When Does Preconditioning Help or Hurt Generalization?

University of Toronto · 東京大学 The University of Tokyo · Columbia University · Vector Institute

Shun-ichi Amari[1], Jimmy Ba[2,3], Roger Grosse[2,3], Xuechen Li[4], Atsushi Nitanda[5,6], Taiji Suzuki[5,6], Denny Wu[2,3], Ji Xu[7]

RIKEN CBS[1], University of Toronto[2], Vector Institute[3], Google Brain[4], University of Tokyo[5], RIKEN AIP[6], Columbia University[7]

Introduction: Preconditioned Gradient Descent

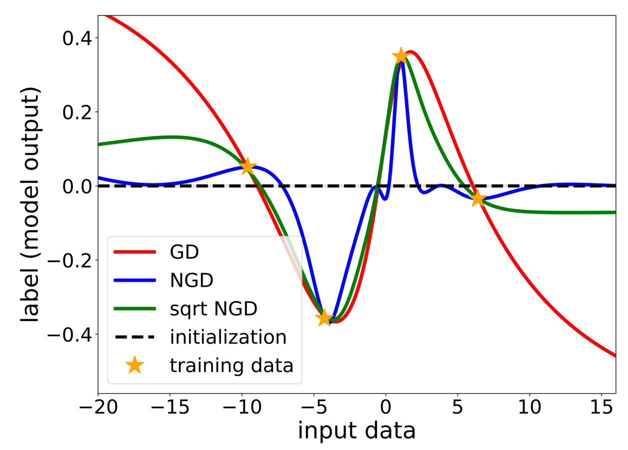
Update Rule: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{P}(t)\nabla_{\boldsymbol{\theta}_t} L(f_{\boldsymbol{\theta}_t}), \quad t = 0, 1, \ldots.$

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

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

Implicit Bias of Preconditioned Updates:

- Modern ML models (e.g. neural nets) are often **overparameterized**.
- Overparameterized models may **interpolate** training data *in different ways*.
- $\boldsymbol{P}$ alters properties of the interpolant.



1D illustration of implicit bias.

Motivation of this work:

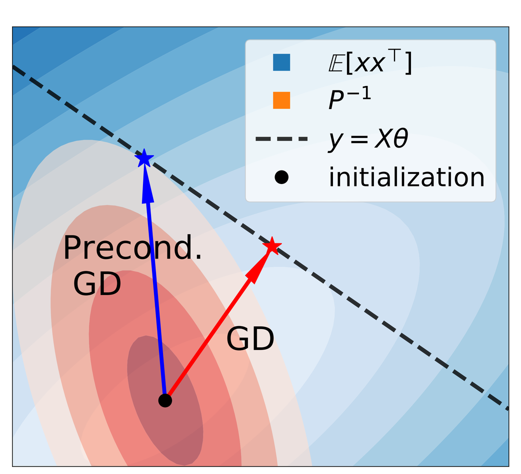
- How does *preconditioning affects generalization* under *interpolation*?
- Can we determine the *optimal preconditioner* for generalization?

Implicit Bias in Least Squares Regression

- **Student-teacher Setup.** $y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}_\star + \varepsilon_i,\ 1 \le i \le n;\ \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{\Sigma}_x \in \mathbb{R}^{d\times d}$.
- **Overparameterized Asymptotics.** $n, d \to \infty,\ d/n \to \gamma \in (1, \infty)$.
- **Update Rule.** Preconditioned gradient descent on *squared loss*:
$$\mathrm{d}\boldsymbol{\theta}(t) = \boldsymbol{P}(t)\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}(t))\mathrm{d}t, \quad \boldsymbol{\theta}(0) = 0.$$

Stationary Solution ($t \to \infty$):

- **Gradient descent:** min ℓ_2-norm interpolant.
- **Preconditioned GD:** for time-invariant and full-rank $\boldsymbol{P}$ ⇒ minimum $\|\boldsymbol{\theta}\|_{\boldsymbol{P}^{-1}}$ interpolant.

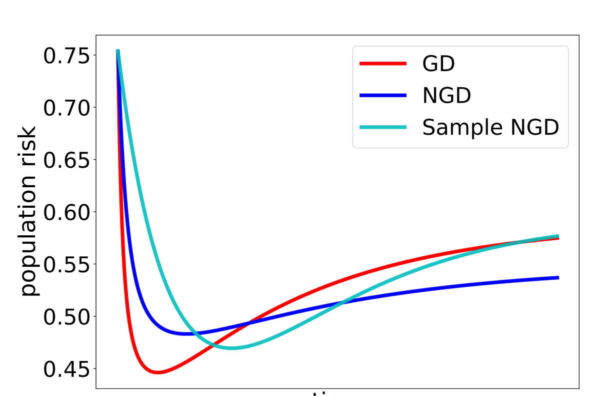


Common Argument: min ℓ_2-norm solution generalizes well; therefore GD is better than preconditioned updates!

Question: Why is the ℓ_2 norm the best measure of generalization?

Noticeable Examples of Preconditioner:

- **Identity:** $\boldsymbol{P} = \boldsymbol{I}_d$ gives the min ℓ_2 norm interpolant (also true for momentum GD and SGD).
- **Population Fisher:** $\boldsymbol{P} = \boldsymbol{F}^{-1} = \boldsymbol{\Sigma}_x^{-1}$ (NGD).
- **Variants of Sample Fisher:** $\boldsymbol{P} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1}$ leads to the same solution as GD.



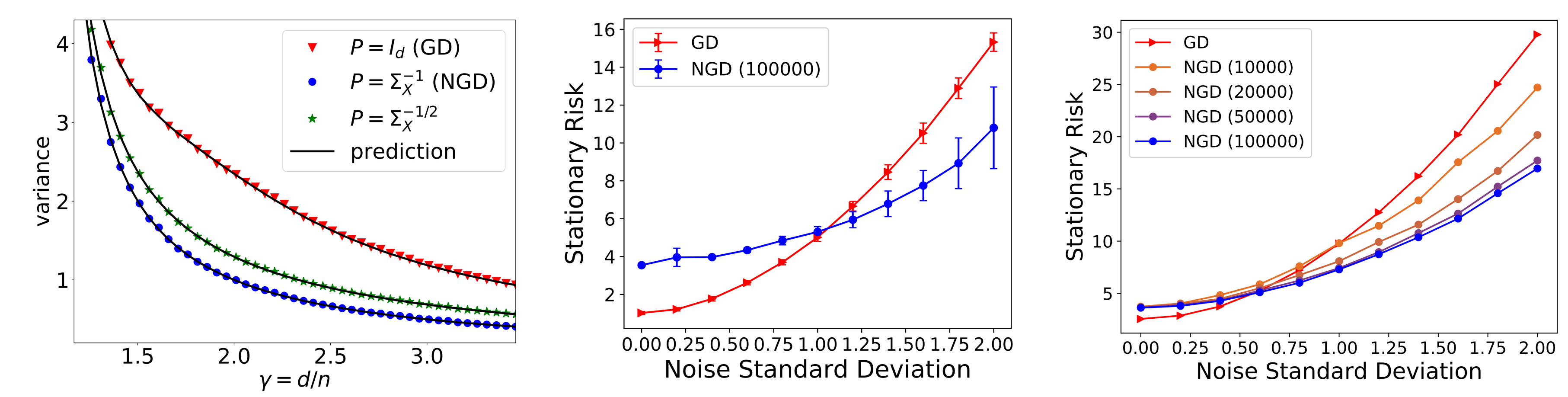
Bias-variance Decomposition of Generalization Error

Thm. (informal). Prediction risk of the min $\|\boldsymbol{\theta}\|_{\boldsymbol{P}^{-1}}$ solution is given as
$$\mathbb{E}(\langle \boldsymbol{x}, \boldsymbol{\theta}_\star\rangle - \langle \boldsymbol{x}, \hat{\boldsymbol{\theta}}_{\boldsymbol{P}^{-1}}\rangle)^2 \xrightarrow{P} m_0' m_0^{-2}\Big(\underbrace{\gamma\mathbb{E}[v_x v_\theta (v_{xp}\cdot m_0 + 1)^{-2}]}_{\text{bias}} + \underbrace{\tilde{\sigma}^2}_{\text{variance}}\Big),$$
where m_0 is the Stieltjes transform of $\frac{1}{n}\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top$ evaluated at $\lambda \to 0_+$.

- **Bias term:** "Difficulty" in learning the *teacher model* $\boldsymbol{\theta}_\star$.
- **Variance term:** "Stability" of learning under *label noise*.

Variance term: NGD is Optimal

Thm. (informal). NGD ($\boldsymbol{P} = \boldsymbol{F}^{-1}$) achieves lowest stationary variance.



(a) Linear regression. (b) 2-layer MLP (MNIST). (c) 2-layer MLP (CIFAR).

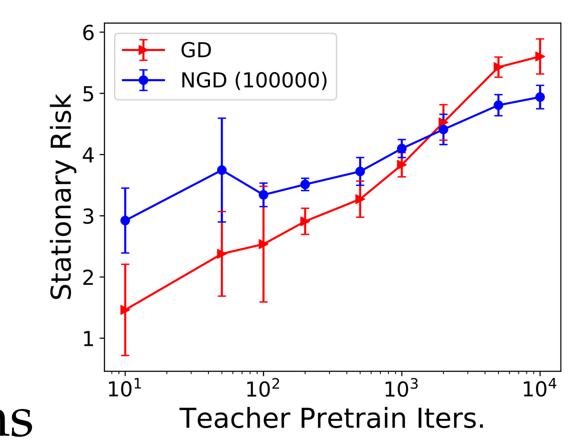
- **(a)(b)**: labels are noisy (risk is *variance-dominated*) ⇒ NGD beneficial.
- **(c)**: this advantage is present only for the **population Fisher.**

Misspecification ≈ Label Noise:

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

Creating Misspecification in Neural Network:

- **Student:** small two-layer MLP.
- **Teacher:** ResNet-20 at varying training epochs



Bias term: No Free Lunch

General Prior: $\mathbb{E}[\boldsymbol{\theta}_\star\boldsymbol{\theta}_\star^\top] = d^{-1}\boldsymbol{\Sigma}_\theta$, i.e. computing the *Bayes risk*.

Thm. (informal). Among all $\boldsymbol{P}$ codiagonalizable with $\boldsymbol{\Sigma}_x$, *bias* is minimized by $\boldsymbol{P} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{U}^\top\boldsymbol{\Sigma}_\theta\boldsymbol{U})\boldsymbol{U}^\top$, where $\boldsymbol{U}$ is the eigenvectors of $\boldsymbol{\Sigma}_x$.

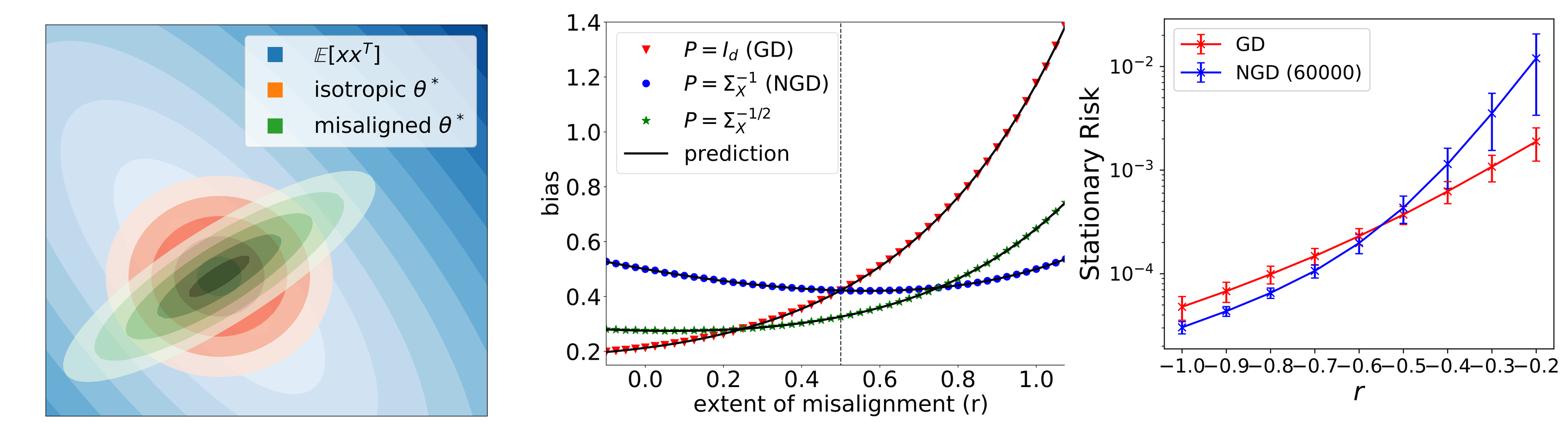
Remark: Setup extends previously assumed *isotropic prior* [Dobriban and Wager 18].

No-free-lunch: The optimal preconditioner depends on the "orientation" of teacher model $\boldsymbol{\theta}_\star$, which is usually not known *a priori*.

Bias Term (continued): Alignment & Source Condition

- GD achieves optimal bias when teacher is **isotropic**: $\boldsymbol{\Sigma}_\theta = \boldsymbol{I}_d$.
- NGD is optimal under **misalignment**: $\boldsymbol{\Sigma}_\theta = \boldsymbol{\Sigma}_x^{-1}$ ("hard" problem).

Remark: We also show that this trend is roughly preserved under **early stopping**.



(a) Intuition of "alignment". (b) Linear regression. (c) 2-layer MLP (MNIST).

Analogy to Source Condition ($\mathbb{E}\|\boldsymbol{\Sigma}_x^{-r/2}\boldsymbol{\theta}_\star\|_2 < \infty$):

Prop. (informal). Consider $\boldsymbol{\Sigma}_\theta = \boldsymbol{\Sigma}_x^{-r}$. Then for some $r^* \in (0, 1)$, NGD achieves lower (higher) bias than GD if and only if $r > (<)\, r^*$.

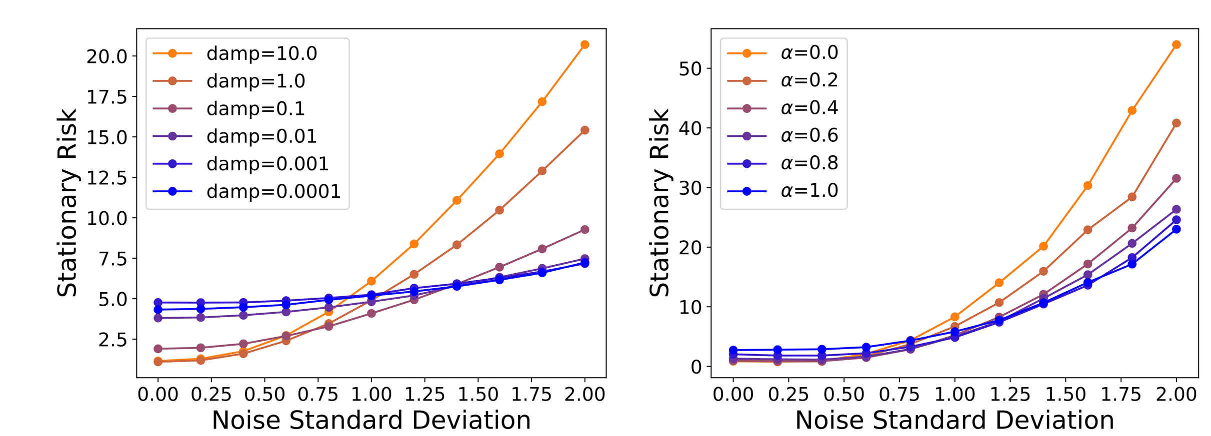
"Interpolating" Between GD and NGD

Question: Is it advantageous to "combine" GD and NGD?

Bias-variance Tradeoff:

- **Additive interp.:** $\boldsymbol{P}_\alpha = (\boldsymbol{\Sigma}_x + \alpha\boldsymbol{I}_d)^{-1}$.
- **Geometric interp.:** $\boldsymbol{P}_\alpha = \boldsymbol{\Sigma}_x^{\alpha-1}$.

- Large α ⇒ GD-like update.
- Small α ⇒ NGD-like update.



Additive interp. (MLP). Geometric interp. (MLP).

Message: At some SNR, *interpolating* between GD and NGD is beneficial.

Fast Decay in Population Risk:

Consider the following preconditioned update in the RKHS.
$$f_t = f_{t-1} - \eta(\Sigma + \alpha I)^{-1}(\hat{\Sigma} f_{t-1} - \hat{S}^* Y), \quad f_0 = 0. \quad f_t \in \mathcal{H}.$$

Remark: Update corresponds to *additive interpolation* between GD and NGD.

Thm. (informal). Preconditioned GD with *properly chosen* α achieves the minimax optimal rate $R(f_t) = \|Sf_t - f^*\|^2_{L_2(P_X)} = \tilde{O}\left(n^{-\frac{2rs}{2rs+1}}\right)$ in $t = \Theta(\log n)$ steps, whereas ordinary GD requires $t = \Theta\left(n^{\frac{2rs}{2rs+1}}\right)$ steps.

Message: Preconditioning can improve the *efficiency of learning*.