Stochastic Gradient Descent in the viewpoint of Graduated Optimization

Anonymous Authors[1]

Abstract

The stochastic gradient descent (SGD) method is popular for solving the non-convex optimization problem in machine learning. This work investigates SGD from the viewpoint of graduated optimization, which is a widely applied approach for non-convex optimization problems. Instead of the actual optimization problem, a series of smoothed optimization problems that can be achieved in various ways are solved in the graduated optimization approach. In this work, a formal formulation of the graduated optimization is provided based on approximate identity, which generalizes the idea of Gaussian smoothing. Also, the asymptotic convergence result is achieved with the techniques in variational analysis. Then, we show that the traditional SGD method can be applied to solve the smoothed optimization problem. The Monte Carlo integration is used to achieve the gradient in the smoothed problem, which can be consistent with distributed computing schemes in real-life applications. From the assumptions on the actual optimization problem, the convergence results of SGD for the smoothed problem can be derived straightforwardly. Numerical examples show the evidence that the graduated optimization approach can provide more accurate training results in certain cases.

1. Introduction

Consider the non-convex optimization problem:

$$\min_x f(x), \quad x \in \Omega,$$

where f is the objective function which is non-convex and smooth enough, Ω is a feasible set. Many machine learning problems and deep neural network training problems can be summarized as the above non-convex optimization problem. The stochastic gradient descent (SGD) method and its variants, such as Adagrad (Duchi et al., 2011) and Adam (Kingma & Ba, 2014), is one of the most important tools in machine learning. The iterative updates in SGD method performs in a way as

$$x_{k+1} = x_k - \alpha_k g(x_k, \xi_k),$$

where α_k is the step size, and g is the stochastic gradient which satisfies some addition assumptions on its expectation and variations. In the classic SGD setting, the noise of the stochastic gradient comes from the dataset ξ_k. In this work, we consider the case when the noise is caused not only by the data, but also by the model x_k of the current iteration. This idea leads to the graduated optimization approach.

The graduated optimization (or named continuation method) is a popular heuristic approach for solving the non-convex optimization problem. A sequence of subproblems with different smoothed objective functions are defined before the optimization process begins. Starting with the smoothest subproblem, the sequence of subproblems are solved sequentially, where the solution of each subproblem serves as the initial value for the following subproblem. Heuristically, this approach may lead to the global minimum solution instead of the local minimum solution for the non-convex optimization problem.

Motivation. The motivation of the graduated optimization approach is demonstrated in Figure 1. If the objective function f is minimized with the initial point x^0, the local minimum around $x = 0.25$ will be achieved as the solution. With the graduated optimization approach, the smoothed objective function f^1 is minimized with the initial value x^0 at the first stage, and its solution is x^1. Next, the smoothed objective function f^2 is minimized with the point x^1 as the initial value. Then, the solution x^2 is used as the initial value of minimizing f^3, and so on. By this construction, we can assume $x^k \to \bar{x}$ which is the global minimum of f as $k \to \infty$. Notice that if we start with a subproblem in which the objective function is not smooth enough, the solution may still be a local minimum. For example, when we minimize f^2 with x^0 as the initial value, the solution

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

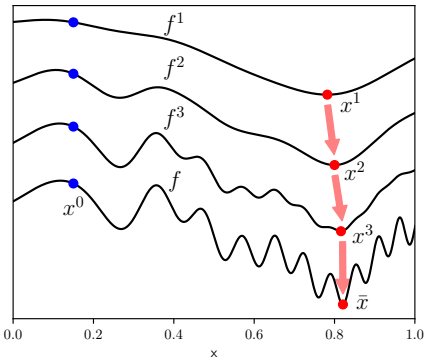

Figure 1. A demonstration of graduated optimization approach.

will still be a local minimum near $x = 0.25$.

Previous work. The works on the graduated optimization approach date back to the 1980s under the name continuation method (Witkin et al., 1987), graduated non-convex (GNC) algorithm (Blake & Zisserman, 1987), and mean field annealing (Yuille, 1989). Later, a theoretical analysis result is provided in the work (Wu, 1996) from an optimization view: under certain second order regularity conditions, for any stationary point of the actual objective function, there exists a continuous curve consisting of the stationary point of the objective function of the subproblems.

Since when it was proposed, the concept of graduated optimization has been successfully applied in computer vision field, in an explicitly or implicitly way (Zerubia & Chellappa, 1993; Nikolova et al., 2010; Mobahi et al., 2012). Also, the graduated optimization approach has been applied to many application scenarios in the machine learning field, such as semi-supervised learning (Chapelle et al., 2006), unsupervised learning (Smith & Eisner, 2004), and ranking (Chapelle & Wu, 2010). In the work (Bengio, 2009), the author suggested some form of continuation in learning which has performed an important role in recent developments in the training of deep architectures (Hinton et al., 2006; Erhan et al., 2009). A comprehensive survey on the development and application of the graduated optimization approach has provided in (Mobahi & Fisher, 2015).

While it is popular in the machine learning field, the theoretical analysis of the graduated optimization is still under development. A bound on the endpoint solution of the continuation method with the Gaussian smoothing was provided in (Mobahi & Fisher III, 2015), with no practical algorithms provided. Later, a special kind of non-convex objective function, (a, σ)-nice function, was studied in (Hazan et al., 2016). In this work, an implicit smoothing is provided by some sampling oracles, and a graduated optimization with gradient oracle algorithm was provided. In addition, the convergence to the global optimum for the (a, σ)-nice function was proved, and the convergence rate of the algorithm was derived. The smoothing effect of the graduated optimization approach is not only studied as an optimization strategy but also used to illustrate why the SGD method performs well in practical machine learning tasks (Kleinberg et al., 2018). Based on this viewpoint, a perturbed SGD algorithm was proposed in (Harshvardhan & Stich, 2021), where a global convergence result was obtained under the PŁ condition. Furthermore, an entropy-SGD method is proposed with a similar smoothing explanation (Chaudhari et al., 2019), where the algorithm was a composition of two nested SGD loops and the Langevin dynamics is used to compute the gradient of the local entropy.

Contribution. In this work, we first provide a formulation of the graduated optimization approach based on the approximate identity, for which Gaussian smoothing is a special case. Then, we derive an asymptotic convergence result of the graduated optimization approach by the techniques in variational analysis with no further requirement of special kinds of functions. We verify that for the smoothed subproblems of graduated optimization when its gradient is obtained by the Monte Carlo random sampling, the assumptions about the original problem can be inherited into the subproblems. In this way, many of the convergence results of solving subproblems with the traditional SGD method can be obtained straightforwardly from the convergence result of the actual optimization problem. Based on this, a multi-layer SGD (ML-SGD) scheme is formally proposed. Synthetic numerical examples demonstrate that the graduated optimization approach provides more accurate training results in low-dimensional settings.

Content. The content of this paper is organized as follows: In Section 2, we propose a formulation of the graduated optimization approach, and the asymptotic convergence result is obtained. The multi-layer SGD scheme is investigated in Section 3. Then, we show that the smoothed subproblem in the graduated optimization approach can be solved with the SGD method and the Monte Carlo random sampling technique. Also, similar convergence results can be achieved for the subproblem under certain assumptions on the actual optimization problem. Two synthetic numerical examples are demonstrated in Section 4. Additional discussion and outlook are provided in Section 5.

2. Formulation of the Graduated Optimization Approach

Rewrite the non-convex optimization problem as

$$(\mathrm{P}) \quad \min_x f(x), \quad \text{such that } x \in \Omega, \tag{1}$$

where $f \in L^1(\mathbb{R}^n)$. Denote the actual optimization problem as (P). The constraint set Ω is nonempty, closed, and

bounded. In practice, we can assume that the Ω as a simple box constraint:

$$\Omega = \{x \in \mathbb{R}^n \mid a \le x^i \le b,\ 1 \le i \le n\},$$

where $a, b \in \mathbb{R}$ and $a \le b$. The constraint set Ω introduced in the above problem is for the convenience of analysis. In this work, we assume that Ω is large enough such that the optimization process of the above problem is in Ω and never reaches the boundary of Ω. Note that we are not assuming Ω is convex. When Ω is convex, other algorithms for the constrained optimization problem can be developed, and we will not discuss it in this work for simplicity.

Definition 2.1 (Approximate identity)**.** An approximate identity is a family of functions $\phi_t \in L^1(\mathbb{R}^n)$ for $t > 0$ such that
(a) $\int_{\mathbb{R}^n} |\phi_t(x)|\, \mathrm{d}x \le A$, where $A > 0$ is a constant independent of t.
(b) $\int_{\mathbb{R}^n} \phi_t(x)\, \mathrm{d}x = 1$, for all $t > 0$.
(c) $\lim_{t\to 0} \int_{|x|\ge\delta} |\phi_t(x)|\, \mathrm{d}x = 0$, for any $\delta > 0$.

Next, we generate an approximate identity with a kernel function ϕ. Given $\phi \in L^1(\mathbb{R}^n)$ with

$$\int_{\mathbb{R}^n} \phi(x)\, \mathrm{d}x = 1.$$

For $t > 0$, let

$$\phi_t(x) = t^{-n}\phi(t^{-1}x).$$

It can be easily checked that $\{\phi_t\}$ is an approximate identity. It is well known that the Gaussian kernel can be formulated as an approximate identity given by

$$\phi_t(x) = (4\pi t)^{-n/2} e^{-|x|^2/4t}. \tag{2}$$

Given an approximate identity $\{\phi_t\}$, a series of smoothed objective functions can be constructed as

$$\tilde{f}^t = f * \phi_t, \tag{3}$$

where $*$ is the convolution operator. Let $\nu = \lceil 1/t \rceil$, ν be a smooth coefficient. Then, define function f^ν as

$$f^\nu = \tilde{f}^{\lceil 1/t \rceil},$$

where $\nu \in \mathbb{N}$. Notice that $\{f^\nu\}$ is a subsequence of $\{\tilde{f}^t\}$. It can be shown that $f^\nu \to f$ uniformly as $\nu \to \infty$ by Proposition A.5. For each f^ν, define the νth smoothed subproblem of (P) as

$$(\mathrm{P}^\nu) \quad \min_x f^\nu(x), \quad \text{such that } x \in \Omega.$$

In such a way, a series of smoothed subproblems (P^ν) of the actual optimization problem are constructed. Then, a formal algorithm for the graduated optimization can be achieved.

Next, we discuss the asymptotic convergence of the graduated optimization approach.

Algorithm 1 Formal graduated optimization algorithm

Input: smooth kernel ϕ; $\nu_1 < \nu_2 < \cdots < \nu_{N_m} < \infty$, initial value x^0
for $m = 1$ **to** N_m **do**
  Construct the kernel function ϕ_{1/ν_m}
  With the initial value x^{m-1}, solve the subproblem (P^{ν_m}). Denote the solution as x^m
end for
Output: x^{N_m}

Definition 2.2 (ε-optimality)**.** The ε-optimal solution of minimizing a proper function f on $\mathbb{R}^n$ can be denoted as a set

$$\varepsilon\text{-}\arg\min f = \{x \mid f(x) \le \inf f + \varepsilon\}.$$

Denote the set of the global minimum of the actual problem (P) as S, and the set of the global minimum of νth subproblem (P^ν) as S^ν, i.e.,

$$S := \arg\min_{x\in\Omega} f(x), \quad S^\nu := \arg\min_{x\in\Omega} f^\nu(x).$$

Also, the ε-optimal solution of function f and f^ν can be denoted respectively as

$$\varepsilon\text{-}S := \varepsilon\text{-}\arg\min f, \quad \varepsilon\text{-}S^\nu := \varepsilon\text{-}\arg\min f^\nu.$$

Our main theorem states the relation between S^ν and S as $\nu \to \infty$.

Theorem 2.3 (Main Theorem)**.** *For the actual optimization problem* (P)*, construct a series of smoothed subproblem* $\{(\mathrm{P}^\nu)\}$*. Then, there exists an index N such that when $\nu \ge N$, the sets S^ν are nonempty and form a bounded sequence with*

$$\limsup_\nu S^\nu \subset S.$$

For any $\varepsilon^\nu \to 0$ and $x^\nu \in \varepsilon^\nu$-$S^\nu$, the sequence $\{x^\nu\}$ is bounded and such that all its cluster points belong to S. If S consists of a unique point $\bar{x}$, then $x^\nu \to \bar{x}$.

The proof of Theorem 2.3 is deferred to Appendix A.

Theorem 2.3 provides an asymptotic convergence result of the graduated optimization approach. For large classes of functions ($L^1(\mathbb{R}^n)$), the convergence of the graduated optimization approach is guaranteed by Theorem 2.3 compared to the previous results in (Wu, 1996; Hazan et al., 2016). Moreover, the kernel functions used to construct the smoothed objective functions are not limited to the Gaussian kernel. However, the above theorem does not provide an accurate convergence rate as in the work (Hazan et al., 2016). How to choose the smooth kernel ϕ and the smooth coefficients ν is still a heuristic decision.

3. Multi-Layer SGD Scheme

This section discusses the multi-layer SGD (ML-SGD) scheme. Based on the graduated optimization approach discussed in the previous section, the optimization process can be formulated as solving a series of sequential subproblems. In many machine learning scenarios, the objection function of the actual optimization problem (P) can be written as

$$f(x) = \int_{\Xi} L(x, \xi)\, \mathrm{d}P(\xi),$$

where ξ is a random variables with distribution P which is supported on $\Xi \subset \mathbb{R}^n$. When we solve the actual optimization problem (P) with the SGD method, at the kth iteration, the stochastic gradient is denoted as $g(x_k, \xi_k)$ with some additional assumptions are satisfied.

Recall the discussion in the previous section, a series of smoothed subproblems are required for the graduated optimization approach. Denote the objective function of the νth subproblem (P^ν) as

$$f^\nu(x) = f * \phi_{1/\nu}(x) = \int_{\mathbb{R}^n} f(y)\phi_{1/\nu}(x - y)\, \mathrm{d}y,$$

where $\phi_{1/\nu}$ is a kernel obtained by the construction of the approximate identity. Then, the gradient of $f^\nu(x)$ is given by the Bochner integral:

$$\nabla f^\nu(x) = \int_{\mathbb{R}^n} \nabla f(y)\phi_{1/\nu}(x - y)\, \mathrm{d}y.$$

Next, apply the SGD method to solve the subproblem (P^ν). Based on the above equation, with the gradient replaced with the stochastic gradient, we denote the stochastic gradient of the smoothed objective function $f^\nu(x)$ at kth iteration as

$$g^\nu(x_k, \xi_k) = \int_{\mathbb{R}^n} g(y, \xi_k)\phi_{1/\nu}(x_k - y)\, \mathrm{d}y. \tag{4}$$

Notice that there is an integral in the above stochastic gradient $g^\nu(x_k, \xi_k)$. One straightforward way is to evaluate the integral with Monte Carlo random sampling. Denote

$$g^\nu_{N_\nu}(X^\nu_k, \xi_k) = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} g((X^\nu_k)_i, \xi_k), \tag{5}$$

where ξ_k is still the same random variable, X^ν_k is a random variable with probability density function $\phi_{1/\nu}(x_k - \cdot)$, which is a kernel function. When the Gaussian kernel (2) is used, $\phi_{1/\nu}(x_k - \cdot)$ is the Gaussian distribution centered at x_k. This is the case of applying Gaussian smoothing. Figure 2 provides a demonstration of gradient evaluation by Monte Carlo random sampling.

Algorithm 2 is in order to apply the SGD method for solving (P^ν) with random sampling.

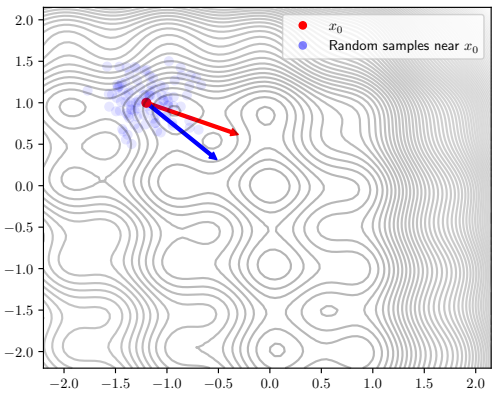

Figure 2. A demonstration of gradient evaluation. The update direction of the actual objective function at x_0 is given by the red arrow. The update direction of the smoothed objective function at x_0 is given by the blue arrow.

Algorithm 2 SGD method solving (P^ν) with random sampling

Input: $\phi_{1/\nu}$, x^ν_0
for $i = 1, 2, \cdots$ **do**
  Generate realizations of random variable ξ_k with P, X^ν_k with $\phi_{1/\nu}(x^\nu_k - \cdot)$
  Compute the stochastic gradient $g^\nu_{N_\nu}(X^\nu_k, \xi_k)$ by equation (5)
  Update $x^\nu_{k+1} = x^\nu_k - \alpha_k g^\nu_{N_\nu}(X^\nu_k, \xi_k)$, where α_k is the stepsize
end for

Recall the updates in the classic SGD method can be written as

$$x_{k+1} = x_k - \alpha_k g(x_k, \xi_k),$$

where ξ_k represents the sample or the batch of samples used in the current iteration. In this case, the randomness of the stochastic gradient is caused by the data. However, in Algorithm 2, the randomness of the stochastic gradient is caused not only by the data but also by the model. Later we show that when certain assumptions are met, as long as the convergence result for solving the actual optimization problem (P) with the SGD method is available, similar conclusions can also be made for solving the subproblem (P^ν) with Algorithm 2.

A formal ML-SGD scheme is provided in Algorithm 3.

Notice that, Algorithm 3 is a formal scheme that provides no guarantee of converging to the global minimum. Indeed, Theorem 2.3 states that the above multi-layer approach can provide a global minimum result. However, we still need to emphasize that the smooth kernel ϕ and smooth coefficients ν are hyperparameters that need to be determined

Algorithm 3 ML-SGD scheme with random sampling

Input: smooth kernel ϕ; $\nu_1 < \nu_2 < \cdots < \nu_{N_m} < \infty$, initial value x^0
for $m = 1$ **to** N_m **do**
  Construct the kernel function ϕ_{1/ν_m}
  With the initial value x^{m-1}, solve the mth subproblem $(\mathcal{P}^{\nu_m})$ by Algorithm 2, denote the solution as x^m
end for
Output: x^{N_m}

heuristically.

3.1. An Example of Convergence Analysis

In this subsection, we show that the convergence result of the subproblem (P^ν) can be directly achieved under the assumptions of the actual problem (P).

First, we discuss the Lipschitz smooth gradient assumption:

Assumption 3.1. Assume $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. The gradient of f, $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant $L > 0$, i.e. for all $x, y \in \mathbb{R}^n$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

It is easy to see that the Lipschitz smoothness of f can be inherited by f^ν.

Lemma 3.2. *For the objective function of νth subproblem f^ν, the gradient $\nabla f^\nu : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant L.*

The proof is deferred to Appendix B.

The following assumptions are made for the stochastic gradient at kth iteration $g(x_k, \xi_k)$.

Assumption 3.3. For any $k \geq 1$, $x \in \Omega$, assume
(a) $\mathbb{E}\left[g(x, \xi_k)\right] = \nabla f(x)$,
(b) $\mathbb{E}\left[\|g(x, \xi_k) - \nabla f(x)\|^2\right] \leq \sigma^2$.

The following lemma states that the stochastic gradient $g^\nu(x_k, \xi_k)$ is an unbiased estimator of $\nabla f^\nu(x_k)$, and the variance of the random variable $\|g^\nu(x_k, \xi_k) - \nabla f^\nu(x_k)\|$ is bounded.

Lemma 3.4. *For the stochastic gradient $g^\nu(x_k, \xi_k)$ defined by equation (4), under Assumption 3.3, we have*
(a) $\mathbb{E}\left[g^\nu(x_k, \xi_k)\right] = \nabla f^\nu(x_k)$,
(b) $\mathbb{E}\left[\|g^\nu(x_k, \xi_k) - \nabla f^\nu(x_k)\|^2\right] \leq \sigma^2$.

The proof is deferred to Appendix B.

The following lemma states that the stochastic gradient $g^\nu_{N_\nu}(X^\nu_k, \xi_k)$ is an unbiased estimator of $\nabla f^\nu(x_k)$, and the variance of the random variable $\left\|g^\nu_{N_\nu}(X^\nu_k, \xi_k) - \nabla f^\nu(x_k)\right\|$ is bounded.

Lemma 3.5. *For the stochastic gradient $g^\nu_{N_\nu}(X^\nu_k, \xi_k)$ in equation (5), $k \geq 1$, $x \in \Omega$, under Assumption 3.3, we have*
(a) $\mathbb{E}\left[g^\nu_{N_\nu}(X^\nu_k, \xi_k)\right] = \nabla f^\nu(x_k)$,
(b) $\mathbb{E}\left[\left\|g^\nu_{N_\nu}(X^\nu_k, \xi_k) - \nabla f^\nu(x_k)\right\|^2\right] \leq \sigma^2 + \frac{\sigma^2 + M}{N_\nu}$, *where* $M = \max_x \|\nabla f(x)\|^2$.

The proof is deferred to Appendix B.

With the above assumptions and lemmas, the stochastic gradient is well-defined for the subproblems. Next, we show that the classical convergence result of the SGD method can be directly achieved for subproblem (P).

Suppose the objective function f^ν satisfies the PŁ inequality, i.e.

$$\frac{1}{2}\|\nabla f^\nu(x)\|^2 \geq \mu\left(f^\nu(x) - f^\nu_*\right), \ \forall x \in \Omega.$$

The following theorem is derived directly from Theorem 4 in (Karimi et al., 2016).

Theorem 3.6. *For the νth subproblem (P^ν), suppose f^ν satisfies the PŁ inequality, and assumptions 3.1 and 3.3 are satisfied. Then, for Algorithm 2,*
(a) Let the stepsize $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$, the convergence rate is

$$\mathbb{E}\left[f^\nu(x_k) - f^\nu_*\right] \leq \frac{L}{2k\mu^2}(\sigma^2 + M)\left(1 + \frac{1}{N_\nu}\right),$$

where $M = \max_x \|\nabla f(x)\|^2$.
(b) With constant stepsize $\alpha_k = \alpha < \frac{1}{2\mu}$, the convergence rate is

$$\mathbb{E}\left[f^\nu(x_k) - f^\nu_*\right] \leq (1 - 2\mu\alpha)^k(f^\nu(x_0) - f^\nu_*) + \frac{L\alpha}{4\mu}(\sigma^2 + M)\left(1 + \frac{1}{N_\nu}\right).$$

The proof is deferred to Appendix B.

Let $N_\nu \to \infty$, this corresponding to the convolution in equation (4) is evaluated accurately. Compared to when N_ν is a finite constant, the convergence rate is $1 + 1/N_\nu$ times slower when the decreasing stepsize is used. The convergence rate is not affected when the constant stepsize is used, but the final bound is $1+1/N_\nu$ times larger. The above theorem suggests that the effect of random sampling on the theoretical convergence results of the SGD method may be slight. For example, when $N_\nu = 10$, $1 + 1/N_\nu = 1.1$, i.e., 10% larger then the case when $N_\nu \to \infty$. However, N_ν gradients need to be evaluated for each iteration which may limit the practical application of the graduated optimization approach. On the other hand, the structure of Algorithm 2 is well suited for distributed computing. The graduated

optimization has the potential of being studied and applied for large-scale distributed computing scenarios, such as federated learning.

4. Synthetic Numerical Experiments

In this section, we provide two synthetic toy experiments to demonstrate that the ML-SGD scheme has the potential to overcome the local minimum solution in a low-dimensional setting.

The first example is

$$\min_x f_1(x), \quad \text{where } f_1(x) = x^2 + 10x\sin(x).$$

This example is used in the work (Harshvardhan & Stich, 2021). The objective function f_1 is highly nonlinear, with two global minima located near ± 4.7 (Harshvardhan & Stich, 2021). The gradient descent algorithm is used to optimize the actual objective function f_1. Algorithm 2 is used to optimize the smoothed objective functions $f_1^{\nu_i}$, where $\nu_1 = 0.1$, $\nu_2 = 0.5$, $\nu_3 = 1$, and $N_\nu = 10$. The standard Gaussian kernel is used. Notice that ν_1 and ν_2 are not integers in this example. This will not affect the theoretical properties of Algorithms 2 and 3 since we can construct a new "base" kernel $\phi^{'} = \phi_{1/\nu_1}$ and then new smooth coefficients can be obtained as $\nu_1^{'} = 1$, $\nu_2^{'} = 5$, and $\nu_3^{'} = 10$. For comparison, we fix the step size with $\alpha = 0.001$, and for each optimization process, the total iteration number is 1000. We generate 1000 random initial values over $[-50, 50]$ following the uniform distribution. First, the gradient descent algorithm is performed to minimize f_1 with the above randomly generated initial values. The result is shown in Figure 3 (a). Next, The minimization result of smoothed objective functions f^{ν_1}, f^{ν_2}, and f^{ν_3} are shown in Figure 3 (b), (c), and (d) respectively. Then, the multi-layer strategy is performed with two layers (ν_1 and ν_3) and three layers (ν_1, ν_2, and ν_3) by Algorithm 3. Since no stochastic gradient is used, we denote this case as ML-GD. The results are shown in Figure 3 (e) and (f) for two and three layers respectively.

As can be seen from subfigure (a), when the actual objective function is solved directly, most of the results are trapped in the local minimum. For smoothed problems, different degrees of smoothness affect the results of optimization. When ν is large, the objective function f^ν is less smooth. Thus, the results in subfigure (b) are more trapped in the local minima than (c) and (d), which are farther away from the global minima. Subfigures (e) and (f) show evidence that the multi-layer structure can indeed improve the optimization result. With the multi-layer strategy, the optimization results are largely improved. Compared to other cases, there are more optimization results that fall into the global minimum in (e) and (f). In addition, by designing a multi-layer structure that transitions more smoothly, the optimization results can be further improved.

The second example is a two-dimensional toy problem formulated as

$$\begin{aligned}
&\min_x \left(f_2(x) := f_R(x) + f_H(x)\right),\\
&\text{where } f_R(x) = 20 + (x_1^2 - 10\cos(2\pi x_1))\\
&\qquad\qquad + (x_2^2 - 10\cos(2\pi x_2)),\\
&\qquad f_H(x) = ((x_1+3)^2 + (x_2+2) - 11)^2\\
&\qquad\qquad + ((x_1+3) + (x_2+2)^2 - 7)^2.
\end{aligned}$$

The global minimum is $(0, 0)$. We simulate the stochastic gradient by letting $g(x_k) = \nabla f_R(x_k)$ or $g(x_k) = \nabla f_H(x_k)$ with equal probability at kth iteration. There are 1000 random initial values generated following the uniform distribution over the rectangle $[-2, 2] \times [-2, 2]$.

The SGD method is used to minimize the actual objective function $f_2(x)$, and the result is shown in Figure 4 (a). Although many local minimum solutions are shown in Figure 4 (a), there are still many solutions around the global minimum. A recent work (Kleinberg et al., 2018) suggests that the traditional SGD algorithm can also be explained by the graduated optimization approach, i.e., the random properties of gradients are equivalent to smoothing the objective function. Let $\nu_1 = 5$, $\nu_2 = 50$, $N_\nu = 10$. Then, Algorithm 2 is used to minimize the smoothed objective function $f_2^{\nu_1}(x)$ and $f_2^{\nu_2}(x)$ with Gaussian kernel. As shown in subfigures (b) and (c), different degrees of smoothness affect the optimization results. When using Algorithm 2 to minimize $f_2^{\nu_1}$, most of the results are located near the global minimum. Based on this observation, we minimize f_2 by Algorithm 3 with $\nu_1 = 5$ and $\nu_2 = 50$ as the first and the second layer. The results are shown in subfigure (d), where most of the results are very close to the global minimum.

5. Discussion and Outlook

In this work, the SGD method is studied with the graduated optimization approach. Under this view of point, the ML-SGD scheme is proposed, and the convergence of graduated optimization is studied. Notice that our main result Theorem 2.3 only provides an asymptotic behavior of graduated optimization, i.e., given the actual problem (P), we can construct a series of smoothed subproblems (P^ν) such that the solutions of the subproblems converge to the solution of the actual problem. Thus, it is still a heuristic decision to choose the kernel ϕ and the smooth parameter ν for a given optimization problem. It is also worth noting that although ML-SGD performs well in low-dimensional situations, the effect on high-dimensional problems still needs to be experimentally verified. Especially for the deep neural network training problems, evidence shows that the transition in the loss landscapes between the chaotic area and nearly convex area is quick (Li et al., 2017). As a result, once the initial value or the model in the current iteration is

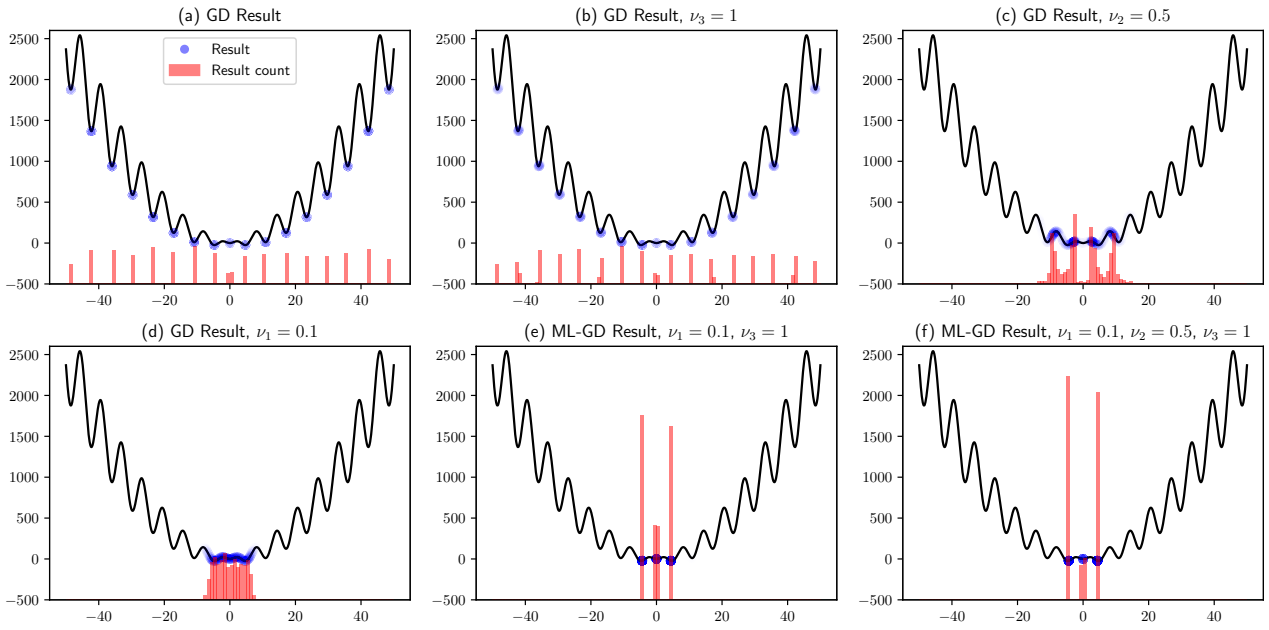

Figure 3. Distribution of the finial results. (a): The gradient descent (GD) algorithm is performed on the objective function $f_1(x)$. (b), (c), (d): Algorithm 2 with accurate gradients instead of stochastic gradients is performed on $f_1(x)$ with different smooth coefficients. (e), (f): two ML-GD schemes are performed on $f_1(x)$ with different smoothing strategy.

in a nearly convex area, the role of graduate optimization may not be obvious. This ensures further research on the use of graduated optimization in practical application scenarios.

On the other hand, the algorithmic structure of graduated optimization through random sampling is particularly suitable for the distributed computing environment of contemporary machine learning tasks. With only minor changes, the graduated optimization approach can be deployed in massively distributed computing problems, which opens up the possibility for further study of the practical role of graduated optimization. In addition, as when we built the ML-SGD algorithm, the idea of the multi-layer structure can be used for other first-order optimization algorithms, such as Adagrad, RMSprop, Adam, etc. Continued research in this direction could lead to more effective training algorithms for large-scale neural network models.

References

Bengio, Y. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

Blake, A. and Zisserman, A. *Visual reconstruction*. MIT press, 1987.

Chapelle, O. and Wu, M. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 13(3):216–235, 2010.

Chapelle, O., Chi, M., and Zien, A. A continuation method for semi-supervised svms. In *Proceedings of the 23rd international conference on Machine learning*, pp. 185–192, 2006.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S., and Vincent, P. The difficulty of training deep architectures

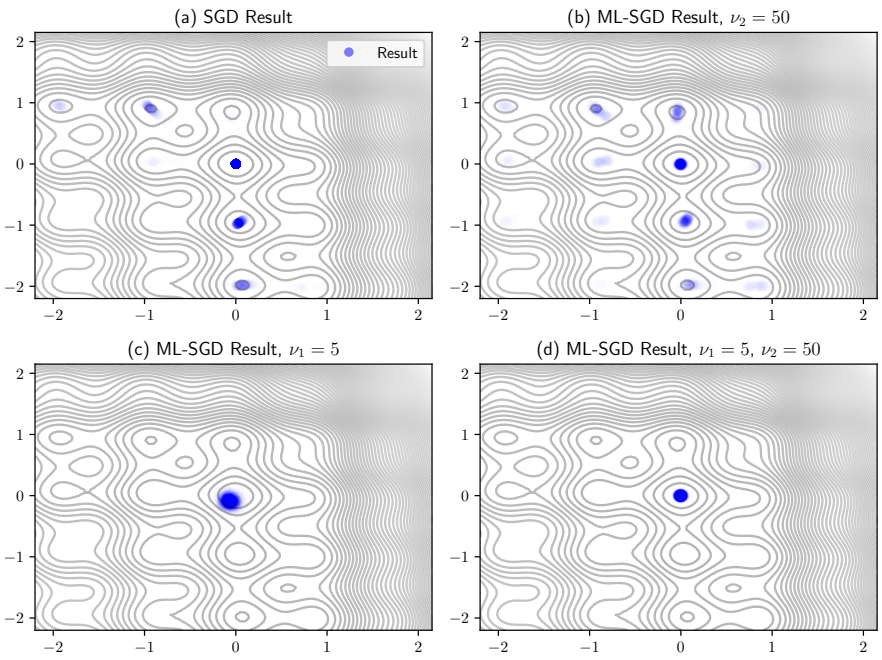

Figure 4. Distribution of the finial results. (a): The SGD method is performed to minimize the objective function $f_2(x)$. (b), (c): Algorithm 2 is performed on $f_2(x)$ with different smooth coefficients. (d): A ML-SGD result on $f_2(x)$.

and the effect of unsupervised pre-training. In *Artificial Intelligence and Statistics*, pp. 153–160. PMLR, 2009.

Folland, G. B. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

Harshvardhan and Stich, S. U. Escaping local minima with stochastic noise. *OPT2021: 13th Annual Workshop on Optimization for Machine Learning*, 2021.

Hazan, E., Levy, K. Y., and Shalev-Shwartz, S. On graduated optimization for stochastic non-convex problems. In *International conference on machine learning*, pp. 1833–1841. PMLR, 2016.

Hinton, G. E., Osindero, S., and Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural computation*, 18 (7):1527–1554, 2006.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kleinberg, B., Li, Y., and Yuan, Y. An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pp. 2698–2707. PMLR, 2018.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 2017.

Mobahi, H. and Fisher, J. W. On the link between gaussian homotopy continuation and convex envelopes. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 43–56. Springer, 2015.

Mobahi, H. and Fisher III, J. A theoretical analysis of optimization by gaussian continuation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Mobahi, H., Zitnick, C. L., and Ma, Y. Seeing through the blur. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1736–1743. IEEE, 2012.

Nikolova, M., Ng, M. K., and Tam, C.-P. Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction. *IEEE Transactions on Image Processing*, 19(12):3073–3088, 2010.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Smith, N. A. and Eisner, J. Annealing techniques for unsupervised statistical language learning. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 486–493, 2004.

Witkin, A., Terzopoulos, D., and Kass, M. Signal matching through scale space. *International journal of computer vision*, 1(2):133–144, 1987.

Wu, Z. The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation. *SIAM Journal on Optimization*, 6(3):748–768, 1996.

Yuille, A. Energy functions for early vision and analog networks. *Biological Cybernetics*, 61(2):115–123, 1989.

Zerubia, J. and Chellappa, R. Mean field annealing using compound gauss-markov random fields for edge detection and image estimation. *IEEE Transactions on neural networks*, 4(4):703–709, 1993.

A. Deferred Proofs of Section 2

A.1. Prerequisite

In this subsection, we briefly review the prerequisite in variational analysis. Most of the contents can be found in the textbook (Rockafellar & Wets, 2009) and (Folland, 1999).

Denote a subset of $\mathbb{N}$ to represent the index of convergence sequences as

$$\mathcal{N}_\infty := \{N \subset \mathbb{N} \mid \mathbb{N}/N \text{ finite}\}.$$

This notation is useful when we representing the index $\nu \to \infty$ for $\nu \in \mathbb{N}$.

Definition A.1 (Eventually level boundedness)**.** A sequence of sets $C^\nu \subset \mathbb{R}^n$ is eventually bounded that for some index set $N \in \mathcal{N}_\infty$ if the set $\cup_{\nu \in N} C^\nu$ is bounded.
A sequence of functions f^ν is eventually level-bounded if for each $\alpha \in \mathbb{R}^n$, the sequence of sets $\mathrm{lev}_{\leq \alpha} f^\nu$ is eventually bounded.

Proposition A.2 ((Rockafellar & Wets, 2009) 7.32 (a))**.** *The sequence $\{f^\nu\}$ is eventually level-bounded if the sequence of sets dom f^ν is eventually bounded.*

Proposition A.3 ((Rockafellar & Wets, 2009) 7.2)**.** *Let f^ν be any sequence of functions on $\mathbb{R}^n$, and let x be any point of $\mathbb{R}^n$. Then, $f^\nu \xrightarrow{e} f$ if and only if at each point x one has*

$$\liminf_\nu f^\nu(x^\nu) \geq f(x) \quad \textit{for every sequence } x^\nu \to x, \tag{6}$$

$$\limsup_\nu f^\nu(x^\nu) \leq f(x) \quad \textit{for some sequence } x^\nu \to x. \tag{7}$$

Proposition A.4 ((Rockafellar & Wets, 2009) 1.6)**.** *For a function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, f is lower semicontinuous on $\mathbb{R}^n$ if and only if the level sets of type $\mathrm{lev}_{\leq \alpha} f$ are closed in $\mathbb{R}^n$.*

The following proposition provides the uniform convergence of $f * \phi_t$.

Proposition A.5 ((Folland, 1999) Theorem 8.14)**.** *Suppose $\phi \in L^1(\mathbb{R}^n)$ and $\int \phi(x)\,\mathrm{d}x = 1$. If f is bounded and uniformly continuous, then $f * \phi_t \to 1f$ uniformly as $t \to 0$.*

The following proposition shows the regularity of the objective functions in the subproblem.

Proposition A.6 ((Folland, 1999), Proposition 8.10)**.** *If $f \in L^1$, $g \in C^k$, and $\partial^\alpha g$ is bounded for $|\alpha| \leq k$, then $f * g \in C^k$ and $\partial^\alpha (f * g) = f * (\partial^\alpha g)$ for $|\alpha| \leq k$.*

The following proposition provides the connection between the uniform convergence and epigraph convergence for the sequence f^ν.

Proposition A.7 ((Rockafellar & Wets, 2009) 7.15)**.** *Consider $f^\nu, f : \mathbb{R}^n \to \bar{\mathbb{R}}$ and a set $X \subset \mathbb{R}^n$. If the functions f^ν are lower semicontinuous (lsc) relative to X and converge uniformly to f on X, then f is lsc relative to X. Also, f^ν epi-converge to f relative to X, written as $f^\nu \xrightarrow{e} f$ relative to X.*

Proposition A.8 ((Rockafellar & Wets, 2009) 7.33 convergence in minimization)**.** *Suppose the sequence $\{f^\nu\}$ is eventually level-bounded, and $f^\nu \xrightarrow{e} f$ with f^ν and f lsc and proper. Then*

$$\inf f^\nu \to \inf f \quad (\textit{finite}),$$

while for ν in some index set $N \in \mathcal{N}_\infty$ the sets $\arg\min f^\nu$ are nonempty and form a bounded sequence with

$$\limsup_\nu (\arg\min f^\nu) \subset \arg\min f.$$

Indeed, for any choice of $\varepsilon^\nu \to 0$ and $x^\nu \in \varepsilon^\nu$-$\arg\min f^\nu$, the sequence $\{x^\nu\}_{\nu \in \mathbb{N}}$ is bounded and such that all its cluster points belong to $\arg\min f$. If $\arg\min f$ consists of a unique point $\bar{x}$, one must actually have $x^\nu \to \bar{x}$.

A.2. Proof of the main theorem

The following lemma provides the epi-convergence result of the objective functions of the subproblems (P^ν).

Lemma A.9. *The objective function f^ν of the subproblem (P_ν) epi-converge to the objective function f of (P) relative to Ω, i.e., $f^\nu \xrightarrow{e} f$ relative to Ω.*

Proof. Since ϕ_t is an approximate identity, by the construction of sequence $\{\tilde{f}^t\}$ (equation (3)), $\tilde{f}^t \to f$ uniformly on Ω as $t \to 0$ by Proposition A.5. Since f is continuous and $\phi_t \in L^1(\mathbb{R}^n)$, $\tilde{f}^t$ is continuous relative to Ω for all $t > 0$ by Proposition A.6. Since f^ν is a subsequence of $\{\tilde{f}^t\}$, the proof is finished by Proposition A.7. □

Proof of the main theorem.

Proof. **1. Construct extended real-valued functions.**

For the objective functions of problem (P) and (P^ν), define extended real-valued functions:

$$g(x) = \begin{cases} f(x), & x \in \Omega, \\ \infty, & x \notin \Omega, \end{cases}$$

$$g^\nu(x) = \begin{cases} f^\nu(x), & x \in \Omega, \\ \infty, & x \notin \Omega, \end{cases}$$

Obviously, problems (P) and (P^ν) are respectively equivalent to

$$(\tilde{\mathrm{P}}) \quad \min_{x\in\mathbb{R}^n} g(x), \quad \text{and} \quad (\tilde{\mathrm{P}}^\nu) \quad \min_{x\in\mathbb{R}^n} g^\nu(x).$$

Also,

$$\arg\min_{x\in\Omega} f(x) = \arg\min_{x\in\mathbb{R}^n} g(x), \tag{8}$$

$$\arg\min_{x\in\Omega} f^\nu(x) = \arg\min_{x\in\mathbb{R}^n} g^\nu(x). \tag{9}$$

2. Eventually level boundedness of $\{g^\nu\}$.

By the construction of g and g^ν, dom $g \subset \Omega$, dom $g^\nu \subset \Omega$. It is easy to see the sequence of sets dom g^ν is a eventually bounded. By Proposition A.2, the sequence of functions $\{g^\nu\}$ is eventually level-bounded.

3. Epi-convergence of $\{g^\nu\}$.

By Lemma A.9, we have $f^\nu \xrightarrow{e} f$ relative to Ω, and equations (6) and (7) hold. Then, we show $g^\nu \xrightarrow{e} g$ by Proposition A.3.

Case 1: $x \notin \Omega$. For any sequence $x^\nu \to x$,

$$\liminf_\nu g^\nu(x^\nu) = \infty, \quad \limsup_\nu g^\nu(x^\nu) = \infty.$$

Also, $g(x) = \infty$, equations (6) and (7) hold.

Case 2: $x \in \Omega$ and $x \notin \partial\Omega$. For any sequence $x^\nu \to x$,

$$\liminf_\nu g^\nu(x^\nu) = \liminf_\nu f^\nu(x^\nu), \quad \limsup_\nu g^\nu(x^\nu) = \limsup_\nu f^\nu(x^\nu).$$

Also, $g(x) = f(x)$. Then, equations (6) and (7) hold.

Case 3: $x \in \partial\Omega$. For any sequence $x^\nu \to x$, consider the case there is a subsequence $x^{\nu_k} \to x$ with $x^{\nu_k} \in \mathbb{R}^n/\Omega$. Then,

$$\liminf_\nu g^\nu(x^\nu) = \infty \geq g(x).$$

For any sequence $x^\nu \to x$ with no subsequence $x^{\nu_k} \to x$ with $x^{\nu_k} \in \mathbb{R}^n/\Omega$,

$$\liminf_\nu g^\nu(x^\nu) = \liminf_\nu f^\nu(x^\nu) = f(x) = g(x).$$

Then, for any sequence $x^\nu \to x$,

$$\liminf_\nu g^\nu(x^\nu) \geq g(x).$$

On the other hand, there exists a sequence $x^\nu \to x$ with all $x^\nu \in \Omega$. In this case,

$$\limsup_\nu g^\nu(x^\nu) = \limsup_\nu f^\nu(x^\nu) \leq f(x) = g(x).$$

Equations (6) and (7) hold. Then, $g^\nu \xrightarrow{e} g$.

4. Continuity of g and g^ν.

By the construction of f^ν and Proposition A.6, f^ν is continuous in Ω. Then, by the construction of g and g^ν, g and g^ν are proper. Since g and g_ν are continuous on a closed set Ω, then the level set $\mathrm{lev}_{\leq\alpha} g$ and $\mathrm{lev}_{\leq\alpha} g_\nu$ are closed for all $\alpha \in \mathbb{R}$. Then g and g_ν are lsc by Proposition A.4.

Notice that by equations (8) and (9), the solution set of (P) and (P_ν) are equivalent to the solution set of $(\tilde{\mathrm{P}})$ and $(\tilde{\mathrm{P}}^\nu)$. Let $S = \arg\min_{x\in\Omega} f(x)$ and $S^\nu = \arg\min_{x\in\Omega} f^\nu(x)$. With all the discussions above, the proof is finished by Proposition A.8.

□

B. Deferred Proofs of Section 3

Proof of Lemma 3.2.

Proof. For any $x, y \in \mathbb{R}^n$,

$$\begin{aligned}
\|f^\nu(x) - f^\nu(y)\| &= \|\phi_{1/\nu}\tilde{*}f(x) - \phi_{1/\nu}\tilde{*}f(y)\| \\
&= \left\| \int_{\mathbb{R}^n} (f(x-z) - f(y-z))\, \phi_{1/\nu}(z)\, \mathrm{d}z \right\| \quad \text{(Bochner integral)} \\
&\leq \int_{\mathbb{R}^n} \left\| (f(x-z) - f(y-z))\, \phi_{1/\nu}(z) \right\| \mathrm{d}z \\
&= \int_{\mathbb{R}^n} \|(f(x-z) - f(y-z))\|\, \phi_{1/\nu}(z)\, \mathrm{d}z \\
&\leq L\|x-y\| \int_{\mathbb{R}^n} \phi_{1/\nu}(z)\, \mathrm{d}z = L\|x-y\|.
\end{aligned}$$

□

Proof of Lemma 3.4.

Proof. (a):

$$\begin{aligned}
\mathbb{E}\left[g^\nu(x_k, \xi_k)\right] &= \mathbb{E}\left[\int_{\mathbb{R}^n} g(y, \xi_k) \phi_{1/\nu}(x_k - y)\, \mathrm{d}y \right] \\
&= \int_{\mathbb{R}^n} \mathbb{E}_{\xi_k}\left[g(y, \xi_k)\right] \phi_{1/\nu}(x_k - y)\, \mathrm{d}y \\
&= \int_{\mathbb{R}^n} \nabla f(y) \phi_{1/\nu}(x_k - y)\, \mathrm{d}y = \nabla f^\nu(x_k).
\end{aligned}$$

(b):

$$
\begin{aligned}
&\mathbb{E}\left[\left\|g^{\nu}(x_k,\xi_k)-\nabla f^{\nu}(x_k)\right\|^2\right]\\
&=\mathbb{E}\left[\left\|\int_{\mathbb{R}^n} g(y,\xi_k)\phi_{1/\nu}(x_k-y)\,\mathrm{d}y-\int_{\mathbb{R}^n}\nabla f(y)\phi_{1/\nu}(x_k-y)\,\mathrm{d}y\right\|^2\right]\\
&=\mathbb{E}\left[\left\|\int_{\mathbb{R}^n}\left(g(y,\xi_k)-\nabla f(y)\right)\phi_{1/\nu}(x_k-y)\,\mathrm{d}y\right\|^2\right]\\
&\leq\mathbb{E}\left[\left(\int_{\mathbb{R}^n}\left\|\left(g(y,\xi_k)-\nabla f(y)\right)\phi_{1/\nu}(x_k-y)\right\|\,\mathrm{d}y\right)^2\right]\ \text{(Bochner integral)}\\
&=\mathbb{E}\left[\left(\int_{\mathbb{R}^n}\left\|g(y,\xi_k)-\nabla f(y)\right\|\phi_{1/\nu}(x_k-y)\,\mathrm{d}y\right)^2\right]\\
&\leq\mathbb{E}\left[\left(\max_y\left(\left\|g(y,\xi_k)-\nabla f(y)\right\|\right)\int_{\mathbb{R}^n}\phi_{1/\nu}(x_k-y)\,\mathrm{d}y\right)^2\right]\\
&=\mathbb{E}\left[\left(\max_y\left(\left\|g(y,\xi_k)-\nabla f(y)\right\|\right)\right)^2\right]\leq\sigma^2.
\end{aligned}
$$

□

Proof of Lemma 3.5.

Proof. (a):

$$
\begin{aligned}
\mathbb{E}\left[g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)\right]&=\mathbb{E}\left[\mathbb{E}\left[g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)|\xi_k\right]\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\frac{1}{N_\nu}\sum_{i=1}^{N_\nu}g((X^{\nu}_k)_i,\xi_k)|\xi_k\right]\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[g(X^{\nu}_k,\xi_k)|\xi_k\right]\right]\\
&=\mathbb{E}\left[\int_{\mathbb{R}^n}g(y,\xi_k)\phi_{1/\nu}(x_k-y)\,\mathrm{d}y\right]\\
&=\mathbb{E}\left[g^{\nu}(x_k,\xi_k)\right]=\nabla f^{\nu}(x_k)\quad\text{by Lemma 3.4 (a).}
\end{aligned}
$$

(b):

$$
\begin{aligned}
\mathbb{E}\left[g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)|\xi_k\right]&=\mathbb{E}\left[\frac{1}{N_\nu}\sum_{i=1}^{N_\nu}g((X^{\nu}_k)_i,\xi_k)\middle|\,\xi_k\right]\\
&=\mathbb{E}\left[\,g(X^{\nu}_k,\xi_k)|\,\xi_k\right]=\int_{\mathbb{R}^n}g(y,\xi_k)\phi_{1/\nu}(x_k-y)\,\mathrm{d}y=g^{\nu}(x_k,\xi_k).
\end{aligned}
$$

Denote $\epsilon(\xi_k)=g^{\nu}(x_k,\xi_k)-\nabla f^{\nu}(x_k)$. By Lemma 3.4 (a) and (b),

$$
\mathbb{E}[\epsilon(\xi_k)]=0,\quad\mathbb{E}[\|\epsilon(\xi_k)\|^2]\leq\sigma^2. \tag{10}
$$

Bias–variance decomposition:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)-\nabla f^{\nu}(x_k)\right\|^2\right]\\
&=\mathbb{E}\left[\left\|g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)-g^{\nu}(x_k,\xi_k)+g^{\nu}(x_k,\xi_k)-\nabla f^{\nu}(x_k)\right\|^2\right]\\
&=\mathbb{E}\left[\left\|g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)-g^{\nu}(x_k,\xi_k)+\epsilon(\xi_k)\right\|^2\right]\\
&=\mathbb{E}\left[\left\|g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)-g^{\nu}(x_k,\xi_k)\right\|^2\right]+2\mathbb{E}\left[\epsilon(\xi_k)^T\left(g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)-g^{\nu}(x_k,\xi_k)\right)\right]+\mathbb{E}\left[\|\epsilon(\xi_k)\|^2\right].
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\left\|g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)-g^{\nu}(x_k,\xi_k)\right\|^2\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\sum_{l=1}^{d}\left(g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)_l-g^{\nu}(x_k,\xi_k)_l\right)^2\middle|\,\xi_k\right]\right]\\
&=\mathbb{E}\left[\sum_{l=1}^{d}\mathbb{V}\left[g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)_l\middle|\,\xi_k\right]\right]=\mathbb{E}\left[\sum_{l=1}^{d}\mathbb{V}\left[\frac{1}{N_\nu}\sum_{i=1}^{N_\nu}g((X^{\nu}_k)_i,\xi_k)_l\middle|\,\xi_k\right]\right]\\
&=\frac{1}{N_\nu}\mathbb{E}\left[\sum_{l=1}^{d}\mathbb{V}\left[g(X^{\nu}_k,\xi_k)_l\middle|\,\xi_k\right]\right]\\
&\le\frac{1}{N_\nu}\mathbb{E}\left[\sum_{l=1}^{d}\mathbb{E}\left[\left(g(X^{\nu}_k,\xi_k)_l\right)^2\middle|\,\xi_k\right]\right]=\frac{1}{N_\nu}\mathbb{E}\left[\mathbb{E}\left[\left\|g(X^{\nu}_k,\xi_k)\right\|^2\middle|\,\xi_k\right]\right]\\
&=\frac{1}{N_\nu}\mathbb{E}\left[\int_{\mathbb{R}^n}\left\|g(y,\xi_k)\right\|^2\phi_{1/\nu}(x_k-y)\,\mathrm{d}y\right]\\
&=\frac{1}{N_\nu}\int_{\mathbb{R}^n}\mathbb{E}\left[\left\|g(y,\xi_k)\right\|^2\right]\phi_{1/\nu}(x_k-y)\,\mathrm{d}y
\end{aligned}
$$

By Assumption 3.3 (b),

$$
\mathbb{E}\left[\left\|g(y,\xi_k)-\nabla f(y)\right\|^2\right]=\mathbb{E}\left[\left\|g(y,\xi_k)\right\|^2\right]-\left\|\nabla f(y)\right\|^2\le\sigma^2.
$$

Then,

$$
\mathbb{E}\left[\left\|g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)-g^{\nu}(x_k,\xi_k)\right\|^2\right]\le\frac{\sigma^2+M}{N_\nu},
$$

where $M=\max_x\|\nabla f(x)\|^2$. We have

$$
\mathbb{E}\left[\left\|g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)-\nabla f^{\nu}(x_k)\right\|^2\right]\le\sigma^2+\frac{\sigma^2+M}{N_\nu}.
$$

□

Proof of Theorem 3.6.

Proof. By the construction of f^ν and Proposition A.6, f^ν is continuous on the closed set Ω. Then subproblem (P^ν) has a non-empty solution set.

By Lemma 3.5 (b),

$$
\mathbb{E}\left[\left\|g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)-\nabla f^{\nu}(x_k)\right\|^2\right]\le\mathbb{E}\left[\left\|g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)\right\|^2\right]-\left\|\nabla f^{\nu}(x_k)\right\|^2\le\sigma^2+\frac{\sigma^2+M}{N_\nu}.
$$

Also,

$$
\begin{aligned}
\left\|\nabla f^{\nu}(x_k)\right\|^2&=\left\|\int_{\mathbb{R}^n}\nabla f(x_k-y)\phi_{1/\nu}(y)\,\mathrm{d}y\right\|^2\\
&\le\left(\int_{\mathbb{R}^n}\left\|\nabla f(x_k-y)\phi_{1/\nu}(y)\right\|\,\mathrm{d}y\right)^2\\
&\le\left(\max_y(\|\nabla f(x_k-y)\|)\int_{\mathbb{R}^n}\phi_{1/\nu}(y)\,\mathrm{d}y\right)^2=\max_x\|\nabla f(x)\|^2=:M.
\end{aligned}
$$

Then,

$$
\mathbb{E}\left[\left\|g^{\nu}_{N_\nu}(X^{\nu}_k,\xi_k)\right\|^2\right]\le\sigma^2+\frac{\sigma^2+M}{N_\nu}+M.
$$

The proof is finished directly by Theorem 4 in (Karimi et al., 2016). □