Learning Poisson Intensities with Pseudo Mirror Descent

Abstract

Learning the intensity functions of a Poisson process when approached by maximizing the likelihood often proves to be computationally challenging, This stems from the positivity constraint on the intensity function which requires expensive projections at each iteration of the optimization algorithm. In this paper, we propose a novel algorithm, pseudo mirror descent, that yields efficient an estimate of intensity functions without performing expensive projections. The algorithm ensures the positivity of the estimate of the intensity by applying a multiplicative update rule. It also guarantees the smoothness of intermediate updates by exploiting pseudo-gradients. We provide a theoretical convergence analysis of the algorithm. Additionally, through simulations, we show that pseudo mirror descent outperforms the state-of-the-art benchmarks for learning Poisson processes both in terms of computational efficiency and prediction accuracy.

1. Introduction

Point processes are the mathematical abstraction used in the modeling and analysis of discrete events that arise in a wide range of applications such as finance (Bacry et al., 2015), social networks (Zhou et al., 2013), computer vision (Ge & Collins, 2009), neuroscience (Quinn et al., 2011), etc.. Yet, learning point processes, or more precisely, their intensity functions, poses a major challenge mainly because the intensity functions must be positive, a constraint that often requires performing projections especially in the nonparametric setting (see discussions in (Yang et al., 2017)). The projection step could be computationally expensive, even for simple Poisson processes.

For a Poisson process over $[0, 1]$, learning the intensity function through maximizing the likelihood requires solving the problem:

$$\min_{x \in \mathcal{H}_+} \int_0^1 x(t)\mathrm{d}t - \sum_{k=1}^{N} \log x(t_k), \tag{1}$$

where $\mathcal{H}$ is some function space and $\mathcal{H}_+$ is the part of $\mathcal{H}$ that contains pointwise positive functions over their support, and $t_1, \ldots, t_N$ are sampled from the Poisson process with intensity $x^*(t)$.

A good learning algorithm that solves (1) must be able to: (i) efficiently enforce the positivity constraint on intensity x, and (ii) guarantee the smoothness of the estimate. Previous work guaranteed the smoothness by requiring $\mathcal{H}$ to be a reproducing kernel Hilbert space with a smooth kernel (Yang et al., 2017; Bagnell & Farahmand, 2015; Flaxman et al., 2017), or by adding regularization on the derivative of the estiamtes (Yuan et al., 2010; Koenker et al.). Meanwhile, the positivity constraint in some work was enforced either by performing projection, which requires solving quadratic programs (Yang et al., 2017; Sarfraz et al., 2010; Wahba, 1990) or by semi-definite relaxations (Bagnell & Farahmand, 2015). However, none of them scales well. Alternatively, the positivity can be ensured by introducing a nonlinear link function, e.g., $x(t) = y^2(t)$, which transforms (1) into an unconstrained problem (Flaxman et al., 2017). However, this could easily break convexity of the problem formulation and consequently prevent obtaining theoretical guarantees. Thus, an efficient algorithm with theoretical guarantees remained elusive.

1.1. Our Contribution

We propose a novel algorithm, pseudo mirror descent, to solve problem (1). Unlike existing approaches, which assume $\mathcal{H}$ is an RKHS, we learn the Poisson intensity over $\mathcal{L}_2([0, 1])$, the space of square integrable functions equipped with inner product $\langle x, y\rangle = \int_{[0,1]} x(t)y(t)\mathrm{d}t$. We restrict x to be continuous, so that the optimization problem is well-defined. Positivity of x is ensured by imposing a multiplicative update inspired by the mirror descent (Nemirovski & Yudin, 1983) as follows:

$$x^{(k+1)} = x^{(k)} \exp\{-\eta_k g^{(k+1)}\},$$

where $x^{(k)}$ is the update at the k-th iteration, and η_k is the step size. Function $g^{(k)}$ is a *pseudo-gradient*, a properly selected smooth function that is closely aligned with the true gradient. The introduction of a pseudo-gradient is necessary as the functional gradient of the maximum likelihood objective contains Dirac's delta function, breaking the continuity of the updates. Generalizing the analysis in (Poljak & Tsypkin, 1973), we provide theoretical guarantees for pseudo mirror descent, showing $\mathcal{O}(1/k)$ convergence in objective value for $\eta_k = \Theta(1/k)$, in parallel to the rates of stochastic gradient descent (Bottou et al., 2018). Numerically, we show that pseudo mirror descent generates fast and competitive performance compared with the state-of-the-art benchmarks from the aforementioned approaches. Last but not least, the pseudo mirror descent algorithm can be easily generalized to handle learning tasks on more complicated point processes, such as spatial Poisson point processes, multivariate Hawkes processes (Hawkes, 1971), etc..

2. Preliminaries on Poisson Processes

A nonhomogeneous Poisson process, $N(t)$, is a counting process whose behavior is determined solely by its intensity function $x^*(t)$. The value $x^*(t)$ describes the average rate of arrival at time t: $\mathbb{E}[\mathrm{d}N(t)] = x^*(t)\mathrm{d}t$; while $\mathbb{E}[N(t)]$ is a Poisson random variable with rate $\int_0^t x^*(\tau)\mathrm{d}\tau$.

For a Poisson process, the negative of its log-likelihood over one sample path that has arrival times $t_1, \ldots, t_N$ is represented in (1). If we take expectation over the sample path, we obtain the negative of the expected log-likelihood:

$$f(x) = \int_0^1 x(t) - x^*(t)\log x(t)\mathrm{d}t, \tag{2}$$

which is the average of the negative log-likelihood evaluated over infinite number of sample paths. Indeed, if we minimize $f(x)$, the first order condition implies $x^*(t)$ to be the optimal solution because $[\nabla f(x)](t) = 1 - x^*(t)/x(t)$.

3. Pseudo Mirror Descent

We present the pseudo mirror descent in Algorithm 1. The core of the algorithm is the multiplicative rule that guarantees positivity of $x^{(k)}$ if $x^{(k-1)}$ is positive. This rule was inspired by the classic mirror descent (Nemirovski & Yudin, 1983), and obtained via solving

$$x^{(k)} = \operatorname*{argmin}_{x\in\mathcal{H}} \left\{ \langle g^{(k)}, x\rangle + \eta_{k-1}^{-1}\Delta_\Phi(x, x^{(k-1)}) \right\}, \tag{3}$$

where $g^{(k)}$ is a function that plays the role of the gradient of objective (1) or (2), and $\Delta_\Phi(x, x^{(k-1)})$ is the Bregman divergence induced by a strongly convex function Φ:

$$\begin{aligned}\Delta_\Phi(x, x^{(k-1)}) = \Phi(x) - \Phi(x^{(k-1)}) - \\ - \langle \nabla\Phi(x^{(k-1)}), x - x^{(k-1)}\rangle.\end{aligned}$$

For the purpose of guaranteeing positivity, we choose specifically $\Phi(x) = \langle x, \log x - 1\rangle$, which yields the multiplicative update rule in Algorithm 1.

A major challenge facing the pseudo mirror descent is the selection of the function $g^{(k)}$ in (3), which, according to the mirror descent algorithm, should either be the functional gradient of (1) if we were to solve the maximum likelihood objective, or the functional gradient of (2) if we were to optimize over the expected log-likelihood. However, each approach has its respective issue: on one hand, in order to obtain the multiplicative update rule, the optimization of (3) must be over $\mathcal{L}_2[0,1]$, but the $\mathcal{L}_2$ gradient of the maximum likelihood objective contains Dirac's delta functions, which makes the update discontinuous; on the other hand, the computation of $\nabla f(x)$ requires the information of $x^*(t)$, which is unavailable in practice. In order to solve the challenge of generating an update that is both positive and continuous, we introduce the concept of *pseudo-gradients*.

3.1. Pseudo-gradients

A function $g^{(k)}$ is a pseudo-gradient with respect to f and Φ if it satisfies

$$\langle \mathbb{E}[g^{(k)}|\mathcal{F}^{(k-1)}], \nabla f_\Phi(\nabla\Phi(x^{(k-1)}))\rangle \geq 0,$$

where $\mathcal{F}^{(k-1)}$ is the minimum σ-algebra generated by the intermediate updates $x^{(0)}, \ldots, x^{(k-1)}$, and $f_\Phi(x) = f(\nabla\Phi^*(x))$ with Φ^* being the Fenchel conjugate of Φ.

Intuitively, a pseudo-gradient is a random function that aligns closely with the direction of the true gradient. The quantity $\nabla f_\Phi(\nabla\Phi(x))$ is a generalization of the gradient, which can be retained upon setting $\Phi(x) = \|x\|_2^2/2$.

For the Poisson process problem of our interest, we have $\nabla\Phi(x) = \log x$, $\nabla\Phi^*(x) = \exp(x)$, and

$$f_\Phi(z) = \int_0^1 [\exp(z)](t)\mathrm{d}t - \int_0^1 x^*(t)z(t)\mathrm{d}t.$$

Hence, $\nabla f_\Phi(\nabla\Phi(x)) = x - x^*$. In practice, one does not have access to the information of x^*. Instead, only samples can be drawn from a Poisson process with intensity function being x^*. Therefore, a natural pseudo-gradient to choose is

$$g(t) = \int_0^1 x(\tau)K(t,\tau)\mathrm{d}\tau - \sum_{i=1}^N K(\tau_i, t),$$

where $K(\cdot,\cdot)$ is a positive definite kernel, and $\tau_1, \ldots, \tau_N$ are arrival times sampled from a Poisson process with intensity $x^*(t)$ over the interval $[0,1]$. Since $\mathbb{E}[g^{(k)}|\mathcal{F}^{(k-1)}]$ is the kernel embedding of $\nabla f_\Phi(\nabla\Phi(x^{(k-1)}))$, we can immediately show that $g^{(k)}$ is a pseudo-gradient.

Algorithm 1 Pseudo Mirror Descent

```
Input: iteration number $T$; step sizes $\{\eta_k\}_{k=0}^T$; negative expected log-likelihood $f$, $\Phi(x) = \langle x, \log x - 1\rangle$.
Initialize $x^{(0)} \in \mathcal{H}_+$, positive and continuous.
for $k = 1$ to $T$ do
    Compute pseudo-gradient $g^{(k)}$.
    $x^{(k)} = x^{(k-1)} \exp\{-\eta_{k-1} g^{(k)}\}$.
end for
Output: $x^{(T)}$.
```

3.2. Asymptotic Convergence

In this section, we analyze the convergence of the proposed pseudo mirror descent algorithm. To proceed, we make the following basic assumptions.

Assumption 1 (General assumptions)**.**

(i) The minimum of f, denoted by f^*, is finite.

(ii) The functional f is Gâteaux differentiable and M-smooth: $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{M}{2}\|y - x\|_2^2, \forall x, y \in \mathcal{H}$.

(iii) $\Phi(x) = \langle x, \log x - 1\rangle$, which is μ-strongly-convex with respect to $\|\cdot\|_1$ when $\|x\|_\infty \leq \mu^{-1}$.

Under the above assumptions, we show that the inner product between the pseudo-gradient and the true gradient converges to 0.

Theorem 1. *Suppose Assumption 1 holds, and for Algorithm 1, suppose that the step sizes satisfy $\eta_k \geq 0$, $\sum_{k=0}^\infty \eta_k = \infty$, and $\sum_{k=0}^\infty \eta_k^2 < \infty$. In addition, suppose $g^{(k)}$'s satisfy, for some sequence λ_k, $\sum_{k=0}^\infty \eta_k^2 \lambda_{k+1} < \infty$, and for universal positive constants K_1 and K_2,*

$$\mathbb{E}[\|g^{(k)}\|_\infty^2 | \mathcal{F}^{(k-1)}] \leq \lambda_k + K_1 f(x^{(k-1)}) + \\ +K_2 \langle \nabla f_\Phi(\nabla \Phi(x^{(k-1)})), \mathbb{E}[g^{(k)} | \mathcal{F}^{(k-1)}]\rangle.$$

Under these assumptions, for $x^{(k)}$ generated by Algorithm 1, $\lim_{k\to\infty} f(x^{(k)})$ exists almost surely, and

$$\liminf_{k\to\infty} \langle \nabla f_\Phi(\nabla \Phi(x^{(k-1)})), \mathbb{E}[g^{(k)} | \mathcal{F}^{(k-1)}]\rangle = 0$$

almost surely.

Remark 2. *When the pseudo-gradient is properly selected, the above theorem further implies that the gradient norm converges to 0. For example, suppose that for Algorithm 1, $g^{(k)}$ satisfies $\mathbb{E}[g^{(k)} | \mathcal{F}^{(k-1)}] = \nabla f_\Phi(\nabla \Phi(x^{(k-1)}))$. Then $\lim_{k\to\infty} \|\nabla f(x^{(k)})\|_2 = 0$ in probability. (See proof in Appendix.)*

We now provide a non-asymptotic convergence result to characterize the accuracy of the estimate under finite number of iterations.

3.3. Non-asymptotic Convergence Analysis

We assume that the original problem satisfies the Polyak-Łojasiewicz condition (Polyak, 1963).

Assumption 2 (Polyak-Łojasiewicz condition)**.** *For any $x \in \mathcal{H}$, suppose there exists $\gamma > 0$, such that*

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \gamma(f(x) - f^*).$$

Indeed, we can verify that the objective function of our interest satisfies Polyak-Łojasiewicz condition with constant ν when $\sup_t x(t) \leq (2\nu)^{-1}$ (See Appendix B for proof).

The above condition is commonly used to guarantee linear convergence of stochastic gradient descent up to a certain distance of the optimal value when constant step size is used (Karimi et al., 2016), or sublinear convergence with diminishing step size. Below, we show a parallel result for pseudo mirror descent.

Theorem 3. *Suppose that Assumptions 1 and 2 hold, and that there exists a universal constant $c_1 > 0$ such that for all $x^{(k)}$ satisfying $f(x^{(k)}) \neq f^*$,*

$$\mathbb{E}[\langle \nabla f_\Phi(\nabla \Phi(x^{(k-1)})), \mathbb{E}[g^{(k)} | \mathcal{F}^{(k-1)}]\rangle] \\ \geq c_1 \mathbb{E}\|\nabla f_\Phi(\nabla \Phi(x^{(k)}))\|_2^2. \quad (4)$$

In addition, suppose there exists constants c_2 and c_3 such that $\mathbb{E}[\|g^{(k)}\|_\infty^2] \leq c_2^2 + c_3^2 \mathbb{E}[\langle \nabla f_\Phi(\nabla \Phi(x^{(k-1)})), \mathbb{E}[g^{(k)} | \mathcal{F}^{(k-1)}]\rangle]$, and that there exists a constant λ_1 such that $\sup_x(\lambda_{\max}(\nabla^2 \Phi(x))) \leq \lambda_1$, where $\lambda_{\max}$ is the largest eigenvalue of $\nabla^2 \Phi(x)$. Under these assumptions, we have

- *Constant step size: choosing $\eta_k \equiv \eta < \min\{\lambda_1^2/(2\gamma c_1), 2M^{-1}\mu c_3^{-2}\}$, we have*

  $$\mathbb{E}[f(x^{(k)}) - f^*] \leq C_0^k [f(x^{(0)}) - f^*] + \frac{M\mu^{-1}\eta^2}{2} c_2^2,$$

  where $C_0 = 1 - 2\gamma c_1 \lambda_1^{-2}\left(\eta - \frac{M\mu^{-1}\eta^2}{2} c_3^2\right)$.

- *Decreasing step size: choosing $\eta_k = \min\{(2k+1)/[\gamma c_1 \lambda_1^{-2}(k+1)^2], M^{-1}\mu c_3^{-2}\}$, we have, for $k \geq M c_3^2 \lambda_1^2/(\gamma c_1 \mu)$,*

  $$\mathbb{E}[f(x^{(k)}) - f^*] \leq \frac{M\mu^{-1}c_2^2}{2\gamma^2 c_1^2 \lambda_1^{-4} k}.$$

4. Numerical Experiment

4.1. Synthetic Dataset

Simulation setup. We set $x^*(t) = \exp(-t)$, and simulated $n = 10^4$ trajectories, each containing a set of arrival times

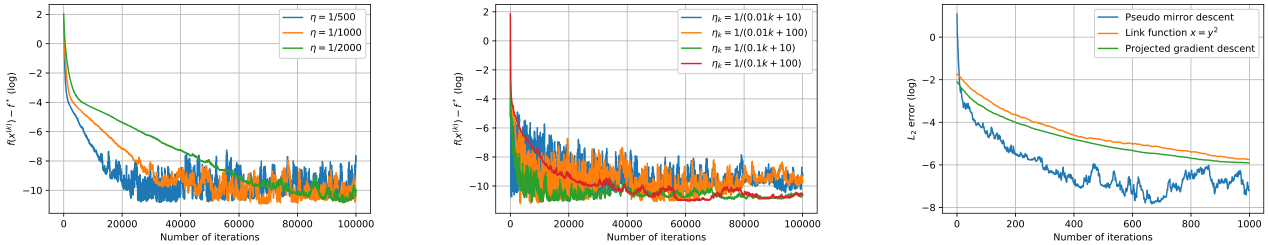

Figure 1. Convergence of objective value for pseudo mirror descent under constant (left) and vanishing (middle) step sizes, as well as the $\mathcal{L}_2$ error of the estimate (right).

in $[0, 1]$. When executing Algorithm 1, the pseudo gradient was computed from a set of 10 randomly chosen trajectories out of the entire pool. The number of iterations was set to 10^5, and the initialization was $x^{(0)}(t) \equiv 10$.

Results. We plot $\log(f(x^{(k)}) - f^*)$ versus the iteration number k in Figure 1. The left-hand side, which uses constant step sizes, shows almost linear convergence at the beginning; the subplot in the middle, which uses diminishing step size, shows sublinear convergence. We also plot the $\mathcal{L}_2$-error of the estimates generated by the pseudo mirror descent, projected gradient descent, and the link function approaches, over the first 1000 iterations, shown on the right-hand side of the figure. The hyper parameters of all algorithms were fine tuned to optimize their respective performances. By comparison, pseudo mirror descent generates superior performance over the benchmarks.

4.2. Real Dataset: Learning London Traffic Accidents

Experiment setup. We tested the performance of pseudo mirror descent on a dataset that contains the traffic accidents between 2005 and 2007 in London. The dataset is available online, [1]and contains more than 60000 entries. Each entry records a traffic accident, including its time of happening, and other information. For our purpose, we modeled the time of occurrence of each accident as random arrivals from a Poisson process, and applied the pseudo mirror descent, and the aforementioned benchmark algorithms to estimate the underlying intensity. For all methods, we set the initialization to be a constant function with value 1 (roughly 1 accident per day), and the step sizes were fine tuned. At each iteration, the gradient was evaluated randomly drawing 10 entries as a minibatch, and using their arrival times to compute the pseudo-gradient $g(t)$ as specified in the previous section. The positive definite kernel used in pseudo-gradient computation was a Gaussian kernel with bandwidth 0.1. The maximum number of iterations was set to 10^4.

Results. The resolution of the arrival time of the data is down to the minute level, and the projected gradient descent method requires computing a Grammian matrix of more than a million entry, which takes more than an hour to compute. By comparison, both the link function and the pseudo mirror descent approaches generated results that resemble the shape of the histogram of the dataset. Both algorithms took less than 1 minute.

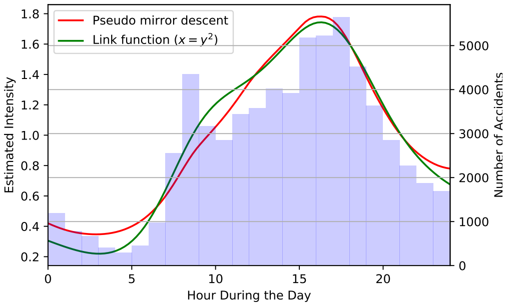

Figure 2. Estimated intensity of a Poisson process modeling the time of occurance of traffic accidents in London.

5. Conclusion

In this paper, we proposed a novel algorithm for estimating the intensity function of Poisson processes. The algorithm, named pseudo mirror descent, is both efficient in the sense that it circumvents the need of performing computationally inefficient projections, and possesses theoretical guarantees. This is achieved by using the multiplicative update rule and pseudo-gradients, which guarantee updates that are both positive and smooth. Simulation on synthetic and real datasets provides numerical proof on the algorithm's convergence, and demonstrates its superior performance in practice.

[1] Dataset available at: https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales

References

Bacry, E., Mastromatteo, I., and Muzy, J.-F. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1 (01):1550005, 2015.

Bagnell, J. A. and Farahmand, A.-m. Learning positive functions in a Hilbert space. 2015.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Flaxman, S., Teh, Y. W., Sejdinovic, D., et al. Poisson intensity estimation with reproducing kernels. *Electronic Journal of Statistics*, 11(2):5081–5104, 2017.

Ge, W. and Collins, R. T. Marked point processes for crowd counting. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2913–2920. IEEE, 2009.

Hawkes, A. G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

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

Koenker, R., Mizera, I., and Yoon, J. What do kernel density estimators optimize? *Journal of Econometric Methods*, 1 (1):15–22.

Nemirovski, A. S. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley, New York, 1983.

Poljak, B. and Tsypkin, Y. Z. Pseudogradient adaptation and training algorithms. *Automation and Remote Control*, 34:45–67, 1973.

Polyak, B. T. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

Quinn, C. J., Coleman, T. P., Kiyavash, N., and Hatsopoulos, N. G. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *Journal of computational neuroscience*, 30(1):17–44, 2011.

Sarfraz, M., Hussain, M. Z., and Nisar, A. Positive data modeling using spline function. *Applied Mathematics and Computation*, 216(7):2036–2049, 2010.

Wahba, G. *Spline models for observational data*, volume 59. SIAM, 1990.

Yang, Y., Etesami, J., He, N., and Kiyavash, N. Online learning for multivariate Hawkes processes. *Advances in Neural Information Processing Systems (NIPS)*, pp. 4937–4946, 2017.

Yuan, M., Cai, T. T., et al. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

Zhou, K., Zha, H., and Song, L. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*, pp. 641–649, 2013.

Appendix

A. Proof of Theorem 1 and its Remark

A.1. Proving Theorem 1

The proof follows the procedure of (Poljak & Tsypkin, 1973). The key to the proof is to show that $f(x^{(k)}) - f^*$ is a semimartingale, and that its limit exists almost surely. Below are the details of the proof.

Since f is M-smooth and Φ is μ-strongly-convex with respect to norm $\|\cdot\|_1$, f_Φ is $M\mu^{-1}$-smooth with respect to $\|\cdot\|_\infty$:

$$\|\nabla f(\nabla \Phi^*(x+y)) - \nabla f(\nabla\Phi^*(x))\|_2 \le M\|\nabla\Phi^*(x+y) - \nabla\Phi^*(x)\|_2 \le M\mu^{-1}\|y\|_\infty,$$

where the two steps use smoothness of f and the smoothness of Φ^*, guaranteed by the strong convexity of Φ. Hence,

$$|f_\Phi(x+y) - f_\Phi(x) - \langle \nabla f_\Phi(x), y\rangle| \le \frac{M\mu^{-1}}{2}\|y\|_\infty^2.$$

Let $x = \nabla\Phi(x^{(k-1)})$ and $y = -\eta_{k-1}g^{(k)}$, we have

$$|f(x^{(k)}) - f(x^{(k-1)}) + \langle \nabla f_\Phi(\nabla\Phi(x^{(k-1)})), \eta_{k-1}g^{(k)}\rangle| \le \frac{M\mu^{-1}\eta_{k-1}^2}{2}\|g^{(k)}\|_\infty^2,$$

which further implies

$$f(x^{(k)}) \le f(x^{(k-1)}) - \eta_{k-1}\langle \nabla f_\Phi(\nabla\Phi(x^{(k-1)})), g^{(k)}\rangle + \frac{M\mu^{-1}\eta_{k-1}^2}{2}\|g^{(k)}\|_\infty^2. \tag{5}$$

Taking conditional expectations on both sides of (5),

$$\begin{aligned}
\mathbb{E}[f(x^{(k)}) - f^*|\mathcal{F}^{(k-1)}] &\le (f^{(k-1)} - f^*) - \eta_{k-1}\langle \nabla f_\Phi(\nabla\Phi(x^{(k-1)})), \mathbb{E}[g^{(k)}|\mathcal{F}^{(k-1)}]\rangle + \\
&+ \frac{M\mu^{-1}\eta_{k-1}^2}{2}\mathbb{E}[\|g^{(k)}\|_\infty^2|\mathcal{F}^{(k-1)}] \\
&\le (f^{(k-1)} - f^*)\left(1 + \frac{K_1 M\mu^{-1}\eta_{k-1}^2}{2}\right) - \\
&- \eta_{k-1}\langle \nabla f_\Phi(\nabla\Phi(x^{(k-1)})), \mathbb{E}[g^{(k)}|\mathcal{F}^{(k-1)}]\rangle\left(1 - \frac{M\mu^{-1}K_2}{2}\eta_{k-1}\right) + \\
&+ \frac{M\mu^{-1}\lambda_k\eta_{k-1}^2}{2} + \frac{M\mu^{-1}\eta_{k-1}^2K_1}{2}f^* \\
&\le (f^{(k-1)} - f^*)\left(1 + \frac{K_1M\mu^{-1}}{2}\eta_{k-1}^2\right) + \frac{M\mu^{-1}\lambda_k\eta_{k-1}^2}{2} + \\
&+ \frac{M\mu^{-1}\eta_{k-1}^2K_1}{2}f^*,
\end{aligned} \tag{6}$$

where the last step holds for sufficiently large k. Let

$$\begin{aligned}
z^{(k)} = (f(x^{(k)}) - f^*)\prod_{\kappa=k}^{\infty}\left(1 + \frac{K_1M\mu^{-1}\eta_\kappa^2}{2}\right) + \\
+ \sum_{\kappa=k}\left(\frac{M\mu^{-1}\lambda_\kappa\eta_k^2}{2} + \frac{M\mu^{-1}\eta_\kappa^2K_1}{2}f^*\right)\prod_{m=\kappa+1}^{\infty}\left(1 + \frac{K_1M\mu^{-1}\eta_m^2}{2}\right).
\end{aligned}$$

Then, we immediately have, upon substituting $z^{(k)}$ into (6),

$$\mathbb{E}[z^{(k)}|\mathcal{F}^{(k-1)}] \leq z^{(k-1)}.$$

Since for any sequence $z^{(k)}$, $f(x^{(k)}) - f^*$ can be uniquely determined (nothing else is random), we can take conditional expectations on both sides of (6) with respect to $z^{(1)}, \ldots, z^{(k-1)}$, and obtain

$$\mathbb{E}[z^{(k)}|z^{(1)}, \ldots, z^{(k-1)}] \leq z^{(k-1)}.$$

This shows that $z^{(k)}$ is a semimartingale, and that $\mathbb{E}z^{(k)} \leq \cdots \leq \mathbb{E}z^{(1)} < \infty$. This implies $\lim_{k\to\infty} z^{(k)}$ exists almost surely, and hence, $\lim_{k\to\infty}(f(x^{(k)}) - f^*)$ exists almost surely, and that $\mathbb{E}[f(x^{(k)}) - f^*]$ are uniformly upper bounded.

Taking unconditional expectations on both sides of (6), we now have

$$\begin{aligned}\mathbb{E}[f(x^{(k)}) - f^*] \leq \mathbb{E}[f(x^{(k-1)}) - f^*]&\left(1 + \frac{K_1 M\mu^{-1}\eta_{k-1}^2}{2}\right) + \frac{M\mu^{-1}\lambda_k\eta_{k-1}^2}{2} - \\ &- \eta_{k-1}\mathbb{E}[\langle \nabla f_\Phi(\nabla\Phi(x^{(k-1)})), \mathbb{E}[g^{(k)}|\mathcal{F}^{(k-1)}]\rangle]\left(1 - \frac{M\mu^{-1}K_2}{2}\eta_{k-1}\right) + \\ &+ \frac{M\mu^{-1}\eta_{k-1}^2 K_1}{2} f^*.\end{aligned}$$

For sufficiently large k, $2 - M\mu^{-1}K_2\eta_k > 0$. Hence, summing both sides from $k = 1$ to ∞, we get

$$\begin{aligned}\mathbb{E}[f(x^{(0)}) - f^*] + \sum_{k=0}^{\infty} \frac{K_1 M\mu^{-1}\eta_k^2}{2}\mathbb{E}[f(x^{(k)}) - f^*] + \sum_{k=0}^{\infty}\left(\frac{M\mu^{-1}\lambda_{k+1}\eta_k^2}{2} + \frac{M\mu^{-1}\eta_k^2 K_1}{2} f^*\right) \geq \\ \sum_{k=0}^{\infty}\eta_k \mathbb{E}[\langle \nabla f_\Phi(\nabla\Phi(x^{(k)})), \mathbb{E}[g^{(k+1)}|\mathcal{F}^{(k)}]\rangle]\left(1 - \frac{M\mu^{-1}K_2}{2}\eta_k\right)\end{aligned}$$

Recall that, for the left-hand side, we have shown that $\mathbb{E}[f(x^{(k)}) - f^*]$ is uniformly bounded, and that, by assumption, $\sum_{k=0}^{\infty}\eta_k^2\lambda_k$ and $\sum_{k=0}^{\infty}\eta_k^2$ are finite. Hence, the left-hand side of the above inequality is finite. In other words,

$$\sum_{k=0}^{\infty}\eta_k \mathbb{E}[\langle \nabla f_\Phi(\nabla\Phi(x^{(k)})), \mathbb{E}[g^{(k+1)}|\mathcal{F}^{(k)}]\rangle]\left(1 - \frac{M\mu^{-1}K_2}{2}\eta_k\right) < \infty.$$

On the other side, we also have $\sum_{k=0}^{\infty}\eta_k = \infty$, $\langle \nabla f_\Phi(\nabla\Phi(x^{(k)})), \mathbb{E}[g^{(k+1)}|\mathcal{F}^{(k)}]\rangle \geq 0$, while for sufficiently large k, $1 - \eta_k M\mu^{-1}K_2/2 \geq \varepsilon > 0$ for some small constant ε. Therefore, there exists a subsequence k_i such that $\langle \nabla f_\Phi(\nabla\Phi(x^{(k)})), \mathbb{E}[g^{(k+1)}|\mathcal{F}^{(k)}]\rangle$ converges in distribution:

$$\lim_{i\to\infty}\mathbb{E}[\langle \nabla f_\Phi(\nabla\Phi(x^{(k_i-1)})), \mathbb{E}[g^{(k_i)}|\mathcal{F}^{(k_i-1)}]\rangle] = 0.$$

Since the sequence converges in distribution to a constant, it also converges in probability, which further implies almost sure convergence of a subsequence:

$$\lim_{j\to\infty}\langle \nabla f_\Phi(\nabla\Phi(x^{(k_{i_j}-1)})), \mathbb{E}[g^{(k_{i_j})}|\mathcal{F}^{(k_{i_j}-1)}]\rangle = 0$$

almost surely. This implies the final result.

A.2. Proving the remark

By the last step in proof in previous part, we have

$$\lim_{k\to\infty}\|\nabla f_\Phi(\nabla\Phi(x^{(k)}))\|_2^2 = 0$$

in probability. By chain rule and boundedness of eigenvalues of $\nabla^2\Phi$, this further implies

$$\lim_{k\to\infty}\|\nabla f(x^{(k)})\|_2^2 = 0$$

in probability.

B. Proving f satisfies the Polyak-Łojasiewicz condition

First notice that

$$\|\nabla f(x)\|_2^2 = \int_0^1 \left(1 - \frac{x^*}{x}(t)\right)^2 \mathrm{d}t,$$

and

$$2\nu(f(x) - f^*) = 2\nu \cdot \left[\int_0^1 (x - x^*)(t)\mathrm{d}t - \int_0^1 x^*(t) \log \frac{x}{x^*}(t)\mathrm{d}t\right].$$

We wish to prove that

$$\int_0^1 \left(1 - \frac{x^*}{x}(t)\right)^2 \mathrm{d}t \geq 2\nu \int_0^1 \left(x - x^* - x^* \log \frac{x}{x^*}\right)(t)\mathrm{d}t.$$

Notice that the left-hand side resembles the form of a χ^2 divergence, whereas the right-hand side resembles the form of a Kullback-Leibler divergence. In fact, when $\sup_{t\in[0,1]} x(t) \leq (2\nu)^{-1}$, we have

$$\begin{aligned}
2\nu \cdot \left[\int_0^1 (x - x^*)(t)\mathrm{d}t - \int_0^1 x^*(t) \log \frac{x}{x^*}(t)\mathrm{d}t\right] &= 2\nu \cdot \left[\int_0^1 (x - x^*)(t)\mathrm{d}t + \int_0^1 x^*(t) \log \frac{x^*}{x}(t)\mathrm{d}t\right] \\
&\leq 2\nu \cdot \left[\int_0^1 (x - x^*)(t)\mathrm{d}t + \int_0^1 x^*(t) \left(\frac{x^*}{x}(t) - 1\right) \mathrm{d}t\right] \\
&= 2\nu \int_0^1 x(t) \left(\frac{x^*}{x}(t) - 1\right)^2 \mathrm{d}t \\
&\leq \int_0^1 \left(\frac{x^*}{x}(t) - 1\right)^2 (t)\mathrm{d}t.
\end{aligned}$$