OPT2023: 15th Annual Workshop on Optimization for Machine Learning

Noise Injection Irons Out Local Minima and Saddle Points

Konstantin Mishchenko KONSTA.MISH@GMAIL.COM
Samsung AI Center, UK

Sebastian U. Stich STICH@CISPA.DE
CISPA, Germany

Abstract

Non-convex optimization problems are ubiquitous in machine learning, especially in Deep Learning. It has been observed in practice, that injecting artificial noise into stochastic gradient descent (SGD) can sometimes improve training and generalization performance.

In this work, we formalize noise injection as a smoothing operator and (review and) derive convergence guarantees of SGD under smoothing. We empirically found that Gaussian smoothing works really well for training two-layer neural networks, but these findings do not translate to deeper nets. We would like to use this contribution to stimulate a discussion in the community to further investigate the impact of noise in training machine learning models.

1. Introduction

Non-convex optimization problems are ubiquitous in deep learning and computer vision. Stochastic gradient descent methods, like SGD are core components for training neural networks. Several works proposed to artificially inject noise into the SGD process for improved generalization [3, 16, 19, 20] in particular also the context of large batch training [8, 13, 25]. The effect of noise injection can formalized as a smoothing operator, which has been well-studied in the optimization literature [17, 18].

In this work, we review some of the classic results on Gaussian smoothing and analyze the convergence of SGD with injected Gaussian noise under weak assumptions. We show that noise injection helps avoid local minima and saddle points if they are due to high-frequency non-convexity. Numerically, we show that SGD with Gaussian noise can outperform other standard training techniques (including SGD with momentum) on the phase retrieval problem. We also run experiments on two-layer neural networks.

Related Work. It has been observed that the gradient noise can help SGD escape saddle points [5] or achieve better generalization [7]. This is often explained by arguing that SGD finds 'flat' minima with favorable generalization properties [9, 10, 12] though 'sharp' minima can also generalize well [4]. Recently, many works proposed to inject artificial noise into SGD training, with the aim to inherit beneficial properties of the noise [3, 8, 13, 16, 19, 20, 25]. However, a rigorous theoretical framework is lacking.

In the optimization community, smoothing has been proposed frequently to facilitate convergence or overcome other optimization difficulties [14, 22]. Theoretical results have been established in [17, 18].

It is well-known that gradient descent is not improvable for general nonconvex optimization [2, 17]. Therefore, additional structural assumptions are needed. In this work, we assume (see below) that the objective function can be decomposed into two components, $f = \hat{f} + \omega$, where ω is a high-frequency nonconvexity-inducing function. Such decompositions appeared, e.g. [15] studied such a decomposition from the perspective of computational inaccuracies, and [1, 21, 24] studied such decompositions with a similar motivation as in this work. The authors of [1] considered the setting closely related to ours with the key difference that they were interested in minimizing $\hat{f}$ with zeroth-order oracle rather than gradient-based minimization of f. They also did not study PL function $\hat{f}$ and instead showed convergence of $\|\nabla \hat{f}\|$ to zero. Recently, [21] studied the problem under a structure similar to ours but imposing smoothness and PL for f (our assumption is different in assuming this for $\hat{f}$). [24] studied the problem under slightly more restrictive assumptions. [23] showed in numerical experiments that many problems are non-smooth, discontinuous, and have high-frequency minima. They claim that smoothing helps with their objectives.

2. Problem and assumptions

Consider the problem of minimizing a non-smooth non-convex function:

$$\min_{x \in \mathbb{R}^d} f(x) = \mathbb{E}\left[f(x;\xi)\right].$$

In general, there is not much we can do as this problem is intractable [17] without further assumption. On the other hand, we face such problems in many applications, most notably when training neural networks. Our main interest is in the situation where there exists an unknown decomposition

$$f(x) = \hat{f}(x) + \omega(x).$$

Here, $\omega(x)$ plays the role of a high-frequency nonconvexity-inducing function. Formally, this property is stated later in Assumption 3. The decomposition might not be unique, so we choose any of them that minimizes the values of $\sup_x |\omega(x)|$ and $\sup_x \|\partial\omega(x)\|$. We also assume that f is lower bounded by some value $f_{\text{inf}} = \inf_{x\in\mathbb{R}^d} f(x)$.

Our assumptions might not be appropriate for the large class of deep neural networks, but we observe good results for shallow networks. This aligns with prior work, for instance, [6] studies the loss landscape of 7 shallow-neural-network training tasks on MNIST and demonstrated that empirically, it is almost convex.

3. General theory for smoothing

We are going to study the following smoothing technique,

$$f_\zeta(x) \overset{\text{def}}{=} \mathbb{E}_{u\sim\mathcal{N}(0,\mathbf{I})}[f(x+\zeta u)], \tag{1}$$

where $\zeta > 0$ is a parameter, $\mathbf{I}$ is the identity matrix, and u is sampled from the standard normal distribution $\mathcal{N}(0,\mathbf{I})$. We define $\hat{f}_\zeta$ and ω_ζ the same way as f_ζ. Notice that it holds by linearity of expectation that $f_\zeta = \hat{f}_\zeta + \omega_\zeta$. The choice of using the normal distribution is not necessary and is only made for simplicity.

Assumption 1 *We assume that function $\hat{f}$ is L-smooth, i.e., for any $x, y \in \mathbb{R}^d$*

$$\|\nabla\hat{f}(x) - \nabla\hat{f}(y)\| \leq L\|x-y\|. \tag{2}$$

Algorithm 1 SGD with noise injections

```
Input: initialization $x_0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, noise coefficients $\{\zeta_k\}_{k=0}^{\infty}$
for $k = 0, 1, 2, \dots$ do
    Sample $\xi_k$ and $u_k \sim \mathcal{N}(0, \mathbf{I})$
    $x_{k+1} = x_k - \gamma \nabla f(x_k + \zeta_k u_k; \xi_k)$
end for
```

In particular, this implies that

$$\hat{f}(y) \leq \hat{f}(x) + \langle \nabla \hat{f}(x), y - x \rangle + \frac{L}{2}\|y - x\|^2. \tag{3}$$

Assumption 2 *We say that the unknown part $\hat{f}$ of f satisfies Polyak-Łojasiewicz inequality if there exists $\mu > 0$ such that*

$$\frac{1}{2}\|\nabla \hat{f}(x)\|^2 \geq \mu(\hat{f}(x) - \hat{f}_{\inf}), \tag{4}$$

where $\hat{f}_{\inf} = \min_x \hat{f}(x)$.

Assumption 3 *We say that the unknown function ω induces high-frequency nonconvexity if it satisfies*

$$|\omega(x)| \leq e_0, \tag{5}$$

$$|\partial\omega(x)| \leq e_1, \tag{6}$$

where $\partial\omega$ denotes a Clarke subdifferential of ω.

4. Convergence of SGD with noise injection

In order to formulate a convergence result for SGD with noise injection, we make the assumption that the stochastic noise is bounded. This is a standard assumption in the literature, but it could also be relaxed [24].

Assumption 4 *The stochastic gradients have bounded variance, i.e., at any $x \in \mathbb{R}^d$*

$$\mathbb{E}\left[\|\nabla f(x;\xi) - \nabla f(x)\|^2\right] \leq \sigma^2. \tag{7}$$

Remark 1 *Instead of Assumption 4, we could also instead assume that the variance of gradients of $\hat{f}$ is bounded by a constant $\hat{\sigma}^2$. From the observation*

$$\begin{aligned}\mathbb{E}\left[\|\nabla f(x;\xi) - \nabla f(x)\|^2\right] &\leq 3\mathbb{E}\left[\|\nabla \hat{f}(x;\xi) - \nabla \hat{f}(x)\|^2\right] + 3\mathbb{E}\left[\|\nabla \omega(x;\xi)\|^2\right] + 3\|\nabla\omega(x)\|^2 \\ &\leq 3\hat{\sigma}^2 + 6e_1^2,\end{aligned}$$

we conclude that this would result in equivalent convergence bounds.

4.1. Additional bias under noise injection

We now state two lemmas that estimate the effect of the noise injection, which is formally given in Algorithm 1. The additional noise is controlled by the parameter ζ (see Equation (1) and Alg. 1).

Proposition 2 (Lemma 3 in [18]) *Let $\hat{f}$ be L-smooth, then for any $\zeta > 0$ and $x \in \mathbb{R}^d$, it holds*

$$\|\nabla \hat{f}_\zeta(x) - \nabla \hat{f}(x)\| \leq \frac{\zeta}{2} L (d+3)^{\frac{3}{2}}. \tag{8}$$

Lemma 3 *The perturbation coming from ω diminishes if we inject more noise,*

$$\|\nabla \omega_\zeta(x)\|^2 \leq \mathbb{E}\left[\|\nabla \omega_\zeta(x)\|^2\right] \leq d \frac{e_0^2}{\zeta^2}. \tag{9}$$

Proof This is a straightforward corollary of the following standard identity (see, e.g., Equation (21) in [18] or Section 9.3 in [17]):

$$\nabla \omega_\zeta(x) = \frac{1}{\zeta} \mathbb{E}\left[\omega(x + \zeta u) u\right].$$

Therefore,

$$\|\nabla \omega_\zeta(x)\|^2 = \frac{1}{\zeta^2} \|\mathbb{E}\left[\omega(x+\zeta u)u\right]\|^2 \leq \frac{1}{\zeta^2} \mathbb{E}\left[\|\omega(x+\zeta u)u\|^2\right] \leq \frac{e_0^2}{\zeta^2} \mathbb{E}\left[\|u\|^2\right] = d\frac{e_0^2}{\zeta^2}, \tag{10}$$

which proves the lemma. ■

Lemma 4 *The noise of the smoothing is bounded as*

$$\mathbb{E}_u\left[\|\nabla f(x+\zeta u) - \nabla f_\zeta(x)\|^2\right] \leq 4dL^2\zeta^2 + 8e_1^2. \tag{11}$$

The proof of this lemma is given in the appendix.

4.2. Convergence of SGD with noise injections

We now prove the convergence of Algorithm 1.

Theorem 5 *Let Assumptions 1, 3 and 2 hold. Then, there exists a stepsize $\gamma \leq \frac{1}{L}$ such that*

$$\mathbb{E}\left[f(x_k) - f_{\inf}\right] \leq \varepsilon + \frac{1}{\mu}\left(\frac{\zeta_k}{2} L(d+3)^{\frac{3}{2}} + d\frac{e_0^2}{\zeta_k^2}\right) + 2e_0.$$

after $k = \mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon} + \frac{L}{\mu}\frac{\sigma^2 + dL^2\zeta_k^2 + e_1^2}{\varepsilon}\right)$ iterations.

On the one hand, this result shows that Algorithm 1 can in general *not* converge. Instead, the final accuracy is limited by the parameters e_0 and the noise ζ_k. This is consistent with other works [18, 24], [21] give a lower bound on quadratic functions.

On the other hand, the result shows that Algorithm 1 will converge to a neighborhood of the minimizer of $\hat{f}$, and thus cannot get stuck at saddle points or local minima of f that are far away.

Notice that if we ignore the complexity and just try to find the value of ζ_k that minimizes the upper bound of Theorem 5, i.e., $\min_\zeta \frac{\zeta}{2} L(d+3)^{\frac{3}{2}} + d\frac{e_0^2}{\zeta^2}$, we obtain $\zeta_k^3 = 2e_0^2 \frac{d}{L(d+3)^{\frac{3}{2}}}$, which means $\zeta_k = \mathcal{O}(\frac{1}{\sqrt[6]{d}})$. Overall, this means convergence to a $\mathcal{O}(\varepsilon + d^{\frac{4}{3}})$ neighborhood after $k = \tilde{\mathcal{O}}(\frac{\sigma^2 + e_1^2 + d^{\frac{2}{3}}}{\varepsilon})$ iterations.

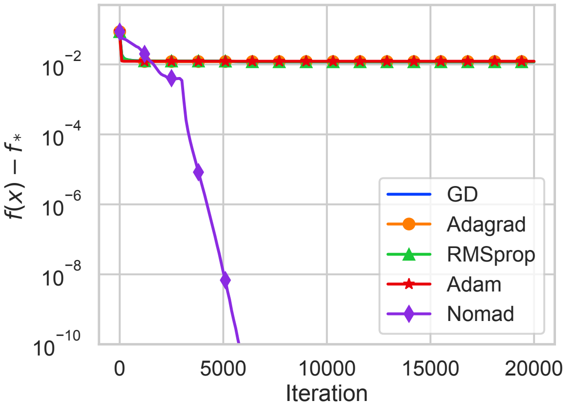

Figure 1: An example of minimizing the phase retrieval loss. As can be seen, only with noise injections as given in Algorithm 2 we can achieve zero loss.

5. Numerical Experiments

To exemplify how one can use noise injections in real-world problems, we consider the phase retrieval problem:

$$\min_x \frac{1}{4}\|(\mathbf{A}x)^2 - b^2\|^2,$$

where $\mathbf{A} \in \mathbb{R}^{n\times d}$ is the data matrix, $b \in \mathbb{R}^n$ is the amplitudes vector, and $(Ax)^2$ is the vector obtained by squaring the elements of $\mathbf{A}x$ coordinate-wise. We randomly generate $\mathbf{A}$ and solution x_* by independently sampling from a standard normal distribution and then set $b = \mathbf{A}x_*$. We tuned the stepsize for each method and used adaptive noise for Nomad. The results are depicted in Figure 1.

6. Discussion and Conclusion

We have proven, that for certain objective functions, Gaussian noise injection is an effective method to overcome optimization difficulties caused by high-frequency noise. Theorem 5 shows that the SGD with noise injection can converge to a neighborhood of the solution, a property that does not hold for vanilla SGD. For our analysis, we assumed the PL property for the underlying regular part $\hat{f}$ of the objective function, but generalization to other settings is possible [see e.g. 24].

One important research direction for making noise injection practical is to make them black-box. Indeed, if we were to choose a single noise coefficient for the whole optimization process, we may end up either with a strong noise that would shift the solution away or with a weak one that is not sufficient to make any difference. Both situations are bad as we want to find a good solution and yet have noise large enough to change the objective.

One of the goals, when we started this project, was to fill the gap in theory and find better explanations for the successes of noise injection that have been reported in other work [11, 19]. However, in numerical tests, we did not find many cases where smoothing outperformed other (standard) DL training methods, especially not for deeper networks.

References

[1] Albert S. Berahas, Liyuan Cao, Krzysztof Choromanski, and Katya Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Foundations of Computational Mathematics*, 22(2):507–560, 2022.

[2] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1):71–120, 2020.

[3] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019 (12):124018, 2019.

[4] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

[5] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

[6] Ian J. Goodfellow, Oriol Vinyals, and Andrew M. Saxe. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*, 2014.

[7] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[8] Kosuke Haruki, Taiji Suzuki, Yohei Hamakawa, Takeshi Toda, Ryuji Sakai, Masahiro Ozawa, and Mitsuhiro Kimura. Gradient noise convolution (GNC): Smoothing loss function for distributed large-batch SGD. *arXiv preprint arXiv:1906.10822*, 2019.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[10] Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

[11] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

[12] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[13] Tao Lin, Lingjing Kong, Sebastian Stich, and Martin Jaggi. Extrapolation for large-batch training in deep learning. In *International Conference on Machine Learning*, pages 6094–6104. PMLR, 2020.

[14] J. Mátyáš. Random optimization. *Automation and Remote control*, 26(2):244–251, 1965.

[15] Jorge J. Moré and Stefan M. Wild. Estimating computational noise. *SIAM Journal on Scientific Computing*, 33(3):1292–1314, 2011.

[16] Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

[17] Arkadij Semenovič Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[18] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[19] Antonio Orvieto, Hans Kersting, Frank Proske, Francis Bach, and Aurelien Lucchi. Anticorrelated noise injection for improved generalization. In *International Conference on Machine Learning*, pages 17094–17116. PMLR, 2022.

[20] Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration. *arXiv preprint arXiv:1706.01905*, 2017.

[21] Boris T. Polyak, Ilia A. Kuruzov, and Fedor S. Stonyakin. Stopping rules for gradient methods for non-convex problems with additive noise in gradient. *arXiv preprint arXiv:2205.07544*, 2022.

[22] L.Ã. Rastrigin. The convergence of the random search method in the extremal control of a many parameter system. *Automaton & Remote Control*, 24:1337–1342, 1963.

[23] H. J. Terry Suh, Max Simchowitz, Kaiqing Zhang, and Russ Tedrake. Do differentiable simulators give better policy gradients? *arXiv preprint arXiv:2202.00817*, 2022.

[24] Harsh Vardhan and Sebastian U. Stich. Tackling benign nonconvexity with smoothing and stochastic gradients. *arXiv preprint arXiv:2202.09052*, 2022.

[25] Wei Wen, Yandan Wang, Feng Yan, Cong Xu, Chunpeng Wu, Yiran Chen, and Hai Li. Smoothout: Smoothing out sharp minima to improve generalization in deep learning. *arXiv preprint arXiv:1805.07898*, 2018.

Appendix A. Adaptive method

Algorithm 2 Nomad (noisy method with adaptive stepsizes)

```
**Input:** initialization $x_0 \in \mathbb{R}^d$, stepsize $\gamma > 0$ (default $10^{-3}$), $\beta_1 \in [0, 1)$ (default 0.9), $\beta_2 \in [0, 1)$ (default 0.999), $\epsilon \geq 0$ (default $10^{-8}$), noise hyper-parameters $\alpha_1, \alpha_2$ (default 0.5)
**for** $k = 0, 1, 2, \dots$ **do**
    Sample $\xi_k$ and $u_k \sim \mathcal{N}(0, \mathbf{I})$
    $\zeta_k = \gamma^{\alpha_2} \frac{(f(x_k)-f_*)^{\alpha}_1}{(v_k+\epsilon)^{\alpha_2/2}}$
    $g_k = \nabla f(x_k + \zeta_k u_k; \xi_k)$
    $m_k = \beta_1 m_{k-1} + (1-\beta_1) g_k$
    $v_k = \beta_2 v_{k-1} + (1-\beta_2) g_k^2$
    $x_{k+1} = x_k - \gamma \frac{m_k}{\sqrt{v_k+\epsilon}}$
**end for**
```

Appendix B. Proof of Lemma 4

Proof [Proof of Lemma 4] By Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}_u \left[\|\nabla f(x+\zeta u) - \nabla f_\zeta(x)\|^2\right] &= \mathbb{E}_u \left[\|\nabla f(x+\zeta u) - \mathbb{E}_v \left[\nabla f(x+\zeta v)\right]\|^2\right] \\
&\leq \mathbb{E}\left[\|\nabla f(x+\zeta u) - \nabla f(x+\zeta v)\|^2\right] \\
&\leq 2\mathbb{E}\left[\|\nabla \hat{f}(x+\zeta u) - \nabla \hat{f}(x+\zeta v)\|^2\right] \\
&\quad + 2\mathbb{E}\left[\|\nabla \omega(x+\zeta u) - \nabla \omega(x+\zeta v)\|^2\right] \\
&\overset{(6)}{\leq} 2\mathbb{E}\left[\|\nabla \hat{f}(x+\zeta u) - \nabla \hat{f}(x+\zeta v)\|^2\right] + 8e_1^2 \\
&\overset{(2)}{\leq} 2L^2\mathbb{E}\left[\|\zeta u - \zeta v\|^2\right] + 8e_1^2.
\end{aligned}$$

Finally, notice that u, v are independent normal variables, so

$$\mathbb{E}\left[\|\zeta u - \zeta v\|^2\right] = \zeta^2 \mathbb{E}\left[\|u\|^2 - 2\langle u, v\rangle + \|v\|^2\right] = 2d\zeta^2.$$

■

Appendix C. Proof of Theorem 5

Proof [Proof of Theorem 5] First, notice that the update rule satisfies, conditioned on x_k, the following:

$$\mathbb{E}\left[x_{k+1}\right] = x_k - \gamma \mathbb{E}\left[\nabla f(x_k + \zeta_k u_k; \xi_k)\right] = x_k - \gamma \mathbb{E}\left[\nabla f(x_k + \zeta_k u_k)\right] = x_k - \gamma \nabla f_{\zeta_k}(x_k). \quad (12)$$

It holds by Assumption 1,

$$\begin{aligned}\mathbb{E}\left[\hat{f}(x_{k+1})\right] &\overset{(3)}{\leq} \hat{f}(x_k) + \mathbb{E}\left[\langle \nabla \hat{f}(x_k), x_{k+1} - x_k\rangle\right] + \frac{L}{2}\mathbb{E}\left[\|x_{k+1} - x_k\|^2\right] \\ &\overset{(12)}{=} \hat{f}(x_k) - \gamma\langle \nabla \hat{f}(x_k), \nabla f_{\zeta_k}(x_k)\rangle + \frac{L\gamma^2}{2}\mathbb{E}\left[\|\nabla f(x_k + \zeta_k u_k; \xi_k)\|^2\right] \\ &\overset{(7)}{\leq} \hat{f}(x_k) - \gamma\langle \nabla \hat{f}(x_k), \nabla f_{\zeta_k}(x_k)\rangle + \frac{L\gamma^2}{2}\mathbb{E}\left[\|\nabla f(x_k + \zeta_k u_k)\|^2\right] + \frac{L\gamma^2}{2}\sigma^2.\end{aligned}$$

Recall that $\nabla f(x_k + \zeta_k u_k)$ is an unbiased estimate of $\nabla f_{\zeta_k}(x_k)$, so

$$\begin{aligned}\mathbb{E}\left[\|\nabla f(x_k + \zeta_k u_k)\|^2\right] &= \|\nabla f_{\zeta_k}(x_k)\|^2 + \mathbb{E}\left[\|\nabla f(x_k + \zeta_k u_k) - \nabla f_{\zeta_k}(x_k)\|^2\right] \\ &\overset{(11)}{\leq} \|\nabla f_{\zeta_k}(x_k)\|^2 + 4dL^2\zeta_k^2 + 8e_1^2.\end{aligned}$$

Using identity $-\langle a, b\rangle = -\frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2 + \frac{1}{2}\|a - b\|^2$ and Lemma 3, we get

$$\begin{aligned}-\langle \nabla \hat{f}(x_k), \nabla f_{\zeta_k}(x_k)\rangle &= -\frac{1}{2}\|\nabla \hat{f}(x_k)\|^2 - \frac{1}{2}\|\nabla f_{\zeta_k}(x_k)\|^2 + \frac{1}{2}\|\nabla \hat{f}(x_k) - \nabla f_{\zeta_k}(x_k)\|^2 \\ &= -\frac{1}{2}\|\nabla \hat{f}(x_k)\|^2 - \frac{1}{2}\|\nabla f_{\zeta_k}(x_k)\|^2 + \frac{1}{2}\|\nabla \hat{f}(x_k) - \nabla \hat{f}_{\zeta_k}(x_k) - \nabla \omega_{\zeta_k}(x_k)\|^2 \\ &\leq -\frac{1}{2}\|\nabla \hat{f}(x_k)\|^2 - \frac{1}{2}\|\nabla f_{\zeta_k}(x_k)\|^2 + \|\nabla \hat{f}(x_k) - \nabla \hat{f}_{\zeta_k}(x_k)\|^2 + \|\nabla \omega_{\zeta_k}(x_k)\|^2.\end{aligned}$$

Proposition 2 in combination with Lemma 3 imply

$$\|\nabla \hat{f}(x_k) - \nabla \hat{f}_{\zeta_k}(x_k)\|^2 + \|\nabla \omega_{\zeta_k}(x_k)\|^2 \leq \frac{\zeta_k}{2}L(d+3)^{\frac{3}{2}} + d\frac{e_0^2}{\zeta_k^2}.$$

Putting the pieces together yields

$$\begin{aligned}\mathbb{E}\left[\hat{f}(x_{k+1})\right] \leq \hat{f}(x_k) - \frac{\gamma}{2}\|\nabla \hat{f}(x_k)\|^2 - \frac{\gamma}{2}\|\nabla f_{\zeta_k}(x_k)\|^2 + \frac{L\gamma^2}{2}\|\nabla f_{\zeta_k}(x_k)\|^2 \\ + \frac{L\gamma^2}{2}(\sigma^2 + 4dL^2\zeta_k^2 + 8e_1^2) + \gamma\left(\frac{\zeta_k}{2}L(d+3)^{\frac{3}{2}} + d\frac{e_0^2}{\zeta_k^2}\right).\end{aligned}$$

Since we assume that $\gamma \leq \frac{1}{L}$, the overall coefficient in front of $\|\nabla f_{\zeta_k}(x_k)\|^2$ is negative and we can drop this term. Next, by the Polyak-Łojasiewicz assumption on $\hat{f}$, we have for the remaining gradient term

$$-\frac{1}{2}\|\nabla \hat{f}(x_k)\|^2 \leq -\mu(\hat{f}(x) - \hat{f}_{\inf}).$$

Thus,

$$\mathbb{E}\left[\hat{f}(x_{k+1}) - \hat{f}_{\inf}\right] \leq (1-\gamma\mu)(\hat{f}(x_k) - \hat{f}_{\inf}) + \frac{L\gamma^2}{2}(\sigma^2 + 4dL^2\zeta_k^2 + 8e_1^2) + \gamma\left(\frac{\zeta_k}{2}L(d+3)^{\frac{3}{2}} + d\frac{e_0^2}{\zeta_k^2}\right).$$

Denote $c_1 = \frac{L}{2}(\sigma^2 + 4dL^2\zeta_k^2 + 8e_1^2)$, $c_2 = \frac{\zeta_k}{2}L(d+3)^{\frac{3}{2}} + d\frac{e_0^2}{\zeta_k^2}$ and $r_k = \mathbb{E}\left[\hat{f}(x_k) - \hat{f}_{\inf}\right]$. We have obtained for arbitrary k that

$$r_{k+1} \leq (1 - \gamma\mu)r_k + \gamma^2 c_1 + \gamma c_2.$$

Recursions of this type are standard and they imply

$$r_k \leq (1-\gamma\mu)^k r_0 + \gamma\frac{c_1}{\mu} + \frac{c_2}{\mu}.$$

Given some $\varepsilon > 0$, to find a sweet spot between the first two terms in the upper bound, we can choose $\gamma = \min\left(\frac{1}{L},\, \frac{\varepsilon}{\mu}\right)$. This choice guarantees after $k = \mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon} + \frac{c_1}{\mu\varepsilon}\right)$ iterations that

$$\mathbb{E}\left[\hat{f}(x_k) - \hat{f}_{\inf}\right] = \mathcal{O}\left(\varepsilon + \frac{c_2}{\mu}\right).$$

Finally, by Assumption 3, we have $\hat{f}_{\inf} = \inf_x \hat{f}(x) \leq \inf_x f(x) + e_0 = f_{\inf} + e_0$, and

$$\mathbb{E}\left[f(x_k) - f_{\inf}\right] \overset{(5)}{\leq} \mathbb{E}\left[\hat{f}(x_k) - \hat{f}_{\inf}\right] + 2e_0.$$

■