

# When Does Preconditioning Help or Hurt Generalization?

---

Shun-ichi Amari<sup>1</sup>, Jimmy Ba<sup>2</sup>, Roger Grosse<sup>2</sup>, Xuechen Li<sup>3</sup>,  
Atsushi Nitanda<sup>4</sup>, Taiji Suzuki<sup>4</sup>, Denny Wu<sup>2</sup>, Ji Xu<sup>5</sup>

<sup>1</sup>RIKEN CBS

<sup>2</sup>University of Toronto and Vector Institute

<sup>3</sup>Google Research, Brain Team

<sup>4</sup>University of Tokyo and RIKEN AIP

<sup>5</sup>Columbia University

**International Conference on Learning Representations 2021**

# Preconditioned Gradient Descent

$$\text{Update rule: } \theta_{t+1} = \theta_t - \eta \mathbf{P}(t) \nabla_{\theta_t} L(f_{\theta_t}), \quad t = 0, 1, \dots$$

Common choices of preconditioner  $\mathbf{P}$  and corresponding algorithm:

- Inverse Fisher information matrix  $\Rightarrow$  natural gradient descent (NGD).
- Certain diagonal matrix  $\Rightarrow$  adaptive gradient methods (e.g. Adagrad, Adam).

**Geometric Intuition:** alleviate the effect of pathological curvature (using 2nd order information) and speed up **optimization**.

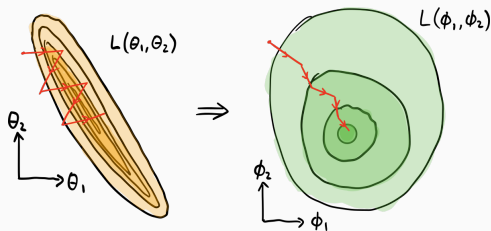


Figure from Xanadu blog post.

**Question:** how does preconditioning affect generalization?

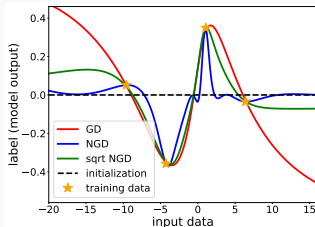
# Motivation: Implicit Bias of Optimizers

In the *online learning* setup, efficient optimization  $\approx$  good generalization.

**This work:** learning a *fixed* dataset, possibly achieving zero training loss.

## Implicit Bias in Interpolants

- Modern machine learning models (e.g. neural nets) are often **overparameterized**.
- Overparameterized models may interpolate training data in *different ways*.
- $P$  affects the properties of the interpolant.



## Motivation of This Work:

- In the *interpolation setting* (i.e. absence of explicit regularization), how does preconditioning influence the generalization performance?

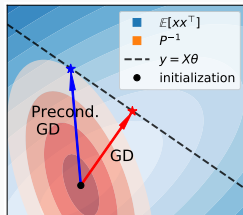
# Implicit Bias in Overparameterized Linear Regression

**Motivating Example:** preconditioned gradient descent (PGD) on the *overparameterized* least squares objective:  $L(\theta) = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\theta\|_2^2$ .

**Stationary Solution** ( $t \rightarrow \infty$ ):

- **Gradient descent:** min  $\ell_2$ -norm solution.
- **Preconditioned GD:** for time-independent and full-rank  $\mathbf{P}$ , min  $\|\theta\|_{\mathbf{P}^{-1}}$  norm solution.

**Common Argument:** min  $\ell_2$ -norm solution generalizes well  $\Rightarrow$  GD ( $\mathbf{P} = \mathbf{I}_d$ ) is better (e.g. [Wilson et al. 2017]).



**Question:** Why is the  $\ell_2$  norm the right measure for generalization?

**Motivation of This Work:**

- In simplified settings, can we determine the *optimal preconditioner* that leads to the lowest generalization error?

# Preconditioned Linear Regression: Problem Setup

- **Data Model:**  $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \Sigma_{\mathbf{x}}$ ;  $\mathbf{X} \in \mathbb{R}^{n \times d}$ ,  $n, d \rightarrow \infty$  and  $d/n \rightarrow \gamma > 1$ .
- **Gradient Update:**  $d\theta(t) = \frac{1}{n}\mathbf{P}(t)\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\theta(t))dt$ ,  $\theta(0) = 0$ .

Consider natural gradient descent (NGD) as an example. Given data distribution and model  $p(\mathbf{X}, y|\theta) = p(\mathbf{X})p(y|f_\theta(\mathbf{X}))$ ,

$$\mathbf{F} = \mathbb{E}[\nabla_\theta \log p(\mathbf{X}, y|\theta) \nabla_\theta \log p(\mathbf{X}, y|\theta)^\top] = -\mathbb{E}[\nabla_\theta^2 \log p(\mathbf{X}, y|\theta)].$$

The NGD update direction is then given by  $\mathbf{F}^{-1}\nabla_\theta L(\mathbf{X}, f_\theta)$ .

**Remark:** for squared loss, the Fisher reduces to  $\mathbb{E}[\mathbf{J}_f^\top \mathbf{J}_f]$  [Martens 2014].

For least squares regression, many preconditioners are *time-invariant*:

- *Sample Fisher (Hessian)*  $\Leftrightarrow$  **sample covariance**  $\mathbf{X}^\top \mathbf{X}/n$ .
- *Population Fisher*  $\Leftrightarrow$  **population covariance**  $\Sigma_{\mathbf{x}}$ .

We thus limit our analysis to *fixed preconditioners*  $\mathbf{P}(t) =: \mathbf{P}$ .

# Stationary Solution of Preconditioned Regression

For positive definite  $\mathbf{P}$ , the gradient flow trajectory is described by

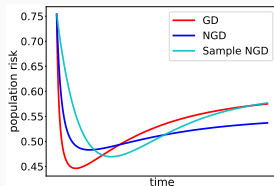
$$\boldsymbol{\theta}_{\mathbf{P}}(t) = \mathbf{P}\mathbf{X}^{\top} \left[ \mathbf{I}_n - \exp\left(-\frac{t}{n}\mathbf{X}\mathbf{P}\mathbf{X}^{\top}\right) \right] (\mathbf{X}\mathbf{P}\mathbf{X}^{\top})^{-1}\mathbf{y},$$

and the stationary solution  $\hat{\boldsymbol{\theta}}_{\mathbf{P}}$  is the  $\min \|\boldsymbol{\theta}\|_{\mathbf{P}^{-1}}$  norm interpolant:

$$\hat{\boldsymbol{\theta}}_{\mathbf{P}} := \lim_{t \rightarrow \infty} \boldsymbol{\theta}_{\mathbf{P}}(t) = \mathbf{P}\mathbf{X}^{\top} (\mathbf{X}\mathbf{P}\mathbf{X}^{\top})^{-1}\mathbf{y} = \arg \min_{\mathbf{X}\boldsymbol{\theta}=\mathbf{y}} \|\boldsymbol{\theta}\|_{\mathbf{P}^{-1}}.$$

## Noticeable examples of preconditioned update:

- **Identity:**  $\mathbf{P} = \mathbf{I}_d$  gives the min  $\ell_2$  norm interpolant (also true for momentum GD and SGD).
- **Population Fisher:**  $\mathbf{P} = \mathbf{F}^{-1} = \Sigma_{\mathbf{x}}^{-1}$ .
- **Sample Fisher:**  $\mathbf{P} = (\mathbf{X}^{\top}\mathbf{X} + \lambda\mathbf{I}_d)^{-1}$  or  $(\mathbf{X}^{\top}\mathbf{X})^{\dagger}$  results in the min  $\ell_2$  norm solution (*same as GD*).



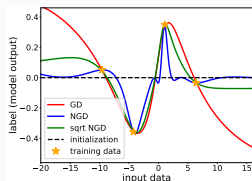
**Remark:** population Fisher can be estimated from extra **unlabeled data**.

# Implicit Bias of Natural Gradient Descent

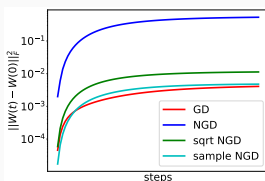
## Starting from zero initialization:

- **GD solution  $\hat{\theta}_I$**  has small parameter norm  $\|\theta\|_2$ .
- **NGD solution  $\hat{\theta}_{F-1}$**  has small function norm  $\mathbb{E}_{p(\mathbf{x})}[f(\mathbf{x})^2] = \|\theta\|_{\Sigma_x}^2$ .
- **Sample Fisher-based updates** behaves similar to **GD**.

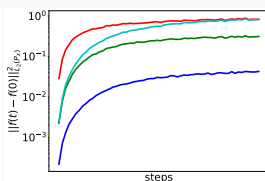
Similar findings also empirically observed in **two-layer neural networks**:



1D illustration (sigmoid).



Parameter difference.



Function difference.

**Question:** How does this difference translate to the generalization performance?

# Bias-variance Decomposition

$$R(\theta) = \underbrace{\mathbb{E}_{P_X}[(f^*(\mathbf{x}) - \mathbf{x}^\top \mathbb{E}_{P_\epsilon}[\theta])^2]}_{B(\theta), \text{ bias}} + \underbrace{\text{tr}(\text{Cov}(\theta)\Sigma_X)}_{V(\theta), \text{ variance}}.$$

- **Bias** depends on the teacher (target function)  $f_*$  and data distribution.
- **Variance** is due to the *label noise* (independent of the teacher model).

**Goal:** determine the optimal preconditioner  $\mathbf{P}$  under different conditions of teacher model (bias) and label noise (variance).

## Precise Asymptotic Risk in Bias-variance Decomposition:

**Thm. (informal).** Under certain conditions, as  $n, d \rightarrow \infty$ ,  $d/n \rightarrow \gamma \in (1, \infty)$ ,

- For positive definite  $\mathbf{P}$ ,  $V(\hat{\theta}_{\mathbf{P}}) \rightarrow \sigma^2 \left( \lim_{\lambda \rightarrow 0_+} \frac{m'(-\lambda)}{m^2(-\lambda)} - 1 \right)$ .
- For linear teacher  $\theta_*$ ,  $B(\hat{\theta}_{\mathbf{P}}) \rightarrow \lim_{\lambda \rightarrow 0_+} \frac{m'(-\lambda)}{m^2(-\lambda)} \mathbb{E} \left[ \frac{v_x v_\theta}{(1 + v_{xp} m(-\lambda))^2} \right]$ .

Where  $m(z) > 0$  is the *Stieltjes transform* of the limiting spectral distribution of  $\mathbf{X} \mathbf{P} \mathbf{X}^\top$ , and  $(v_x, v_\theta, v_{xp})$  relates to the eigenvalues of  $\mathbf{P}$ ,  $\Sigma_x$ , and  $\mathbb{E}[\theta_* \theta_*^\top]$ .

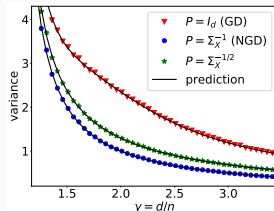


# Variance Term: NGD is Optimal

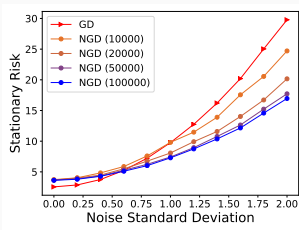
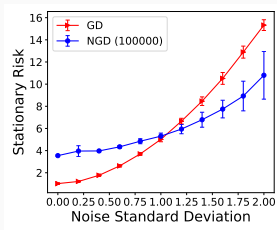
**Thm.** Among all positive definite  $P$ , the variance is minimized by **NGD**:  $P = F^{-1}$ .

**Message:** when labels are noisy (risk is dominated by variance), NGD is beneficial.

**Remark:** Note that **population Fisher** is required.



## Two-layer MLP: student-teacher setup (distillation)



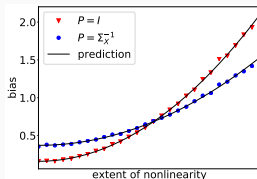
- **Left:** NGD (population Fisher) achieves lower risk under large label noise.
- **Right:** sample Fisher (i.e. less unlabeled data used) behaves like GD.

# Misspecification $\approx$ Label Noise

**Misspecified Model:**  $f_*(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\theta}_* + f_*^c(\mathbf{x})$ ; the residual  $f_*^c$  cannot be learned by the student model.

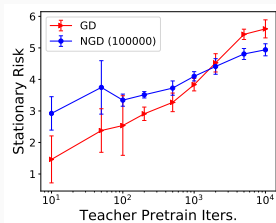
**Intuition:**  $f_*^c$  is “similar” to additive label noise.

**Message:** **NGD** is beneficial under misspecification.

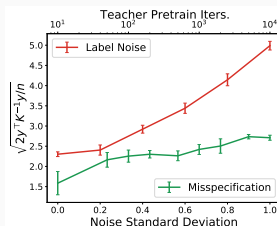


## Misspecification in Neural Networks

- **Student:** two-layer MLP; **Teacher:** ResNet-20 at varying training epochs.
- **Heuristic measure of misspecification:**  $\sqrt{\mathbf{y}^\top \mathbf{K}^{-1} \mathbf{y} / n}$ , where  $\mathbf{K}$  is the *neural tangent kernel* (NTK) matrix of the student.



Misspecification on CIFAR-10.

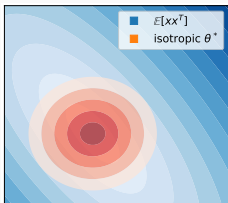


Measure of misspecification.

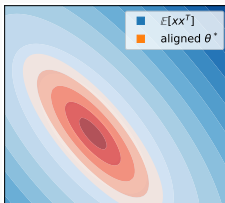
# Bias Term: Well-specified Case

**Well-specified Model:**  $f_*(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\theta}_*$ , with general prior  $\mathbb{E}[\boldsymbol{\theta}_* \boldsymbol{\theta}_*^\top] = \boldsymbol{\Sigma}_\theta$ .

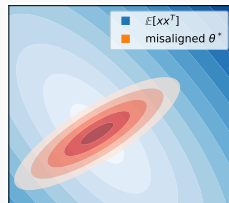
- Setup extends previously assumed *isotropic prior* [Dobriban and Wager 18].
- **Alignment** between  $\boldsymbol{\Sigma}_x$  and  $\boldsymbol{\Sigma}_\theta$  relates to the **source condition** in RKHS.



Isotropic (previous work).



Aligned (easy problem).

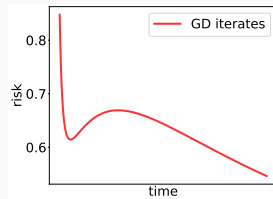


Misaligned (hard problem).

## “Surprises” under General Setup:

- Gradient descent may lead to prediction risk **non-monotonic in time**, even if  $\sigma = 0$ .

**Remark:** when  $\boldsymbol{\Sigma}_x$  or  $\boldsymbol{\Sigma}_\theta$  is isotropic, the bias term is always *monotonically decreasing* through time.



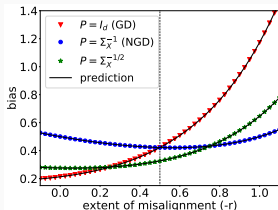
# Well-specified Bias (continued)

**Theorem (informal).** Among all positive definite  $\mathbf{P}$  codiagonalizable with  $\Sigma_x$ , the stationary *bias* is minimized by  $\mathbf{P} = \mathbf{U} \text{diag}(\mathbf{U}^\top \Sigma_\theta \mathbf{U}) \mathbf{U}^\top$ .

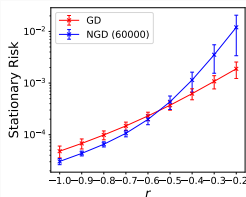
**No-free-lunch:** the optimal  $\mathbf{P}$  is usually not known *a priori*:

- **GD** generalizes better when target is **isotropic**  $\Sigma_\theta = \mathbf{I}_d$ .
- **NGD** is optimal under **misalignment**  $\Sigma_\theta = \Sigma_x^{-1}$  (“hard” problem).

**Prop. (source condition).** When  $\Sigma_\theta = \Sigma_x^r$ , there exists a transition point  $r^* \in (-1, 0)$  s.t. **GD** achieves lower (higher) bias than **NGD** iff  $r > (<) r^*$ .



Linear regression.



Two-layer MLP (MNIST).

**Misalignment in MLP:** Construct the teacher parameters in the small eigen-directions of the student's Fisher. Large  $|r| \Rightarrow$  more misaligned.

# Bias-variance Tradeoff: Interpolating between $P$

The optimal  $P$  for the *bias* and *variance* are in general **different**.

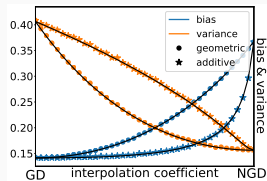
**Question:** how can we trade in one of bias/variance for the other?

**Example:** Consider  $\Sigma_\theta = I_d$ ,  $\Sigma_x \neq I_d$ , and the following interpolation schemes:

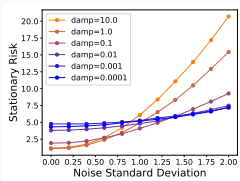
- **Additive:**  $P_\alpha = (\alpha \Sigma_x + (1-\alpha)I_d)^{-1}$ , corresponds to the *damped inverse*.
- **Geometric:**  $P_\alpha = \Sigma_x^{-\alpha}$ , covers the “conservative” *square-root scaling*.

**Proposition (informal).** The stationary bias/variance is *monotonically* increasing/decreasing w.r.t.  $\alpha$  in a certain range between 0 and 1.

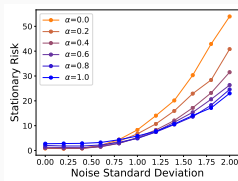
$\Rightarrow$  At certain SNR, **interpolating** between GD and NGD is beneficial.



Monotonicity of bias/variance.



Additive interpolation (MLP).



Geometric interpolation (MLP).

# Bias-variance Tradeoff: Early Stopping

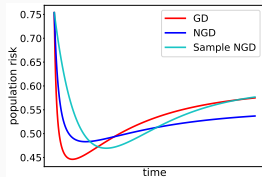
We have thus far only looked at the stationary solution ( $t \rightarrow \infty$ ).

**Question:** what about algorithmic regularization such as *early stopping*?

**Proposition.** Define the optimal early-stopping bias  $B^{\text{opt}}(\theta) = \inf_{t \geq 0} B(\theta(t))$ .

1. When  $\Sigma_{\theta} = \Sigma_x^{-1}$  (misaligned),  $B^{\text{opt}}(\theta_P) \geq B^{\text{opt}}(\theta_{F-1})$ .
2. When  $\Sigma_{\theta} = I_d$  (isotropic),  $B^{\text{opt}}(\theta_I) \leq B^{\text{opt}}(\theta_{F-1})$ .
3. The variance  $V(\theta_P(t))$  *monotonically increases* through time.

- (3) suggests that early stopping is beneficial when data is noisy (due to reduction of variance).
- (1-2) suggests that early stopping may not alter the comparison of the well-specified bias (between GD and NGD).



**Question:** What about the **early stopping time**, i.e. number of steps (efficiency) needed to achieve the *optimal population risk*?

# RKHS Regression: Fast Decay of Population Risk

**Aim to show:** preconditioning  $\Rightarrow$  efficient reduction of *population risk*.

- **Model:**  $y_i = f^*(x_i) + \varepsilon_i$ .  $S : \mathcal{H} \rightarrow L_2(P_X)$ .  $\Sigma = S^*S$ ;  $L = SS^*$ .
- **Optimization:**  $f_t = f_{t-1} - \eta(\Sigma + \alpha I)^{-1}(\hat{\Sigma}f_{t-1} - \hat{S}^*Y)$ ,  $f_0 = 0$ .  $f_t \in \mathcal{H}$ .

**Remark:** the population Fisher corresponds to the *covariance operator*  $\Sigma$ . The update is thus an **additive interpolation** between GD and NGD.

## Assumptions:

- **Source Condition:**  $\exists r \in (0, \infty)$  s.t.  $f^* = L^r h^*$  for some  $h^* \in L_2(P_X)$ .
- **Capacity Condition:**  $\exists s > 1$  s.t.  $\text{tr}(\Sigma^{1/s}) < \infty$  and  $2r + s^{-1} > 1$ .
- **Regularity of RKHS:**  $\exists \mu \in [s^{-1}, 1]$ ,  $C_\mu > 0$  s.t.  $\sup_x \|\Sigma^{1/2-1/\mu} K_x\|_{\mathcal{H}} \leq C_\mu$ .

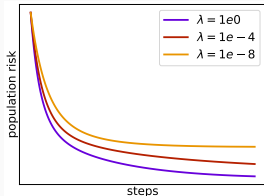
**Remark:** *source condition* relates to the previously discussed alignment:  
large  $r \Rightarrow$  smoother teacher model, i.e. "easier" problem; vice versa.

# Fast Decay of Population Risk (continued)

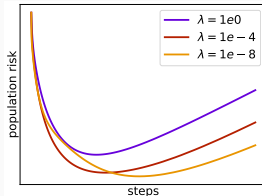
**Theorem (informal).** Given  $\mu \leq 2r$  or  $r \geq 1/2$ , for sufficiently large  $n$ , preconditioned update with  $\alpha = n^{-\frac{2s}{2rs+1}}$  achieves the minimax optimal convergence rate  $R(f_t) = \|Sf_t - f^*\|_{L_2(P_X)}^2 = \tilde{O}\left(n^{-\frac{2rs}{2rs+1}}\right)$  in  $t = \Theta(\log n)$  steps, whereas ordinary gradient descent requires  $t = \Theta\left(n^{\frac{2rs}{2rs+1}}\right)$  steps.

**Remark:** similar to the role of *momentum* [Pagliana and Rosasco 2019].

- The optimal interpolation coefficient  $\alpha$  and stopping time  $t$  are chosen to *balance the bias and variance*.
- $\alpha$  **increases** with  $r$  – NGD is advantageous for “hard” problems.



$r = 3/4$  (“easy” problem).



$r = 1/4$  (“hard” problem).



# Discussion and Conclusion

## Overparameterized Least Squares Regression:

- Identified factors that impact the generalization of ridgeless interpolant.
  - NGD is advantageous under *noisy labels* or *misaligned* (“hard”) problem.
- Discussed how bias-variance tradeoff can be realized.

**RKHS Regression:** preconditioned update achieves minimax optimal rate in much fewer steps (i.e. faster decay in population risk).

**Neural Networks:** empirical trends matching our theoretical analysis.

## Future directions:

- Understand time-varying preconditioners (e.g. adaptive methods)
- Characterize additional factors (gradient noise, step size, etc.)
- Combine analysis with explicit regularization.

Companion work: Wu, D. and Xu, J. (2020). On the Optimal Weighted  $\ell_2$  Regularization in Overparameterized Linear Regression. *NeurIPS 2020*.

## Additional Reference

- Amari, S.I., 1998. Natural gradient works efficiently in learning.
- Rubio, F. and Mestre, X., 2011. Spectral convergence for a general class of random matrices.
- Martens, J., 2014. New insights and perspectives on the natural gradient method.
- Wilson, A.C., Roelofs, R., Stern, M., Srebro, N. and Recht, B., 2017. The marginal value of adaptive gradient methods in machine learning.
- Dobriban, E. and Wager, S., 2018. High-dimensional asymptotics of prediction: Ridge regression and classification.
- Jacot, A., Gabriel, F. and Hongler, C., 2018. Neural tangent kernel: Convergence and generalization in neural networks.
- Arora, S., Du, S.S., Hu, W., Li, Z. and Wang, R., 2019. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks.
- Pagliana, N. and Rosasco, L., 2019. Implicit Regularization of Accelerated Methods in Hilbert Spaces.