On the Optimal Weighted $\ell_2$ Regularization in Overparameterized Linear Regression

Denny Wu\textsuperscript{1} and Ji Xu\textsuperscript{2}

\textsuperscript{1}University of Toronto and Vector Institute
\textsuperscript{2}Columbia University

Conference on Neural Information Processing Systems 2020
Motivating Example – Ridge Regression Estimator: given feature matrix $X \in \mathbb{R}^{n \times d}$ and response $y \in \mathbb{R}^{n}$, estimate the true parameters via

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

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

- **Intuition (classical):** more overparameterized model (larger $\gamma$) $\Rightarrow$ more regularization required (larger $\lambda$).

- **Reality:** without regularization ($\lambda \to 0$), the population risk may decrease as $\gamma$ increases.

**Message:** *estimators in the overparameterized 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.
Implicit Regularization of Overparameterization

**One explanation:** overparameterization $\Rightarrow$ *implicit* $\ell_2$ regularization.

**Example:** Let $y_i = x_i^\top \theta^* + \epsilon_i$, where $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$. Let $\gamma = d/n > 1$ and $\hat{\theta}$ be the minimum $\ell_2$ norm solution,

$$
E[\|\hat{\theta}\|_2^2 | X] \to \|\theta^*\|_2^2 / \gamma + \text{Var}(\epsilon)/(\gamma - 1), \quad \text{as } n, d \to \infty
$$

which is a *decreasing function* of $\gamma$.

**Intuition (non-rigorous):** larger $\gamma \approx$ stronger (implicit) $\ell_2$ 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.
• **Data model:** \( y_i = x_i^\top \theta_\star + \varepsilon_i, 1 \leq i \leq n; \ x_i \in \mathbb{R}^d. \)

• **Estimator:** generalized ridge regression

\[
\hat{\theta}_\lambda = (X^\top X + \lambda \Sigma_w)^{\dagger} X^\top y.
\]

• **Goal:** characterize the prediction risk

\[
R(\hat{\theta}_\lambda) = \mathbb{E}_{\tilde{x}, \tilde{\varepsilon}, \theta_\star}(\tilde{y} - \tilde{x}^\top \hat{\theta}_\lambda)^2.
\]

**Remark:** When \( \lambda \geq 0, \ \hat{\theta}_\lambda = \arg \min_\theta \sum_{i=1}^n (y_i - x_i^\top \theta)^2 + \lambda \theta^\top \Sigma_w \theta. \)

**Basic Assumptions (A1):**

• **Proportional Asymptotics:** \( n, d \to \infty, \ d/n \to \gamma \in (1, \infty). \)

• **Random Design:** \( x_i = z_i \Sigma_x^{1/2} / \sqrt{n}, \ z_i \text{i.i.d.} \sim P_z \) with zero-mean and bounded 12th moment. \( \mathbb{E}[\varepsilon] = 0, \ \text{Var}(\varepsilon) = \sigma^2. \)

• **General Prior:** \( \mathbb{E}[\theta_\star \theta_\star^\top] = \Sigma_\theta. \) Note that this assumption covers both *deterministic* and *random* \( \theta_\star. \)
Motivation: Generalized Ridge Regression

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

The formulation covers:

- **Standard ridge regression**: $\Sigma_w = I_d$.
- **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* $\Sigma_w$ for the prediction risk?
- Can we show the *benefit of weighted shrinkage* over other approaches?

- I. Loshchilov, F. Hutter, Decoupled weight decay regularization.
Motivation: Anisotropic Prior

For standard ridge regression, $\lambda$ is **provably non-negative** under
- Isotropic signal $\Sigma_\theta = I_d$ [Dobriban and Wager 2018].
- Isotropic data $\Sigma_x = I_d$ [Hastie et al. 2019].

**Motivation of This Work:**
- Can we precisely characterize the “negative ridge” phenomenon?

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

**Motivation of This Work:**
- How does the *alignment* between $\Sigma_x$ and $\Sigma_\theta$ ($\alpha$ in source condition) affect the optimal regularization strength $\lambda$?

- **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 $\Sigma_x$ and $\Sigma_\theta$, the prediction risk can be highly **non-monotonic w.r.t.** $\gamma$, i.e. the level of overparameterization.

**Remark:** when $\Sigma_x$ 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 $\Sigma_x$ or $\Sigma_\theta$ is isotropic, the bias term is *monotonically decreasing* through time.
Alignment between Feature and Signal

(A2) **Converging Eigenvalues:** empirical distributions of \((d_{x/w}, d_{w\theta})\) jointly converge to bounded r.v. \((\nu_{x/w}, \nu_{w\theta})\), where \(\nu_{x/w} \geq c_l > 0\), \(d_{w\theta} = \text{diag}\left(U_{x/w} \Sigma_w^{1/2} \Sigma_{\theta} \Sigma_w^{1/2} U_{x/w}^\top\right)\), and \(d_{x/w}\) and \(U_{x/w}\) are eigenvalues and eigenvectors of \(\Sigma_w^{-1/2} \Sigma_x \Sigma_w^{-1/2}\).

**Intuition:** when \(\Sigma_w = I_d\) (i.e., standard ridge regression),
- \(d_{x/w}\) (or \(\nu_{x/w}\)): eigenvalues of \(\Sigma_x\).
- \(d_{w\theta}\) (or \(\nu_{w\theta}\)): projection of target \(\beta^*_\star\) onto eigenvectors of \(\Sigma_x\).

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

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

$$
\tilde{E}\left(\tilde{y} - \tilde{x}^\top \hat{\theta}_\lambda \right)^2 \xrightarrow{p} \frac{m'(-\lambda)}{m^2(-\lambda)} \left( \gamma E\left[\frac{\nu_{x/w} v_{w\theta}(\nu_{x/w} \cdot m(-\lambda) + 1)^{-2}}{\nu_{x/w}}\right] + \tilde{\sigma}^2 \right),
$$

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

- Regularization *suppresses* the double descent peak [Krogh and Hertz 1992].
- Weighted regularization often dominates standard isotropic shrinkage (red).
When is Optimal $\lambda_{opt}$ Negative?

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

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

**Example:** Consider $\Sigma_{\theta} = \Sigma_{\xi}^r$, then for the *bias* term $\lambda_{opt} \gtrless 0$ iff $r \gtrless 0$.

**Remark:** for the *variance* term $\lambda_{opt}$ is always **non-negative**.
Comparison with previous works: when $\Sigma_x = I_d$ or $\Sigma_\theta = I_d$,

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

Our findings under more general setup: given $\Sigma_w = I_d$,

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

Bias-variance Tradeoff: as $\sigma$ increases, the variance term eventually dominates, and $\lambda_{opt}$ becomes positive.
Properties of $\lambda_{opt}$ and the Optimal Risk

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

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

**Implicit $\ell_2$ Regularization:**

Consider $\Sigma_w = I_d$ and $\Sigma_{\theta} = \Sigma_{x}^{\alpha}$.

Note that larger $\alpha \Rightarrow$ more aligned problem.

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

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

**Prop. (informal).** Given $\Sigma_{\theta} \propto \frac{1}{d} I_d$ and $\Sigma_w = I_d$, the **optimally regularized** prediction risk $R(\lambda_{opt})$ is an increasing function of $\gamma \in (0, \infty)$.

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

---

**Graphical Representation:**

- **Graph 1:** Comparison of $\lambda_{opt}$ for different $\alpha$ values ($\alpha = -2, -1, 0, 1, 2$) at $\gamma = 0$ and $\gamma = 5$ for noiseless and SNR $\xi = 5$.

- **Graph 2:** Variation of $R(\lambda)$ with $\gamma^{-1} = n/p$ for $\lambda = 0$ and $\lambda_{opt} = \gamma \sigma^2/c$. 

---

12
Optimal Weighting Matrix $\Sigma_w$

Questions we aim to address:

- What is the optimal $\Sigma_w$ that minimizes $\lim_{\lambda \to 0} R(\hat{\theta}_\lambda)$?
- What is the optimal $\Sigma_w$ that minimizes $\min_{\lambda} R(\hat{\theta}_\lambda)$?
- What is the best $\Sigma_w$ we can construct when knowledge on the true parameters $\theta_*$ is not available?

- **(A3) Codiagonalizability:** $\Sigma_x = UD_x U^\top$ and $\Sigma_w = UD_w U^\top$, where $U \in \mathbb{R}^{d \times d}$ is orthogonal, and $D_x = \text{diag}(d_x)$, $D_w = \text{diag}(d_w)$.

- **(A4) Converging Eigenvalues:** empirical distributions of $(d_x, \tilde{d}_\theta, d_{x/w})$ jointly converge to non-negative randomly variables $(\upsilon_x, \upsilon_\theta, \upsilon_{x/w})$ that are both upper- and lower-bounded away from 0, where $\tilde{d}_\theta = \text{diag}(U^\top \Sigma_\theta U)$.

Remark: when $\Sigma_\theta$ is codiagonalizable with $\Sigma_x$, $d_\theta$ corresponds to its eigenvalues.
Optimal $\Sigma_w$: Ridgeless Interpolant $\lambda \to 0$

**Ridgeless Limit:** taking $\lambda \to 0$ yields the min $\|\theta\|_{\Sigma_w}$-norm interpolant.

Recall the **bias-variance decomposition** of the prediction risk:

$$\text{Bias: } \frac{m'(0)}{m^2(0)} \cdot \gamma \mathbb{E} \left( \frac{v_x v_\theta}{v_{x/w} \cdot m(0) + 1} \right)^2; \quad \text{Variance: } \frac{m'(0)}{m^2(0)} \cdot \tilde{\sigma}^2.$$  

- **Variance** is due to the *label noise* and does not depend on $\Sigma_\theta$.
- **Bias** depends on both the true parameters and the data distribution.

**Theorem.** Among all $\Sigma_w$ satisfying (A3-4), at the ridgeless limit $\lambda = 0$,

- **Variance** term is minimized by $\Sigma_w = \Sigma_x$.
- **Bias** term is minimized by $\Sigma_w = (U \text{ diag}(\bar{d}_\theta) U^\top)^{-1}$.

**Remark:** for the bias term, the optimal minimum $\|\theta\|_{\Sigma_w}$-norm interpolant also outperforms any principal component regression (PCR) estimator.
Optimal $\Sigma_w$: Generalized Ridge Estimator

Thm. $\Sigma_w^{-1} = U \text{diag}(\tilde{d}_\theta) U^\top$ is optimal among all $\Sigma_w$ satisfying (A3-4).

- Matches the *maximum a posteriori* estimate.
- Requires knowledge of $\Sigma_\theta$ (not practical).

**Question:** is there a reasonable $\Sigma_w$ based on $\Sigma_x$, which can be estimated from *unlabeled data*?

**Coro.** $\Sigma_w^{-1} = f(\Sigma_x)$ is optimal among all $\Sigma_w$ only *depending on* $\Sigma_x$, where $f(v_x) = \mathbb{E}[v_\theta|v_x]$ applies to the eigenvalues.

- **Heuristic:** approximate $f$ with polynomial function and cross-validate the parameters.
- When $\mathbb{E}[v_\theta|v_x] = \mathbb{E}[v_\theta]$, $\Sigma_w = I_d$ is reasonable.

Illustration of optimal $\Sigma_w$. Proposed heuristic.
By analyzing *generalized ridge regression* under general setup,

- We determined the sign of the optimal ridge regularization.
  - **Negative ridge** can be beneficial under **aligned** ("easy") problem.
- We characterized the optimal **explicit regularization** $\Sigma_w$.

**Future Directions:**
- Estimate $\Sigma_w$ 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).

**Companion work:** Amari et al. (2020). When Does Preconditioning Help or Hurt Generalization? *ICLR 2021*. 
• Xu, J. and Hsu D., 2019. On the number of variables to use in principal component regression.