# Obtaining Adjustable Regularization for Free via Iterate Averaging

**Jingfeng Wu** [1] **Vladimir Braverman** [1] **Lin F. Yang** [2]

## Abstract

Regularization for optimization is a crucial technique to avoid overfitting in machine learning. In order to obtain the best performance, we usually train a model by tuning the regularization parameters. It becomes costly, however, when a single round of training takes significant amount of time. Very recently, Neu & Rosasco (2018) show that if we run stochastic gradient descent (SGD) on linear regression problems, then by averaging the SGD iterates properly, we obtain a regularized solution. It left open whether the same phenomenon can be achieved for other optimization problems and algorithms. In this paper, we establish an averaging scheme that *provably* converts the iterates of SGD on an arbitrary strongly convex and smooth objective function to its regularized counterpart with an *adjustable* regularization parameter. Our approaches can be used for accelerated and preconditioned optimization methods as well. We further show that the same methods work empirically on more general optimization objectives including neural networks. In sum, we obtain *adjustable* regularization *for free* for a large class of optimization problems and resolve an open question raised by Neu & Rosasco (2018).

## 1. Introduction

Regularization for optimization is a key technique for avoiding over-fitting in machine learning and statistics (Grandvalet & Bengio, 2005; Krogh & Hertz, 1992; Tibshirani, 1996; Tikhonov & Arsenin, 1977). The effects of explicit regularization methods, i.e., an extra regularization term added to the vanilla objective, are well studied, e.g., ridge regression (Tikhonov & Arsenin, 1977), LASSO (Tibshirani, 1996) and entropy regularization (Grandvalet & Bengio, 2005). Despite the great benefits of adopting explicit regularization, it could cause a huge computational burden to search for the optimal hyperparameter associated with the extra regularization term, especially for large-scale machine learning problems (Devlin et al., 2018; He et al., 2016; Silver et al., 2017).

In another line of research, people recognize and utilize the implicit regularization caused by certain components in machine learning algorithms, e.g., initialization (He et al., 2015; Hu et al., 2020), batch normalization (Ioffe & Szegedy, 2015; Cai et al., 2018), iterate averaging (Bach & Moulines, 2013; Jain et al., 2018; Neu & Rosasco, 2018), and optimizer such as gradient descent (GD) (Gunasekar et al., 2018; Soudry et al., 2018; Suggala et al., 2018). The regularization effect usually happens along the process of training the model and/or requires little post-computation. A great deal of evidence indicates that such a implicit bias plays a crucial role for the generalization abilities in many modern machine learning models (Zhang et al., 2016; Zhu et al., 2018; Wilson et al., 2017; Soudry et al., 2018). However, the implicit regularization is often a fixed effect and lacks the flexibility to be adjusted. To fully utilize it, we need a thorough understanding about the mechanism of the implicit regularization.

Among all the efforts spent on understanding and utilizing the implicit regularization, the work on bridging iterate averaging with explicit regularization (Neu & Rosasco, 2018) is extraordinarily appealing. In particular, Neu & Rosasco (2018) show that for linear regression, one can achieve $\ell_2$-regularization effect *for free* by simply taking geometrical averaging over the optimization path generated by stochastic gradient descent (SGD), which costs little additional computation. More interestingly, the regularization is *adjustable*, i.e., the solution biased by the regularizer in arbitrary strength can be obtained by iterate averaging using the corresponding weighting scheme. In a nutshell, this regularization approach has advantages over both the implicit regularization methods for being adjustable, and the explicit regularization methods for being cheap to tune.

Nevertheless, Neu & Rosasco (2018) only provide a method and its analysis for linear regression optimized by SGD. However linear regression itself is a rather restricted optimization objective. A nature question arises:

*Can we obtain "free" and "adjustable" regularization for*

[1]Johns Hopkins University, Baltimore, MD, USA [2]University of California, Los Angeles, CA, USA. Correspondence to: Jingfeng Wu <uuujf@jhu.edu>, Lin F. Yang <linyang@ee.ucla.edu>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Vienna, Austria, PMLR 119, 2020. 

*broader objective functions and optimization methods?*

In this work, we answer this question positively from the following aspects:

1. For linear regression, we analyze the regularization effects of averaging the optimization paths of SGD as well as preconditioned SGD, with adaptive learning rates. The averaged solutions achieve effects of $\ell_2$-regularization and generalized $\ell_2$-regularization respectively, in an adjustable manner. Similar results hold for kernel ridge regression as well.

2. We show that for Nesterov's accelerated stochastic gradient descent, the iterate averaged solution can also realize $\ell_2$-regularization effect by a modified averaging scheme. This resolves an open question raised by Neu & Rosasco (2018).

3. Beside linear regression, we study the regularization effects of iterate averaging for strongly convex and smooth loss functions, hence establishing a provable approach for obtaining nearly *free* and *adjustable* regularization for a broad class of functions.

4. Empirical studies on both synthetic and real datasets verify our theory. Moreover, we test iterate averaging with modern *deep neural networks* on CIFAR-10 and CIFAR-100 datasets, and the proposed approaches *still* obtain effective and adjustable regularization effects with little additional computation, demonstrating the broad applicability of our methods.

Our analysis is motivated from continuous approximation based on differential equations. When the learning rate tends to zero, the discrete algorithmic iterates tends to be the continuous path of an ordinary differential equation (ODE), on which we can establish a continuous version of our theory. We then discretize the ODE and generalize the theory to that of finite step size. This technique is of independent interests since it can be applied to analyze other comprehensive optimization problems as well (Su et al., 2014; Hu et al., 2017a; Li et al., 2017; Yang et al., 2018; Shi et al., 2019). Our results, in addition to the linear regression result in (Neu & Rosasco, 2018), illustrate the promising application of iterate averaging to obtain *adjustable* regularization *for free*.

## 2. Preliminaries

Let $\{(x_i, y_i) \in \mathbb{R}^{d\times 1}\}_{i=1}^n$ be the training data and $w \in \mathbb{R}^d$ be the parameters to be optimized. The goal is to minimize a lower bounded loss function $L(w)$

$$\min_w L(w) := \frac{1}{n}\sum_{i=1}^n \ell(x_i, y_i, w). \qquad (\mathcal{P}_1)$$

One important example is linear regression under the square loss where $L(w) = \frac{1}{2n}\sum_{i=1}^n \left\|w^\top x_i - y_i\right\|_2^2$. The optimization problem often involves an explicit regularization term

$$\min_{\hat{w}} L(\hat{w}) + \lambda R(\hat{w}), \qquad (\mathcal{P}_2)$$

where $R(\hat{w})$ is a regularizer and $\lambda$ is the associated hyperparameter. For example, the $\ell_2$-regularizer is $R(\hat{w}) = \frac{1}{2}\left\|\hat{w}\right\|_2^2$. Given an iterative algorithm, e.g., SGD, an optimization path is generated by running the algorithm. With a little abuse of notations, we use $\{w_k\}_{k=0}^\infty$ and $\{\hat{w}_k\}_{k=0}^\infty$ to represent the optimization paths for the unregularized problem ($\mathcal{P}_1$) and the regularized problem ($\mathcal{P}_2$), respectively. Sometimes we write $\hat{w}_k$ with a script as $\hat{w}_{k,\lambda}$ to emphasize its dependence on the hyperparameter $\lambda$. We use $\eta_k$ and $\gamma_k$ to denote the learning rates for training the unregularized and regularized objectives respectively. For simplicity we always initialize the iterative algorithms from zero, i.e., $w_0 = \hat{w}_0 = 0$.

**Iterate averaging** The core idea in this work is a technique called *iterate averaging*. Given a series of parameters $\{w_k\}_{k=0}^\infty$, a *weighting scheme* $\{p_k\}_{k=0}^\infty$ is defined as a probability distribution associated to the series, i.e., $p_k \geq 0$, $\sum_{k=0}^\infty p_k = 1$. Its accumulation is denoted as $P_k = \sum_{i=0}^k p_i$, where $\lim_{k\to\infty} P_k = 1$. Since a weighting scheme and its accumulation identifies each other by $p_k = P_k - P_{k-1}$ for $k \geq 1$, we also call $\{P_k\}_{k=0}^\infty$ a weighting scheme. Then the iterate averaged parameters are

$$\tilde{w}_k = {P_k}^{-1}\textstyle\sum_{i=0}^k p_i w_i, \quad k \geq 0.$$

Various kinds of averaging schemes (for the SGD optimization path) have been studied before. Theoretically, arithmetic averaging is shown to bring better convergence (Bach & Moulines, 2013; Lakshminarayanan & Szepesvari, 2018); tail-averaging is analyzed by Jain et al. (2018); and Neu & Rosasco (2018) discuss geometrically averaging and its regularization effect for SGD and linear regression. Empirically, arithmetic averaging is also shown to be helpful for modern deep neural networks (Izmailov et al., 2018; Zhang et al., 2019; Granziol et al., 2020). Inspired by Neu & Rosasco (2018), in this work we explore in depth the regularization effect induced by iterate averaging for various kinds of optimization algorithms and loss functions.

**Stochastic gradient descent** The optimization problem ($\mathcal{P}_1$) is often solved by stochastic gradient descent (SGD): at every iteration, a mini-batch is sampled uniformly at random, and then the parameters are updated according to the gradient of the loss estimated using the mini-batch. For simplicity we let the batch size be 1. Then with learning rate $\eta_k > 0$, SGD takes the following update:

$$w_{k+1} = w_k - \eta_k \nabla \ell(x_k, y_k, w_k). \qquad (1)$$

Similarly, for the regularized problem ($\mathcal{P}_2$), with learning rate $\gamma_k > 0$, SGD takes update:

$$\hat{w}_{k+1} = \hat{w}_k - \gamma_k \left( \nabla \ell(x_k, y_k, \hat{w}_k) + \lambda \nabla R(\hat{w}_k) \right). \quad (2)$$

For linear regression problem and fixed learning rates, Neu & Rosasco (2018) discuss the geometrically averaging over the SGD iterates (1). They show that by doing so one obtains the solution of the $\ell_2$-regularized problem ($\mathcal{P}_2$) where $R(\hat{w}) = \frac{1}{2}\|\hat{w}\|_2^2$ for arbitrary hyperparameter $\lambda$. In this work, we analyze a much broader class of algorithms and functions. In particular, we establish adjustable $\ell_2$-regularization effect for (i) SGD with adaptive learning rate, (ii) kernel ridge regression (Mohri et al., 2018), and (iii) general strongly convex and smooth loss functions.

**Preconditioned stochastic gradient descent** We also study iterate averaging for preconditioned stochastic gradient descent (PSGD). Given a positive definite matrix $Q$ as the preconditioning matrix and $\eta_k$ as the learning rate, the PSGD takes following update to optimize problem ($\mathcal{P}_1$):

$$w_{k+1} = w_k - \eta_k Q^{-1} \nabla \ell(x_k, y_k, w_k), \quad (3)$$

Similarly, the regularized problem ($\mathcal{P}_2$) can be solved by PSGD with learning rate $\gamma_k > 0$ as:

$$\hat{w}_{k+1} = \hat{w}_k - \gamma_k Q^{-1} \left( \nabla \ell(x_k, y_k, \hat{w}_k) + \lambda \nabla R(\hat{w}_k) \right). \quad (4)$$

We remark that PSGD unifies several important algorithms as natural gradient descent and Newton's method at special cases where the curvature matrices can be replaced by constant matrices (Martens, 2014; Dennis Jr & Schnabel, 1996; Bottou & Bousquet, 2008).

For linear regression problems, we will show that geometrically averaging the PSGD iterates (3) leads to a solution biased by the *generalized $\ell_2$-regularizer*, i.e., the solution of problem ($\mathcal{P}_2$) with $R(w) = \frac{1}{2} w^\top Q w = \frac{1}{2}\|w\|_Q^2$. The obtained regularization is adjustable, too.

**Nesterov's accelerated stochastic gradient descent** In problem ($\mathcal{P}_1$), suppose the loss function $L(w)$ is $\alpha$-strongly convex. Let $\eta > 0$ be the learning rate and $\tau = \frac{1-\sqrt{\eta\alpha}}{1+\sqrt{\eta\alpha}}$, then the Nesterov's accelerated stochastic gradient descent (NSGD) takes update (Nesterov, 1983; Su et al., 2014; Yang et al., 2018):

$$\begin{aligned} w_{k+1} &= v_k - \eta \nabla \ell(x_k, y_k, v_k), \\ v_k &= w_k + \tau(w_k - w_{k-1}). \end{aligned} \quad (5)$$

Now we consider the regularized problem ($\mathcal{P}_2$) with the $\ell_2$-regularizer, $R(\hat{w}) = \frac{1}{2}\|\hat{w}\|_2^2$. The objective function then becomes $(\alpha + \lambda)$-strongly convex. Let $\gamma > 0$ be the learning rate and $\hat{\tau} = \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1+\sqrt{\gamma(\alpha+\lambda)}}$, then the NSGD takes update:

$$\begin{aligned} \hat{w}_{k+1} &= \hat{v}_k - \gamma \left( \nabla \ell(x_k, y_k, \hat{v}_k) + \lambda \hat{v}_k \right), \\ \hat{v}_k &= \hat{w}_k + \hat{\tau}(\hat{w}_k - \hat{w}_{k-1}). \end{aligned} \quad (6)$$

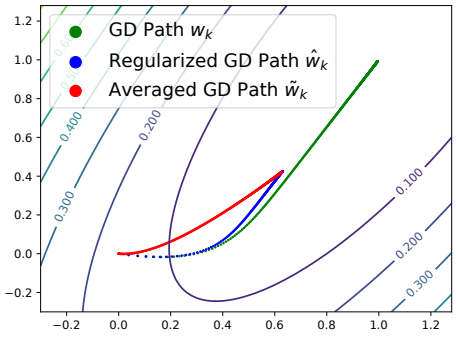



*Figure 1.* A 2-D demonstration of the effect of an averaged SGD path (Theorem 1). Green dots: the vanilla GD path $w_k$; blue dots: the regularized GD path $\hat{w}_k$; red dots: the averaged GD path $\tilde{w}_k$. The red dots converge to the blue ones.

It is proposed as an open question by Neu & Rosasco (2018) whether or not adjustable regularization can be obtained by averaging the NSGD optimization path. Our work offers an affirmative answer by showing that for linear regression, one can perform iterate averaging over the NSGD path to obtain the $\ell_2$-regularized solution as well.

## 3. The free and adjustable regularization induced by iterate averaging

In this section, we show that adjustable regularization effects can be obtained for "free" via *iterate averaging* for: (i) different SGD schemes, e.g., linear regression or kernel ridge regression with adaptive learning rates; (ii) PSGD; (iii) NSGD; (iv) arbitrary strongly convex and smooth loss functions. Not limited to SGD with fixed learning rate and linear regression, our results manifest the broader potential of employing iterate averaging to obtain regularization that can be tuned with little computation overhead.

Our analysis is motivated from continuous differential equations, which is postponed to Section A of Supplementary Materials due to space limitation. In the following we present our results in discrete cases.

### 3.1. The effect of an averaged SGD path

We first introduce a generalized averaging scheme for the SGD algorithm. Unlike the method in (Neu & Rosasco, 2018), our approach works even with adaptive learning rates. Specifically, given a learning rate schedule and a regularization parameter $\lambda$, we compute a weighting scheme

for averaging a stored SGD path. Then the averaged solution converges to the regularized solution with hyperparameter $\lambda$. Theorem 1 formally justifies our method.

**Theorem 1** (The effect of an averaged SGD path)**.** *Consider loss function $L(w) = \frac{1}{2n}\sum_{i=1}^{n}\left\|w^\top x_i - y_i\right\|_2^2$, and regularizer $R(w) = \frac{1}{2}\|w\|_2^2$. Let $\alpha$ and $\beta$ be such that $L(w)$ is $\alpha$-strongly convex[1] and $\beta$-smooth. Let $\{w_k\}_{k=0}^{\infty}$ and $\{\hat{w}_k\}_{k=0}^{\infty}$ be the SGD paths for the vanilla loss function $L(w)$ with learning rate $\eta_k$, and the regularized loss function $L(\hat{w}) + \lambda R(\hat{w})$ with learning rate $\gamma_k$, respectively. Suppose $1 - \lambda\gamma_k = \gamma_k/\eta_k$, $\eta_k \in (\eta, 1/\beta)$, $\eta > 0$ and $\gamma := \eta/(1+\lambda\eta)$. Let*

$$P_k := \textstyle\sum_{i=0}^{k} p_i = 1 - \Pi_{i=0}^{k}(\gamma_i/\eta_i).$$

*Then for $\tilde{w}_k = P_k^{-1}\sum_{i=0}^{k} p_i w_i$ we have*

*1. $P_k \cdot \mathbb{E}[\tilde{w}_k] = \mathbb{E}[\hat{w}_k] - (1 - P_k)\cdot\mathbb{E}[w_k]$.*

*2. Both $\mathbb{E}[w_k]$ and $\mathbb{E}[\hat{w}_k]$ converge. Moreover, we have $\|\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k]\|_2 \le \mathcal{O}((1-\lambda\gamma)^k)$.*

*3. If the gradient noise $\epsilon_k = \nabla\ell(x_k, y_k, w) - \nabla L(w)$ has uniformly bounded variance $\mathbb{E}[\|\epsilon_k\|_2^2] \le \sigma^2$, then for $k$ large enough, with probability at least $1-\delta$ we have[2]*

$$\|P_k\tilde{w}_k - P_k\mathbb{E}[\tilde{w}_k]\|_2 \le \epsilon,$$

*where $\epsilon = \frac{\sigma}{\gamma(\lambda+\alpha)(\lambda+\beta)^2}\cdot\sqrt{\frac{\lambda}{\delta\gamma(2-\lambda\gamma)}}$.*

The proof is left in Supplementary Materials, Section C.1. A 2-D illustration for Theorem 1 is presented in Figure 1.

Theorem 1 guarantees the method of obtaining adjustable $\ell_2$-regularization for free via iterate averaging. Specifically, we first collect an SGD path $\{w_k\}_{k=0}^{\infty}$ for $L(w)$ under a learning rate schedule $\eta_k$ (it can be chosen in a broad range); then for a regularization parameter $\lambda$, we compute an averaging scheme $\{p_k\}_{k=0}^{\infty}$ that converts the collected SGD path to the regularized solution, $\hat{w}_\infty$. Note that the learning rate schedule $\gamma_k$ is only for analysis and does not need to be known.

Specially, when the learning rates are constants, i.e., $\eta_i = \eta$ and $\gamma_i = \gamma$, the first two conclusions in Theorem 1 recover the Proposition 1 and Proposition 2 in (Neu & Rosasco, 2018). Besides, the third claim in Theorem 1 characterizes the deviation of the averaged solution, which relies on the models, learning rates, and the regularization parameter, etc. And empirical studies in Section 4.2 do suggest that such a deviation is sufficiently small that it does not affect the induced regularization effect.

*Remark.* We emphasize that the method of Neu & Rosasco (2018) only applies to SGD with constant learning rate. Moreover, their theory only guarantees the averaged solution has *convergence in expectation*, which is not very useful since the averaged solution might not converge to the regularized solution almost surely, not even in probability (a.k.a. weak convergence) (see Section 4.2). Nevertheless, our theory carefully characterizes the deviation between the averaged solution and the regularized solution.

More interestingly, Theorem 1.1 shows that this method is also applicable to kernel ridge regression (in the dual space).

**Theorem 1.1.** *Let $K \in \mathbb{R}^{n\times n}$ be a kernel, $K(i,j) = \phi(x_i)^\top\phi(x_j)$, where $\phi : \mathbb{R}^d \to \mathcal{H}$ is the kernel map. Consider kernel ridge regression*

$$\min_{\alpha\in\mathbb{R}^n} L(\alpha, \lambda) := \frac{1}{2}\|y - K\alpha\|_2^2 + \frac{\lambda}{2}\alpha^\top K\alpha$$

*where $y = (y_1, \ldots, y_n)^T$ is the labels and $\alpha \in \mathbb{R}^n$ is the dual parameter. Let $\{\alpha_k\}_{k=0}^{\infty}$ and $\{\hat{\alpha}_k\}_{k=0}^{\infty}$ be the GD paths for the loss $L(\alpha, \lambda)$ with learning rate $\eta_k$, and the loss $L(\hat{\alpha}, \hat{\lambda})$ with generalized learning rate $\gamma_k$, respectively. Suppose $\hat{\lambda} > \lambda$, $\gamma_k = \eta_k\left(I + (\hat{\lambda} - \lambda)\eta_k K\right)^{-1}$. Let*

$$P_k := \textstyle\sum_{i=0}^{k} p_i = 1 - \Pi_{i=0}^{k}\left(\gamma_i/\eta_i\right).$$

*Then for $\tilde{\alpha}_k = P_k^{-1}\sum_{i=0}^{k} p_i\alpha_i$ we have*

*1. $P_k\tilde{\alpha}_k = \hat{\alpha}_k - (1 - P_k)\alpha_k$.*

*2. Both $\alpha_k$ and $\hat{\alpha}_k$ converge provided suitable learning rates. Moreover, we have $\|\hat{\alpha}_k - \tilde{\alpha}_k\|_2 \le \mathcal{O}(C^k)$ where $C \in (0,1)$ is a constant decided by $K$, $\hat{\lambda} - \lambda$ and $\eta_k$.*

### 3.2. The effect of an averaged PSGD path

In practice, we usually need many different regularizers. And one important class of them is the *generalized $\ell_2$-regularizers*, i.e., $R(w) := \frac{1}{2}w^\top Qw$ for some positive definite matrix $Q$. But it is painful to adjust its regularization parameter $\lambda$ by re-training the model. Luckily, we show that the solution biased by such a regularizer can also be obtained for "free" by averaging the optimization path of PSGD. Our result is formally presented in the next theorem.

**Theorem 2** (The effect of an averaged PSGD path)**.** *Consider loss function $L(w) = \frac{1}{2n}\sum_{i=1}^{n}\left\|w^\top x_i - y_i\right\|_2^2$, and regularizer $R(w) = \frac{1}{2}w^\top Qw$, where $Q$ is a positive definite matrix. Let $\alpha$ and $\beta$ be such that $\alpha Q \preceq \Sigma =$*

[1] The strong convexity assumption does not limit the application of our method. For a convex but not strongly convex loss $L(w)$, we can instead collect an optimization path of $L(w) + \lambda_0\|w\|_2^2$ for some small $\lambda_0$, which is then strongly convex, and then we apply Theorem 1 to obtain the regularized solutions for a different $\lambda$. Similar arguments apply to the theorems afterwards as well.

[2] In this high probability result, the confidence parameter $\delta$ appears in a polynomial order, $\frac{1}{\sqrt{\delta}}$. However this is only due to the assumption of bounded variance of the noise and an application of Chebyshev's inequality. It is straightforward to obtain a logarithm dependence on $\delta$ by assuming the sub-Gaussianity of the noise and applying Hoeffding's inequality. Similar arguments apply to the theorems afterwards as well.

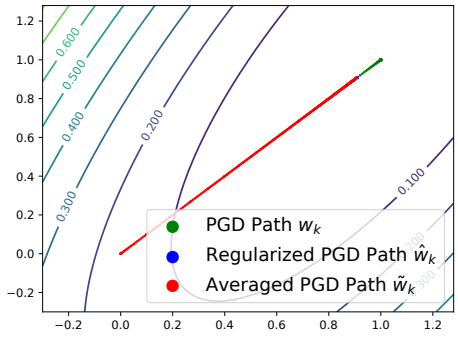



*Figure 2.* A 2-D demonstration of the effect of an averaged PSGD path (Theorem 2). The Hessian is used as the preconditioning matrix. Green dots: the vanilla PGD path $w_t$; blue dots: the regularized PGD path $\hat{w}_t$; red dots: the averaged PGD path $\tilde{w}_t$. The red dots converge to the blue ones.

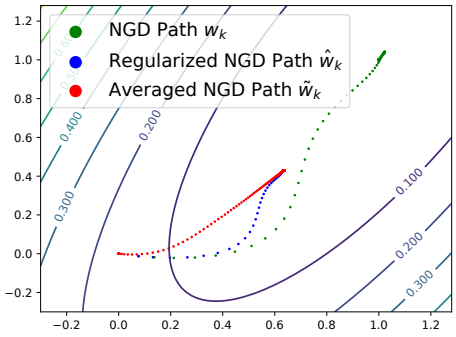



*Figure 3.* A 2-D demonstration of the effect of an averaged NSGD path (Theorem 3). Green dots: the vanilla NGD path $w_t$; blue dots: the regularized NGD path $\hat{w}_t$; red dots: the averaged NGD path $\tilde{w}_t$. The red dots converge to the blue ones.

$n^{-1}\sum_{i=1}^n x_i x_i^\top \preceq \beta Q$. *With $Q$ as the preconditioning matrix, let $\{w_k\}_{k=0}^\infty$ and $\{\hat{w}_k\}_{k=0}^\infty$ be the PSGD paths for the vanilla loss function $L(w)$ with learning rate $\eta_k$, and the regularized loss function $L(\hat{w}) + \lambda R(\hat{w})$ with learning rate $\gamma_k$, respectively. Suppose $1 - \lambda\gamma_k = \gamma_k/\eta_k$, $\eta_k \in (\eta, 1/\beta)$, $\eta > 0$ and $\gamma := \eta/(1+\lambda\eta)$. Let*

$$P_k := \sum_{i=0}^k p_i = 1 - \Pi_{i=0}^k(\gamma_i/\eta_i).$$

*Then for $\tilde{w}_k = P_k^{-1}\sum_{i=0}^k p_i w_i$ we have*

*1. $P_k \cdot \mathbb{E}[\tilde{w}_k] = \mathbb{E}[\hat{w}_k] - (1-P_k)\cdot\mathbb{E}[w_k]$.*

*2. Both $\mathbb{E}[w_k]$ and $\mathbb{E}[\hat{w}_k]$ converge. Moreover, we have $\|\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k]\|_2 \le \mathcal{O}((1-\lambda\gamma)^k)$.*

*3. If the noise $\epsilon_k = Q^{-1}(\nabla\ell(x_k, y_k, w) - \nabla L(w))$ has uniform bounded variance $\mathbb{E}[\|\epsilon_k\|_2^2] \le \sigma^2$, then for $k$ large enough, with probability at least $1-\delta$ we have*

$$\|P_k\tilde{w}_k - P_k\mathbb{E}[\tilde{w}_k]\|_2 \le \epsilon,$$

*where $\epsilon = \frac{\sigma\|Q\|_2}{\gamma(\lambda+\alpha)(\lambda+\beta)^2}\sqrt{\frac{\lambda}{\delta\gamma(2-\lambda\gamma)}}$.*

The proof is left in Supplementary Materials, Section C.3. A 2-D illustration for Theorem 2 is presented in Figure 2.

The importance of Theorem 2 is two-folds. On the one hand, averaging the PSGD path has an effect as the generalized $\ell_2$-regularizer. And as before, this induced regularization is both adjustable and costless. The considered PSGD algorithm applies to natural gradient descent and Newton's method in certain circumstances where the curvature matrices can be replaced by constant matrices (Martens, 2014; Dennis Jr & Schnabel, 1996; Bottou & Bousquet, 2008). On the other hand, to obtain a desired type of generalized $\ell_2$-regularization effect, we should store and average a PSGD path with the corresponding preconditioning matrix as indicated in Theorem 2, instead of using a SGD path.

### 3.3. The effect of an averaged NSGD path

In this part, we show how to obtain adjustable regularization effect by applying averaging schemes on the NSGD path.

**Theorem 3** (The effect of an averaged NSGD path)**.** *Consider loss function $L(w) = \frac{1}{2n}\sum_{i=1}^n \left\|w^\top x_i - y_i\right\|_2^2$, and regularizer $R(w) = \frac{1}{2}\|w\|_2^2$. Let $\alpha$ and $\beta$ be such that $L(w)$ is $\alpha$-strongly convex and $\beta$-smooth. Let $\{w_k\}_{k=0}^\infty$ and $\{\hat{w}_k\}_{k=0}^\infty$ be the NSGD paths for the vanilla loss function $L(w)$ with learning rate $\eta$, and the regularized loss function $L(\hat{w}) + \lambda R(\hat{w})$ with learning rate $\gamma$, respectively. Suppose $1 - \lambda\gamma = \gamma/\eta$, $\eta \in (0, 1/\beta)$. Let*

$$P_k := \sum_{i=0}^k p_k = 1 - \frac{\gamma}{\eta}\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^{k-1}.$$

*Then for $\tilde{w}_k = P_k^{-1}\sum_{i=0}^k p_i w_i$ we have*

*1. $P_k \cdot \mathbb{E}[\tilde{w}_k] = \mathbb{E}[\hat{w}_k] - (1-P_k)\cdot\mathbb{E}[w_k]$.*

*2. $\mathbb{E}[w_k]$ and $\mathbb{E}[\hat{w}_k]$ converge. And $\|\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k]\|_2 \le \mathcal{O}\left(C^k\right)$, where $C = \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \in (0,1)$.*

*3. If the gradient noise $\epsilon_k = \nabla\ell(x_k, y_k, w) - \nabla L(w)$ has uniformly bounded variance $\mathbb{E}[\|\epsilon_k\|_2^2] \le \sigma^2$, then for $k$ large enough, with probability at least $1-\delta$ we have*

$$\|P_k\tilde{w}_k - P_k\mathbb{E}[\tilde{w}_k]\|_2 \le \epsilon,$$

*where $\epsilon$ depends on $\sigma, \alpha, \beta, \eta, \gamma$.*

The proof and the exact value of $\epsilon$ are given in Supplementary Materials, Section C.4. A 2-D illustration for Theorem 3 is presented in Figure 3.

Theorem 3 affirmatively answers an open question raised by Neu & Rosasco (2018): there exists an averaging scheme for NSGD to achieve $\ell_2$-regularization in arbitrary strength. In addition to the results for averaging SGD, Theorem 3 provides us wider choices of applicable optimizers for obtaining adjustable $\ell_2$-regularization effect by iterate averaging.

### 3.4. The effect of an averaged GD path for strongly convex and smooth loss functions

In this section, we show that the iterate averaging methods work for not only simple optimization objectives like least square, but also a much broader set of loss functions. In fact, we show that any strongly convex and smooth function admits an iterate averaging scheme, which brings $\ell_2$-regularization effect in a tunable manner. More formally, in the problems ($\mathcal{P}_1$) and ($\mathcal{P}_2$), let $L(w)$ be $\alpha$-strongly convex and $\beta$-smooth, and $R(w) := \frac{1}{2}\|w\|_2^2$ be the $\ell_2$-regularizer. For the sake of representation, we focus on gradient descent (GD) with constant learning rate applied on the loss functions. Similar arguments can also be applied for SGD, PSGD and NSGD. The GD takes update

$$w_{k+1} = w_k - \eta \nabla L(w_k),$$
$$\hat{w}_{k+1,\lambda} = \hat{w}_{k,\lambda} - \gamma(\nabla L(\hat{w}_{k,\lambda}) + \lambda \hat{w}_{k,\lambda}),$$

for optimizing problems ($\mathcal{P}_1$) and ($\mathcal{P}_2$), respectively. Let $b = -\nabla L(w_0) = -\nabla L(0)$. Let us denote two iterations

$$u_{k+1} - u_k = -\eta(\alpha u_k - b), \quad v_{k+1} - v_k = -\eta(\beta v_k - b),$$

where $u_0 = v_0 = 0$. Consider an averaging scheme $P_k = \sum_{i=0}^k p_i = 1 - (\gamma/\eta)^{k+1}$. Let $\tilde{u}_k = {P_k}^{-1} \sum_{i=0}^k p_i u_i$, $\tilde{v}_k = {P_k}^{-1} \sum_{i=0}^k p_i v_i$, and $\tilde{w}_k = P_k^{-1} \sum_{i=0}^k p_i w_i$. Then the next theorem characterizes the regularization effect of a averaged GD path for general strongly convex and smooth loss functions.

**Theorem 4** (The effect of an averaged GD path for strongly convex and smooth loss functions)**.** *Without loss of generality, assume the unique minimum $w_*$ of $L(w)$ satisfies $w_* > w_0 = 0$ entry-wisely. Suppose $1/(2\beta - \alpha) < \eta < 1/\beta$, $0 < \gamma < \eta/(\eta(\beta - \alpha) + 1)$. Then for hyperparameters*

$$\lambda_1 = 1/\gamma - 1/\eta + \beta - \alpha, \quad \lambda_2 = 1/\gamma - 1/\eta + \alpha - \beta,$$

*we have*

*1. $\hat{w}_{k,\lambda_1} + (1 - P_k)(\tilde{v}_k - v_k) \le \tilde{w}_k \le \hat{w}_{k,\lambda_2} + (1 - P_k)(\tilde{u}_k - u_k)$, where the "$\le$" is defined entry-wisely.*

*2. $u_k, \tilde{u}_k, v_k, \tilde{v}_k, \hat{w}_{k,\lambda_1}, \hat{w}_{k,\lambda_2}$ converge. Moreover let $m = (\hat{w}_{\infty,\lambda_2} + \hat{w}_{\infty,\lambda_1})/2$, $d = (\hat{w}_{\infty,\lambda_2} - \hat{w}_{\infty,\lambda_1})/2$ and $C = \max\{(1 - \gamma(\alpha + \lambda_1), (1 - \gamma(\alpha + \lambda_2), \frac{\gamma}{\eta}\} \in (0, 1)$, then $\|\tilde{w}_k - m\|_2 \le \|d\|_2 + \mathcal{O}(C^k)$.*

The proof is left in Supplementary Materials, Section C.5 .

According to Theorem 4, for strongly convex and smooth objectives, the averaged GD path $\{\tilde{w}_k\}_{k=0}^\infty$ lies in the area between two regularized GD paths, $\{\hat{w}_{k,\lambda_1}\}_{k=0}^\infty$ and $\{\hat{w}_{k,\lambda_2}\}_{k=0}^\infty$. Furthermore, $\tilde{w}_k$ converges to a hyper cube whose diagonal vertices are defined by $\hat{w}_{\infty,\lambda_1}$ and $\hat{w}_{\infty,\lambda_2}$. In this way for this class of loss functions, averaging the GD path has an "approximate" $\ell_2$-regularization effect that is in between two $\ell_2$-regularizers with hyperparameters as $\lambda_1$ and $\lambda_2$ respectively. In addition, $\lambda_1$ and $\lambda_2$ can be adjusted through changing the weighting scheme. Finally, we note that $\|d\|_2 = \mathcal{O}(\beta - \alpha)$, thus when the objective is quadratic, we have $\alpha = \beta$ and $d = 0$ and thus the "approximate" $\ell_2$-regularization effect becomes the exact $\ell_2$-regularization by Theorem 4.

We therefore conjecture that, generally, *for arbitrary loss functions and iterative optimizers, an iterate averaging scheme admits a specific yet unknown regularization effect.* Indeed, our experiments in the next section empirically verifies such an effect by performing iterate averaging on deep neural networks, which are highly comprehensive.

## 4. Experiments

In this section we present our empirical studies. The detailed setups are explained in Supplementary Materials, Section D.

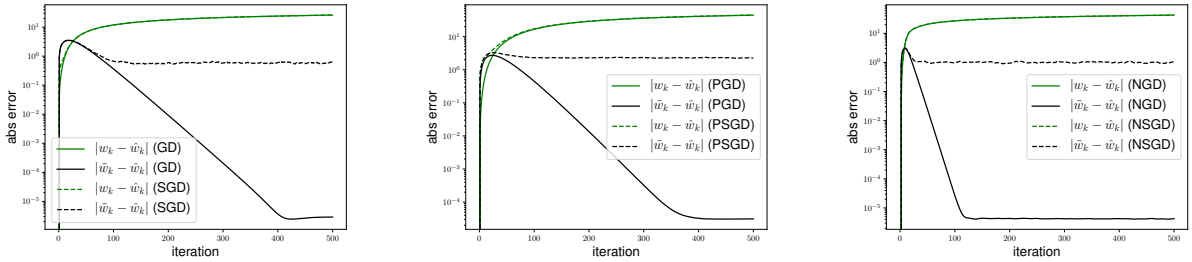
### 4.1. Two dimensional demonstration

We first introduce a two dimensional toy example to demonstrate the regularization effect of iterate averaging. The vanilla loss function is quadratic with a unique minimum at $(1, 1)$, as shown in Figure 1∼3. For the purpose of demonstration we only run deterministic algorithms with constant learning rates. We plot the trajectories of the concerned optimizers for learning the vanilla loss function and the regularized loss function, as well as the averaged solutions. All of the optimizers start iterations from zero.

In Figure 1, the green and the blue dots represent the GD paths for optimizing the vanilla/regularized loss functions respectively, while the red dots are the path of iterate averaged solutions. We observe that the red dots do converge to the blue ones, indicating the averaged solution has the same effect of an $\ell_2$-regularizer, as suggested by Theorem 1. Similarly the phenomenon holds for averaging the NGD path, as indicated in Figure 3. In Figure 2, the preconditioning matrix is set to be the Hessian. And as predicted by Theorem 2, the averaged solution converges to the solution biased by a generalized $\ell_2$-regularizer.

### 4.2. Real data verification

We then present experiments on the MNIST dataset.




(a) GD and SGD (b) PGD and PSGD (c) NGD and PNGD

*Figure 4.* Linear regression on MNIST dataset. X-axis: iteration; y-axis: the absolute approximation error in logarithmic scale. Green lines represent $\|w_t - \hat{w}_t\|_1$ and black lines represent $\|\tilde{w}_t - \hat{w}_t\|_1$, where $w_t$, $\hat{w}_t$ and $\tilde{w}_t$ are the unregularized path, the regularized path and the iterate averaged path, respectively. Solid lines and dashed lines are the results obtained by running deterministic and stochastic algorithms respectively. For deterministic algorithms, the error between $\tilde{w}_t$ and $\hat{w}_t$ converges to zero. For stochastic algorithms, the error between $\tilde{w}_t$ and $\hat{w}_t$ remains small.

**Linear regression** Firstly, we study linear regression under quadratic loss functions and the regularization effects caused by averaging the optimization paths of (S)GD, P(S)GD and N(S)GD. The learning rates are set to be constant. For P(S)GD, we set the preconditioning matrix as the Hessian, which is known as the Newton's method.

Our theories predict that averaging the (S)GD and N(S)GD paths leads to the solutions biased by $\ell_2$-regularizers (Theorem 1, 3), while averaging the P(S)GD path introduces an effect of the generalized $\ell_2$-regularization (Theorem 2). To verify the predictions, we generate the paths of the averaged solutions $\tilde{w}_k$ and the regularized solutions $\hat{w}_k$, and then compute the approximation errors between them. The results are plotted in Figure 4.

In Figure 4 (a), the solid lines clearly indicate that the averaged solution converges to the regularized solution when running GD, which also corresponds to the convergence in expectation in SGD cases, as predicted by Theorem 1. For SGD, however, the dashed lines in Figure 4 (a) show that there is a small error between the averaged solution and the regularized solution. The error exists since the convergence of the averaged solution does not hold in probability. Luckily, the error would not grow large as the deviation of the averaged solution is controllable by Theorem 1. Hence by comparing the dashed green and black lines, we see that averaging the SGD path still leads to an effect of $\ell_2$-regularization ignoring a tolerable error.

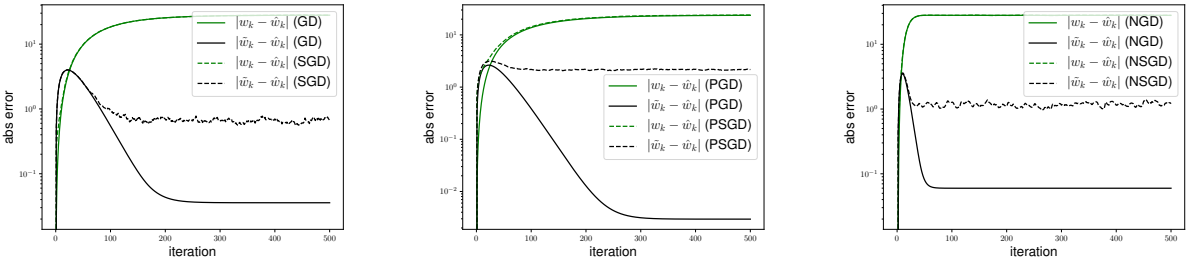
Figure 4 (b) shows the results for PGD and PSGD. Again, averaging the PGD path causes a perfect generalized $\ell_2$-regularization effect, and there is a small gap for averaging the optimization path with noise. These support Theorem 2.

The results related to NGD and NSGD are shown in Figure 4 (c). Again, for the deterministic algorithm, the solid lines manifest the convergence between the averaged solution and the regularized solution, verifying our Theorem 3. And the dashed lines once more suggest the stochastic algorithm causes a tolerable approximation error.

**Logistic regression** Next we set the loss function $L(w)$ to be the logistic regression objective with a small $\ell_2$-regularizer, which is then strongly convex and smooth, as required by Theorem 4. We firstly generate the unregularized paths and perform iterate averaging over them. Next, since it is impossible to visualize a high dimensional cubic with vertices decided by Theorem 4, instead we set $\lambda = 1/\gamma - 1/\eta$, and add an extra regularization term with this particular hyperparameter to obtain the regularized paths. Lastly we measure the errors between the averaged solutions and the regularized solutions to verify the effect of iterate averaging applied on strongly convex and smooth loss functions. The learning rates are set to be constant. The approximation errors are plotted in Figure 5.

In Figure 5 (a), the solid black line measures the error between the averaged GD path and the regularized GD path, and indeed the error is bounded and small as predicted by Theorem 4; the dashed black line is the result obtained by running SGD, which suggests that the approximation error, though increases a little due to randomness, is still small.

For completeness, we also test P(S)GD and N(S)GD with results shown in Figure 5 (b) and (c). For P(S)GD, we use the Hessian in linear regression experiments as the preconditioning matrix (since the Hessian of the logistic loss varies during training). Figure 5 (b) and (c) show that the averaged solutions approximately achieve the generalized/vanilla $\ell_2$-regularization effects respectively. And for the stochastic optimization paths, the approximation errors between the averaged paths and the regularized paths increase by a small amount due to the randomness of the algorithms.




(a) GD and SGD (b) PGD and PSGD (c) NGD and NSGD

*Figure 5.* Logistic regression on MNIST dataset. X-axis: iteration; y-axis: the absolute approximation error in logarithmic scale. Green lines represent $\|w_t - \hat{w}_t\|_1$ and black lines represent $\|\tilde{w}_t - \hat{w}_t\|_1$, where $w_t$, $\hat{w}_t$ and $\tilde{w}_t$ are the unregularized path, the regularized path and the iterate averaged path, respectively. Solid lines and dashed lines are the results obtained by running deterministic and stochastic algorithms respectively. For both deterministic and stochastic algorithms, we see that the error between $\tilde{w}_t$ and $\hat{w}_t$ has a small upper bound. Moreover, the error bounds for the paths generated by stochastic algorithms are relatively bigger.

*Table 1.* CIFAR-10 and CIFAR-100 experiments

| Dataset | CIFAR-10 | | CIFAR-100 |
|---|---|---|---|
| Model | VGG-16 | ResNet-18 | ResNet-18 |
| Accuracy after training (%) | 92.54 ±0.22 | 94.54 ±0.04 | 75.62 ±0.16 |
| Accuracy after averaging (%) | **93.18** ±0.06 | **94.72** ±0.04 | **76.24** ±0.05 |
| Time of training | $\sim$ 4.5h | $\sim$ 8.3h | $\sim$ 8.3h |
| Time of averaging[3] | $\sim$ 47s | $\sim$ 56s | $\sim$ 58s |

### 4.3. Application in deep neural networks

Lastly, we study the benefits of using iterate averaging in modern deep neural networks.

We train VGG-16 (Simonyan & Zisserman, 2014) and ResNet-18 (He et al., 2016) on CIFAR-10 and CIFAR-100 datasets, with standard tricks including batch normalization, data augmentation, learning rate decay and weight decay. All experiments are repeated three times to obtain the mean and deviation. The running times are measured by performing the experiments using a single GPU K80. The models are trained for 300 epochs using SGD. We perform "epoch averaging" using the 240 checkpoints saved from the 61st to the 300th epoch. The first 60 epochs are skipped since the models in the early phase are extremely unstable. After averaging the parameters, we apply a trick proposed by Izmailov et al. (2018) to handle the batch normalization statistics which are not trained by SGD. Specifically, we make a forward pass on the training data to compute the activation statistics for the batch normalization layers. For the choice of averaging scheme, we test standard geometric distribution with success probability $p \in \{0.9999, 0.999, 0.99, 0.9\}$.

[3]The time of averaging contains the time of IO and fixing BN, which takes the major overhead. For example, in CIFAR-10 and VGG-16 experiments, IO takes $\sim$ 22s, fixing BN takes $\sim$ 18s, while performing averaging and evaluation take merely $\sim$ 7s.

The results are shown in Table 1. We see that (i) averaging the SGD path does improve performance since it introduce an implicit regularization by our understanding; (ii) obtaining such regularization by iterate averaging is computationally cheap. It only takes *a few seconds* to test a hyperparameter of the averaging scheme. In contrast, several hours are required to test a hyperparameter for traditional explicit regularization since it requires re-training the model. Finally, we emphasize that the space cost of our method is also affordable. In fact, in our experiments, we perform epoch-wise averaging instead of iterate-wise averaging, thus we only need to store a few hundreds of the checkpoints.

## 5. Discussion

**$\ell_1$-regularization** Notice that all our results obtain $\ell_2$-type regularization effects. A natural follow-up question would be whether or not there is an averaging scheme that acts as an $\ell_1$-regularizer. However, we here provide some evidence that this question is relatively hard. As illustrated in Figure 6, even for simple quadratic loss, the $\ell_1$-regularized solutions could lie outside of the convex hull of a SGD path. Therefore, any averaging scheme with positive weights fails to obtain such $\ell_1$-regularized solutions.

**Infinite width neural network** Recent works suggest that a sufficient wide neural network trained by SGD behaves like a quadratic model, i.e., the neural tangent kernel (NTK) (Jacot et al., 2018; Arora et al., 2019; Cao & Gu, 2019). Nonetheless, the NTK approximation fails when there is an explicit $\ell_2$-regularizer (Wei et al., 2019). Since our results hold for kernel ridge regression, we conjecture that iterate averaging could be a potential approach to achieve $\ell_2$-regularization for the NTK regime. We leave

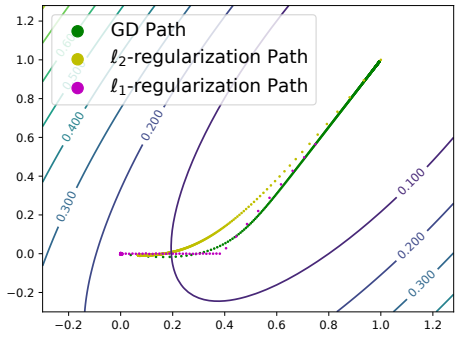



*Figure 6.* A 2-D demonstration of the $\ell_1$-regularization path. Green dots: the vanilla GD path $w_t$; yellow dots: the $\ell_2$-regularization path $\hat{w}_{\lambda,\ell_2}$; purple dots: the $\ell_1$-regularization path $\hat{w}_{\lambda,\ell_1}$. There exist $\ell_1$-regularized solutions outside of the convex hull of the GD path, while all of the $\ell_2$-regularized solutions are inside of that.

further investigation of this issue in future works.

## 6. Conclusions

In this work, we establish averaging schemes for various optimization methods and objective functions to obtain adjustable $\ell_2$-type regularization effects, i.e., SGD with preconditioning and adaptive learning rate schedules, Nesterov's accelerated stochastic gradient descent, and strongly convex and smooth objective functions. Particularly, we resolve an open question in (Neu & Rosasco, 2018). The method of achieving regularization by iterate averaging requires little computation. It is further shown experimentally that iterate averaging even benefits practical deep learning models. Our theoretical and empirical results demonstrate the potential of adopting iterate averaging to obtain adjustable regularization for free in a much broader class of optimization methods and objective functions.

## Acknowledgement

This research is supported in part by NSF CAREER grant 1652257, ONR Award N00014-18-1-2364 and the Lifelong Learning Machines program from DARPA/MTO.

## References

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019.

Bach, F. and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pp. 773–781, 2013.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Bottou, L. and Bousquet, O. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pp. 161–168, 2008.

Cai, Y., Li, Q., and Shen, Z. A quantitative analysis of the effect of batch normalization on gradient descent. *arXiv preprint arXiv:1810.00122*, 2018.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210*, 2019.

Clark, D. S. Short proof of a discrete gronwall inequality. *Discrete applied mathematics*, 16(3):279–281, 1987.

Dennis Jr, J. E. and Schnabel, R. B. *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16. Siam, 1996.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Grandvalet, Y. and Bengio, Y. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pp. 529–536, 2005.

Granziol, D., Wan, X., and Roberts, S. Iterate averaging helps: An alternative perspective in deep learning. *arXiv preprint arXiv:2003.01247*, 2020.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. doi: 10.1109/cvpr.2016.90. URL http://dx.doi.org/10.1109/CVPR.2016.90.

Hu, W., Li, C. J., Li, L., and Liu, J.-G. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017a.

Hu, W., Li, C. J., and Su, W. On the global convergence of a randomly perturbed dissipative nonlinear oscillator. *arXiv preprint arXiv:1712.05733*, 2017b.

Hu, W., Xiao, L., and Pennington, J. Provable benefit of orthogonal initialization in optimizing deep linear networks. *arXiv preprint arXiv:2001.05992*, 2020.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Jain, P., Kakade, S., Kidambi, R., Netrapalli, P., and Sidford, A. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18, 2018.

Krogh, A. and Hertz, J. A. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pp. 950–957, 1992.

Lakshminarayanan, C. and Szepesvari, C. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pp. 1347–1355, 2018.

Li, Q., Tai, C., et al. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2101–2110. JMLR. org, 2017.

Martens, J. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Nesterov, Y. E. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, 1983.

Neu, G. and Rosasco, L. Iterate averaging as regularization for stochastic gradient descent. *arXiv preprint arXiv:1802.08009*, 2018.

Shi, B., Du, S. S., Su, W., and Jordan, M. I. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, pp. 5745–5753, 2019.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Su, W., Boyd, S., and Candes, E. A differential equation for modeling nesterovs accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pp. 2510–2518, 2014.

Suggala, A., Prasad, A., and Ravikumar, P. K. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, pp. 10608–10619, 2018.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tikhonov, A. N. and Arsenin, V. Y. *Solutions of ill-posed problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York, 1977. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pp. 9709–9721, 2019.

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pp. 4148–4158, 2017.

Yang, L., Arora, R., Zhao, T., et al. The physical systems behind optimization algorithms. In *Advances in Neural Information Processing Systems*, pp. 4372–4381, 2018.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhang, M., Lucas, J., Ba, J., and Hinton, G. E. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, pp. 9593–9604, 2019.

Zhou, X. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.

Zhu, Z., Wu, J., Yu, B., Wu, L., and Ma, J. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

## A. Continuous analysis

To motivate our proofs for the theorems in main text, let us first elaborate the continuous cases. Then we will extend our analysis to the discrete circumstances. One can safely skip this part and go directly to Section C for the missing proofs in main text, which is self-consistent.

**Continuous optimization paths** To ease notations and preliminaries, in this part we only discuss gradient descent (GD) and Nesterov's accelerated gradient descent (NGD), and their strong continuous approximation via ordinary differential equations (ODEs). For SGD and NSGD, existing works show that there are weak continuous approximation by stochastic differential equations (SDEs) (Hu et al., 2017a;b; Li et al., 2017). Our analysis can be extended to SDEs, but we believe it serves better to motivate our discrete proofs by focusing on ODEs.

We consider loss $L(w)$ and $\ell_2$-regularizer $R(w) = \frac{1}{2}\|w\|_2^2$. Let the learning rate $\eta \to 0$, the path of $L(w)$ optimized by GD converges to the following ODE (Yang et al., 2018)

$$\mathrm{d}w_t = -\nabla L(w_t)\mathrm{d}t.$$

Similarly the continuous GD optimization path of regularized loss admits

$$\mathrm{d}\hat{w}_t = -\left(\nabla L(\hat{w}_t) + \lambda \hat{w}_t\right)\mathrm{d}t.$$

As for NGD, Su et al. (2014); Yang et al. (2018) show if the loss is $\alpha$-strongly convex, then the NGD optimization path converges to

$$w_t'' + 2\sqrt{\alpha} w_t' + L'(w_t) = 0.$$

Since $\hat{L}(\hat{w}) = L(\hat{w}) + \frac{\lambda}{2}\|\hat{w}\|_2^2$ is $(\alpha+\lambda)$-strongly convex, the NGD path of the regularized loss satisfies

$$\hat{w}_t'' + 2\sqrt{\alpha+\lambda}\hat{w}_t' + L'(\hat{w}_t) + \lambda \hat{w}_t = 0.$$

**Continuous weighting scheme** We define the continuous weighting scheme as

$$p_t \geq 0, \quad t \geq 0, \quad P_t = \int_0^t p(s)\mathrm{d}s, \quad \lim_{t\to\infty} P_t = 1.$$

**Lemma 1.** *Given two continuous dynamic $x_t$, $\hat{x}_t$, $t \geq 0$. Let $\tilde{x}_t = P_t^{-1}\int_0^t p_s x_s \mathrm{d}s$. Suppose $x_0 = \hat{x}_0 = 0$. If the continuous weighting scheme $P_t$ satisfies*

$$\mathrm{d}\hat{x}_t = (1-P_t)\mathrm{d}x_t, \quad t \geq 0,$$

*then we have*

$$P_t(x_t - \tilde{x}_t) = x_t - \hat{x}_t, \quad t \geq 0,$$

*and*

$$\hat{x}_t - \tilde{x}_t = (1-P_t)(x_t - \tilde{x}_t), \quad t \geq 0.$$

*Proof.* By definition we have for $t \geq 0$,

$$\begin{aligned}
\tilde{x}_t &= P_t^{-1}\int_0^t p_s x_s \mathrm{d}s = P_t^{-1}\left(x_s P_s|_0^t - \int_0^t P_s \mathrm{d}x_s\right) = x_t - P_t^{-1}\int_0^t P_s \mathrm{d}x_s \\
&= x_t - P_t^{-1}\left(x_t - \int_0^t (1-P_s)\mathrm{d}x_s\right) = x_t - P_t^{-1}\left(x_t - \int_0^t \mathrm{d}\hat{x}_s\right) \\
&= x_t - P_t^{-1}\left(x_t - \hat{x}_t\right).
\end{aligned}$$

Thus

$$P_t(x_t - \tilde{x}_t) = x_t - \hat{x}_t,$$

and

$$\hat{x}_t - \tilde{x}_t = x_t - P_t(x_t - \tilde{x}_t) - \tilde{x}_t = (1-P_t)(x_t - \tilde{x}_t).$$

□

### A.1. Continuous Theorem 1

Consider linear regression problem $L(w) = \frac{1}{2n}\sum_{i=1}^{n}\left\|w^\top x_i - y_i\right\|_2^2 = \frac{1}{2}w^\top \Sigma w - w^\top a + \text{const}$, and $\ell_2$-regularizer $R(w) = \frac{1}{2}\|w\|_2^2$. Assume the initial condition $w_0 = \hat{w}_0 = 0$, then the GD dynamics for the unregularized and regularized losses are

$$\begin{aligned}\mathrm{d}w_t &= -\left(\Sigma w_t - a\right)\mathrm{d}t, \quad w_0 = 0,\\ \mathrm{d}\hat{w}_t &= -\left(\Sigma \hat{w}_t - a + \lambda \hat{w}_t\right)\mathrm{d}t, \quad \hat{w}_0 = 0.\end{aligned}$$

The ODEs are solved by

$$w_t = \left(I - e^{-\Sigma t}\right)\Sigma^{-1}a, \quad \hat{w}_t = \left(I - e^{-(\Sigma+\lambda I)t}\right)(\Sigma+\lambda I)^{-1}a.$$

Now let the continuous weighting scheme be

$$P_t = 1 - e^{\lambda t},$$

then we have

$$\mathrm{d}\hat{w}_t = (1-P_t)\mathrm{d}w_t,$$

thus by Lemma 1 we obtain

$$\hat{w}_t - \tilde{w}_t = (1-P_t)(w_t - \tilde{w}_t),$$

which proves the continuous version of Theorem 1.

### A.2. Continuous Theorem 3

Consider linear regression problem $L(w) = \frac{1}{2n}\sum_{i=1}^{n}\left\|w^\top x_i - y_i\right\|_2^2 = \frac{1}{2}w^\top \Sigma w - w^\top a + \text{const}$, and $\ell_2$-regularizer $R(w) = \frac{1}{2}\|w\|_2^2$. Assume the initial condition $w_0 = w_0' = 0$ and $\hat{w}_0 = \hat{w}_0' = 0$. Then the unregularized and regularized NGD dynamics are

$$w_t'' + 2\sqrt{\alpha}w_t' + \Sigma w_t - a = 0, \quad w_0 = w_0' = 0, \tag{7}$$

$$\hat{w}_t'' + 2\sqrt{\alpha+\lambda}\hat{w}_t' + (\Sigma+\lambda)\hat{w}_t - a = 0, \quad \hat{w}_0 = \hat{w}_0' = 0. \tag{8}$$

We first solve the order-2 ODE Eq. (7) in the canonical way, and then obtain the solution of Eq. (8) similarly. To do so, let's firstly ignore the constant term and solve the homogenous ODE of Eq. (7), and obtain two general solutions of the homogenous equation as

$$w_{t,1} = e^{\sqrt{\alpha}t}\cos\sqrt{\Sigma-\alpha}t, \quad w_{t,2} = e^{\sqrt{\alpha}t}\sin\sqrt{\Sigma-\alpha}t.$$

Then we guess a particular solution of Eq. (7) as $w_{t,0} = \Sigma^{-1}a$. Thus the general solution of ODE (7) can be decomposed as $w_t = \lambda_1 w_{t,1} + \lambda_2 w_{t,2} + w_{t,0}$. Consider the initial conditions $w_0 = w_0' = 0$, we obtain $\lambda_1 = -\Sigma^{-1}a$, $\lambda_2 = -\Sigma^{-1}a\sqrt{(\Sigma-\alpha)^{-1}\alpha}$. Thus the solution of Eq. (7) is

$$\begin{aligned}w_t &= \Sigma^{-1}a\left(1 - e^{-\sqrt{\alpha}t}\cos\sqrt{\Sigma-\alpha}t - \sqrt{\alpha(\Sigma-\alpha)^{-1}}e^{-\sqrt{\alpha}t}\sin\sqrt{\Sigma-\alpha}t\right),\\ w_t' &= a\sqrt{(\Sigma-\alpha)^{-1}}e^{-\sqrt{\alpha}t}\sin\sqrt{\Sigma-\alpha}t.\end{aligned} \tag{9}$$

Repeat these procedures, Eq. (9) is solved by

$$\begin{aligned}\hat{w}_t &= (\Sigma+\lambda)^{-1}a\left(1 - e^{-\sqrt{\alpha+\lambda}t}\cos\sqrt{\Sigma-\alpha}t - \sqrt{(\alpha+\lambda)(\Sigma-\alpha)^{-1}}e^{-\sqrt{\alpha+\lambda}t}\sin\sqrt{\Sigma-\alpha}t\right),\\ \hat{w}_t' &= a\sqrt{(\Sigma-\alpha)^{-1}}e^{-\sqrt{\alpha+\lambda}t}\sin\sqrt{\Sigma-\alpha}t.\end{aligned} \tag{10}$$

Now let the continuous weighting scheme be

$$P_t = 1 - e^{-(\sqrt{\alpha+\lambda}-\sqrt{\lambda})t},$$

then we have

$$\mathrm{d}\hat{w}_t = (1-P_t)\mathrm{d}w_t,$$

thus by Lemma 1 we obtain

$$\hat{w}_t - \tilde{w}_t = (1-P_t)(w_t - \tilde{w}_t),$$

which proves the continuous version of Theorem 3.

### A.3. Continuous Theorem 4

Consider an $\alpha$-strongly convex and $\beta$-smooth loss function $L(w)$, and $\ell_2$-regularizer. Without loss of generality assume the minimum of $L(w)$ satisfies $w_* > w_0 = 0$. Then by Lemma 3 we have

$$\alpha w - b \leq \nabla L(w) \leq \beta w - b, \quad \forall w \in (0, w_*),$$

where $b = -\nabla L(0)$, and "$\leq$" is defined entry-wisely. We study the continuous optimization paths caused by GD.

Consider the following three dynamics:

$$\mathrm{d}w_t = -\nabla L(w_t)\mathrm{d}t, \quad \mathrm{d}u_t = -(\alpha u_t - b)\mathrm{d}t, \quad \mathrm{d}v_t = -(\beta v_t - b)\mathrm{d}t, \quad w_0 = u_0 = v_0 = 0.$$

By the comparison theorem of ODEs (Gronwall's inequality), and solution of linear ODEs, we claim that for all $t > 0$,

$$v_t \leq w_t \leq u_t, \quad u_t = \frac{b}{\alpha}(1 - e^{-\alpha t}), \quad v_t = \frac{b}{\beta}(1 - e^{-\beta t}). \tag{11}$$

In a similar manner, for the following three dynamics of regularized loss:

$$\mathrm{d}\hat{w}_{t,\lambda} = -(\nabla L(\hat{w}_{t,\lambda}) + \lambda \hat{w}_{t,\lambda})\mathrm{d}t, \qquad \mathrm{d}\hat{u}_{t,\lambda} = -((\lambda+\alpha)\hat{u}_{t,\lambda} - b)\mathrm{d}t, \quad \mathrm{d}\hat{v}_{t,\lambda} = -((\lambda+\beta)\hat{v}_{t,\lambda} - b)\mathrm{d}t,$$

where $\hat{w}_{0,\lambda} = \hat{u}_{0,\lambda} = \hat{v}_{0,\lambda} = 0$. Similarly we have for all $t > 0$,

$$\hat{v}_{t,\lambda} \leq \hat{w}_{t,\lambda} \leq \hat{u}_{t,\lambda}, \quad \hat{u}_{t,\lambda} = \frac{b}{\lambda+\alpha}(1 - e^{-(\lambda+\alpha)t}), \quad \hat{v}_{t,\lambda} = \frac{b}{\lambda+\beta}(1 - e^{-(\lambda+\beta)t}).$$

For the continuous weighting scheme

$$P_t = 1 - e^{-\zeta t}, \quad p_t = \zeta e^{-\zeta t}, \quad t \geq 0, \quad \zeta > 0,$$

the averaged solution is defined as $\tilde{w}_t = P_t^{-1}\int_0^t p_t w_t \mathrm{d}t = w_t - P_t^{-1}\int_0^t P_s \mathrm{d}w_s$, similar there are $\tilde{u}_t, \tilde{v}_t$. Thanks to Eq. (11) and $p_t$ being non-negative, we have $\tilde{v}_t \leq \tilde{w}_t \leq \tilde{u}_t$. Let

$$\lambda_1 = \zeta + \beta - \alpha, \quad \lambda_2 = \zeta + \alpha - \beta,$$

then

$$\begin{aligned} P_t(u_t - \tilde{u}_t) = \int_0^t P_s \mathrm{d}u_s = \int_0^t (1 - e^{-(\lambda_2+\beta-\alpha)s}) b e^{-\alpha s}\mathrm{d}t = b\int_0^t e^{-\alpha s} - e^{-(\beta+\lambda_2)s}\mathrm{d}s \\ = b\left(\frac{1}{\alpha}(1 - e^{-\alpha t}) - \frac{1}{\lambda_2+\beta}(1 - e^{-(\lambda_2+\beta)t})\right) = u_t - \hat{v}_{t,\lambda_2}. \end{aligned}$$

Thus

$$\tilde{w}_t - \hat{w}_{t,\lambda_2} \leq \tilde{u}_t - \hat{v}_{t,\lambda_2} = \tilde{u}_t - u_t + P_t(u_t - \tilde{u}_t) = (1 - P_t)(\tilde{u}_t - u_t).$$

Similarly, since

$$\begin{aligned} P_t(v_t - \tilde{v}_t) = \int_0^t P_s \mathrm{d}v_s = \int_0^t (1 - e^{-(\lambda_1-\beta+\alpha)s}) b e^{-\beta s}\mathrm{d}t = b\int_0^t e^{-\beta s} - e^{-(\alpha+\lambda_1)s}\mathrm{d}s \\ = b\left(\frac{1}{\beta}(1 - e^{-\beta t}) - \frac{1}{\lambda_1+\alpha}(1 - e^{-(\lambda_1+\alpha)t})\right) = v_t - \hat{u}_{t,\lambda_1}, \end{aligned}$$

we can obtain a lower bound as

$$\tilde{w}_t - \hat{w}_{t,\lambda_1} \geq \tilde{v}_t - \hat{u}_{t,\lambda_1} = \tilde{v}_t - v_t + P_t(v_t - \tilde{v}_t) = (1 - P_t)(\tilde{v}_t - v_t).$$

These inequalities give us

$$\hat{w}_{t,\lambda_1} + (1 - P_t)(\tilde{v}_t - v_t) \leq \tilde{w}_t \leq \hat{w}_{t,\lambda_2} + (1 - P_t)(\tilde{u}_t - u_t),$$

which proves the continuous version of Theorem 4.

## B. Technical Lemmas

**Lemma 2.** *Consider two series* $\{x_k\}_{k=0}^{\infty}$, $\{\hat{x}_k\}_{k=0}^{\infty}$, *and a weighting scheme* $\{p_k\}_{k=0}^{\infty}$ *such that* $\sum_{k=0}^{\infty} p_k = 1,\ p_k \geq 0$, $P_k = \sum_{i=1}^{k} p_i$. *Let* $\tilde{x}_k := P_k^{-1} \sum_{i=0}^{k} p_i x_i$. *Suppose* $x_0 = \hat{x}_0 = 0$. *Suppose the weighting scheme* $P_k$ *satisfies*

$$\hat{x}_{k+1} - \hat{x}_k = (1 - P_k)(x_{k+1} - x_k), \quad k \geq 0.$$

*Then we have*

$$P_k(x_k - \tilde{x}_k) = x_k - \hat{x}_k, \quad k \geq 0,$$

*and*

$$\hat{x}_k - \tilde{x}_k = (1 - P_k)\,(x_k - \tilde{x}_k), \quad k \geq 0.$$

*More generally, the weighting scheme* $\{p_k\}_{k=0}^{\infty}$ *could be a series of positive semi-definite matrix where*

$$\lim_{k \to +\infty} P_k = I, \quad 0 \preceq P_k \preceq I, \quad p_k = P_k - P_{k-1}.$$

*Proof.* By definition we know $p_0 = P_0,\ p_k = P_k - P_{k-1},\ k \geq 1$, and

$$\begin{aligned} P_k \tilde{x}_k =& \sum_{i=1}^{k} p_i x_i = \sum_{i=1}^{k} (P_i - P_{i-1}) x_i = \sum_{i=1}^{k} P_i x_i - \sum_{i=1}^{k} P_{i-1} x_i \\ =& P_k x_k + \sum_{i=1}^{k} P_{i-1} x_{i-1} - \sum_{i=1}^{k} P_{i-1} x_i = P_k x_k - \sum_{i=1}^{k} P_{i-1}(x_i - x_{i-1}). \end{aligned}$$

Therefore

$$\begin{aligned} P_k(x_k - \tilde{x}_k) =& \sum_{i=1}^{k} P_{i-1}(x_i - x_{i-1}) = \sum_{i=1}^{k} (x_i - x_{i-1}) - \sum_{i=1}^{k} (1 - P_{i-1})(x_i - x_{i-1}) \\ =& x_k - \sum_{i=1}^{k} (1 - P_{i-1})(x_i - x_{i-1}). \end{aligned}$$

Now use the assumption, we obtain

$$P_k(x_k - \tilde{x}_k) = x_k - \sum_{i=1}^{k} (\hat{x}_i - \hat{x}_{i-1}) = x_k - \hat{x}_k, \quad k \geq 1.$$

Thus we have

$$\hat{x}_k - \tilde{x}_k = x_k - P_k(x_k - \tilde{x}_k) - \tilde{x}_k = (1 - P_k)\,(x_k - \tilde{x}_k), \quad k \geq 1.$$

One can directly verify that the above equation also holds for $k = 0$, which concludes our proof. □

**Lemma 3.** *Let* $x \in \mathbb{R}$. *Let* $f(x)$ *be* $\alpha$*-strongly convex and* $\beta$*-smooth,* $0 < \alpha \leq \beta$. *Let* $f(x)$ *be lower bounded, then* $x_* = \arg\min_{x \in \mathbb{R}} f(x)$ *exists. Consider GD with learning rate* $\eta \in (0, \frac{1}{\beta})$, *the optimization path* $\{x_k\}_{k=0}^{+\infty}$ *is given by*

$$x_{k+1} = x_k - \eta \nabla f(x_k).$$

*If* $x_0 < x_*$, *then we have*

1. *For all* $k > 0$, $x_k \in (x_0, x_*)$.
2. *For all* $x \in (x_0, x_*)$, *we have* $\beta(x - x_*) \leq \nabla f(x) \leq \alpha(x - x_*)$.
3. *For all* $x \in (x_0, x_*)$, *we have* $\alpha(x - x_0) + \nabla f(x_0) \leq \nabla f(x) \leq \beta(x - x_0) + \nabla f(x_0)$.

*Similarly if* $x_0 > x_*$, *then we have*

1. *For all* $k > 0$, $x_k \in (x_*, x_0)$.
2. *For all* $x \in (x_*, x_0)$, *we have* $\alpha(x - x_*) \le \nabla f(x) \le \beta(x - x_*)$.
3. *For all* $x \in (x_*, x_0)$, *we have* $\beta(x - x_0) + \nabla f(x_0) \le \nabla f(x) \le \alpha(x - x_0) + \nabla f(x_0)$.

*Proof.* We only prove Lemma 3 in case of $x_0 < x_*$. The other case is true in a similar manner.

To prove the first conclusion we only need to show that $x_0 < x_1 < x_*$, then recursively we obtain $x_0 < x_1 < \cdots < x_k < x_*$.

Note that $\nabla f(x_*) = 0$. Since $f(x)$ is $\alpha$-strongly convex and $\beta$-smooth, we have (Zhou, 2018)

$$\alpha(x - y)^2 \le (\nabla f(x) - \nabla f(y))(x - y) \le \beta(x - y)^2.$$

Thus $\alpha(x_* - x_0)^2 \le -\nabla f(x_0)(x_* - x_0) \le \beta(x_* - x_0)^2$. Now by the assumption that $x_0 < x_*$, we obtain $0 < \alpha(x_* - x_0) \le -\nabla f(x_0) \le \beta(x_* - x_0)$. Hence

$$\begin{aligned} x_1 &= x_0 - \eta \nabla f(x_0) > x_0 \\ x_1 &= x_0 - \eta \nabla f(x_0) < x_0 + \eta\beta(x_* - x_0) < x_0 + x_* - x_0 < x_*. \end{aligned}$$

To prove the second conclusion, recall that $\alpha(x_* - x)^2 \le -\nabla f(x)(x_* - x) \le \beta(x_* - x)^2$, thus for $x \in (x_0, x_*)$, we obtain $\alpha(x_* - x) \le -\nabla f(x) \le \beta(x_* - x)$.

As for the third conclusion, since $\alpha(x - x_0)^2 \le (\nabla f(x) - \nabla f(x_0))(x - x_0) \le \beta(x - x_0)^2$, thus for $x \in (x_0, x_*)$, we obtain $\alpha(x - x_0) + \nabla f(x_0) \le \nabla f(x) \le \beta(x - x_0) + \nabla f(x_0)$. which completes our proof. □

## C. Missing proofs in main text

### C.1. Proof of Theorem 1

*Proof.* The first part of the theorem is an extension of Proposition 1 and Proposition 2 in (Neu & Rosasco, 2018). Beyond the analysis of constant learning rate in (Neu & Rosasco, 2018), we show the corresponding results for adaptive learning rates.

Recall the SGD updates for linear regression problem

$$w_{k+1} = w_k - \eta_k(x_{k+1}x_{k+1}^\top w_k - x_{k+1}y_{k+1}), \quad w_0 = 0.$$

Let

$$\Sigma = \mathbb{E}_x[xx^\top], \quad a = \mathbb{E}_{x,y}[xy], \quad w_* = \Sigma^{-1}a, \quad \epsilon_k = (\Sigma w_k - a) - (x_{k+1}x_{k+1}^\top w_k - x_{k+1}y_{k+1}),$$

where $\epsilon_k$ is the gradient noise, and $\mathbb{E}_{k+1}[\epsilon_k] = 0$. Under these notations we have

$$w_{k+1} = w_k - \eta_k(\Sigma w_k - a) + \eta_k\epsilon_k = w_k - \eta_k\Sigma(w_k - w_*) + \eta_k\epsilon_k, \quad w_0 = 0. \tag{12}$$

Similarly for linear regression with $\ell_2$-regularization, SGD takes update

$$\hat{w}_{k+1} = \hat{w}_k - \gamma_k(x_{k+1}x_{k+1}^\top \hat{w}_k - x_{k+1}y_{k+1} + \lambda\hat{w}_k), \quad \hat{w}_0 = 0.$$

Let

$$\hat{w}_* = (\Sigma + \lambda I)^{-1}a,$$

then

$$\hat{w}_{k+1} = \hat{w}_k - \gamma_k(\Sigma\hat{w}_k - a + \lambda\hat{w}_k) + \gamma_k\epsilon_k = \hat{w}_k - \gamma_k(\Sigma + \lambda I)(\hat{w}_k - \hat{w}_*) + \gamma_k\epsilon_k, \quad \hat{w}_0 = 0. \tag{13}$$

**Expectations** First let us compute the expectations. For Eq. (12), after taking expectation at time $k+1$, we have

$$\mathbb{E}_{k+1}[w_{k+1}] = w_k - \eta_k \Sigma(w_k - w_*).$$

Then recursively taking expectation at time $k, \ldots, 1$, we obtain

$$\mathbb{E}[w_{k+1}] = \mathbb{E}[w_k] - \eta_k \Sigma(\mathbb{E}[w_k] - w_*), \quad \mathbb{E}[w_0] = w_0 = 0.$$

Solving the above recurrence relation we have

$$\mathbb{E}[w_k] - w_* = \Pi_{i=0}^{k-1}(I - \eta_i \Sigma)(w_0 - w_*), \quad w_0 = 0, \quad w_* = \Sigma^{-1} a,$$

hence

$$\mathbb{E}[w_{k+1}] - \mathbb{E}[w_k] = -\Pi_{i=0}^{k-1}(I - \eta_i \Sigma)\eta_k \Sigma(w_0 - w_*) = \Pi_{i=0}^{k-1}(I - \eta_i \Sigma)\eta_k a, \quad \mathbb{E}[w_0] = 0.$$

In a same way we can solve Eq. (13) in expectation and obtain

$$\mathbb{E}[\hat{w}_{k+1}] - \mathbb{E}[\hat{w}_k] = \Pi_{i=0}^{k-1}(I - \gamma_i(\Sigma + \lambda I))\gamma_k a, \quad \mathbb{E}[\hat{w}_0] = 0.$$

Notice that the weighting scheme is defined by

$$P_k = 1 - \Pi_{i=0}^{k}(1 - \lambda\gamma_i),$$

and $1 - \lambda\gamma_i = \frac{\gamma_i}{\eta_i}$, we can directly verify that

$$\mathbb{E}[\hat{w}_{k+1}] - \mathbb{E}[\hat{w}_k] = (1 - P_k)(\mathbb{E}[w_{k+1}] - \mathbb{E}[w_k]).$$

Thus by Lemma 2, we know that

$$P_k \mathbb{E}[\tilde{w}_k] = \mathbb{E}[\hat{w}_k] - (1 - P_k)\mathbb{E}[w_k], \quad k \geq 0.$$

Hence the first conclusion holds.

**Convergence** By assumptions we know $0 < \eta \leq \eta_i < \frac{1}{\beta} \leq \frac{1}{\lambda_{\max}}$, where $\lambda_{\max}$ is the largest eigenvalue of $\Sigma$. Thus

$$\|\mathbb{E}[w_k] - w_*\|_2 \leq \left\|\Pi_{i=0}^{k-1}(I - \eta_i \Sigma)\right\|_2 \cdot \|w_0 - w_*\|_2 \leq \left\|(I - \eta\Sigma)^k\right\|_2 \cdot \|w_0 - w_*\|_2 \to 0,$$

and $\lim_{k \to +\infty} \mathbb{E}[w_k] = w_* = \Sigma^{-1} a$.

In a similar manner, since $\gamma_i = \frac{\eta_i}{1 + \eta_i \lambda}$ and $0 < \eta \leq \eta_i < \frac{1}{\beta} \leq \frac{1}{\lambda_{\max}}$, we have $0 < \frac{\eta}{1+\lambda\eta} = \gamma \leq \gamma_i < \frac{1}{\beta + \lambda} \leq \frac{1}{\lambda_{\max} + \lambda}$. Thus

$$\|\mathbb{E}[\hat{w}_k] - \hat{w}_*\|_2 \leq \left\|\Pi_{i=0}^{k-1}(I - \gamma_i(\Sigma + \lambda I))\right\|_2 \cdot \|\hat{w}_0 - \hat{w}_*\|_2 \leq \left\|(I - \gamma(\Sigma + \lambda I))^k\right\|_2 \cdot \|\hat{w}_0 - \hat{w}_*\|_2 \to 0,$$

and $\lim_{k \to +\infty} \mathbb{E}[\hat{w}_k] = \hat{w}_* = (\Sigma + \lambda I)^{-1} a$.

On the other hand, by the first conclusion we know

$$\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k] = (1 - P_k)(\mathbb{E}[w_k] - \mathbb{E}[\tilde{w}_k]).$$

Since $\mathbb{E}[w_k]$ converges, $\mathbb{E}[\tilde{w}_k] = P_k^{-1} \sum_{i=1}^{k} p_i \mathbb{E}[w_i]$ is bounded. Therefore

$$\|\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k]\|_2 = (1 - P_k) \|\mathbb{E}[w_k] - \mathbb{E}[\tilde{w}_k]\|_2 = \mathcal{O}(1 - P_k) = \mathcal{O}(\Pi_{i=0}^{k}(1 - \lambda\gamma_i)) \leq \mathcal{O}((1 - \lambda\gamma)^k).$$

Hence the second claim is true.

**Variance** Now we turn to analyze the deviation of the averaged solution. From Eq. (12), we can recursively obtain

$$w_i = \mathbb{E}[w_i] + \xi_i, \quad \xi_i = \sum_{j=0}^{i-1} \Pi_{h=j+1}^{i-1}(I - \eta_h \Sigma)\eta_j \epsilon_j,$$

where we abuse the notation and let $\Pi_{h=i}^{i-1}(I - \eta_h\Sigma) = I$.

Now applying iterate averaging with respect to $p_i = \lambda\gamma_i \Pi_{h=0}^{i-1}(1 - \lambda\gamma_h)$, we have

$$P_k \tilde{w}_k = \sum_{i=1}^{k} p_i w_i = \sum_{i=1}^{k} p_i \mathbb{E}[w_i] + \sum_{i=1}^{k} p_i \xi_i = P_k \mathbb{E}[\tilde{w}_k] + \sum_{i=1}^{k} p_i \xi_i.$$

We turn to calculate the noise term $\sum_{i=1}^{k} p_i \xi_i$. Note that in every step, all of the matrices can be diagonalized simultaneously, thus they commute, similarly hereinafter.

$$\begin{aligned}
&\sum_{i=1}^{k} p_i \xi_i = \sum_{i=1}^{k} p_i \left( \sum_{j=0}^{i-1} \Pi_{h=j+1}^{i-1}(I - \eta_h\Sigma)\eta_j \epsilon_j \right) \\
=&\sum_{j=0}^{k-1} \left( \sum_{i=j+1}^{k} p_i \Pi_{h=j+1}^{i-1}(I - \eta_h\Sigma)\eta_j \right) \epsilon_j \\
=&\sum_{j=0}^{k-1} \left( \sum_{i=j+1}^{k} \lambda\gamma_i \Pi_{h=0}^{i-1}(1 - \lambda\gamma_h) \Pi_{h=j+1}^{i-1}(I - \eta_h\Sigma)\eta_j \right) \epsilon_j \\
=&\sum_{j=0}^{k-1} \left( \sum_{i=j+1}^{k} \lambda\gamma_i \left( \Pi_{h=0}^{j-1}(1 - \lambda\gamma_h) \right) \left( \Pi_{h=j+1}^{i-1}(1 - \lambda\gamma_h)(I - \eta_h\Sigma) \right) ((1 - \lambda\gamma_j)\eta_j) \right) \epsilon_j \\
=&\sum_{j=0}^{k-1} \left( \left( \Pi_{h=0}^{j-1}(1 - \lambda\gamma_h) \right) \left( \sum_{i=j+1}^{k} \lambda\gamma_i \Pi_{h=j+1}^{i-1} \left( I - \gamma_h(\Sigma + \lambda I) \right) \right) \gamma_j \right) \epsilon_j \\
=&\sum_{j=0}^{k-1} A_j \epsilon_j,
\end{aligned}$$

where $A_j = \gamma_j \left( \Pi_{h=0}^{j-1}(1 - \lambda\gamma_h) \right) \left( \sum_{i=j+1}^{k} \lambda\gamma_i \Pi_{h=j+1}^{i-1} \left( I - \gamma_h(\Sigma + \lambda I) \right) \right)$. Recall that $\epsilon_0, \epsilon_1 \ldots, \epsilon_k$ is a martingale difference sequence, then $\sum_{i=1}^{k} p_i \xi_i = \sum_{j=0}^{k-1} A_j \epsilon_j$ is a martingale. Thus

$$\operatorname{Tr} \operatorname{Var} \left[ \sum_{i=1}^{k} p_i \xi_i \right] = \operatorname{Tr} \operatorname{Var} \left[ \sum_{j=0}^{k-1} A_j \epsilon_j \right] = \sum_{j=0}^{k-1} \operatorname{Tr} \operatorname{Var} \left[ A_j \epsilon_j \right],$$

where "Var" is the covariance of a random vector. and "Tr" is the trace of a matrix.

Next we bound each term in the summation as

$$\operatorname{Tr} \operatorname{Var} \left[ A_j \epsilon_j \right] = \operatorname{Tr} \mathbb{E} \left[ (A_j \epsilon_j)(A_j \epsilon_j)^\top \right] = \mathbb{E} \left[ \|A_j \epsilon\|_2^2 \right] \leq \|A_j\|_2^2 \cdot \mathbb{E} \left[ \|\epsilon\|_2^2 \right] \leq \sigma^2 \|A_j\|_2^2 .$$

And we remain to bound $\|A_j\|_2^2$. Remember that $\eta \le \eta_h \le \frac{1}{\beta},\ \gamma \le \gamma_h \le \frac{1}{\lambda+\beta}$, we have

$$\begin{aligned}
\|A_j\|_2^2 &= \left\| \gamma_j \left( \Pi_{h=0}^{j-1}(1-\lambda\gamma_h) \right) \left( \sum_{i=j+1}^{k} \lambda\gamma_i \Pi_{h=j+1}^{i-1} \left( I - \gamma_h(\Sigma+\lambda I) \right) \right) \right\|_2^2 \\
&\le \left\| \frac{1}{\lambda+\beta} \left( (1-\lambda\gamma)^j \right) \left( \sum_{i=j+1}^{k} \frac{\lambda}{\lambda+\beta} \left( I - \gamma(\Sigma+\lambda I) \right)^{i-j-1} \right) \right\|_2^2 \\
&= \left\| \frac{\lambda}{(\lambda+\beta)^2} \left( (1-\lambda\gamma)^j \right) \left( \sum_{i=0}^{k-j-1} \left( I - \gamma(\Sigma+\lambda I) \right)^i \right) \right\|_2^2 \\
&\le \left( \frac{\lambda}{(\lambda+\beta)^2} \left( (1-\lambda\gamma)^j \right) \left( \sum_{i=0}^{k-j-1} \left( 1 - \gamma(\alpha+\lambda) \right)^i \right) \right)^2 \\
&\le \left( \frac{\lambda}{(\lambda+\beta)^2} \left( (1-\lambda\gamma)^j \right) \left( \frac{1}{\gamma(\alpha+\lambda)} \right) \right)^2 \\
&= \frac{\lambda^2}{\gamma^2(\lambda+\alpha)^2(\lambda+\beta)^4} (1-\lambda\gamma)^{2j}.
\end{aligned}$$

The second equality holds because $\alpha \le \lambda_{min}(\Sigma)$.

Based on previous discussion we have

$$\begin{aligned}
&\operatorname{Tr}\operatorname{Var}\left[ \sum_{i=1}^{k} p_i \xi_i \right] = \sum_{j=0}^{k-1} \operatorname{Tr}\operatorname{Var}\left[ A_j \epsilon_j \right] \le \sum_{j=0}^{k-1} \sigma^2 \|A_j\|_2^2 \\
\le &\sum_{j=0}^{k-1} \frac{\lambda^2\sigma^2}{\gamma^2(\lambda+\alpha)^2(\lambda+\beta)^4} (1-\lambda\gamma)^{2j} \le \frac{\lambda^2\sigma^2}{\gamma^2(\lambda+\alpha)^2(\lambda+\beta)^4} \frac{1}{1-(1-\lambda\gamma)^2} \\
= &\frac{\lambda\sigma^2}{\gamma^3(2-\lambda\gamma)(\lambda+\alpha)^2(\lambda+\beta)^4}.
\end{aligned}$$

Now by multivariate Chebyshev's inequality, we have

$$\mathbb{P}\left( \left\| \sum_{i=1}^{k} p_i \xi_i \right\|_2 \ge \epsilon \right) \le \frac{\operatorname{Tr}\operatorname{Var}\left[ \sum_{i=1}^{k} p_i \xi_i \right]}{\epsilon^2} \le \frac{\lambda\sigma^2}{\epsilon^2\gamma^3(2-\lambda\gamma)(\lambda+\alpha)^2(\lambda+\beta)^4} = \quad \delta.$$

That is, with probability at least $1-\delta$, we have

$$\|P_k\tilde{w}_k - P_k\mathbb{E}[\tilde{w}_k]\|_2 = \left\| \sum_{i=1}^{k} p_i \xi_i \right\|_2 \le \epsilon,$$

where

$$\epsilon = \frac{\sigma}{\gamma(\lambda+\alpha)(\lambda+\beta)^2} \sqrt{\frac{\lambda}{\delta\gamma(2-\lambda\gamma)}}.$$

This completes our proof. □

### C.2. Proof of Theorem 1.1

*Proof.* The derivation of kernel ridge regression can be found in (Mohri et al., 2018). We consider the following loss function of the dual problem

$$L(\alpha, \lambda) = \frac{1}{2}\|y - K\alpha\|_2^2 + \frac{\lambda}{2}\alpha^\top K\alpha,$$

where $y = (y_1, \ldots, y_n)^T$ is the label set. Then GD takes update

$$\alpha_{k+1} = \alpha_k - \eta_k \left(K^2\alpha_k - Ky + \lambda K\alpha_k\right), \quad \alpha_0 = 0.$$

Let $\alpha_* = (K + \lambda I)^{-1}y$, then

$$\alpha_{k+1} - \alpha_* = \left(I - \eta_k(K^2 + \lambda K)\right)(\alpha_k - \alpha_*),$$

thus

$$\alpha_{k+1} - \alpha_* = \Pi_{i=0}^{k}\left(I - \eta_i(K^2 + \lambda K)\right)(\alpha_0 - \alpha_*),$$

and

$$\alpha_{k+1} - \alpha_k = \Pi_{i=0}^{k-1}\left(I - \eta_i(K^2 + \lambda K)\right)\cdot \eta_k(K^2 + \lambda K)\cdot(K + \lambda I)^{-1}y = \Pi_{i=0}^{k-1}\left(I - \eta_i(K^2 + \lambda K)\right)\eta_k K y.$$

Similarly for $\hat{\alpha}_k$, i.e., the GD path for $L(\hat{\alpha}, \hat{\lambda})$ with learning rate $\gamma_k$, we have

$$\hat{\alpha}_{k+1} - \hat{\alpha}_k = \Pi_{i=0}^{k-1}\left(I - \gamma_i(K^2 + \hat{\lambda}K)\right)\gamma_k K y.$$

We emphasize that the generalized learning rate $\gamma_k = \left(I + (\hat{\lambda} - \lambda)\eta_k K\right)^{-1}\eta_k$ commutes with $K$. And

$$I - \gamma_k(\hat{\lambda} - \lambda)K = \frac{\gamma_k}{\eta_k}.$$

Thus for the generalized weighting scheme $P_K = 1 - \Pi_{i=0}^{k}(\gamma_i/\eta_i)$ we have

$$\begin{aligned}
&(1 - P_k)(\alpha_{k+1} - \alpha_k) = \Pi_{i=0}^{k-1}\left(\frac{\gamma_i}{\eta_i}\left(I - \eta_i(K^2 + \lambda K)\right)\right)\frac{\gamma_k}{\eta_k}\eta_k K y \\
=&\Pi_{i=0}^{k-1}\left(\frac{\gamma_i}{\eta_i} - \gamma_i(K^2 + \lambda K)\right)\gamma_k K y = \Pi_{i=0}^{k-1}\left(I - \gamma_i(\hat{\lambda} - \lambda)K - \gamma_i(K^2 + \lambda K)\right)\gamma_k K y \\
=&\Pi_{i=0}^{k-1}\left(I - \gamma_i(K^2 + \hat{\lambda}K)\right)\gamma_k K y = \hat{\alpha}_{k+1} - \hat{\alpha}_k.
\end{aligned}$$

Therefore by Lemma 2 we have

$$P_k\tilde{\alpha}_k = \hat{\alpha}_k - (1 - P_k)\alpha_k.$$

Let $\lambda_{\max}$ and $\lambda_{\min}$ be the maximal and minimal eigenvalue of $K$ respectively. Then if

$$\eta \le \eta_k \le \max\left\{\frac{1}{\lambda_{\max}(\lambda_{\max} + \lambda)}, \frac{1}{\lambda_{\max}(\lambda_{\max} + 2\hat{\lambda} - \lambda)}\right\}, \quad \gamma = \left(I + (\hat{\lambda} - \lambda)\eta K\right)^{-1}\eta,$$

we have

$$\eta(K^2 + \lambda K) \preceq \eta_k(K^2 + \lambda K) \prec I, \quad \gamma(K^2 + \hat{\lambda}K) \preceq \gamma_k(K^2 + \hat{\lambda}K) \prec I,$$

which guarantees the convergence of $\alpha_k$ and $\hat{\alpha}_k$. Hence both $\alpha_k$ and $\tilde{\alpha}_k$ are bounded. And the convergence rate is given by

$$\|\hat{\alpha}_k - \tilde{\alpha}_k\|_2 = \|(1 - P_k)(\alpha_k - \tilde{\alpha}_k)\|_2 = \mathcal{O}\left(\|1 - P_k\|_2\right) \le \mathcal{O}\left(\|\gamma/\eta\|_2^k\right) = \mathcal{O}\left((1 + (\hat{\lambda} - \lambda)\eta\lambda_{\min})^{-k}\right).$$

□

### C.3. Proof of Theorem 2

*Proof.* Let us consider changing of variable $v_k = Q^{\frac{1}{2}}w_k$, then

$$\begin{aligned}
&v_{k+1} = Q^{\frac{1}{2}}w_{k+1} = Q^{\frac{1}{2}}w_k - \eta_k Q^{-\frac{1}{2}}(x_k x_k^\top w_k - x_k y_k) \\
=&Q^{\frac{1}{2}}w_k - \eta_k(Q^{-\frac{1}{2}}x_k x_k^T Q^{-\frac{1}{2}}Q^{\frac{1}{2}}w_k - Q^{-\frac{1}{2}}x_k y_k) \\
=&v_k - \eta_k(Q^{-\frac{1}{2}}x_k x_k^\top Q^{-\frac{1}{2}}v_k - Q^{-\frac{1}{2}}x_k y_k).
\end{aligned}$$

Similarly let $\hat{v}_k = Q^{\frac{1}{2}}\hat{w}_k$, then

$$\begin{aligned}&\hat{v}_{k+1} = Q^{\frac{1}{2}}\hat{w}_{k+1} = Q^{\frac{1}{2}}\hat{w}_k - \gamma_k Q^{-\frac{1}{2}}(x_k x_k^\top \hat{w}_k - x_k y_k - \lambda Q\hat{w}_k)\\ =&Q^{\frac{1}{2}}\hat{w}_k - \gamma_k(Q^{-\frac{1}{2}}x_k x_k^\top Q^{-\frac{1}{2}}Q^{\frac{1}{2}}\hat{w}_k - Q^{-\frac{1}{2}}x_k y_k - \lambda Q^{\frac{1}{2}}\hat{w}_k)\\ =&\hat{v}_k - \gamma_k(Q^{-\frac{1}{2}}x_k x_k^\top Q^{-\frac{1}{2}}\hat{v}_k - Q^{-\frac{1}{2}}x_k y_k - \lambda\hat{v}_k).\end{aligned}$$

Let us denote

$$\Sigma = \mathbb{E}_x[xx^T], \quad a = \mathbb{E}_{x,y}[xy], \quad w_* = \Sigma^{-1}a, \quad \hat{w}_* = (\Sigma+\lambda I)^{-1}a, \quad \epsilon_k = (\Sigma w_k - a) - (x_{k+1}x_{k+1}^\top w_k - x_{k+1}y_{k+1}),$$

and correspondingly,

$$\Lambda = Q^{-\frac{1}{2}}\Sigma Q^{-\frac{1}{2}}, \quad b = Q^{-\frac{1}{2}}a, \quad v_* = Q^{-\frac{1}{2}}w_*, \quad \hat{v}_* = Q^{-\frac{1}{2}}\hat{w}_*, \quad \iota_k = Q^{-\frac{1}{2}}\epsilon_k.$$

Under these notations we have

$$v_{k+1} = v_k - \eta_k(\Lambda v_k - b) + \eta_k\iota_k, \quad v_0 = 0. \tag{14}$$

and

$$\hat{v}_{k+1} = \hat{v}_k - \gamma_k(\Lambda\hat{v}_k - b + \lambda\hat{v}_k) + \gamma_k\iota_k, \quad \hat{v}_0 = 0. \tag{15}$$

We can see that Eq. (14) and Eq. (15) are exactly what we have studied in Theorem 1. Also by assumption we know

$$\alpha I \preceq \Lambda \preceq \beta I.$$

Thus by Theorem 1 we have the following conclusions:

1. In expectation for any $k > 0$,
$$P_k\mathbb{E}[\tilde{v}_k] = \mathbb{E}[\hat{v}_k] - (1-P_k)\mathbb{E}[v_k].$$
2. Both $\mathbb{E}[v_k]$ and $\mathbb{E}[\hat{v}_k]$ converge. And there exists a constant $K$ such that for all $k > K$,
$$\|\mathbb{E}[\hat{v}_k] - \mathbb{E}[\tilde{v}_k]\|_2 \le \mathcal{O}((1-\lambda\gamma)^k).$$
Hence the limitation of $\mathbb{E}[\tilde{v}_k]$ exists and $\lim_{k\to\infty}\mathbb{E}[\tilde{v}_k] = \lim_{k\to\infty}\mathbb{E}[\hat{v}_k]$.
3. If the noise $\iota_k$ has uniform bounded variance
$$\mathbb{E}[\|\tilde{\iota}_k\|_2^2] \le \|Q\|_2\sigma^2, \quad \forall k.$$
Then for $k$ large enough, with probability at least $1-\delta$, we have
$$\|P_k\tilde{v}_k - P_k\mathbb{E}[\tilde{v}_k]\|_2 \le \epsilon,$$
where
$$\epsilon = \frac{\|Q\|_2^{\frac{1}{2}}\sigma}{\gamma(\lambda+\alpha)(\lambda+\beta)^2}\sqrt{\frac{\lambda}{\delta\gamma(2-\lambda\gamma)}}.$$

Now let $w_k = Q^{-\frac{1}{2}}v_k$, $\hat{w}_k = Q^{-\frac{1}{2}}\hat{v}_k$, then $\tilde{w}_k = \frac{1}{P_k}\sum_{i=1}^k p_i w_i = Q^{-\frac{1}{2}}\frac{1}{P_k}\sum_{i=1}^k p_i v_i = Q^{-\frac{1}{2}}\tilde{v}_k$. Hence we have

1. In expectation for any $k > 0$,
$$P_k\mathbb{E}[\tilde{w}_k] = \mathbb{E}[\hat{w}_k] - (1-P_k)\mathbb{E}[w_k].$$
2. Both $\mathbb{E}[w_k]$ and $\mathbb{E}[\hat{w}_k]$ converge. And there exists a constant $K$ such that for all $k > K$,
$$\|\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k]\|_2 \le \mathcal{O}((1-\lambda\gamma)^k).$$
Hence the limitation of $\mathbb{E}[\tilde{w}_k]$ exists and $\lim_{k\to\infty}\mathbb{E}[\tilde{w}_k] = \lim_{k\to\infty}\mathbb{E}[\hat{w}_k]$.

3. If the PSGD noise $Q^{-1}\epsilon_k$ has uniform bounded variance

$$\mathbb{E}[\left\|Q^{-1}\epsilon_i\right\|_2^2] \leq \sigma^2, \quad \forall i.$$

Then for $k$ large enough, with probability at least $1-\delta$, we have

$$\|P_k\tilde{w}_k - P_k\mathbb{E}[\tilde{w}_k]\|_2 \leq \epsilon,$$

where

$$\epsilon = \frac{\sigma \left\|Q^{-\frac{1}{2}}\right\|_2 \cdot \left\|Q^{\frac{1}{2}}\right\|_2}{\gamma(\lambda+\alpha)(\lambda+\beta)^2}\sqrt{\frac{\lambda}{\delta\gamma(2-\lambda\gamma)}} \leq \frac{\sigma\|Q\|_2}{\gamma(\lambda+\alpha)(\lambda+\beta)^2}\sqrt{\frac{\lambda}{\delta\gamma(2-\lambda\gamma)}}.$$

Hence our claims are proved.

□

### C.4. Proof of Theorem 3

*Proof.* First, provided $0 < \eta < \frac{1}{\beta} < \frac{1}{\alpha}$ and $\gamma = \frac{1}{\frac{1}{\eta}+\lambda}$, we have

$$\frac{\eta\alpha}{\alpha+\lambda} = \frac{1}{\frac{1}{\eta}+\frac{\lambda}{\eta\alpha}} < \frac{1}{\frac{1}{\eta}+\lambda} = \gamma < \frac{1}{\beta+\lambda} \leq \frac{1}{\alpha+\lambda}.$$

Therefore $0 < \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} < 1$, and

$$P_k = 1 - \frac{\gamma}{\eta}\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^{k-1}, \quad p_k = P_k - P_{k-1},$$

is a well defined weighting scheme, i.e., $P_k$ is non-negative, non-decreasing and $\lim_{k\to\infty} P_k = 1$.

Recall the NSGD updates for linear regression problem

$$w_{k+1} = v_k - \eta(x_{k+1}x_{k+1}^\top v_k - x_{k+1}y_{k+1}), \quad v_k = w_k + \tau(w_k - w_{k-1}), \quad w_0 = w_1 = 0,$$

where $\tau = \frac{1-\sqrt{\eta\alpha}}{1+\sqrt{\eta\alpha}}$.

Let

$$\Sigma = \mathbb{E}_x[xx^\top], \quad a = \mathbb{E}_{x,y}[xy], \quad \epsilon_k = (\Sigma v_k - a) - (x_{k+1}x_{k+1}^\top v_k - x_{k+1}y_{k+1}),$$

where $\epsilon_k$ is the gradient noise, and $\mathbb{E}_{k+1}[\epsilon_k] = 0$. Under these notations we have

$$w_{k+1} = v_k - \eta(\Sigma v_k - a) + \eta\epsilon_k, \quad v_k = w_k + \tau(w_k - w_{k-1}), \quad w_0 = w_1 = 0.$$

Thus

$$w_{k+1} = (1+\tau)(1-\eta\Sigma)w_k - \tau(1-\eta\Sigma)w_{k-1} + \eta a + \eta\epsilon_k, \quad w_0 = w_1 = 0. \tag{16}$$

Similarly for the linear regression with $\ell_2$-regularization, NSGD takes update

$$\hat{w}_{k+1} = \hat{v}_k - \gamma\left((x_{k+1}x_{k+1}^T + \lambda)\hat{v}_k - x_{k+1}y_{k+1}\right), \quad \hat{v}_k = \hat{w}_k + \hat{\tau}(\hat{w}_k - \hat{w}_{k-1}), \quad \hat{w}_0 = \hat{w}_1 = 0,$$

where $\hat{\tau} = \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1+\sqrt{\gamma(\alpha+\lambda)}}$.

And we have

$$\hat{w}_{k+1} = (1+\hat{\tau})\left(1-\gamma(\Sigma+\lambda)\right)\hat{w}_k - \hat{\tau}\left(1-\gamma(\Sigma+\lambda)\right)\hat{w}_{k-1} + \gamma a + \gamma\epsilon_k, \quad \hat{w}_0 = \hat{w}_1 = 0. \tag{17}$$

**Expectation** First let us compute the expectations. Let $z_k = \mathbb{E}[w_{k+1}] - \mathbb{E}[w_k],\ \hat{z}_k = \mathbb{E}[\hat{w}_{k+1}] - \mathbb{E}[\hat{w}_k]$, we aim to show that

$$(1 - P_k)z_k = \hat{z}_k, \quad k \geq 0. \tag{18}$$

Then according to Lemma 2, we prove the first conclusion in Theorem 3.

We begin with solving $z_k$.

For Eq. (16), taking expectation with respect to the random mini-batch sampling procedure, we have

$$\mathbb{E}[w_{k+1}] = (1+\tau)(1-\eta\Sigma)\mathbb{E}[w_k] - \tau(1-\eta\Sigma)\mathbb{E}[w_{k-1}] + \eta a, \quad \mathbb{E}[w_0] = \mathbb{E}[w_1] = 0.$$

Thus $z_k = \mathbb{E}[w_{k+1}] - \mathbb{E}[w_k]$ satisfies

$$z_{k+1} = (1+\tau)(1-\eta\Sigma)z_k - \tau(1-\eta\Sigma)z_{k-1}, \quad z_0 = 0, \quad z_1 = \eta a. \tag{19}$$

Without loss of generality, let us assume $\Sigma$ is diagonal in the following. Otherwise consider its eigenvalue decomposition $\Sigma = U\Lambda U^T$, and replace $z_k$ with $U^\top z_k$. All of the operators in the following are defined entry-wisely.

Eq. (19) defines a homogeneous linear recurrence relation with constant coefficients, which could be solved in a standard manner. Let

$$A = (1+\tau)(1-\eta\Sigma) = \frac{2(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}, \quad B = -\tau(1-\eta\Sigma) = \frac{-(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}},$$

then the characteristic function of Eq. (19) is

$$r^2 - Ar - B = 0. \tag{20}$$

Since $\Sigma$ is diagonal, $0 < \eta < \frac{1}{\alpha}$, and $\alpha$ is no greater than the smallest eigenvalue of $\Sigma$, we have

$$A^2 + 4B = \frac{4\eta(1-\eta\Sigma)(\alpha-\Sigma)}{(1+\sqrt{\eta\alpha})^2} \leq 0.$$

Thus the characteristic function (20) has two conjugate complex roots $r_1$ and $r_2$ (they might be equal). Suppose $r_{1,2} = s \pm ti$. Then the solution of Eq. (19) can be written as

$$z_k = 2(-B)^{\frac{k}{2}}\left(E\cos(\theta k) + F\sin(\theta k)\right), \quad k \geq 0,$$

where $E$ and $F$ are constants decided by initial conditions $z_0 = 0,\ z_1 = \eta a$, and $\theta$ satisfies

$$\cos\theta = \frac{s}{\sqrt{s^2+t^2}}, \quad \sin\theta = \frac{t}{\sqrt{s^2+t^2}}, \quad r_{1,2} = s \pm ti.$$

Since $2s = r_1 + r_2 = A, s^2 + t^2 = r_1\dot{r}_2 = -B$, we have

$$\cos\theta = \frac{A}{2\sqrt{-B}} = \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}}, \quad \sin\theta = \frac{\sqrt{-4B-A^2}}{2\sqrt{-B}} = \sqrt{\frac{\eta(\Sigma-\alpha)}{1-\eta\alpha}}.$$

Because $z_0 = 0, z_1 = \eta a$, we know that

$$E = 0, \quad 2F = \frac{\eta a}{(-B)^{\frac{1}{2}}\sin\theta}.$$

Thus

$$z_k = \frac{\eta a}{\sin\theta}(-B)^{\frac{k-1}{2}}\sin(\theta k), \quad k \geq 0. \tag{21}$$

where

$$B = \frac{-(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}, \quad \cos\theta = \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}}, \quad \sin\theta = \sqrt{\frac{\eta(\Sigma-\alpha)}{1-\eta\alpha}}.$$

One can directly verify that Eq. (21) solves the recurrence relation (19).

Then we solve $\hat{z}_k$.

Similarly treat Eq. (17), we know $\hat{z}_k = \mathbb{E}[\hat{w}_{k+1}] - \mathbb{E}[\hat{w}_k]$ satisfies

$$\hat{z}_{k+1} - (1+\hat{\tau})\left(1-\gamma(\Sigma+\lambda)\right)\hat{z}_k + \hat{\tau}\left(1-\gamma(\Sigma+\lambda)\right)\hat{z}_{k-1} = 0, \quad \hat{z}_0 = 0, \quad \hat{z}_1 = -\gamma a.$$

Repeat the calculation, we obtain

$$\hat{z}_k = \frac{\gamma a}{\sin\hat{\theta}}(-\hat{B})^{\frac{k-1}{2}}\sin(\hat{\theta}k), \quad k \geq 0,$$

where

$$\hat{B} = \frac{-\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)\left(1-\gamma(\Sigma+\lambda)\right)}{1+\sqrt{\gamma(\alpha+\lambda)}},$$
$$\cos\hat{\theta} = \sqrt{\frac{1-\gamma(\Sigma+\lambda)}{1-\gamma(\alpha+\lambda)}}, \quad \sin\hat{\theta} = \sqrt{\frac{\gamma(\Sigma-\alpha)}{1-\gamma(\alpha+\lambda)}}.$$

Finally we verify the sufficient condition in Lemma 2 (Eq. (18)).

First we show that if $1-\lambda\gamma = \frac{\gamma}{\eta}$, we have $\theta \equiv \hat{\theta} \pmod{2\pi}$. To see this, we only need to verify that $\cos\hat{\theta} = \cos\theta,\ \sin\hat{\theta} = \sin\theta$. This is because

$$\cos\hat{\theta} = \sqrt{\frac{1-\gamma\lambda-\gamma\Sigma}{1-\gamma\lambda-\gamma\alpha}} = \sqrt{\frac{\frac{\gamma}{\eta}-\gamma\Sigma}{\frac{\gamma}{\eta}-\gamma\alpha}} = \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}} = \cos\theta;$$
$$\sin\hat{\theta} = \sqrt{\frac{\gamma(\Sigma-\alpha)}{1-\gamma\lambda-\gamma\alpha}} = \sqrt{\frac{\gamma(\Sigma-\alpha)}{\frac{\gamma}{\eta}-\gamma\alpha}} = \sqrt{\frac{\eta(\Sigma-\alpha)}{1-\eta\alpha}} = \sin\theta.$$

Therefore we have

$$z_k = \frac{\eta a}{\sin\theta}(-B)^{\frac{k-1}{2}}\sin(\theta k), \quad \hat{z}_k = \frac{\gamma a}{\sin\theta}(-\hat{B})^{\frac{k-1}{2}}\sin(\theta k).$$

Since

$$1 - P_k = \frac{\gamma}{\eta}\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^{k-1}, \quad \frac{\gamma}{\eta} = 1-\lambda\gamma,$$

we have

$$\frac{\eta}{\gamma}(1-P_k)(-B)^{\frac{k-1}{2}} = \left(\frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2}{(1-\sqrt{\eta\alpha})^2}\cdot\frac{(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}\right)^{\frac{k-1}{2}}$$
$$= \left(\frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2(1-\eta\Sigma)}{1-\eta\alpha}\right)^{\frac{k-1}{2}} = \left(\frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2(1-\gamma(\Sigma+\lambda))}{1-\gamma(\alpha+\lambda)}\right)^{\frac{k-1}{2}}$$
$$= \left(\frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)(1-\gamma(\Sigma+\lambda))}{1+\sqrt{\gamma(\alpha+\lambda)}}\right)^{\frac{k-1}{2}} = (-\hat{B})^{\frac{k-1}{2}}.$$

Thus $(1-P_k)z_k = \hat{z}_k$. And according to Lemma 2, we have

$$\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k] = (1-P_k)\left(\mathbb{E}[w_k] - \mathbb{E}[\tilde{w}_k]\right), \quad k \geq 0.$$

Hence the first conclusion holds.

**Convergence** Since $L(w)$ is $\beta$-smooth, and the corresponding learning rate $\eta < \frac{1}{\beta}$, $\mathbb{E}[w_k]$ converges (Beck & Teboulle, 2009). Similarly, $\hat{L}(\hat{w}) = L(\hat{w}) + \frac{\lambda}{2}\|\hat{w}\|_2^2$ is $(\beta+\lambda)$-smooth, and the corresponding learning rate $\gamma = \frac{1}{\frac{1}{\eta}+\lambda} < \frac{1}{\beta+\lambda}$,

thus $\mathbb{E}[\hat{w}_k]$ converges (Beck & Teboulle, 2009). Specially for linear regression, these can be also verified by noticing that $0 < -B < 1$ because $\eta < \frac{1}{\beta}$ and

$$\sum_{i=1}^{k} |z_i| = \sum_{i=1}^{k} \left| \frac{\eta a}{\sin\theta} (-B)^{\frac{i-1}{2}} \sin(\theta i) \right| \leq \sum_{i=1}^{k} \left| \frac{\eta a}{\sin\theta} (-B)^{\frac{i-1}{2}} \right| < +\infty,$$

i.e., the right hand side of the above series converge, which implies that $\mathbb{E}[w_k] = \sum_{i=1}^{k} z_i$ converges absolutely, hence it converges. In a same manner $\mathbb{E}[\hat{w}_k]$ converges. Thus there exist constants $M$ and $K$ such that for all $k > K$, $\|\mathbb{E}[w_k]\|_2 \leq M$, $\|\mathbb{E}[\hat{w}_k]\|_2 \leq M$. Hence

$$\|\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k]\|_2 = (1 - P_k) \|\mathbb{E}[w_k] - \mathbb{E}[\hat{w}_k]\|_2 \leq \frac{\gamma}{\eta} C^{k-1} \cdot 2M = \mathcal{O}(C^k),$$

where $C = \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \in (0, 1)$, thus by taking limitation in both sides we obtain

$$\lim_{k\to\infty} \mathbb{E}[\tilde{w}_k] = \lim_{k\to\infty} \mathbb{E}[\hat{w}_k],$$

Hence the second conclusion holds.

**Variance** Next we turn to analyze the deviation of the averaged solution.

Let $w_i = \mathbb{E}[w_i] + \xi_i$. Based on Eq. (16), we first prove that

$$\xi_i = \sum_{j=1}^{i-1} a_{i-j} \eta \epsilon_j, \quad i \geq 1, \tag{22}$$

where

$$a_{k+1} = A a_k + B a_{k-1}, \quad a_0 = 0, \quad a_1 = 1. \tag{23}$$

We prove Eq. (22) by mathematical induction.

For $i = 1, 2$, by Eq. (16) we know $\xi_1 = w_1 - \mathbb{E}[w_1] = 0$ and $\xi_2 = w_2 - \mathbb{E}[w_2] = \eta\epsilon_1$, thus Eq. (22) holds. Now suppose Eq. (22) holds for $i - 1$ and $i$, then we consider $i + 1$. In Eq. (16), since $\xi_i = w_i - \mathbb{E}[w_i]$, taking difference we have

$$\xi_{i+1} = A\xi_i + B\xi_{i-1} + \eta\epsilon_i.$$

Now combining the induction assumptions we have

$$\begin{aligned}
\xi_{i+1} =& A \sum_{j=1}^{i-1} a_{i-j} \eta \epsilon_j + B \sum_{j=1}^{i-2} a_{i-j-1} \eta \epsilon_j + \eta \epsilon_i \\
=& \sum_{j=1}^{i-2} (A a_{i-j} + B a_{i-j-1}) \eta \epsilon_j + A a_1 \eta \epsilon_{i-1} + \eta \epsilon_i \\
=& \sum_{j=1}^{i-2} a_{i-j+1} \eta \epsilon_j + a_2 \eta \epsilon_{i-1} + a_1 \eta \epsilon_i \\
=& \sum_{j=1}^{i} a_{i-j+1} \eta \epsilon_j.
\end{aligned}$$

Thus by mathematical induction Eq. (22) is true for all $i \geq 1$.

Similarly to solve $z_k$, we can solve the recurrence relation Eq. (23) and obtain

$$a_k = \frac{1}{\sin\theta} (-B)^{\frac{k-1}{2}} \sin(\theta k), \quad k \geq 0, \tag{24}$$

where

$$B = \frac{-(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}, \quad \cos\theta = \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}}, \quad \sin\theta = \sqrt{\frac{\eta(\Sigma-\alpha)}{1-\eta\alpha}}.$$

Thus

$$\begin{aligned}\sqrt{-B} =& \sqrt{\frac{(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}} = (1-\sqrt{\eta\alpha})\sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}},\\ \frac{1}{\sin\theta} =& \sqrt{\frac{1-\eta\alpha}{\eta(\Sigma-\alpha)}} \preceq \sqrt{\frac{1-\eta\alpha}{\eta(\lambda_{\min}-\alpha)}} I,\end{aligned}$$

where $\lambda_{\min}$ is the smallest eigenvalue of $\Sigma$.

Now apply iterate averaging with respect to

$$p_i = P_i - P_{i-1} = \frac{\gamma}{\eta}\left(\frac{\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}}{1-\sqrt{\eta\alpha}}\right)\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^{i-2},$$

we have

$$P_k\tilde{w}_k = \sum_{i=1}^{k} p_i w_i = \sum_{i=1}^{k} p_i \mathbb{E}[w_i] + \sum_{i=1}^{k} p_i\xi_i = P_k\mathbb{E}[\tilde{w}_k] + \sum_{i=1}^{k} p_i\xi_i.$$

We turn to calculate the noise term $\sum_{i=1}^{k} p_i\xi_i$. Note that in every step, all of the matrices can be diagonalized simultaneously, thus they commute, similarly hereinafter.

$$\begin{aligned}\sum_{i=1}^{k} p_i\xi_i =& \sum_{i=1}^{k} p_i \sum_{j=1}^{i-1} a_{i-j}\eta\epsilon_j\\ =& \sum_{j=1}^{k-1}\left(\sum_{i=j+1}^{k} p_i a_{i-j}\right)\eta\epsilon_j\\ =& \sum_{j=1}^{k-1} A_j\epsilon_j,\end{aligned}$$

where $A_j = \eta\sum_{i=j+1}^{k} p_i a_{i-j}$. Recall that $\epsilon_0, \epsilon_1 \ldots, \epsilon_k$ is a martingale difference sequence, $\sum_{i=1}^{k} p_i\xi_i = \sum_{j=0}^{k-1} A_j\epsilon_j$ is a martingale. Thus

$$\operatorname{Tr}\operatorname{Var}\left[\sum_{i=1}^{k} p_i\xi_i\right] = \operatorname{Tr}\operatorname{Var}\left[\sum_{j=1}^{k-1} A_j\epsilon_j\right] = \sum_{j=1}^{k-1}\operatorname{Tr}\operatorname{Var}\left[A_j\epsilon_j\right].$$

Next we bound each term in the summation as

$$\operatorname{Tr}\operatorname{Var}\left[A_j\epsilon_j\right] = \operatorname{Tr}\mathbb{E}\left[(A_j\epsilon_j)(A_j\epsilon_j)^\top\right] = \mathbb{E}\left[\|A_j\epsilon\|_2^2\right] \le \|A_j\|_2^2 \cdot \mathbb{E}\left[\|\epsilon\|_2^2\right] \le \sigma^2\|A_j\|_2^2.$$

And we remain to bound $\|A_j\|_2^2$:

$$
\begin{aligned}
\|A_j\|_2^2 =& \left\| \eta \sum_{i=j+1}^{k} p_i a_{i-j} \right\|_2^2 \\
=& \left\| \frac{\gamma}{\sin\theta} \frac{\sqrt{\gamma(\alpha+\lambda)} - \sqrt{\eta\alpha}}{1-\sqrt{\eta\alpha}} \sum_{i=j+1}^{k} \left( \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \right)^{i-2} (-B)^{\frac{i-j-1}{2}} \sin(\theta(i-j)) \right\|_2^2 \\
\le& \left\| \frac{\gamma}{\sin\theta} \frac{\sqrt{\gamma(\alpha+\lambda)} - \sqrt{\eta\alpha}}{1-\sqrt{\eta\alpha}} \sum_{i=j+1}^{k} \left( \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \right)^{i-2} \left( (1-\sqrt{\eta\alpha}) \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}} \right)^{i-j-1} \right\|_2^2 \\
\le& \left( \frac{\gamma}{\sin\theta} \frac{\sqrt{\gamma(\alpha+\lambda)} - \sqrt{\eta\alpha}}{1-\sqrt{\eta\alpha}} \sum_{i=j+1}^{k} \left( \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \right)^{i-2} (1-\sqrt{\eta\alpha})^{i-j-1} \right)^2 \\
=& \left( \frac{\gamma}{\sin\theta} \frac{\sqrt{\gamma(\alpha+\lambda)} - \sqrt{\eta\alpha}}{1-\sqrt{\eta\alpha}} (1-\sqrt{\eta\alpha})^{1-j} \sum_{i=j+1}^{k} \left( 1-\sqrt{\gamma(\alpha+\lambda)} \right)^{i-2} \right)^2 \\
\le& \left( \gamma \sqrt{\frac{1-\eta\alpha}{\eta(\lambda_{\min}-\alpha)}} \cdot \frac{\sqrt{\gamma(\alpha+\lambda)} - \sqrt{\eta\alpha}}{\left(1-\sqrt{\eta\alpha}\right)^j} \cdot \frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^{j-1}}{\sqrt{\gamma(\alpha+\lambda)}} \right)^2 \\
=& \frac{\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)} - \sqrt{\eta\alpha}\right)^2}{\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2} \left( \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \right)^{2j}.
\end{aligned}
$$

The first inequality is because $\sin(\theta(i-j)) \le 1$, and the second inequality is because $\alpha < \lambda_{\min}(\Sigma)$.

Based on previous discussion we have

$$
\begin{aligned}
&\operatorname{Tr}\operatorname{Var}\left[ \sum_{i=1}^{k} p_i \xi_i \right] = \sum_{j=1}^{k-1} \operatorname{Tr}\operatorname{Var}\left[ A_j \epsilon_j \right] \le \sum_{j=1}^{k-1} \sigma^2 \|A_j\|_2^2 \\
\le& \sum_{j=1}^{k-1} \frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)} - \sqrt{\eta\alpha}\right)^2}{\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2} \left( \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \right)^{2j} \\
\le& \frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)} - \sqrt{\eta\alpha}\right)^2}{\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2} \cdot \frac{\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^2}{1-\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^2} \\
\le& \frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)} - \sqrt{\eta\alpha}\right)^2}{\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2} \cdot \frac{\left(1-\sqrt{\gamma(\alpha+\lambda)}\right)^2}{\left(2-\sqrt{\eta\alpha}-\sqrt{\gamma(\alpha+\lambda)}\right)\left(\sqrt{\gamma(\alpha+\lambda)} - \sqrt{\eta\alpha}\right)} \\
=& \frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)} - \sqrt{\eta\alpha}\right)}{\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(2-\sqrt{\eta\alpha}-\sqrt{\gamma(\alpha+\lambda)}\right)}.
\end{aligned}
$$

Now by multivariate Chebyshev's inequality, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^{k} p_i\xi_i\right\|_2 \geq \epsilon\right) \leq \frac{\operatorname{Tr}\operatorname{Var}[\sum_{i=1}^{k} p_i\xi_i]}{\epsilon^2} \leq \frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}\right)}{\epsilon^2\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(2-\sqrt{\eta\alpha}-\sqrt{\gamma(\alpha+\lambda)}\right)} =: \delta.$$

That is, with probability at least $1-\delta$, we have

$$\|P_k\tilde{w}_k - P_k\mathbb{E}[\tilde{w}_k]\|_2 = \left\|\sum_{i=1}^{k} p_i\xi_i\right\|_2 \leq \epsilon,$$

where

$$\epsilon = \sqrt{\frac{\sigma^2\gamma(1-\eta\alpha)\left(\sqrt{\gamma(\alpha+\lambda)}-\sqrt{\eta\alpha}\right)}{\delta\eta(\lambda_{\min}-\alpha)(\alpha+\lambda)\left(2-\sqrt{\eta\alpha}-\sqrt{\gamma(\alpha+\lambda)}\right)}}.$$

This completes our proof.

□

### C.5. Proof of Theorem 4

*Proof.* We will prove a stronger version of Theorem 4 by showing the conclusions hold for any 1-dim projection direction $v_1 \in \mathbb{R}^d$. Concisely, given a unit vector $v_1 \in \mathbb{R}^d$, we can extend it to a group of orthogonal basis, $v_1, v_2, \ldots, v_d$. For $w \in \mathbb{R}^d$, we denote its decomposition as

$$w = w^{(1)}v_1 + w^{(2)}v_2 + \cdots + w^{(d)}v_d, \quad w^{(i)} \in \mathbb{R}.$$

Define $h(w^{(1)}) = L(w) = L(w^{(1)}v_1 + \cdots + w^{(d)}v_d)$, then $\nabla h(w^{(1)}) = v_1^\top \nabla L(w)$. Now for one step of GD,

$$w_{k+1} = w_k - \eta\nabla L(w_k),$$

by multiplying $v_1$ in both sides, we obtain

$$w_{k+1}^{(1)} = v_1^\top w_{k+1} = v_1^\top w_k - \eta v_1^\top \nabla L(w_k) = w_k^{(1)} - \eta\nabla h(w_k^{(1)}). \tag{25}$$

We turn to study GD along direction $v_1$ by analyzing Eq. (25).

Firstly $h(w^{(1)})$ is $\alpha$-strongly convex, $\beta$-smooth and lower bounded since $L(w)$ is $\alpha$-strongly convex, $\beta$-smooth, and lower bounded. Let $w_*$ be the unique minimum of $L(w)$, then $w_*^{(1)} = v_1^\top w_*$ is the minimum of $h(w^{(1)})$. Without loss of generality, assume

$$w_*^{(1)} > 0 = w_0^{(1)}.$$

Then by Lemma 3, we know the optimization path of Eq. (25) lies between $(0, w_*^{(1)})$, and for any $v \in (0, w_*^{(1)})$, we have

$$\alpha v - b \leq \nabla h(v) \leq \beta v - b, \quad b = -\nabla h(0).$$

Thus for Eq. (25) we have

$$w_{k+1}^{(1)} - w_k^{(1)} = -\eta\nabla h(w_k^{(1)}) \leq -\eta(\alpha w_k^{(1)} - b),$$
$$w_{k+1}^{(1)} - w_k^{(1)} = -\eta\nabla h(w_k^{(1)}) \geq -\eta(\beta w_k^{(1)} - b).$$

Define the following dynamics:

$$u_{k+1}^{(1)} - u_k^{(1)} = -\eta(\alpha u_k^{(1)} - b), \quad v_{k+1}^{(1)} - v_k^{(1)} = -\eta(\beta v_k^{(1)} - b), \quad u_0^{(1)} = v_0^{(1)} = 0.$$

By the discrete Gronwall's inequality (Clark, 1987), we have

$$v_k^{(1)} \leq w_k^{(1)} \leq u_k^{(1)}.$$

Furthermore, $u_k^{(1)}$ and $v_k^{(1)}$ satisfy two first order recurrence relations respectively, thus they can be solved by

$$u_k^{(1)} = \eta \sum_{i=1}^{k} (1-\eta\alpha)^{i-1} b, \quad v_k^{(1)} = \eta \sum_{i=1}^{k} (1-\eta\beta)^{i-1} b.$$

Since $\eta < \frac{1}{\beta} \le \frac{1}{\alpha}$, $u_k^{(1)}$ and $v_k^{(1)}$ converge. And $w_k^{(1)}$ also converges since $h(\cdot)$ is $\beta$-smooth convex and $\eta < \frac{1}{\beta}$.

In a same way, for the regularized path,

$$\hat{w}_{k+1,\lambda}^{(1)} = \hat{w}_{k,\lambda}^{(1)} - \gamma(\nabla h(\hat{w}_{k,\lambda}^{(1)}) + \lambda \hat{w}_{k,\lambda}^{(1)}), \quad \hat{w}_{0,\lambda}^{(1)} = 0,$$

we have

$$\begin{aligned}
\hat{w}_{k+1,\lambda}^{(1)} - \hat{w}_{k,\lambda}^{(1)} &= -\gamma(\nabla h(\hat{w}_{k,\lambda}^{(1)}) + \lambda \hat{w}_{k,\lambda}^{(1)}) \le -\gamma\left((\alpha+\lambda)\hat{w}_{k,\lambda}^{(1)} - b\right), \\
\hat{w}_{k+1,\lambda}^{(1)} - \hat{w}_{k,\lambda}^{(1)} &= -\gamma(\nabla h(\hat{w}_{k,\lambda}^{(1)}) + \lambda \hat{w}_{k,\lambda}^{(1)}) \ge -\gamma\left((\beta+\lambda)\hat{w}_{k,\lambda}^{(1)} - b\right).
\end{aligned}$$

Consider the following dynamics:

$$\hat{u}_{k+1,\lambda}^{(1)} - \hat{u}_{k,\lambda}^{(1)} = -\gamma((\alpha+\lambda)\hat{u}_{k,\lambda}^{(1)} - b), \quad \hat{v}_{k+1,\lambda}^{(1)} - \hat{v}_{k,\lambda}^{(1)} = -\gamma((\beta+\lambda)\hat{v}_{k,\lambda}^{(1)} - b),$$

where $\hat{u}_{0,\lambda}^{(1)} = \hat{v}_{0,\lambda}^{(1)} = 0$. Then by the discrete Gronwall's inequality (Clark, 1987) and the solution of the first order recurrence relation we obtain

$$\hat{v}_{k,\lambda}^{(1)} \le \hat{w}_{k,\lambda}^{(1)} \le \hat{u}_{k,\lambda}^{(1)}, \quad \hat{u}_{k,\lambda}^{(1)} = \gamma \sum_{i=1}^{k} (1-\gamma(\alpha+\lambda))^{i-1} b, \quad \hat{v}_{k,\lambda}^{(1)} = \gamma \sum_{i=1}^{k} (1-\gamma(\beta+\lambda))^{i-1} b.$$

Now we turn to bound the iterate averaged solution. Consider

$$\lambda_1 = \frac{1}{\gamma} - \frac{1}{\eta} + \beta - \alpha, \quad \lambda_2 = \frac{1}{\gamma} - \frac{1}{\eta} + \alpha - \beta,$$

since $\beta \ge \alpha$ and $0 < \gamma < \frac{1}{\beta - \alpha + 1/\eta}$ we know $\lambda_1 \ge \lambda_2 > 0$. Notice that

$$0 < \gamma(\alpha+\lambda_2) \le \{\gamma(\alpha+\lambda_1),\ \gamma(\beta+\lambda_2)\} \le \gamma(\beta+\lambda_1) = 1 - \gamma(-\frac{1}{\eta} + 2\beta - \alpha) < 1,$$

where the last inequality is because $\eta > \frac{1}{2\beta - \alpha}$. Thus $\hat{u}_{k,\lambda_1}^{(1)}$, $\hat{u}_{k,\lambda_2}^{(1)}$, $\hat{v}_{k,\lambda_1}^{(1)}$, $\hat{v}_{k,\lambda_2}^{(1)}$ converge. Further $\hat{w}_{k,\lambda_1}$ and $\hat{w}_{k,\lambda_2}$ also converge since $\gamma < \frac{1}{\beta+\lambda_1} \le \frac{1}{\beta+\lambda_2}$ and the corresponding regularized losses are $(\beta+\lambda_1)$ and $(\beta+\lambda_2)$-smooth, respectively.

Next let us consider the weighting scheme $P_k = 1 - \left(\frac{\gamma}{\eta}\right)^{k+1}$, which is well defined since $0 < \gamma < \frac{1}{\beta - \alpha + 1/\eta} \le \eta$.

One can directly verify that $\tilde{u}_k^{(1)} = \frac{1}{P_k}\sum_{i=1}^{k} p_i u_i^{(1)}$, $\tilde{v}_k^{(1)} = \frac{1}{P_k}\sum_{i=1}^{k} p_i v_i^{(1)}$ converge, and

$$(1-P_k)(u_{k+1}^{(1)} - u_k^{(1)}) = \hat{v}_{k+1,\lambda_2}^{(1)} - \hat{v}_{k,\lambda_2}^{(1)}, \quad (1-P_k)(v_{k+1}^{(1)} - v_k^{(1)}) = \hat{u}_{k+1,\lambda_1}^{(1)} - \hat{u}_{k,\lambda_1}^{(1)}.$$

Thus according to Lemma 2 we have

$$P_k(u_k^{(1)} - \tilde{u}_k^{(1)}) = u_k^{(1)} - \hat{v}_{k,\lambda_2}^{(1)}, \quad P_k(v_k^{(1)} - \tilde{v}_k^{(1)}) = v_k^{(1)} - \hat{u}_{k,\lambda_1}^{(1)}.$$

Therefore

$$\begin{aligned}
\tilde{w}_k^{(1)} - \hat{w}_{k,\lambda_2}^{(1)} &\le \tilde{u}_k^{(1)} - \hat{v}_{k,\lambda_2}^{(1)} = \tilde{u}_k^{(1)} - u_k^{(1)} + P_k(u_k^{(1)} - \tilde{u}_k^{(1)}) = (1-P_k)(\tilde{u}_k^{(1)} - u_k^{(1)}), \\
\tilde{w}_k^{(1)} - \hat{w}_{k,\lambda_1}^{(1)} &\ge \tilde{v}_k^{(1)} - \hat{u}_{k,\lambda_1}^{(1)} = \tilde{v}_k^{(1)} - v_k^{(1)} + P_k(v_k^{(1)} - \tilde{v}_k^{(1)}) = (1-P_k)(\tilde{v}_k^{(1)} - v_k^{(1)}),
\end{aligned}$$

which implies that

$$\hat{w}_{k,\lambda_1}^{(1)} + (1-P_k)(\tilde{v}_k^{(1)} - v_k^{(1)}) \le \tilde{w}_k^{(1)} \le \hat{w}_{k,\lambda_2}^{(1)} + (1-P_k)(\tilde{u}_k^{(1)} - u_k^{(1)}). \tag{26}$$

Note that $u_k^{(1)},\ \tilde{u}_k^{(1)},\ v_k^{(1)},\ \tilde{v}_k^{(1)},\ \hat{w}_{k,\lambda_1}^{(1)},\ \hat{w}_{k,\lambda_2}^{(1)}$ converge, therefore there is a constant $M$ controlling their $\ell_2$-norm. Define $m_k^{(1)} = (\hat{w}_{k,\lambda_2}^{(1)} + \hat{w}_{k,\lambda_1}^{(1)})/2$, $d_k^{(1)} = (\hat{w}_{k,\lambda_2}^{(1)} - \hat{w}_{k,\lambda_1}^{(1)})/2$. Recall that $\hat{w}_{k,\lambda_1}^{(1)}$ are the GD optimization path of a $(\alpha+\lambda_1)$-strongly convex and $(\beta+\lambda_1)$-smooth loss, thus $\hat{w}_{k,\lambda_1}^{(1)}$ converges in rate $\mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right)$. Similarly $\hat{w}_{k,\lambda_2}^{(1)}$ converges in rate $\mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right)$. Thus triangle inequality we have

$$\left\|m_k^{(1)} - m^{(1)}\right\|_2 \le \frac{1}{2}\left\|\hat{w}_{k,\lambda_2}^{(1)} - \hat{w}_{\infty,\lambda_2}^{(1)}\right\|_2 + \frac{1}{2}\left\|\hat{w}_{k,\lambda_1}^{(1)} - \hat{w}_{\infty,\lambda_1}^{(1)}\right\|_2 \le \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) + \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right).$$

$$\left\|d_k^{(1)} - d^{(1)}\right\|_2 \le \frac{1}{2}\left\|\hat{w}_{k,\lambda_2}^{(1)} - \hat{w}_{\infty,\lambda_2}^{(1)}\right\|_2 + \frac{1}{2}\left\|\hat{w}_{k,\lambda_1}^{(1)} - \hat{w}_{\infty,\lambda_1}^{(1)}\right\|_2 \le \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) + \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right).$$

By Eq. (26) we obtain

$$\begin{aligned}
\tilde{w}_k^{(1)} - m_k^{(1)} &\le d_k^{(1)} + (1-P_k)(\tilde{u}_k^{(1)} - u_k^{(1)}) \le d_k^{(1)} + 2M\left(\frac{\gamma}{\eta}\right)^{k+1} \\
&\le d^{(1)} - d^{(1)} + d_k^{(1)} + \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right) \\
&\le d^{(1)} + \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) + \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right) + \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right),
\end{aligned}$$

and

$$\begin{aligned}
\tilde{w}_k^{(1)} - m_k^{(1)} &\ge d_k^{(1)} + (1-P_k)(\tilde{v}_k^{(1)} - v_k^{(1)}) \ge d_k^{(1)} - 2M\left(\frac{\gamma}{\eta}\right)^{k+1} \\
&\ge d^{(1)} - d^{(1)} + d_k^{(1)} - \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right) \\
&\ge d^{(1)} - \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) - \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right) - \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right).
\end{aligned}$$

Thus

$$\left\|\tilde{w}_k^{(1)} - m_k^{(1)}\right\|_2 \le d^{(1)} + \mathcal{O}(C^k), \quad C = \max\{(1-\gamma(\alpha+\lambda_1), (1-\gamma(\alpha+\lambda_2), \frac{\gamma}{\eta}\}.$$

In conclusion we have

$$\left\|\tilde{w}_k^{(1)} - m^{(1)}\right\|_2 \le \left\|\tilde{w}_k^{(1)} - m_k^{(1)}\right\|_2 + \left\|m_k^{(1)} - m^{(1)}\right\|_2 \le d^{(1)} + \mathcal{O}(C^k).$$

□

## D. Experiments setups

The experiments are conducted using one GPU K80 and `PyTorch 1.3.1`.

### D.1. Two dimensional toy example

The loss function is

$$L(w) = \frac{1}{2}(w-w_*)^\top \Sigma (w-w_*), \quad w_* = (1,1)^\top, \quad \Sigma = U\,\mathrm{Diag}\,(0.1,1)\,U^T,$$
$$U = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \theta = \frac{\pi}{3}.$$

All the algorithms are initiated from zero. The learning rate for the unregularized problem is $\eta = 0.1$. The hyperparameter for the vanilla/generalized $\ell_2$-regularization is $\lambda = 0.1$. And the learning rate for the regularized problem is $\gamma = \frac{1}{\lambda + 1/\eta}$. The preconditioning matrix is set to be $Q = \Sigma$. We run the algorithms for $500$ iterations. For NGD and NSGD, we set the strongly convex coefficient to be $\alpha = 0.05$.

### D.2. MNIST dataset

**Dataset** http://yann.lecun.com/exdb/mnist/

**Linear regression** The image data is scaled to $[0, 1]$. The label data is one-hotted. The loss function is standard linear regression under squared loss, without bias term, $L(w) = \frac{1}{2n}\sum_{i=1}^{n} \left\|w^T x_i - y_i\right\|_2^2$. All the algorithms are initiated from zero. The learning rate for the unregularized problem is $\eta = 0.01$. The hyperparameter for the vanilla/generalized $\ell_2$-regularizer is $\lambda = 4.0$. And the learning rate for the regularized problem is $\gamma = \frac{1}{\lambda + 1/\eta}$. The preconditioning matrix is set to be $Q = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$. The batch size for the stochastic algorithms are $b = 500$. We run the algorithms for $500$ iterations. For NGD and NSGD, we set the strongly convex coefficient to be $\alpha = 1.0$.

**Logistic regression** The image data is scaled to $[0, 1]$. The label data is one-hotted. The loss function is standard logistics regression loss plus an $\ell_2$-regularization term, $L(w) = \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{KL}}(y_i \,||\, \sigma(w^\top x_i)) + \frac{\lambda_0}{2}\|w\|_2^2$, where $\sigma(x)$ is the softmax function and $\lambda_0 = 1.0$. All the algorithms are initiated from zero. The learning rate for the unregularized problem is $\eta = 0.01$. The hyperparameter for the vanilla/generalized $\ell_2$-regularizer is $\lambda = 4.0$. And the learning rate for the regularized problem is $\gamma = \frac{1}{\lambda + 1/\eta}$. The preconditioning matrix is set to be $Q = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$. The batch size for the stochastic algorithms are $b = 500$. We run the algorithms for $500$ iterations. For NGD and NSGD, we set the strongly convex coefficient to be $\alpha = 1.0$.

### D.3. CIFAR-10 and CIFARR-100 datasets

**Datasets** https://www.cs.toronto.edu/~kriz/cifar.html

**VGG-16 on CIFAR-10** The image data is scaled to $[0, 1]$ and augmented by horizontally flipping and randomly cropping. The label data is one-hotted. The model is standard VGG-16 with batch normalization. We train the model with vanilla SGD for $300$ epochs. The batch size is $100$. The learning rate is $0.1$, and decreased by ten times at epoch $150$ and $250$. The weight decay is set to be $5 \times 10^{-4}$.

After finishing the SGD training process, we average the checkpoints from $61$ to $300$ epoch with standard geometric distribution. We test the success probability $p \in \{0.9999, 0.999, 0.99, 0.9\}$. And the best one is $0.99$.

**ResNet-18 on CIFAR-10** The image data is scaled to $[0, 1]$ and augmented by horizontally flipping and randomly cropping. The label data is one-hotted. The model is standard ResNet-18. We train the model with vanilla SGD for $300$ epochs. The batch size is $100$. The learning rate is $0.1$, and decreased by ten times at epoch $150$ and $250$. The weight decay is set to be $5 \times 10^{-4}$.

After finishing the SGD training process, we average the checkpoints from $61$ to $300$ epoch with standard geometric distribution. We test the success probability $p \in \{0.9999, 0.999, 0.99, 0.9\}$. And the best one is $0.99$.

**ResNet-18 on CIFAR-100** The image data is scaled to $[0, 1]$ and augmented by horizontally flipping and randomly cropping. The label data is one-hotted. The model is standard ResNet-18. We train the model with vanilla SGD for $300$ epochs. The batch size is $100$. The learning rate is $0.1$, and decreased by ten times at epoch $150$ and $250$. The weight decay is set to be $5 \times 10^{-4}$.

After finishing the SGD training process, we average the checkpoints from $61$ to $300$ epoch with standard geometric distribution. We test the success probability $p \in \{0.9999, 0.999, 0.99, 0.9\}$. And the best one is $0.99$.

**Additional experiments for deep nets without weight decay** For ResNet-18 trained on CIFAR-10, without weight decay, and with the other setups the same, vanilla SGD has $92.95\%$ test accuracy, and our method has $93.21\%$ test accuracy. This result is consistent with the results presented in the main text.