.

Preconditioned Gradient Flow

Gradient flow:

$$\dot{\theta}(t) = \frac{1}{n} P X^\top (y - X\theta(t))$$

Stationary solution:

$$\hat{\theta}_P = P X^\top (X P X^\top)^{-1} y$$

Interpretation:

$$\hat{\theta}_P = \arg\min_\theta \|\theta\|_{P^{-1}} \quad \text{s.t. } X\theta = y$$

Preconditioning changes the implicit norm.

Examples of P

- $P = I \rightarrow$ Gradient Descent (GD)
- $P = \Sigma_X^{-1} \rightarrow$ Natural Gradient (NGD)
- $P = (X^\top X + \lambda I)^{-1} \rightarrow$ Sample Fisher (same stationary solution as GD)

Key distinction:

- Sample Fisher $\neq$ Population Fisher

Population Risk

Population risk:

$$R(\theta) = \mathbb{E}[(x^\top \theta - x^\top \theta^*)^2]$$

Bias–variance decomposition:

$$R = B + V$$

$$B = \mathbb{E}[(x^\top \mathbb{E}[\hat{\theta}] - x^\top \theta^*)^2]$$

$$V = \mathrm{tr}(\mathrm{Cov}(\hat{\theta})\Sigma_X)$$

Exact asymptotics computed as $n, d \to \infty$.

Main Result 1: Variance

Theorem:
Among admissible preconditioners,

$$V(\hat{\theta}_P) \geq \frac{\sigma^2}{\gamma - 1}$$

Equality iff:

$$P = \Sigma_X^{-1}$$

Conclusion:
Natural gradient minimizes variance.

If label noise dominates $\rightarrow$ NGD wins.

Main Result 2: Bias

Assume random effects prior:

$$\theta^* \sim \mathcal{N}(0, d^{-1}\Sigma_\theta)$$

Bias minimized when:

$$P = \Sigma_\theta$$

Special cases:

- $\Sigma_\theta = I \rightarrow$ GD optimal
- $\Sigma_\theta = \Sigma_X^{-1} \rightarrow$ NGD optimal

Bias depends on signal alignment.

Signal–Data Alignment

Alignment matters:

- If signal is isotropic $\rightarrow$ GD has smaller bias
- If signal lies in small-variance directions $\rightarrow$ NGD better

Intuition:
NGD rescales directions by inverse feature covariance.

Misspecification

Suppose:

$$y_i = x_i^\top \theta^* + x_{c,i}^\top \theta_c + \varepsilon_i$$

Then:

$$\text{Misspecified bias} \propto V(\hat{\theta}_P) + 1$$

Misspecification behaves like additional label noise.

Thus NGD helps under misspecification.

Three Governing Factors

Regime	Winner
High noise	NGD
Misspecified model	NGD
Clean isotropic signal	GD
Misaligned signal	NGD

No universally optimal optimizer.

Interpolating Preconditioners

Consider interpolation:

$$P_\alpha = (1-\alpha)I + \alpha\Sigma_X^{-1}$$

As α increases:

- Variance $\downarrow$ monotonically
- Bias $\uparrow$ (for isotropic signal)

Optimal α balances bias and variance.

Early Stopping

Variance increases monotonically in time.

Thus:

- Early stopping reduces variance
- GD may outperform NGD at early times

Stationary comparison $\neq$ early-time comparison.

Extension to RKHS

Kernel regression with covariance operator Σ.
Preconditioned update:

$$f_t = f_{t-1} - \eta(\Sigma + \alpha I)^{-1}(\hat{\Sigma} f_{t-1} - \hat{S}^* Y)$$

With optimal α:

$$R(f_t) = \tilde{O}\left(n^{-\frac{2rs}{2rs+1}}\right)$$

in $O(\log n)$ steps.
GD requires polynomially many steps.
Preconditioning accelerates population risk decay.

Neural Network Experiments

Experiments on MNIST and CIFAR-10:

- Increasing label noise $\rightarrow$ NGD wins
- Increasing teacher complexity $\rightarrow$ NGD wins
- Synthetic misalignment $\rightarrow$ NGD wins
- Interpolation can outperform both

Linear theory qualitatively matches neural nets.

Conceptual Takeaways

1. Preconditioning changes implicit regularization.
2. Variance always favors NGD.
3. Bias depends on signal geometry.
4. Misspecification $\approx$ label noise.
5. Interpolation can be optimal.

Open Problems

- Adaptive methods (Adam, Adagrad)
- Beyond squared loss
- Finite-sample regimes
- Population vs. sample Fisher gap

Conclusion

Optimization geometry $\neq$ Generalization geometry.

Preconditioning can help or hurt depending on:

- Noise
- Misspecification
- Alignment

Thank you.

EMORY UNIVERSITY

Conceptual Summary

- Variance always favors NGD.
- Bias depends on signal alignment.
- Misspecification behaves like noise.
- Interpolation balances bias–variance.
- Early stopping changes the comparison.

EMORY UNIVERSITY

Gradient Descent

Goal: minimize empirical risk (training error) $L(f_{\mathbf{w}})$.

- $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{P}_t^{-1} \nabla_{\mathbf{w}_t} L(f_{\mathbf{w}_t})$
  - $\mathbf{P} = \mathbf{I}$: ordinary gradient descent
  - $\mathbf{P} = \mathbf{F}$: natrual gradient descent where $\mathbf{F}$ is Fisher information matrix.
  - $\mathbf{P} = \mathbf{H}$: Newton's method, where $\mathbf{H}$ is the Hessian
  - $\mathbf{P} = \mathbf{D}$: ADAM AdaGrad
- $\mathbf{w}_{t+1} = \mathbf{w}_t + \beta_t (\mathbf{w}_t - \mathbf{w}_{t-1}) - \eta_t \nabla L (\mathbf{w}_t + \gamma_t (\mathbf{w}_t - \mathbf{w}_{t-1}))$
  - $\beta_t = \gamma_t$: Nesterov
  - $\gamma_t = 0$: Polyak's Heavy Ball

EMORY UNIVERSITY

Mirror Descent

Given a strongly convex AND differentiable potential ψ $\mathbf{w}_{t+1} = \text{argmin}_{\mathbf{w}\in\mathcal{W}}\, \eta_t \langle \mathbf{w}, \nabla L(\mathbf{w}_t)\rangle + D_\psi(\mathbf{w}, \mathbf{w}_t)$, where

$$D_\psi(\mathbf{w}, \mathbf{w}') = \psi(\mathbf{w}) - \psi(\mathbf{w}') - \langle \nabla\psi(\mathbf{w}'), \mathbf{w} - \mathbf{w}'\rangle$$

is the *Bregman divergence* wrt ψ, and $\mathcal{W}$ is some constraint set for parameters $\mathbf{w}$.
If $\mathcal{W} = \mathbb{R}^d$, we have unconstrained optimization, and the update is equivalent to

$$\nabla\psi(\mathbf{w}_{t+1}) = \nabla\psi(\mathbf{w}_t) - \eta_t \nabla L(\mathbf{w}_t)$$