Explicit and Implicit Regularization in Overparameterized Least Squares Regression

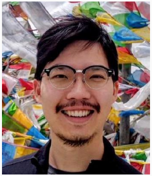
Denny Wu

University of Toronto
Vector Institute for Artificial Intelligence
https://www.cs.toronto.edu/~dennywu/

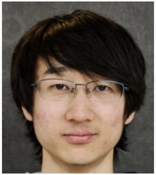
Introduction

- Wu, D. and Xu, J., "On the optimal weighted ℓ_2 regularization in overparameterized linear regression." **NeurIPS 2020**.
- Amari, S., Ba, J., Grosse, R., Li, X., Nitanda, A., Suzuki, T., Wu, D., and Xu, J., "When does preconditioning help or hurt generalization?" **ICLR 2021**.

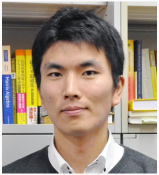

Shun-ichi Amari

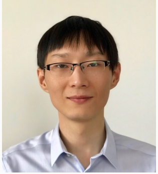

Jimmy Ba


Roger Grosse


Xuechen Li


Atsushi Nitanda


Taiji Suzuki


Ji Xu

Introduction

Task: given n training samples and p parameters to be estimated, characterize the **generalization performance** of the empirical risk minimizer.

- **Classical Large-sample Limit:** $n \to \infty$ under fixed p.
- **Proportional Asymptotic Limit:** $n, p \to \infty,\ p/n \to (0, \infty)$.

Why do we care about the proportional limit?

- Modern machine learning systems are often **overparameterized**.
- Many interesting phenomena can be precisely analyzed in this regime.

This Talk: least squares regression in the overparameterized regime:

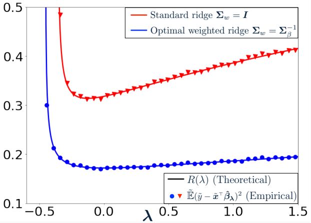
- (generalized) ridge regression: what is the optimal *explicit* regularization?
- (weighted) ridgeless interpolant: what is the optimal *implicit* regularization?

On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression

Denny Wu and Ji Xu.

(NeurIPS 2020)

- Rigorous explanation of the observation that the optimal λ in ridge regression can be **negative**.
- Characterization of the *optimal* weighted shrinkage under overparameterization.



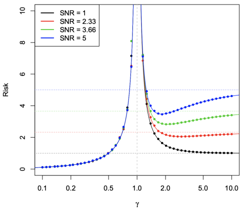
Surprises in Overparameterized Least Squares Regression

Motivating Example – Ridge Regression: given feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and response $\boldsymbol{y} \in \mathbb{R}^n$, estimate the true parameters via

$$\hat{\theta} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

What happens in the overparameterized regime, i.e. $\gamma = d/n > 1$?

- **Intuition (classical):** more overparameterized model (larger γ) $\Rightarrow$ more regularization required (larger λ).
- **Reality:** without regularization ($\lambda \to 0$), the population risk may **decrease** as γ increases.



Message: *estimators in the overparamterized regime can generalize* (in the absence of explicit regularization)

- M. Belkin, D. Hsu, S. Ma, S. Mandal. *Reconciling modern machine learning and the bias-variance trade-off.*
- T. Hastie, A. Montanari, S. Rosset, R. Tibshirani. *Surprises in high-dimensional ridgeless interpolation.*

Implicit Regularization of Overparameterization

One explanation: overparameterization $\Rightarrow$ *implicit ℓ_2 regularization* (?)

Example: Let $y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}_\star + \varepsilon_i$, where $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_d)$. Let $\gamma = d/n > 1$ and $\hat{\boldsymbol{\theta}}$ be the minimum ℓ_2 norm solution,

$$\mathbb{E}[\|\hat{\boldsymbol{\theta}}\|_2^2 | \boldsymbol{X}] \to \|\boldsymbol{\theta}_\star\|_2^2/\gamma + \mathrm{Var}(\varepsilon)/(\gamma - 1), \quad \text{as } n, d \to \infty$$

which is a **decreasing function** of γ.

Rough intuition: larger $\gamma \approx$ stronger (implicit) ℓ_2 regularization.

Question: Can optimal regularization be ***negative*** ($\lambda < 0$) when $d > n$?

- **Empirically?** Yes! “Negative ridge” phenomenon [Kobak et al. 2020].
- **Theoretically?** Not yet! Requires more general setup (this work).

- Kobak et al. 2020. *Optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization.*

Problem Setup and Assumptions

- **Data model:** $y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}_\star + \varepsilon_i$, $1 \le i \le n$; $\boldsymbol{x}_i \in \mathbb{R}^d$.
- **Estimator:** generalized ridge regression

$$\hat{\boldsymbol{\theta}}_\lambda = \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{\Sigma}_w\right)^\dagger \boldsymbol{X}^\top \boldsymbol{y}.$$

- **Goal:** characterize the prediction risk $R(\hat{\boldsymbol{\theta}}_\lambda) = \mathbb{E}_{\tilde{x},\tilde{\epsilon},\boldsymbol{\theta}_\star}(\tilde{y} - \tilde{\boldsymbol{x}}^\top \hat{\boldsymbol{\theta}}_\lambda)^2$.

Remark: When $\lambda \ge 0$, $\hat{\boldsymbol{\theta}}_\lambda = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2 + \lambda \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_w \boldsymbol{\theta}$.

Basic Assumptions (A1):

- **Proportional Asymptotics:** $n, d \to \infty$, $d/n \to \gamma \in (1, \infty)$.
- **Random Design:** $\boldsymbol{x}_i = \boldsymbol{z}_i \boldsymbol{\Sigma}_x^{1/2}/\sqrt{n}$, $\boldsymbol{z}_i \overset{\text{i.i.d.}}{\sim} P_z$ with zero-mean and bounded 12th moment. $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$.
- **General Prior:** $\mathbb{E}[\boldsymbol{\theta}_* \boldsymbol{\theta}_*^\top] = \boldsymbol{\Sigma}_\theta$. Note that this assumption covers both *deterministic* and *random* $\boldsymbol{\theta}_*$.

Motivation: Generalized Ridge Regression

- Known formulation, but analysis under **overparameterization** lacking.
- For $\lambda > 0$, equivalent to Gaussian prior with **general covariance** on $\hat{\theta}$.

The formulation covers:

- **Standard ridge regression**: $\mathbf{\Sigma}_w = \boldsymbol{I}_d$.
- **Principal Component Regression (PCR)**: discard lower eigendirections by applying large penalty.
- **Algorithms in Deep Learning**: connection to decoupled weight decay and elastic weight consolidation.

Motivation of This Work:

- What is the *optimal weighting matrix* $\mathbf{\Sigma}_w$ for the prediction risk?
- Can we show the *benefit of weighted shrinkage* over other approaches?

- I. Loshchilov, F. Hutter, *Decoupled weight decay regularization*.
- Kirkpatrick et al. 2017. *Overcoming catastrophic forgetting in neural networks*.

Motivation: Anisotropic Prior

For standard ridge regression, λ is **provably non-negative** under

- Isotropic signal $\boldsymbol{\Sigma}_\theta = \boldsymbol{I}_d$ [Dobriban and Wager 2018].
- Isotropic data $\boldsymbol{\Sigma}_x = \boldsymbol{I}_d$ [Hastie et al. 2019].

Motivation of This Work:

- Can we precisely characterize the *"negative ridge" phenomenon*?

Relation between $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\theta$ is analogous to the **source condition** in RKHS literature: $\mathbb{E}\|\boldsymbol{\Sigma}_x^{-\alpha/2}\boldsymbol{\theta}_*\| < \infty$.

Motivation of This Work:

- How does the *alignment* between $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\theta$ (α in source condition) affect the optimal regularization strength λ?

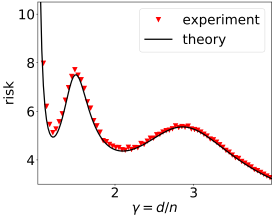
- **Concurrent work:** Richards, D., Mourtada, J. and Rosasco, L., 2020. *Asymptotics of Ridge (less) Regression under General Source Condition*.

Benefit of General Setup

"Multiple Descent" Risk Curve

- By manipulating $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_\theta$, the prediction risk can be highly **non-monotonic w.r.t.** γ, i.e. level of overparameterization.

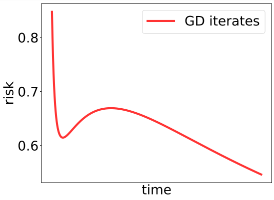
Remark: when $\mathbf{\Sigma}_x$ is isotropic, the risk *does not* exhibit multiple peaks for $\gamma > 1$.



Epoch-wise Double Descent

- Gradient descent (flow) on the least squares objective may lead to prediction risk **non-monotonic in time**, even if $\sigma = 0$.

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



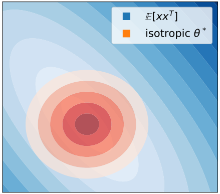
Alignment between Feature and Signal

(A2) Converging Eigenvalues: empirical distributions of $(\boldsymbol{d}_{x/w}, \boldsymbol{d}_{w\theta})$ jointly converge to bounded r.v. $(\upsilon_{x/w}, \upsilon_{w\theta})$, where $\upsilon_{x/w} \geq c_l > 0$, $\boldsymbol{d}_{w\theta} = \text{diag}\left(\boldsymbol{U}_{x/w}\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{\Sigma}_\theta\boldsymbol{\Sigma}_w^{1/2}\boldsymbol{U}_{x/w}^\top\right)$, and $\boldsymbol{d}_{x/w}$ and $\boldsymbol{U}_{x/w}$ are eigenvalues and eigenvectors of $\boldsymbol{\Sigma}_w^{-1/2}\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_w^{-1/2}$.

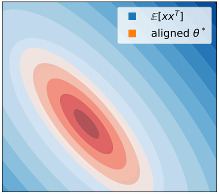
Intuition: when $\boldsymbol{\Sigma}_w = \boldsymbol{I}_d$ (i.e., standard ridge regression),

- $\boldsymbol{d}_{x/w}$ (or $\upsilon_{x/w}$): eigenvalues of $\boldsymbol{\Sigma}_x$.
- $\boldsymbol{d}_{w\theta}$ (or $\upsilon_{w\theta}$): projection of target $\boldsymbol{\beta}_*$ onto eigenvectors of $\boldsymbol{\Sigma}_x$.

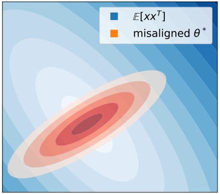
Definition of Alignment: For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, we say $\boldsymbol{a}$ is aligned (misaligned) with $\boldsymbol{b}$ when $a_i \geq a_j$ iff $b_i \gtreqless b_j$ for all i,j.



Isotropic (previous work).



Aligned (easy problem).



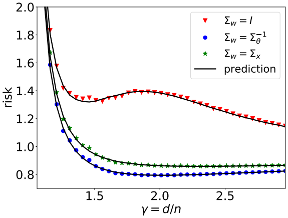
Misaligned (hard problem).

Characterization of Prediction Risk

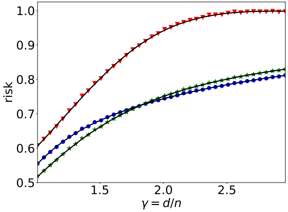
Thm. Under (A1-2), the asymptotic prediction risk $R(\hat{\theta}_\lambda)$ is given as

$$\tilde{\mathbb{E}}\left(\tilde{y} - \tilde{\boldsymbol{x}}^\top \hat{\boldsymbol{\theta}}_\lambda\right)^2 \xrightarrow{p} \frac{m'(-\lambda)}{m^2(-\lambda)} \Big(\underbrace{\gamma \mathbb{E}[\upsilon_{x/w} \upsilon_{w\theta} (\upsilon_{x/w} \cdot m(-\lambda) + 1)^{-2}]}_{\textit{bias}} + \underbrace{\tilde{\sigma}^2}_{\textit{variance}} \Big),$$

$\forall \lambda > -c_0$, where $c_0 = (\sqrt{\gamma} - 1)^2 c_l$, and $m(z) > 0$ is the *Stieltjes transform* of the limiting distribution of the eigenvalues of $\boldsymbol{X}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{X}^\top$.



$\lambda = 0$.



$\lambda = 0.1$.

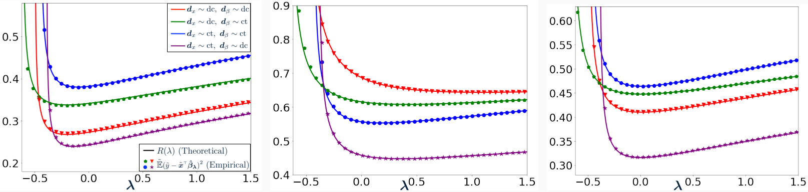
- Regularization *suppresses* the double descent peak [Krogh and Hertz 1992].
- Weighted regularization often dominates standard isotropic shrinkage (red).

When is Optimal λ_{opt} Negative?

Theorem (informal). When the risk is dominated by the *bias* term,

- $\underline{\lambda_{\text{opt}} < 0}$ when $\boldsymbol{d}_{x/w}$ is **aligned** with $\boldsymbol{d}_{w\theta}$.
- $\underline{\lambda_{\text{opt}} > 0}$ when $\boldsymbol{d}_{x/w}$ is **misaligned** with $\boldsymbol{d}_{w\theta}$.
- $\underline{\lambda_{\text{opt}} = 0}$ when the order is **random**, i.e. $\mathbb{E}[\upsilon_{w\theta}|\upsilon_{wx}] \overset{a.s.}{=} \mathbb{E}[\upsilon_{w\theta}]$.

Example: Consider $\boldsymbol{\Sigma}_\theta = \boldsymbol{\Sigma}_x^r$, then for the *bias* term $\lambda_{\text{opt}} \gtreqless 0$ iff $r \gtreqless 0$.



(a) Aligned, noiseless **(b)** Misaligned, noiseless **(c)** Random, noiseless

Remark: for the *variance* term λ_{opt} is always **non-negative**.

When is Optimal λ_{opt} Negative?

Comparison with previous works: when $\boldsymbol{\Sigma}_x = \boldsymbol{I}_d$ or $\boldsymbol{\Sigma}_\theta = \boldsymbol{I}_d$,

- $\lambda_{\text{opt}} = 0$ if $\sigma = 0$, i.e. *interpolation is optimal* when label is clean.
- $\lambda_{\text{opt}} > 0$ if $\sigma > 0$, i.e. *positive regularization* is required for noisy data.

Our findings under more general setup: given $\boldsymbol{\Sigma}_w = \boldsymbol{I}_d$,

- **Negative** λ is beneficial when features are useful (*"easy" problem*); consequently, interpolation can be optimal even if $\sigma > 0$.
- **Positive** λ is beneficial under misalignment (*"hard" problem*), even in the *absence of label noise* ($\sigma = 0$).

Bias-variance Tradeoff: as σ increases, the variance term eventually dominates, and λ_{opt} becomes positive.

Properties of λ_{opt} and the Optimal Risk

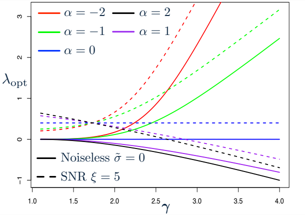
Proposition: when $\gamma < 1$, λ_{opt} is always *non-negative* under (A1-2).

Message: "negative ridge" is a **unique** feature of *overparameterization*.

Implicit ℓ_2 Regularization:

Consider $\boldsymbol{\Sigma}_w = \boldsymbol{I}_d$ and $\boldsymbol{\Sigma}_\theta = \boldsymbol{\Sigma}_x^\alpha$.
Note that larger $\alpha \Rightarrow$ more *aligned* problem.

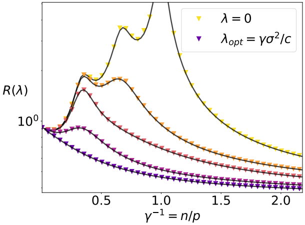
- When $\alpha > 0$ (aligned), λ_{opt} **decreases** as γ increases; vice versa.



Monotonicity of Optimal Risk $R(\lambda_{\text{opt}})$:

Prop. (informal). Given $\boldsymbol{\Sigma}_\theta \propto \frac{1}{d}\boldsymbol{I}_d$ and $\boldsymbol{\Sigma}_w = \boldsymbol{I}_d$, the *optimally regularized* prediction risk $R(\lambda_{\text{opt}})$ is an **increasing** function of $\gamma \in (0, \infty)$.

Message: Optimal ridge regularization (purple) can *suppress multiple descent*.



Optimal Weighting Matrix Σ_w

Questions we aim to address:

- What is the optimal $\boldsymbol{\Sigma}_w$ that minimizes $\min_\lambda R(\hat{\boldsymbol{\theta}}_\lambda)$?
- What is the best $\boldsymbol{\Sigma}_w$ we can construct when knowledge on the true parameters $\boldsymbol{\theta}_*$ is *not available*?

- **(A3) Codiagonalizability:** $\boldsymbol{\Sigma}_x = \boldsymbol{U}\boldsymbol{D}_x\boldsymbol{U}^\top$ and $\boldsymbol{\Sigma}_w = \boldsymbol{U}\boldsymbol{D}_w\boldsymbol{U}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ is orthogonal, and $\boldsymbol{D}_x = \mathrm{diag}(\boldsymbol{d}_x), \boldsymbol{D}_w = \mathrm{diag}(\boldsymbol{d}_w)$.
- **(A4) Converging Eigenvalues:** the empirical distributions of $(\boldsymbol{d}_x, \bar{\boldsymbol{d}}_\theta, \boldsymbol{d}_{x/w})$ jointly converge to non-negative randomly variables $(\upsilon_x, \upsilon_\theta, \upsilon_{x/w})$ upper- and lower-bounded away from 0, in which we defined $\bar{\boldsymbol{d}}_\theta = \mathrm{diag}\big(\boldsymbol{U}^\top\boldsymbol{\Sigma}_\theta\boldsymbol{U}\big)$.

Remark: when $\boldsymbol{\Sigma}_\theta$ is also codiagonalizable with $\boldsymbol{\Sigma}_x$, $\bar{\boldsymbol{d}}_\theta$ corresponds to its eigenvalues, i.e. $\boldsymbol{\Sigma}_\theta = \boldsymbol{U}\boldsymbol{D}_\theta\boldsymbol{U}^\top$ and $\mathrm{diag}(\boldsymbol{D}_\theta) = \bar{\boldsymbol{d}}_\theta$.

Optimal Weighting Matrix Σ_w (continued)

Thm. $\boldsymbol{\Sigma}_w^{-1} = \boldsymbol{U}\,\mathrm{diag}(\bar{\boldsymbol{d}}_\theta)\,\boldsymbol{U}^\top$ is optimal among all $\boldsymbol{\Sigma}_w$ satisfying (A3-4).

- Matches the *maximum a posteriori* estimate.
- Requires knowledge of $\boldsymbol{\Sigma}_\theta$ (not practical).

Question: is there a reasonable $\boldsymbol{\Sigma}_w$ based on $\boldsymbol{\Sigma}_x$, which can be estimated from *unlabeled data*?

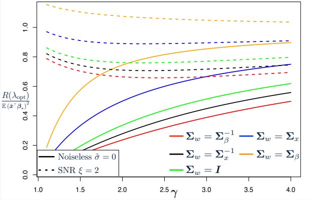
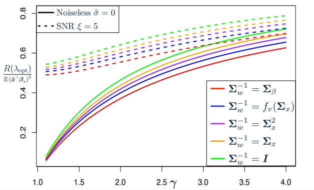


Illustration of optimal $\boldsymbol{\Sigma}_w$.

Coro. $\boldsymbol{\Sigma}_w^{-1} = f(\boldsymbol{\Sigma}_x)$ is optimal among all $\boldsymbol{\Sigma}_w$ only *depending on* $\boldsymbol{\Sigma}_x$, where $f(\upsilon_x) = \mathbb{E}[\upsilon_\theta|\upsilon_x]$ applies to the eigenvalues.

- **Heuristic:** approximate f with polynomial function and cross-validate the parameters.
- When $\mathbb{E}[\upsilon_\theta|\upsilon_x] = \mathbb{E}[\upsilon_\theta]$, $\boldsymbol{\Sigma}_w = \boldsymbol{I}_d$ is reasonable.



Proposed heuristic.

Discussion and Conclusion

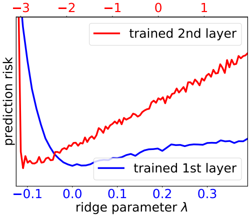
By analyzing *generalized ridge regression* under general setup,

- We determine the sign of the optimal ridge regularization.
  - **Negative ridge** can be beneficial under **aligned** ("easy") problem.
- We characterize the optimal **explicit regularization** $\boldsymbol{\Sigma}_w$.

Future Directions:

- Estimate $\boldsymbol{\Sigma}_w$ based on training samples.
- Extend result to more complicated models, e.g. random features model and neural net.

Remark: benefit of negative regularization is also empirically observed in RF model (red).



two-layer neural network.

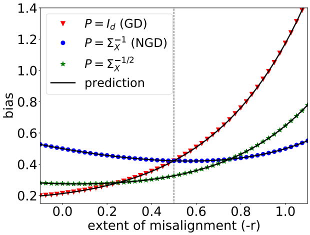
Question: what about **implicit regularization**, i.e. $\lambda \to 0$?

When Does Preconditioning Help or Hurt Generalization?

Shun-ichi Amari, Jimmy Ba, Roger Grosse, Xuechen Li,
Atsushi Nitanda, Taiji Suzuki, Denny Wu, Ji Xu.

(ICLR 2021)

- Precise error analysis of preconditioned least squares regression (*ridgeless*) in the overparameterized regime.
- Empirical validation of theoretical findings in neural networks.



Preconditioned Gradient Descent

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

Common choices of preconditioner $\boldsymbol{P}$ 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**.

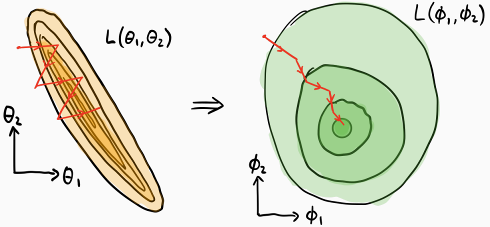


Figure from Xanadu blog post.

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

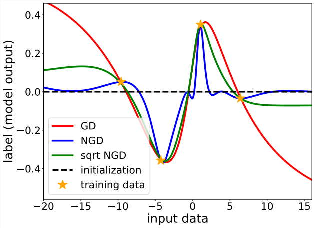
Motivation: Implicit Bias of Optimizers

In the *online learning* setup, efficient optimization ≈ 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*.
- $\boldsymbol{P}$ affects the properties of the interpolant.



Motivation of This Work:

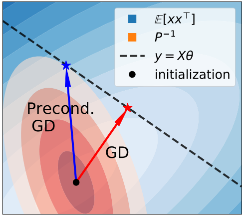
- In the *interpolation setting* (i.e. absence of explicit regularization), how does preconditioning influence the generalization performance?

Implicit Bias in Overparameterized Linear Regression

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

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

- **Gradient descent:** min ℓ_2-norm solution.
- **Preconditioned GD:** for time-independent and full-rank $\boldsymbol{P}$, min $\|\boldsymbol{\theta}\|_{\boldsymbol{P}^{-1}}$ norm solution.



Common Argument: min ℓ_2-norm solution generalizes well $\Rightarrow$ GD ($\boldsymbol{P} = \boldsymbol{I}_d$) is better (e.g. [Wilson et al. 2017]).

Question: Why is the ℓ_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:** $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{\Sigma}_x$; $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, $n, d \to \infty$ and $d/n \to \gamma > 1$.
- **Gradient Update:** $\mathrm{d}\boldsymbol{\theta}(t) = \frac{1}{n}\boldsymbol{P}(t)\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}(t))\mathrm{dt}, \ \ \boldsymbol{\theta}(0) = 0.$

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

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

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

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

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

- *Sample Fisher (Hessian)* $\Leftrightarrow$ **sample covariance** $\boldsymbol{X}^\top\boldsymbol{X}/n$.
- *Population Fisher* $\Leftrightarrow$ **population covariance** $\boldsymbol{\Sigma}_x$.

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

Stationary Solution of Preconditioned Regression

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

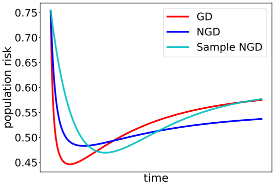
$$\boldsymbol{\theta_P}(t) = \boldsymbol{PX}^\top \left[\boldsymbol{I}_n - \exp\left(-\frac{t}{n}\boldsymbol{XPX}^\top\right)\right](\boldsymbol{XPX}^\top)^{-1}\boldsymbol{y},$$

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

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

Noticeable examples of preconditioned update:

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



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

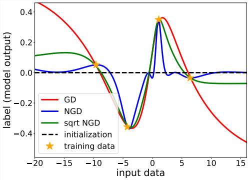
For parametric approximations see talk this afternoon!

Implicit Bias of Natural Gradient Descent

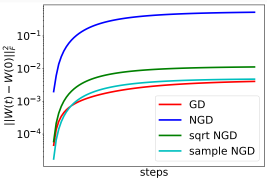
Starting from zero initialization:

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

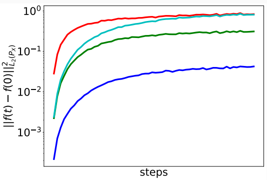
Similar findings also empirically observed in simple *neural networks*:



1D illustration.



Parameter difference.



Function difference.

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

Bias-variance Decomposition

- **Student-teacher setup:** labels are generated by a *teacher model* (target function) with additive noise: $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$.
- **Goal:** determine the optimal preconditioner $\boldsymbol{P}$ under different conditions of label noise and teacher model.

Key observation: $\lim_{\lambda \to 0} \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{P}^{-1}\right)^\dagger \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{P}\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top)^{-1} \boldsymbol{y}$.

⇒ It suffices to analyze the **ridgeless limit** of *generalized ridge regression*.

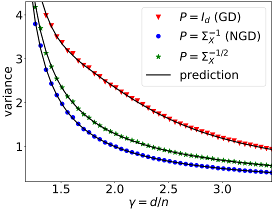
Bias-variance Decomposition:

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

- **Variance** term is due to the *label noise* (independent to the teacher).
- **Bias** term only depends on the teacher model and data distribution.

Variance Term: NGD is Optimal

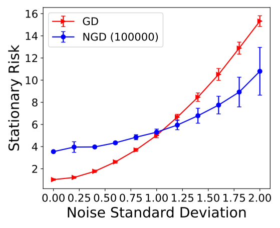
Thm. Given (A1-2), the variance is minimized by NGD: $\boldsymbol{P} = \boldsymbol{F}^{-1} = \boldsymbol{\Sigma}_x^{-1}$.

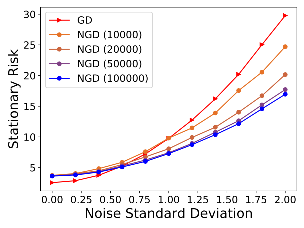


Message: when labels are noisy (risk is dominated by variance), NGD is beneficial.

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

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





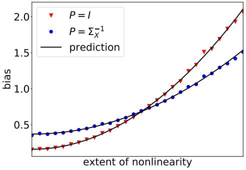
- **Left:** NGD (population Fisher) achieves lower risk under large label noise.
- **Right:** sample Fisher (i.e. less unlabeled data used) behaves like GD.

Misspecification ≈ Label Noise

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

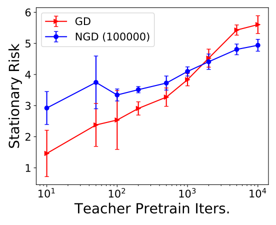
Intuition: f_*^c is "similar" to additive label noise.

Message: NGD is beneficial under misspecification.

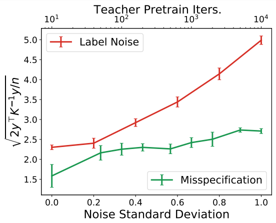


Misspecification in Neural Networks

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



Misspecification on CIFAR-10.



Measure of misspecification.

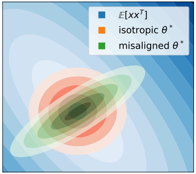
Bias Term: the Well-specified Case

Well-specified Model: $f_*(\boldsymbol{x}) = \boldsymbol{x}^\top \theta_*$. **General prior:** $\mathbb{E}[\theta_* \theta_*^\top] = \Sigma_\theta$.

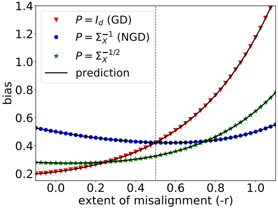
Thm. Under (A1,3,4), the bias is minimized by $\boldsymbol{P} = \boldsymbol{U} \operatorname{diag}\left(\boldsymbol{U}^\top \Sigma_\theta \boldsymbol{U}\right) \boldsymbol{U}^\top$.

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

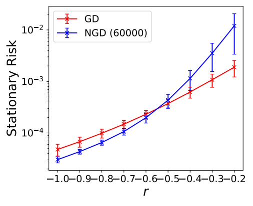
- GD generalizes better when target is **isotropic** $\Sigma_\theta = \boldsymbol{I}_d$.
- NGD is optimal under **misalignment** $\Sigma_\theta = \Sigma_x^{-1}$.



Example (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 when $r > (<)\, r^*$.



Linear regression.



Two-layer MLP (MNIST).

Bias-variance Tradeoff: Interpolating between P

The optimal $\boldsymbol{P}$ for the *bias* and *variance* are in general **different**.

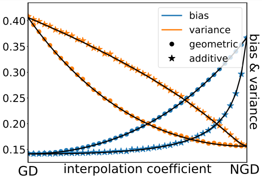
Question: how can we trade in one of bias/variance for the other?

Example: Consider $\boldsymbol{\Sigma}_\theta = \boldsymbol{I}_d$, $\boldsymbol{\Sigma}_x \neq \boldsymbol{I}_d$, and the following interpolation schemes:

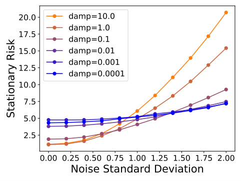
- **Additive:** $\boldsymbol{P}_\alpha = (\alpha\boldsymbol{\Sigma}_x + (1-\alpha)\boldsymbol{I}_d)^{-1}$, corresponds to the *damped inverse*.
- **Geometric:** $\boldsymbol{P}_\alpha = \boldsymbol{\Sigma}_x^{-\alpha}$, covers the "conservative" *square-root scaling*.

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

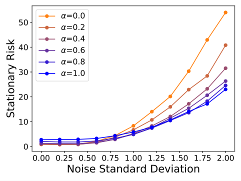
$\Rightarrow$ At certain SNR, **interpolating** between GD and NGD is beneficial.



Monotonicity of bias/variance.



Additive interpolation (MLP).



Geometric interpolation (MLP).

Bias-variance Tradeoff: Early Stopping

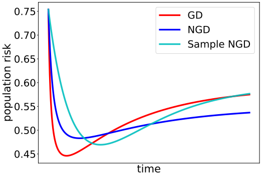
We have thus far only looked at the stationary solution ($t \to \infty$).

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

Proposition (informal). Define $B^{\mathrm{opt}}(\boldsymbol{\theta}) = \inf_{t \geq 0} B(\boldsymbol{\theta}(t))$. Under (A1-4),

1. the variance $V(\boldsymbol{\theta_P}(t))$ *monotonically increases* through time.
2. when $\boldsymbol{\Sigma}_\theta = \boldsymbol{\Sigma}_x^{-1}$ (misaligned), $B^{\mathrm{opt}}(\boldsymbol{\theta_P}) \geq B^{\mathrm{opt}}(\boldsymbol{\theta_{F^{-1}}})$.
3. when $\boldsymbol{\Sigma_\theta} = \boldsymbol{I}_d$ (isotropic), $B^{\mathrm{opt}}(\boldsymbol{\theta_I}) \leq B^{\mathrm{opt}}(\boldsymbol{\theta_{F^{-1}}})$.

- (1) suggests that early stopping is beneficial when data is noisy (due to reduction of variance).
- (2-3) 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:** $y_i = f^*(\boldsymbol{x}_i) + \varepsilon_i$. $S : \mathcal{H} \to L_2(P_X)$. $\Sigma = S^* S$; $L = SS^*$.
- **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* Σ. 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. $\mathrm{tr}\left(\Sigma^{1/s}\right) < \infty$ and $2r + s^{-1} > 1$.
- **Regularity of RKHS:** $\exists \mu \in [s^{-1}, 1], C_\mu > 0$ s.t. $\sup_{\boldsymbol{x}} \|\Sigma^{1/2 - 1/\mu} K_{\boldsymbol{x}}\|_{\mathcal{H}} \leq C_\mu$.

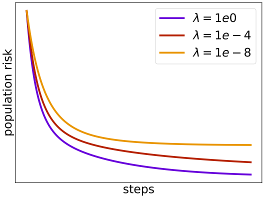
Remark: *source condition* relates to the previously discussed **alignment**:
Large $r \Rightarrow$ smoother teacher model, i.e. "easier" problem; vice versa.

Fast Decay of Population Risk (continued)

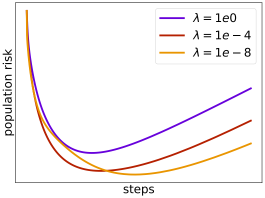
Theorem (informal). Given $\mu \leq 2r$ or $r \geq 1/2$, for sufficiently large n, preconditioned update with $\alpha = n^{-\frac{2s}{2rs+1}}$ achieves the minimax optimal convergence rate $R(f_t) = \|Sf_t - f^*\|^2_{L_2(P_X)} = \tilde{O}\left(n^{-\frac{2rs}{2rs+1}}\right)$ in $t = \Theta(\log n)$ steps, 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 α and stopping time t are chosen to *balance the bias and variance*.
- α **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 (step size, explicit regularization, etc.)

Caution: properties of linear or kernel model *may not* translate to neural network...

See talks this afternoon!

Additional Reference

- Krogh and Hertz 1992. *A simple weight decay can improve generalization*.
- Amari 1998. *Natural gradient works efficiently in learning*.
- Rubio and Mestre 2011. *Spectral convergence for a general class of random matrices*.
- Martens 2014. *New insights and perspectives on the natural gradient method*.
- Wilson et al. 2017. *The marginal value of adaptive gradient methods in machine learning*.
- Dobriban and Wager 2018. *High-dimensional asymptotics of prediction: Ridge regression and classification*.
- Jacot et al. 2018. *Neural tangent kernel: Convergence and generalization in neural networks*.
- Chizat and Bach 2018. *On Lazy Training in Differentiable Programming*.
- Arora et al. 2019. *Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks*.
- Xu and Hsu 2019. *On the number of variables to use in principal component regression*.
- Mei and Montanari 2019. *The generalization error of random features regression: Precise asymptotics and double descent curve*.
- Yang, G. and Hu, E. J., 2020. *Feature learning in infinite-width neural networks*.