When Does Preconditioning Help or Hurt Generalization?

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

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

International Conference on Learning Representations 2021

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 $\Rightarrow$ 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**.

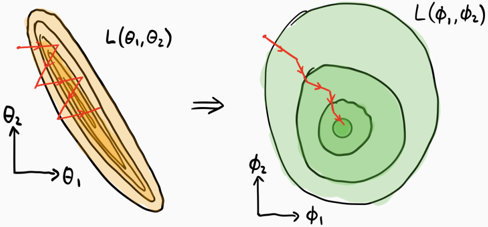


Figure from Xanadu blog post.

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

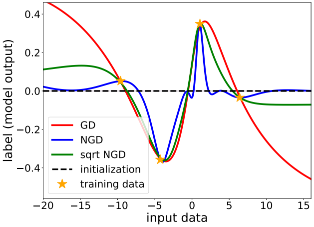
Motivation: Implicit Bias of Optimizers

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:

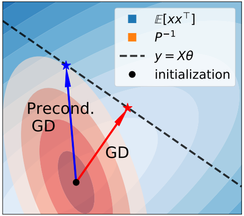
- In the *interpolation setting* (i.e. absence of explicit regularization), how does preconditioning influence the generalization performance?

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:** $\mathrm{d}\boldsymbol{\theta}(t) = \frac{1}{n}\boldsymbol{P}(t)\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}(t))\mathrm{d}t, \;\; \boldsymbol{\theta}(0) = 0$.

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

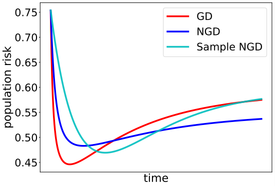
$$\boldsymbol{\theta}_{\boldsymbol{P}}(t) = \boldsymbol{P}\boldsymbol{X}^\top \left[\boldsymbol{I}_n - \exp\left(-\frac{t}{n}\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top\right)\right](\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top)^{-1}\boldsymbol{y},$$

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

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

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*).

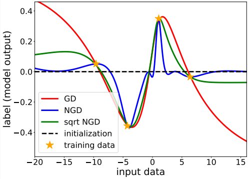


Remark: population Fisher can be estimated from extra **unlabeled data**.

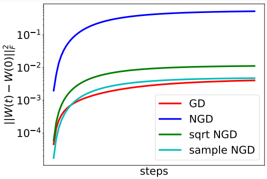
Starting from zero initialization:

- 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.

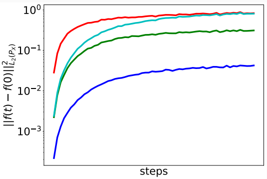
Similar findings also empirically observed in **two-layer neural networks**:



1D illustration (sigmoid).



Parameter difference.



Function difference.

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

Bias-variance Decomposition

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

- **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 $\boldsymbol{P}$ under different conditions of teacher model (bias) and label noise (variance).

Precise Asymptotic Risk in Bias-variance Decomposition:

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

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

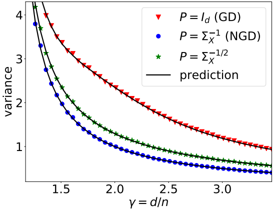
Where $m(z) > 0$ is the *Stieltjes transform* of the limiting spectral distribution of $\boldsymbol{XPX}^\top$, and $(\upsilon_x, \upsilon_\theta, \upsilon_{xp})$ relates to the eigenvalues of $\boldsymbol{P}$, $\boldsymbol{\Sigma}_x$, and $\mathbb{E}[\boldsymbol{\theta}_* \boldsymbol{\theta}_*^\top]$.

Variance Term: NGD is Optimal

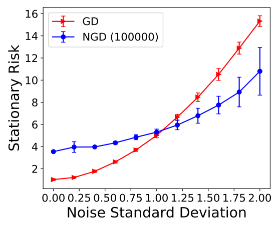
Thm. Among all positive definite $\boldsymbol{P}$, the *variance* is minimized by NGD: $\boldsymbol{P} = \boldsymbol{F}^{-1}$.

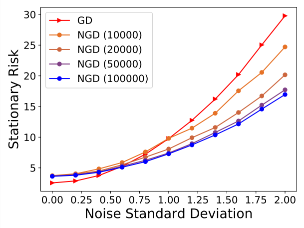
Message: when labels are noisy (risk is dominated by variance), NGD is beneficial.

Remark: Note that population Fisher is required.



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





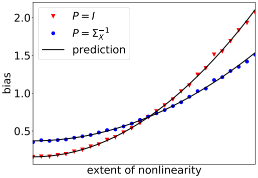
- Left: NGD (population Fisher) achieves lower risk under large label noise.
- Right: sample Fisher (i.e. less unlabeled data used) behaves like GD.

Misspecification ≈ 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 model.

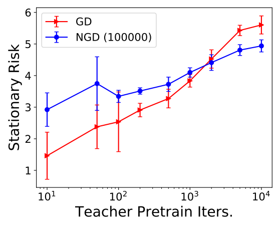
Intuition: f_*^c is "similar" to additive label noise.

Message: NGD is beneficial under misspecification.

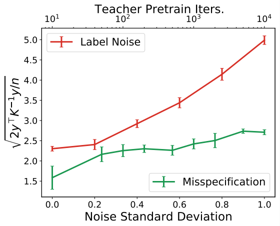


Misspecification in Neural Networks

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



Misspecification on CIFAR-10.

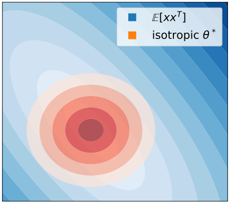


Measure of misspecification.

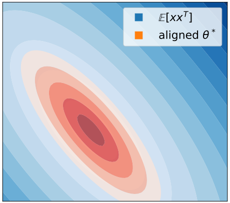
Bias Term: Well-specified Case

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

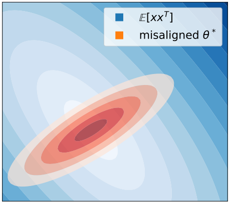
- Setup extends previously assumed *isotropic prior* [Dobriban and Wager 18].
- Alignment between $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\theta$ relates to the source condition in RKHS.



Isotropic (previous work).



Aligned (easy problem).

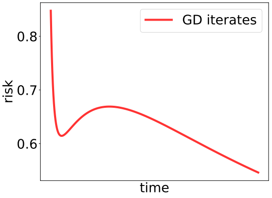


Misaligned (hard problem).

"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.



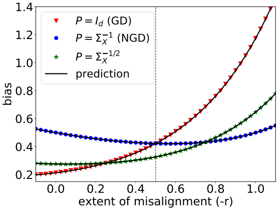
Well-specified Bias (continued)

Theorem (informal). Among all positive definite $\boldsymbol{P}$ codiagonalizable with $\boldsymbol{\Sigma}_x$, the stationary *bias* is minimized by $\boldsymbol{P} = \boldsymbol{U}\,\mathrm{diag}\big(\boldsymbol{U}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{U}\big)\boldsymbol{U}^\top$.

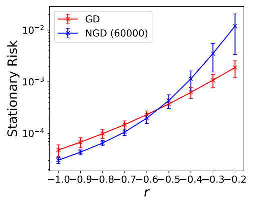
No-free-lunch: the optimal $\boldsymbol{P}$ is usually not known *a priori*:

- GD generalizes better when target is isotropic $\boldsymbol{\Sigma}_\theta = \boldsymbol{I}_d$.
- NGD is optimal under misalignment $\boldsymbol{\Sigma}_\theta = \boldsymbol{\Sigma}_x^{-1}$ ("hard" problem).

Prop. (source condition). When $\underline{\boldsymbol{\Sigma}_\theta = \boldsymbol{\Sigma}_x^r}$, 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 $\boldsymbol{P}$ for the *bias* and *variance* are in general **different**.

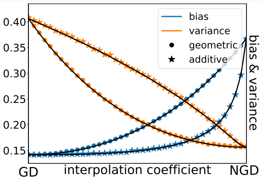
Question: how can we trade in one of bias/variance for the other?

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

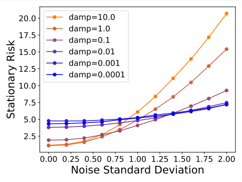
- Additive: $\boldsymbol{P}_\alpha = (\alpha\boldsymbol{\Sigma}_x + (1-\alpha)\boldsymbol{I}_d)^{-1}$, corresponds to the *damped inverse*.
- Geometric: $\boldsymbol{P}_\alpha = \boldsymbol{\Sigma}_x^{-\alpha}$, covers the "conservative" *square-root scaling*.

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

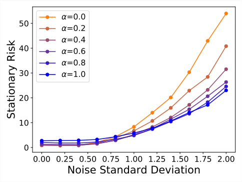
$\Rightarrow$ At certain SNR, **interpolating** between GD and NGD is beneficial.



Monotonicity of bias/variance.



Additive interpolation (MLP).



Geometric interpolation (MLP).

Bias-variance Tradeoff: Early Stopping

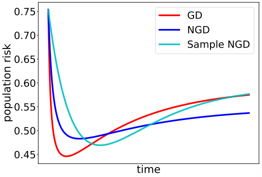
We have thus far only looked at the stationary solution ($t \to \infty$).

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

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

1. When $\boldsymbol{\Sigma}_\theta = \boldsymbol{\Sigma}_x^{-1}$ (misaligned), $B^{\mathrm{opt}}(\boldsymbol{\theta}_{\boldsymbol{P}}) \geq B^{\mathrm{opt}}(\boldsymbol{\theta}_{\boldsymbol{F}^{-1}})$.
2. When $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{I}_d$ (isotropic), $B^{\mathrm{opt}}(\boldsymbol{\theta}_{\boldsymbol{I}}) \leq B^{\mathrm{opt}}(\boldsymbol{\theta}_{\boldsymbol{F}^{-1}})$.
3. The variance $V(\boldsymbol{\theta}_{\boldsymbol{P}}(t))$ *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:** $y_i = f^*(\boldsymbol{x}_i) + \varepsilon_i$. $S : \mathcal{H} \to L_2(P_X)$. $\Sigma = S^*S$; $L = SS^*$.
- **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. $\sup_{\boldsymbol{x}} \|\Sigma^{1/2-1/\mu} K_{\boldsymbol{x}}\|_{\mathcal{H}} \le C_\mu$.

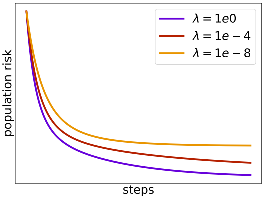
Remark: *source condition* relates to the previously discussed alignment:
large $r \Rightarrow$ smoother teacher model, i.e. "easier" problem; vice versa.

Fast Decay of Population Risk (continued)

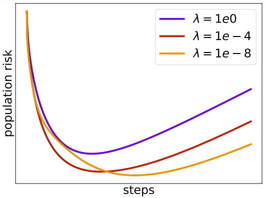
Theorem (informal). Given $\mu \leq 2r$ or $r \geq 1/2$, for sufficiently large n, preconditioned update with $\alpha = n^{-\frac{2s}{2rs+1}}$ achieves the minimax optimal convergence rate $R(f_t) = \|Sf_t - f^*\|^2_{L_2(P_X)} = \tilde{O}\left(n^{-\frac{2rs}{2rs+1}}\right)$ in $t = \Theta(\log n)$ steps,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

Companion work: Wu, D. and Xu, J. (2020). On the Optimal Weighted ℓ_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.