A Further Examples of LRM Schemes

Example 5. The classic *Stochastic Gradient Langevin Dynamics* [57] iterates as

$$x_{k+1} = x_k - \gamma_{k+1} \widetilde{\nabla} f(x_k) + \sqrt{2\gamma_{k+1}}\, \xi_{k+1}, \tag{SGLD}$$

where $\widetilde{\nabla} f$ is the gradient of the negative log-likelihood of a random batch of the data. (SGLD) fits the LRM template by setting $U_{k+1} \coloneqq \widetilde{\nabla} f(x_k) - \nabla f(x_k)$, and $b_{k+1} \coloneqq 0$. □

Example 6. The *Proximal Langevin Algorithm* [9, 46, 59] is defined via

$$x_{k+1} = x_k - \gamma_{k+1} \nabla f(x_{k+1}) + \sqrt{2\gamma_{k+1}}\, \xi_{k+1}. \tag{PLA}$$

This algorithm is implicit, and it is assumed that one can solve (PLA) for x_{k+1}. By setting $b_{k+1} \coloneqq \nabla f(x_{k+1}) - \nabla f(x_k)$ and $U_{k+1} \coloneqq 0$, we see that this algorithm also follows the LRM template. □

B Additional Related Work

Our paper studies the behavior of a wide range of Langevin-based sampling algorithms proposed in the literature in the asymptotic setting under minimal assumptions. This allows us to give last-iterate guarantees in Wasserstein distance. As stressed in Section 1, our goal is *not* to provide non-asymptotic rates in this general setting as the problem is inherently NP-Hard. However, given more assumptions and structures on the potential f, there is a plethora of works which prove convergence rates for the last iterates in Wasserstein distance. In this appendix, we provide additional backgraound for these works and the methods used in the literature.

A powerful framework for quantifying the global discretization error of a numerical algorithm is the mean-square analysis framework [40]. This framework furnishes a general recipe for controlling short and long-term integration errors. For sampling, this framework has been applied to prove convergence rates for Langevin Monte-Carlo (the Euler-Maruyama discretization of (LD)) in the strongly-convex setting [32, 34]. Similar to our work, the convergence obtained in these works is last-iterate and in Wasserstein distance. One of the essential ingredients in the latter work is the contraction property of the SDE, which is ensured by the strong convexity assumption. This, in turn, implies strong non-asymptotic convergence guarantees.

It is an interesting future direction to study the combination of the Mean-Squared analysis together with the Picard process and its applicability to more sophisticated algorithms (such as LRM schemes with bias and noise), as well as non-convex potentials.

As explained in Section 3, one of the main themes in proving error bounds for sampling is the natural relation between sampling and optimization in the Wasserstein space. This point of view, when applied to strongly-convex potentials, has produced numerous non-asymptotic guarantees; see [14, 18] for a recent account and the references therein. Note that strong convexity is crucial for the analysis used in the aforementioned work. Moreover, the error bounds for biased and noisy discretizations do *not* decrease with the step-size or iteration count; see [18, Theorem 4, Eqn. (14)]. This means that while the bound is non-asymptotic, it does not automatically result in an asymptotic convergence. Finally, we stress that these approaches are orthogonal to our techniques: We view a sampling algorithm as a (noisy and biased) discretization of a dynamical system (and not necessarily a gradient flow), and use tools from dynamical system theory to provide asymptotic convergence results.

C Proofs for Section 4

C.1 Proof of Theorem 1

In this appendix, we bring the detailed proof of Theorem 1. Recall that we interpolate the iterates of the LRM scheme $\{x_k\}$ as

$$X_t = x_k + (t - \tau_k)\{v(x_k) + \mathbb{E}[Z_{k+1} \mid \mathscr{F}_t]\} + \sigma(x_k)\,(B_t - B_{\tau_k}). \tag{3}$$

Moreover, for a fixed $t > 0$, we considered the Brownian motion $B_s^{(t)} = B_{t+s} - B_t$, and constructed two important processes: the Langevin flow defined via

$$\mathrm{d}\Phi_s^{(t)} = v(\Phi_s^{(t)})\,\mathrm{d}s + \sigma(\Phi_s^{(t)})\,\mathrm{d}B_s^{(t)}, \quad \Phi_0^{(t)} = X_t, \tag{12}$$

and the Picard process (6) constructed as

$$Y_s^{(t)} = X_t + \int_0^s v(X_{t+u})\,\mathrm{d}u + \int_0^s \sigma(X_{t+u})\,\mathrm{d}B_u^{(t)}. \tag{6}$$

Let us fix $T > 0$, and for $s \in [0, T]$ decompose the distance between the interpolation and the Langevin flow as

$$\tfrac{1}{2}\|X_{t+s} - \Phi_s^{(t)}\|^2 \le \|Y_s^{(t)} - \Phi_s^{(t)}\|^2 + \|X_{t+s} - Y_s^{(t)}\|^2, \tag{7}$$

where we have used $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$. We now bound each term of this decomposition. Notice that due to the synchronous coupling of the processes, the Brownian motion cancels out in the differences.

The first term controls how close the Picard process is to the Langevin flow, and is bounded in the following lemma.

Lemma 3. *For fixed $t, T > 0$ and $0 \le s \le T$, the distance of the Picard process and the Langevin flow is bounded as*

$$\|Y_s^{(t)} - \Phi_s^{(t)}\|^2 \le 2(T+1)L^2 \int_0^s \|\Phi_u^{(t)} - X_{t+u}\|^2\,\mathrm{d}u.$$

Proof. By the auxiliary Lemma 4 below, Lipschitzness of v, σ, Itô isometry (see, e.g., [62]) and $s \le T$, we have

$$\begin{aligned}
\mathbb{E}\|Y_s^{(t)} - \Phi_s^{(t)}\|^2 &= \mathbb{E}\left\|\int_0^s v(\Phi_u^{(t)}) - v(X_{t+u})\,\mathrm{d}u + \int_0^s \sigma(\Phi_u^{(t)}) - \sigma(X_{t+u})\,\mathrm{d}B_u^{(t)}\right\|^2 \\
&\le 2s\int_0^s \mathbb{E}\left\|v(\Phi_u^{(t)}) - v(X_{t+u})\right\|^2 \mathrm{d}u + 2\mathbb{E}\int_0^s \left\|\sigma(X_{t+u}) - \sigma(\Phi_u^{(t)})\right\|_F^2 \mathrm{d}u \\
&\le 2(T+1)L^2 \int_0^s \mathbb{E}\|\Phi_u^{(t)} - X_{t+u}\|^2\,\mathrm{d}u.
\end{aligned}$$

■

For the rest of the proof, we need to define the continuous-time piecewise-constant processes $\overline{X}(\tau_k + s) = X_k$, $\overline{\gamma}(\tau_k + s) = \gamma_{k+1}$, $\overline{Z}(\tau_k + s) = Z_{k+1}$, and $Z(\tau_k + s) = \mathbb{E}[Z_{k+1} \mid \mathscr{F}_{\tau_k + s}]$, for $0 \le s < \gamma_{k+1}$. Also, let $m(t) = \sup\{k \ge 0 : \tau_k \le t\}$ so that $\tau_{m(t)} \le t < \tau_{m(t)+1}$.

To bound the second term in (7), we have seen that

$$\begin{aligned}
X_{t+s} - Y_s^{(t)} = &\int_t^{t+s} v(\overline{X}(u))\,\mathrm{d}u - \int_0^s v(X_{t+u})\,\mathrm{d}u \\
&+ \int_t^{t+s} \sigma(\overline{X}(u))\,\mathrm{d}B_u - \int_0^s \sigma(X_{t+u})\,\mathrm{d}B_u^{(t)} \\
&+ \Delta_Z(t, s),
\end{aligned}$$

where $\Delta_Z(t, s)$ plays the role of accumulated noise and bias from time t to $t + s$, and is defined as

$$\Delta_Z(t, s) \coloneqq \sum_{i=n}^{k-1} \gamma_{i+1} Z_{i+1} + (t + s - \tau_k)\mathbb{E}[Z_{k+1} \mid \mathscr{F}_{t+s}] - (t - \tau_n)\mathbb{E}[Z_{n+1} \mid \mathscr{F}_t], \tag{13}$$

with $k = m(t + s)$ and $n = m(t)$. We therefore have

$$\begin{aligned}
\mathbb{E}\|X_{t+s} - Y_s^{(t)}\|^2 &\le 3\mathbb{E}\left\|\int_t^{t+s} v(X_u) - v(\overline{X}(u))\,\mathrm{d}u\right\|^2 \\
&\quad + 3\mathbb{E}\left\|\int_t^{t+s} \sigma(X_u) - \sigma(\overline{X}(u))\,\mathrm{d}B_u\right\|^2 + 3\mathbb{E}\|\Delta_Z(t, s)\|^2 \\
&\le 3s\int_t^{t+s} \mathbb{E}\left\|v(X_u) - v(\overline{X}(u))\right\|^2 \mathrm{d}u \\
&\quad + 3\mathbb{E}\int_t^{t+s} \left\|\sigma(X_u) - \sigma(\overline{X}(u))\right\|_F^2 \mathrm{d}u + 3\mathbb{E}\|\Delta_Z(t, s)\|^2 \\
&\le 3(s+1)L^2 \int_t^{t+s} \mathbb{E}\|X_u - \overline{X}(u)\|^2\,\mathrm{d}u + 3\mathbb{E}\|\Delta_Z(t, s)\|^2.
\end{aligned} \tag{14}$$

For bounding the term inside the integral, we have

$$\begin{aligned}\mathbb{E}\|X_u - \overline{X}(u)\|^2 &= \mathbb{E}\|(u-\tau_{m(u)})\{v(\overline{X}(u)) + Z(u)\} + \sigma(\overline{X}(u))\,(B_u - B_{\tau_{m(u)}})\|^2 \\ &\le 4\overline{\gamma}(u)^2\Big(\mathbb{E}\|v(\overline{X}(u))\|^2 + \mathbb{E}\|Z(u)\|^2\Big) + 2\overline{\gamma}(u)\,\mathbb{E}\,\mathrm{tr}\Big(\sigma(\overline{X}(u))^\top \sigma(\overline{X}(u))\Big).\end{aligned}$$

We have used the fact that

$$\begin{aligned}\mathbb{E}\|\sigma(\overline{X}(u))\,(B_u - B_{\tau_{m(u)}})\|^2 &= \mathbb{E}\Big((B_u - B_{\tau_{m(u)}})^\top \sigma(\overline{X}(u))^\top \sigma(\overline{X}(u))\,(B_u - B_{\tau_{m(u)}})\Big) \\ &= \mathbb{E}\mathrm{tr}\Big(\sigma(\overline{X}(u))^\top \sigma(\overline{X}(u))\,(B_u - B_{\tau_{m(u)}})(B_u - B_{\tau_{m(u)}})^\top\Big) \\ &= \mathbb{E}\Big[\mathbb{E}[\mathrm{tr}(\sigma(\overline{X}(u))^\top \sigma(\overline{X}(u))\,(B_u - B_{\tau_{m(u)}})(B_u - B_{\tau_{m(u)}})^\top) \mid \mathscr{F}_{\tau_{m(u)}}]\Big] \\ &= (u-\tau_{m(u)})\mathbb{E}\Big[\mathrm{tr}(\sigma(\overline{X}(u))^\top \sigma(\overline{X}(u)))\Big]\end{aligned}$$

Notice that since conditional expectation is a projection in L^2, we have $\mathbb{E}\|Z(u)\|^2 \le \mathbb{E}\|\overline{Z}(u)\|^2$. Using this fact, along with boundedness of $\sigma(\cdot)$ by C_σ, and Lemma 2 we get

$$\begin{aligned}\mathbb{E}\Big[\|X_u - \overline{X}(u)\|^2\Big] &\le 4\overline{\gamma}(u)^2\Big(\mathbb{E}\|v(\overline{X}(u))\|^2 + \mathbb{E}\|\overline{Z}(u)\|^2\Big) + 2\overline{\gamma}(u)\,\mathbb{E}\,\mathrm{tr}\Big(\sigma(\overline{X}(u))^\top \sigma(\overline{X}(u))\Big) \\ &\le 4\overline{\gamma}(u)^2\mathbb{E}\|v(\overline{X}(u))\|^2 + 8\overline{\gamma}(u)^2\sigma^2 + 4\overline{\gamma}(u)^2\,O(\overline{\gamma}(u)) + 2C_\sigma\overline{\gamma}(u) \le C\overline{\gamma}(u),\end{aligned}$$

for some constant $C > 0$. Plugging this estimate into (14) after taking expectation yields

$$\begin{aligned}\mathbb{E}\Big[\|X_{t+s} - Y_s^{(t)}\|^2\Big] &\le 3(s+1)L^2C\int_t^{t+s}\overline{\gamma}(u)\,\mathrm{d}u + 3\mathbb{E}\|\Delta_Z(t,s)\|^2 \\ &\le 3(s+1)sL^2C\sup_{u\in[t,t+s]}\overline{\gamma}(u) + 3\mathbb{E}\|\Delta_Z(t,s)\|^2 \\ &\le 3(T+1)^2L^2C\sup_{u\in[t,t+T]}\overline{\gamma}(u) + 3\sup_{u\in[0,T]}\mathbb{E}\|\Delta_Z(t,u)\|^2\end{aligned}$$

Taking supremum over $s \in [0,T]$ and noticing that the right-hand-side is independent of s and $\gamma_k \to 0$, together with Lemma 1 yields

$$\begin{aligned}A_t &:= \sup_{0\le s\le T}\mathbb{E}\Big[\|X_{t+s} - Y_s^{(t)}\|^2\Big] \\ &\le 3(T+1)^2L^2C\sup_{t\le u\le t+T}\overline{\gamma}(u) + 3\sup_{0\le u\le T}\mathbb{E}\big[\|\Delta_Z(t,u)\|^2\big] \\ &\to 0 \quad \text{as } t\to\infty,\end{aligned} \tag{15}$$

showing that the Picard process gets arbitrary close to the original interpolation, as $t \to \infty$.

Let us return to the decomposition (7). By taking expectation and using (8) and (15) we obtain

$$\begin{aligned}\mathbb{E}\Big[\|X_{t+s} - \Phi_s^{(t)}\|^2\Big] &\le 2(T+1)L^2\int_0^s \mathbb{E}\Big[\|X_{t+u} - \Phi_u^{(t)}\|^2\Big]\,\mathrm{d}u + 2A_t \\ &\le 2A_t\exp\Big(s(T+1)L^2\Big) \\ &\le 2A_t\exp((T+1)^2L^2),\end{aligned}$$

where in the last line we have used the Grönwall lemma. Thus,

$$\lim_{t\to\infty}\sup_{s\in[0,T]}\mathbb{E}\Big[\|X_{t+s} - \Phi_s^{(t)}\|^2\Big] = 0.$$

Recall that the Wasserstein distance between X_{t+s} and $\Phi_s^{(t)}$ is the infimum over all possible couplings between them, having the correct marginals. As $\Phi_s^{(t)}$ has the same marginal as the Langevin diffusion started from X_t at time s, and the synchronous coupling of the interpolation and the Langevin flow produces a specific coupling between them, we directly get

$$W_2(X_{t+s}, \Phi_s^{(t)}) \le \mathbb{E}\Big[\|X_{t+s} - \Phi_s^{(t)}\|^2\Big]^{\frac{1}{2}},$$

which implies

$$\lim_{t\to\infty}\sup_{s\in[0,T]} W_2(X_{t+s}, \Phi_s^{(t)}) = 0,$$

as desired. ■

C.2 Auxiliary Lemmas

Lemma 1. *Suppose Assumptions 1–3 hold. Then, for any fixed $T > 0$ we have*

$$\lim_{t\to\infty} \sup_{0\le s\le T} \mathbb{E}\|\Delta_Z(t,s)\|^2 = 0.$$

Proof. Define Δ_b and Δ_U the same way as in (13). By Cauchy-Schwarz we have

$$\begin{aligned}
&\|\Delta_b(t,s)\|^2 \\
&\le \left(\sum_{i=n}^{k-1} \gamma_{i+1}\|b_{i+1}\| + (t+s-\tau_k)\|\mathbb{E}[b_{k+1} \mid \mathscr{F}_{t+s}]\| + (t-\tau_n)\|\mathbb{E}[b_{n+1} \mid \mathscr{F}_t]\|\right)^2 \\
&\le (2\gamma_{n+1}+s)\left(\sum_{i=n}^{k-1} \gamma_{i+1}\|b_{i+1}\|^2 + (t+s-\tau_k)\|\mathbb{E}[b_{k+1} \mid \mathscr{F}_{t+s}]\|^2 + (t-\tau_n)\|\mathbb{E}[b_{n+1} \mid \mathscr{F}_t]\|^2\right),
\end{aligned}$$

where the last inequality comes from $\sum_{i=n}^{k-1}\gamma_{i+1} \le s$, $t+s-\tau_k \le \gamma_{k+1}$, $t-\tau_n \le \gamma_{n+1}$, and $\gamma_{k+1} \le \gamma_{n+1}$. Noticing that conditional expectation is a contraction in L^2 and letting $k' = m(t+T)$, we get

$$\sup_{0\le s\le T} \mathbb{E}\left[\|\Delta_b(t,s)\|^2\right] \le (2+T)\left(\sum_{i=n}^{k'-1} \gamma_{i+1}\mathbb{E}\|b_{i+1}\|^2 + \sup_{n\le j\le k'+1} \gamma_{j+1}\mathbb{E}\|b_{j+1}\|^2 + \gamma_{n+1}\mathbb{E}\|b_{n+1}\|^2\right)$$

Now, invoking Lemma 2 yields

$$\begin{aligned}
\sup_{0\le s\le T} \mathbb{E}\left[\|\Delta_b(t,s)\|^2\right] &\le C(2+T)\left(\sum_{i=n}^{k'-1} \gamma_{i+1}^2 + \sup_{n\le j\le k'+1} \gamma_{j+1}^2 + \gamma_{n+1}^2\right) \\
&\le C(2+T)\left(\sum_{i=n}^{k'-1} \gamma_{i+1}^2 + 2\gamma_{n+1}^2\right) \\
&\le C(2+T)(T+2\gamma_{n+1}) \sup_{0\le s\le T} \overline{\gamma}(t+s).
\end{aligned}$$

As $t \to \infty$, the last quantity vanishes, since $\gamma_n \to 0$.

For the noise we have

$$\begin{aligned}
\|\Delta_U(t,s)\|^2 &\le 2\left\|\sum_{i=n}^{k-1} \gamma_{i+1}U_{i+1}\right\|^2 + 4\|(t+s-\tau_k)\mathbb{E}[U_{k+1} \mid \mathscr{F}_{t+s}]\|^2 + 4\|(t-\tau_n)\mathbb{E}[U_{n+1} \mid \mathscr{F}_t]\|^2 \\
&\le 2\left\|\sum_{i=n}^{k-1} \gamma_{i+1}U_{i+1}\right\|^2 + 4\gamma_{k+1}^2\|U_{k+1}\|^2 + 4\gamma_{n+1}^2\|U_{n+1}\|^2.
\end{aligned}$$

Taking expectations and then sup, we get

$$\sup_{0\le s\le T} \mathbb{E}\left[\|\Delta_U(t,s)\|^2\right] \le 2 \sup_{n+1\le k\le m(t+T)} \mathbb{E}\left\|\sum_{i=n}^{k-1} \gamma_{i+1}U_{i+1}\right\|^2 + 4\gamma_{k+1}^2\sigma^2 + 4\gamma_{n+1}^2\sigma^2.$$

Since $\{U_i\}$ is a martingale difference sequence, we have that $\left\{\sum_{i=n}^{k-1}\gamma_{i+1}U_{i+1}\right\}_{k>n}$ is a martingale. Thus, by the boundedness of the second moments of U_i, we get

$$\mathbb{E}\left\|\sum_{i=n}^{k-1} \gamma_{i+1}U_{i+1}\right\|^2 = \sum_{i=n}^{k-1} \gamma_{i+1}^2\mathbb{E}\|U_{i+1}\|^2 \le \sigma^2 \sum_{i=n}^{k-1} \gamma_{i+1}^2.$$

Hence,

$$\lim_{n\to\infty} \sup\left\{\mathbb{E}\|\sum_{i=n}^{k-1} \gamma_{i+1}U_{i+1}\|^2 : n < k \le m(\tau_n+T)\right\} \le \lim_{n\to\infty} \sigma^2 \sum_{i=n}^{\infty} \gamma_{i+1}^2 = 0.$$

■

Lemma 2. *Let $\{x_k\}_{k\in\mathbb{N}}$ be the iterates of* (LRM) *and suppose Assumptions 1–3 hold. Then, $\mathbb{E}\|x_k\|^2 = O(1/\gamma_{k+1})$. This in turn implies $\mathbb{E}\|v(x_k)\|^2 = O(1/\gamma_{k+1})$ and $\mathbb{E}\|b_{k+1}\|^2 = O(\gamma_{k+1})$.*

Proof. Without loss of generality, suppose v has a stationary point at 0. We repeatedly use the fact that $\mathbb{E}\|v(x_k)\|^2 \le L^2\mathbb{E}\|x_k\|^2$. Moreover, by Assumption 1 we have $\langle v(x), x\rangle \le C_v(\|x\|+1)$, and $\|\sigma(x)\|_F^2 \le C_\sigma$.

Define $a_k := \mathbb{E}\|x_k\|^2$. We have

$$\begin{aligned} a_{k+1} - a_k &= \gamma_{k+1}^2\mathbb{E}\|v(x_k) + Z_{k+1}\|^2 + \gamma_{k+1}\mathbb{E}\|\sigma(x_k)\xi_{k+1}\|^2 + 2\gamma_{k+1}\mathbb{E}\langle x_k, v(x_k)+Z_{k+1}\rangle \\ &\quad + 2\gamma_{k+1}^{1/2}\mathbb{E}\langle x_k, \sigma(x_k)\xi_{k+1}\rangle + 2\gamma_{k+1}^{3/2}\mathbb{E}\langle v(x_k)+Z_{k+1}, \sigma(x_k)\xi_{k+1}\rangle \\ &\le 2L^2\gamma_{k+1}^2 a_k + 2\gamma_{k+1}^2\mathbb{E}\|Z_{k+1}\|^2 + \gamma_{k+1}C_\sigma + 2\gamma_{k+1}C_v(\sqrt{a_k}+1) + 2\gamma_{k+1}\sqrt{a_k}\sqrt{\mathbb{E}\|Z_{k+1}\|^2} \\ &\quad + 2\gamma_{k+1}^{3/2}\sqrt{C_\sigma}\sqrt{\mathbb{E}\|Z_{k+1}\|^2} \end{aligned} \tag{16}$$

By Assumption 3, there is some $C_b > 0$ such that $\mathbb{E}\|b_{k+1}\|^2 \le C_b(\gamma_{k+1}^2 a_k + \gamma_{k+1})$, and we have

$$\mathbb{E}\|Z_{k+1}\|^2 \le 2\mathbb{E}\|b_{k+1}\|^2 + 2\mathbb{E}\|U_{k+1}\|^2 \le 2C_b(\gamma_{k+1}^2 a_k + \gamma_{k+1}) + 2\sigma^2. \tag{17}$$

Moreover, as $\sqrt{p+q} \le \sqrt{p} + \sqrt{q}$, we have

$$\sqrt{\mathbb{E}\|Z_{k+1}\|^2} \le \sqrt{2C_b}(\gamma_{k+1}\sqrt{a_k} + \sqrt{\gamma_{k+1}}) + \sqrt{2}\sigma. \tag{18}$$

Plugging the bounds from (17) and (18) into (16) gives

$$\begin{aligned} a_{k+1} - a_k &\le 2L^2\gamma_{k+1}^2 a_k + 4C_b\gamma_{k+1}^4 a_k + 4C_b\gamma_{k+1}^3 + 4\gamma_{k+1}^2\sigma^2 \\ &\quad + \gamma_{k+1}C_\sigma + 2\gamma_{k+1}C_v\sqrt{a_k} + 2\gamma_{k+1}C_v \\ &\quad + 2\sqrt{2C_b}\gamma_{k+1}^2 a_k + 2\sqrt{2C_b}\gamma_{k+1}^{3/2}\sqrt{a_k} + 2\sqrt{2}\sigma\gamma_{k+1}\sqrt{a_k} \\ &\quad + 2\sqrt{2C_bC_\sigma}\gamma_{k+1}^{5/2}\sqrt{a_k} + 2\sqrt{2C_bC_\sigma}\gamma_{k+1}^2 + 2\gamma_{k+1}^{3/2}\sqrt{2C_\sigma}\sigma \\ &=: P\gamma_{k+1}^2\, a_k + Q\gamma_{k+1}\sqrt{a_k} + R\gamma_{k+1}, \end{aligned} \tag{19}$$

where

$$\begin{aligned} P &= 2L^2 + 4C_b\gamma_{k+1}^2 + 2\sqrt{2C_b} \\ Q &= 2C_v + 2\sqrt{2C_b}\sqrt{\gamma_{k+1}} + 2\sqrt{2}\sigma + 2\sqrt{2C_b}\gamma_{k+1} + 2\sqrt{2C_bC_\sigma}\gamma_{k+1}^{3/2} \\ R &= 4C_b\gamma_{k+1}^2 + 4\gamma_{k+1}\sigma^2 + C_\sigma + 2C_v + 2\sqrt{2C_bC_\sigma}\gamma_{k+1} + 2\gamma_{k+1}^{1/2}\sqrt{2C_\sigma}\sigma. \end{aligned}$$

The exact values of P, Q, and R are irrelevant, and we only need upper bounds for them. Assuming that $\gamma_{k+1} < 1$ for all k, we replace the three quantities by

$$\begin{aligned} P &= 2L^2 + 4C_b + 2\sqrt{2C_b} \\ Q &= 2C_v + 2\sqrt{2C_b} + 2\sqrt{2}\sigma + 2\sqrt{2C_b} + 2\sqrt{2C_bC_\sigma} \\ R &= 4C_b + 4\sigma^2 + C_\sigma + 2C_v + 2\sqrt{2C_bC_\sigma} + 2\sqrt{2C_\sigma}\sigma. \end{aligned} \tag{20}$$

Now, define $h_k = \gamma_{k+1}^2 a_k$. The recursion (19) in terms of h_k becomes

$$h_{k+1} \le h_k(1 + P\gamma_{k+1}^2)\frac{\gamma_{k+2}^2}{\gamma_{k+1}^2} + \sqrt{h_k}Q\gamma_{k+2}^2 + R\gamma_{k+1}\gamma_{k+2}^2.$$

We now prove that there exists some $M > 0$ so that $h_k \le M\gamma_{k+1}$ by induction. Suppose it is the case for k, and we prove it for $k+1$. Using the induction hypothesis we get

$$\begin{aligned} h_{k+1} &\le M\gamma_{k+1}(1 + P\gamma_{k+1}^2)\frac{\gamma_{k+2}^2}{\gamma_{k+1}^2} + \sqrt{M\gamma_{k+1}}Q\gamma_{k+2}^2 + R\gamma_{k+1}\gamma_{k+2}^2 \\ &= M(1 + P\gamma_{k+1}^2)\frac{\gamma_{k+2}^2}{\gamma_{k+1}} + \sqrt{M}\,Q\sqrt{\gamma_{k+1}}\gamma_{k+2}^2 + R\gamma_{k+1}\gamma_{k+2}^2 \end{aligned}$$

For the last to be less than $M\gamma_{k+2}$, we have to verify

$$M(1 + P\gamma_{k+1}^2)\frac{\gamma_{k+2}}{\gamma_{k+1}} + \sqrt{M}Q\sqrt{\gamma_{k+1}}\gamma_{k+2} + R\gamma_{k+1}\gamma_{k+2} \le M$$

or equivalently,

$$M\left(\frac{\gamma_{k+2}}{\gamma_{k+1}} + P\gamma_{k+1}\gamma_{k+2} - 1\right) + \sqrt{M}Q\sqrt{\gamma_{k+1}}\gamma_{k+2} + R\gamma_{k+1}\gamma_{k+2} \le 0.$$

This is a quadratic equation in $\sqrt{M}$, and for this inequality to hold, we prove that the leading coefficient is negative, and the largest root is bounded above by some constant not depending on n.

Negativity of the leading coefficient is equivalent to

$$\frac{\gamma_{k+2}}{\gamma_{k+1}} + P\gamma_{k+1}\gamma_{k+2} < 1,$$

which is implied by our assumption on the step size.

The larger root of the equation is

$$\begin{aligned}
&\frac{\left(-4\gamma_{k+1}^2\gamma_{k+2}^2 PR + \gamma_{k+1}\gamma_{k+2}(\gamma_{k+2}Q^2 + 4R) - 4R\gamma_{k+2}^2\right)^{1/2} + \sqrt{\gamma_{k+1}}\gamma_{k+2}Q}{2(1 - \gamma_{k+1}\gamma_{k+2}P - \gamma_{k+2}/\gamma_{k+1})} \\
&< \frac{\sqrt{\gamma_{k+1}}\gamma_{k+2}Q + \sqrt{R\gamma_{k+1}\gamma_{k+2}}}{(1 - \gamma_{k+1}\gamma_{k+2}P - \gamma_{k+2}/\gamma_{k+1})} \\
&\le \frac{\sqrt{\gamma_{k+1}}\gamma_{k+1}Q + \sqrt{R}\gamma_{k+1}}{(1 - \gamma_{k+1}\gamma_{k+2}P - \gamma_{k+2}/\gamma_{k+1})}.
\end{aligned}$$

By our assumption on the step size that

$$\frac{\gamma_{k+2}}{\gamma_{k+1}} + P\gamma_{k+1}\gamma_{k+2} < 1 - \gamma_{k+1},$$

we get that the larger root is smaller than

$$\frac{\sqrt{\gamma_{k+1}}\gamma_{k+1}Q + \sqrt{R}\gamma_{k+1}}{\gamma_{k+1}} = \sqrt{\gamma_{k+1}}Q + \sqrt{R} < Q + \sqrt{R}.$$

Letting $M := Q + \sqrt{R}$ gives the desired result.

The second argument of the lemma follows from Assumption 3 and the first result of the lemma. ■

Lemma 4. *For a vector valued function $g \in L^2(\mathbb{R}; \mathbb{R}^d)$, one has*

$$\left\|\int_0^s g(u)\,du\right\|^2 \le \left(\int_0^s \|g(u)\|\,du\right)^2 \le s\int_0^s \|g(u)\|^2\,du.$$

D Proofs for Section 5

D.1 Proof of Theorem 3

For brevity, let us write $\mathscr{F}_k$ instead of $\mathscr{F}_{\tau_k}$. Opening up $\|x_{k+1}\|^2 = \|x_k + \gamma_{k+1}\{v(x_k) + Z_{k+1}\} + \sqrt{\gamma_{k+1}}\sigma(x_k)\,\xi_{k+1}\|^2$ and ignoring every term that is zero-mean under $\mathbb{E}[\cdot\,|\,\mathscr{F}_k]$, we get

$$\begin{aligned}
\mathbb{E}[\|x_{k+1}\|^2 \,|\, \mathcal{F}_k] &= \mathbb{E}\Big[\|x_k\|^2 + 2\gamma_{k+1}\langle x_k, v(x_k) + Z_{k+1}\rangle \\
&\qquad + \gamma_{k+1}^2\|v(x_k) + Z_{k+1}\|^2 + \gamma_{k+1}\|\sigma(x_k)\xi_{k+1}\|^2 + 2\gamma_{k+1}^{\frac{3}{2}}\langle\sigma(x_k)\xi_{k+1}, b_{k+1}\rangle \,\Big|\, \mathcal{F}_k\Big] \\
&\le \|x_k\|^2 + 2\gamma_{k+1}(\langle x_k, v(x_k)\rangle + C_\sigma/2) + 2\gamma_{k+1}^2\|v(x_k)\|^2 \\
&\qquad + \mathbb{E}\Big[2\gamma_{k+1}^2\|Z_{k+1}\|^2 + 2\gamma_{k+1}\langle x_k, Z_{k+1}\rangle + 2\gamma_{k+1}^{\frac{3}{2}}\langle\sigma(x_k)\xi_{k+1}, b_{k+1}\rangle \,\Big|\, \mathcal{F}_k\Big] \\
&\le \|x_k\|^2 + 2\gamma_{k+1}\left(\langle x_k, v(x_k)\rangle + C_\sigma/2 + \gamma_{k+1}^{\frac{1}{2}}C_\sigma/4\right) + 2\gamma_{k+1}^2\|v(x_k)\|^2 \qquad (21) \\
&\qquad + \mathbb{E}\big[2\gamma_{k+1}^2\|Z_{k+1}\|^2|\mathcal{F}_k\big] + \gamma_{k+1}^{\frac{3}{2}}\mathbb{E}\big[\|b_{k+1}\|^2|\mathcal{F}_k\big] + 2\mathbb{E}[\gamma_{k+1}\langle x_k, b_{k+1}\rangle|\mathcal{F}_k].
\end{aligned}$$

Recalling (5) in Assumption 3, we have for some $C > 0$

$$\mathbb{E}\|Z_{k+1}\|^2 \leq 2\sigma^2 + 2C\left(\gamma_{k+1}^2\mathbb{E}\|v(x_k)\|^2 + \gamma_{k+1}\right) \tag{22}$$

Without loss of generality, assume $\gamma_k \leq 1$ and $\mathbb{E}\|x_k\|^2 \geq 1$ (so that $\left(\mathbb{E}\|x_k\|^2\right)^2 \geq \mathbb{E}\|x_k\|^2$) for all k. Then, $\|v(x_k)\|^2 \leq L^2\|x_k\|^2$, together with Assumption 4 and the Cauchy-Schwartz inequality on the last term of (21), implies

$$\begin{aligned}
\mathbb{E}\|x_{k+1}\|^2 \leq \mathbb{E}\|x_k\|^2 &- 2\alpha\gamma_{k+1}\mathbb{E}\|x_k\|^2 + 2\gamma_{k+1}\left(\beta + C_\sigma + \frac{1}{2}\gamma_{k+1}^{\frac{1}{2}}C_\sigma\right) + 2L^2\gamma_{k+1}^2\mathbb{E}\|x_k\|^2 \\
&+ 2\gamma_{k+1}^2\left[2\sigma^2 + 2C\left(L^2\gamma_{k+1}^2\mathbb{E}\|x_k\|^2 + \gamma_{k+1}\right)\right] \\
&+ \gamma_{k+1}^{\frac{3}{2}}C\left(L^2\gamma_{k+1}^2\mathbb{E}\|x_k\|^2 + \gamma_{k+1}\right) \\
&+ 2\gamma_{k+1}\sqrt{C}\sqrt{L^2\gamma_{k+1}^2\left(\mathbb{E}\|x_k\|^2\right)^2 + \gamma_{k+1}\mathbb{E}\|x_k\|^2} \\
\leq \mathbb{E}\|x_k\|^2&(1 - C_1\gamma_{k+1} + C_2\gamma_{k+1}^{\frac{3}{2}}) + C_3\gamma_{k+1}
\end{aligned}$$

for some constants C_1, C_2, C_3 depending on $L, C, \sigma, \alpha, \beta$, and d. Since $\gamma_k \longrightarrow 0$, there exist $\tilde{\alpha}, \tilde{\beta} > 0$ and k_0 such that, for all $k \geq k_0$,

$$\mathbb{E}\|x_{k+1}\|^2 \leq \mathbb{E}\|x_k\|^2(1 - \tilde{\alpha}\gamma_{k+1}) + \tilde{\beta}\gamma_{k+1}, \quad 1 - \tilde{\alpha}\gamma_{k+1} > 0.$$

A simple induction yields

$$\sup_k \mathbb{E}\|x_k\|^2 \leq \max\left\{\frac{\tilde{\beta}}{\tilde{\alpha}}, \mathbb{E}\|x_{k_0}\|^2\right\}$$

which concludes the proof. ■

D.2 Proof of Theorem 4 for Constant Diffusion

Before proceeding, we need a lemma which can be distilled from [20, Proposition 8]:

Lemma 5. *Suppose ∇f is L-Lipschitz. Fix $x \in \mathbb{R}^d$ and $\gamma > 0$, let $\tilde{x}^+ = x - \gamma\nabla f(x) + \sqrt{2\gamma}\xi$. Then*

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\langle\nabla f(x), \tilde{x}^+ - x\rangle + \frac{L}{4}\|\tilde{x}^+ - x\|^2\right)\right] \leq (1 - \gamma L)^{-d/2}e^{-\frac{\gamma}{4}\|\nabla f(x)\|^2}. \tag{23}$$

Let $\tilde{x}_{k+1} := x_k - \gamma_{k+1}\nabla f(x_k) + \sqrt{2\gamma_{k+1}}\,\xi_{k+1}$ so that $x_{k+1} - x_k = \tilde{x}_{k+1} - x_k - \gamma_{k+1}(U_{k+1} + b_{k+1})$. Conditioned on $x_k, U_{k+1}, U'_{k+1}, \xi'_{k+1}$, and using the L-Lipschitzness of ∇f, we get

$$\begin{aligned}
&e^{-\frac{1}{2}f(x_k)}\mathbb{E}e^{\frac{1}{2}f(x_{k+1})} \\
&\leq \mathbb{E}\exp\left(\frac{1}{2}\langle\nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{4}\|x_{k+1} - x_k\|^2\right) &(24)\\
&\leq \mathbb{E}\exp\left\{\frac{1}{2}\langle\nabla f(x_k), \tilde{x}_{k+1} - x_k\rangle - \frac{1}{2}\langle\nabla f(x_k), \gamma_{k+1}U_{k+1}\rangle\right. &(25)\\
&\qquad\left. - \frac{1}{2}\langle\nabla f(x_k), \gamma_{k+1}b_{k+1}\rangle + \frac{L}{2}\|\tilde{x}_{k+1} - x_k\|^2 + L\gamma_{k+1}^2\|U_{k+1}\|^2 + L\gamma_{k+1}^2\|b_{k+1}\|^2\right\}. &(26)
\end{aligned}$$

Let $\delta \in (0, 1)$. Since

$$\begin{aligned}
-\frac{1}{2}\langle\nabla f(x_k), \gamma_{k+1}U_{k+1}\rangle &\leq \gamma_{k+1}^{2-\delta}\|\nabla f(x_k)\|^2 + \gamma_{k+1}^{\delta}\|U_{k+1}\|^2, \\
-\frac{1}{2}\langle\nabla f(x_k), \gamma_{k+1}b_{k+1}\rangle &\leq \gamma_{k+1}^2\|\nabla f(x_k)\|^2 + \|b_{k+1}\|^2,
\end{aligned}$$

we have

$$e^{-\frac{1}{2}f(x_k)}\mathbb{E}e^{\frac{1}{2}f(x_{k+1})} \tag{27}$$

$$\le \mathbb{E}\exp\left\{\frac{1}{2}\langle\nabla f(x_k), \tilde{x}_{k+1}-x_k\rangle + \frac{L}{2}\|\tilde{x}_{k+1}-x_k\|^2\right. \tag{28}$$

$$\left. + \left(\gamma_{k+1}^{2-\delta}+\gamma_{k+1}^2\right)\|\nabla f(x_k)\|^2 + \left(L\gamma_{k+1}^2+\gamma_{k+1}^{\delta}\right)\|U_{k+1}\|^2 + \left(L\gamma_{k+1}^2+1\right)\|b_{k+1}\|^2\right\}. \tag{29}$$

Invoking (11) an denoting $c' \triangleq \left(L\gamma_{k+1}^2+1\right)\cdot c$, we get

$$e^{-\frac{1}{2}f(x_k)}\mathbb{E}e^{\frac{1}{2}f(x_{k+1})} \le e^{A_k}\cdot\mathbb{E}\exp\left\{\frac{1}{2}\langle\nabla f(x_k), \tilde{x}_{k+1}-x_k\rangle + \frac{L}{2}\|\tilde{x}_{k+1}-x_k\|^2 + c'\cdot\gamma_{k+1}\|\xi_{k+1}\|^2\right\}, \tag{30}$$

where,

$$\begin{aligned} A_k \triangleq &\left(\gamma_{k+1}^{2-\delta}+\gamma_{k+1}^2+c'\gamma_{k+1}^2\right)\|\nabla f(x_k)\|^2 \\ &+\left(L\gamma_{k+1}^2+\gamma_{k+1}^{\delta}\right)\|U_{k+1}\|^2 \\ &+c'\left(\gamma_{k+1}^2\|U'_{k+1}\|^2+\gamma_{k+1}\|\xi'_{k+1}\|^2\right). \end{aligned} \tag{31}$$

Recalling that $\sqrt{2\gamma_{k+1}}\xi_{k+1} = \tilde{x}_{k+1}-x_k+\gamma_{k+1}\nabla f(x_k)$, we have $\gamma_{k+1}\|\xi_{k+1}\|^2 \le \|\tilde{x}_{k+1}-x_k\|^2 + \gamma_{k+1}^2\|\nabla f(x_k)\|^2$, and thus

$$e^{-\frac{1}{2}f(x_k)}\mathbb{E}e^{\frac{1}{2}f(x_{k+1})} \le e^{A'_k}\cdot\mathbb{E}\exp\left\{\frac{1}{2}\langle\nabla f(x_k), \tilde{x}_{k+1}-x_k\rangle + \left(\frac{L}{2}+c'\right)\|\tilde{x}_{k+1}-x_k\|^2\right\}, \tag{32}$$

where $A'_k = A_k + c'\gamma_{k+1}^2\|\nabla f(x_k)\|^2$. Lemma 5 then implies

$$e^{-\frac{1}{2}f(x_k)}\mathbb{E}e^{\frac{1}{2}f(x_{k+1})} \le e^{A''_k}\cdot(1-\gamma_{k+1}L')^{-\frac{d}{2}} \tag{33}$$

where $A''_k = A'_k - \frac{\gamma_{k+1}}{4}\|\nabla f(x_k)\|^2$.

We now take the expectation over $x_k, U_{k+1}, U'_{k+1}, \xi'_{k+1}$ (in other words, we are now only conditioning on x_k). Set $\epsilon \triangleq (1-\gamma_{k+1}L')^{-\frac{1}{2}} - 1 > 0$. Since $U_{k+1}, U'_{k+1}, \xi'_{k+1}$ are sub-Gaussian and since $\gamma_k \to 0$, for k sufficiently large we have

$$\mathbb{E}A''_k \le (1+\epsilon)\cdot\exp\left[\left(-\frac{\gamma_{k+1}}{4}+\gamma_{k+1}^{2-\delta}+\gamma_{k+1}^2+c'\gamma_{k+1}^2+c'\gamma_{k+1}^2\right)\|\nabla f(x_k)\|^2\right] \tag{34}$$

$$\le (1+\epsilon)\cdot e^{-\frac{\gamma_{k+1}}{8}\|\nabla f(x_k)\|^2}. \tag{35}$$

To summarize, we have shown that, conditioned on x_k,

$$e^{-\frac{1}{2}f(x_k)}\mathbb{E}e^{\frac{1}{2}f(x_{k+1})} \le (1-\gamma_{k+1}L')^{-\frac{d+1}{2}}e^{-\frac{\gamma_{k+1}}{8}\|\nabla f(x_k)\|^2}. \tag{36}$$

A simple induction à la [20, Lemma 1 & Proposition 8] then concludes the proof. ■

D.3 Proof of Theorem 4 for Mirror Langevin

Here, we bring the proof of Theorem 4 for the case of Example 4 and without noise. The proof for the noisy case is the same as in Appendix D.2.

Define

$$x^+ = x - \gamma\nabla f\circ\nabla\phi^*(x) + \sqrt{2\gamma}(\nabla^2\phi^*(x)^{-1})^{1/2}\xi,$$

where ξ is a standard Gaussian random variable. Let $U(x) = f(\nabla\phi^*(x))$. For a fixed x, we have

$$\mathbb{E}e^{\frac{1}{2}U(x^+)-\frac{1}{2}U(x)} = \frac{1}{(2\pi)^{d/2}}\int\exp\left(\frac{1}{2}U(x^+)-\frac{1}{2}U(x)-\frac{\|\xi\|^2}{2}\right)d\xi$$

Notice that we have

$$\xi = \frac{1}{\sqrt{2\gamma}} (\nabla^2 \phi^*(x))^{1/2} \big(x^+ - x + \gamma \nabla f \circ \nabla \phi^*(x)\big)$$

which implies

$$d\xi = (\sqrt{2\gamma})^{-d} \sqrt{\det \nabla^2 \phi^*(x)}\, dx^+$$

Thus, the integral, after the change of variable from ξ to x^+ becomes

$$\frac{1}{C} \int \exp\left(\frac{1}{2} U(x^+) - \frac{1}{2} U(x) - \frac{1}{4\gamma} \| (\nabla^2 \phi^*(x))^{1/2} \big(x^+ - x + \gamma \nabla f \circ \nabla \phi^*(x)\big) \|^2 \right) dx^+ \tag{37}$$

with $C = (4\pi\gamma)^{d/2} \sqrt{\det \nabla^2 \phi^*(x)^{-1}}$. Now we use the smoothness of f:

$$\begin{aligned} U(x^+) - U(x) &= f(\nabla \phi^*(x^+)) - f(\nabla \phi^*(x)) \\ &\leq \langle \nabla^2 \phi^*(x) \nabla f(\nabla \phi^*(x)), x^+ - x \rangle + \frac{L}{2} \|x^+ - x\|^2 \end{aligned}$$

On the other hand, we have

$$\begin{aligned} &\| (\nabla^2 \phi^*(x))^{1/2} \big(x^+ - x + \gamma \nabla f \circ \nabla \phi^*(x)\big) \|^2 \\ &\quad = \| (\nabla^2 \phi^*(x))^{1/2} (x^+ - x) \|^2 + \gamma^2 \| (\nabla^2 \phi^*(x))^{1/2} \nabla f(\nabla \phi^*(x)) \|^2 \\ &\qquad + 2\gamma \langle \nabla^2 \phi^*(x) \nabla f \nabla \phi^*(x), x^+ - x \rangle \end{aligned}$$

Notice that in (37), the colored terms cancel out, and what we are left with is

$$\begin{aligned} &\mathbb{E} e^{\frac{1}{2} U(x^+) - \frac{1}{2} U(x)} \\ &\leq \frac{1}{C} \int \exp\left(\frac{L}{4} \|x^+ - x\|^2 - \frac{1}{4\gamma} \| (\nabla^2 \phi^*(x))^{1/2} (x^+ - x) \|^2 - \frac{\gamma}{4} \| (\nabla^2 \phi^*(x))^{1/2} \nabla f(\nabla \phi^*(x)) \|^2 \right) dx^+ \end{aligned}$$

As, by our assumption, $\nabla^2 \phi^*$ is bounded from above and below, we get the exact form as in Lemma 5. The rest of the proof is the same as in Appendix D.2. ■

D.4 Proof of Proposition 1

In this section, we prove that Examples 1–6 satisfy our bias conditions, which, as we have seen in Section 5, implies Proposition 1. For brevity, we write $\mathscr{F}_k$ for $\mathscr{F}_{\tau_k}$.

§ Proof for Example 1. For randomized mid-point method, by replacing $\widetilde{\nabla} f(x_k)$ and $\widetilde{\nabla} f(x_{k+1/2})$ with $\nabla f(x_k) + U'_{k+1}$ and $\nabla f(x_{k+1/2}) + U_{k+1}$ respectively, we have

$$\begin{aligned} x_{k+1/2} &= x_k - \gamma_{k+1} \alpha_{k+1} \{ \nabla f(x_k) + U'_{k+1} \} + \sqrt{2\gamma_{k+1} \alpha_{k+1}} \xi'_{k+1}, \\ x_{k+1} &= x_k - \gamma_{k+1} \{ \nabla f(x_{k+1/2}) + U_{k+1} \} + \sqrt{2\gamma_{k+1}} \xi_{k+1}, \end{aligned}$$

where $\{\alpha_k\}$ are i.i.d. and uniformly distributed in $[0, 1]$, $\{U_k\}$ and $\{U'_k\}$ are noises in evaluating ∇f at the corresponding points, and ξ_k, ξ'_k are independent standard Gaussians.

Notice that the Lipschitzness of ∇f, and the fact that $\alpha_k \leq 1$ implies that the bias term $b_{k+1} := \nabla f(x_{k+1/2}) - \nabla f(x_k)$ satisfies

$$\begin{aligned} \mathbb{E}[\|b_{k+1}\|^2 \mid \mathscr{F}_k] &\leq L^2 \mathbb{E}[\|x_{k+1/2} - x_k\|^2 \mid \mathscr{F}_k] \\ &\leq L^2 \left(\gamma_{k+1}^2 \mathbb{E}[\|\nabla f(x_k) + U'_{k+1}\|^2 \mid \mathscr{F}_k] + 2\gamma_{k+1} d \right) \\ &\leq 2L^2 \gamma_{k+1}^2 \|\nabla f(x_k)\|^2 + 2L^2 \gamma_{k+1}^2 \sigma^2 + 2L^2 d \gamma_{k+1} \\ &= O(\gamma_{k+1}^2 \|\nabla f(x_k)\|^2 + \gamma_{k+1}). \end{aligned}$$

§ Proof for Example 2. Recall that the new algorithm Optimistic Randomized Mid-Point Method has the iterates

$$x_{k+1/2} = x_k - \gamma_{k+1}\alpha_{k+1}\widetilde{\nabla} f(x_{k-\frac{1}{2}}) + \sqrt{2\gamma_{k+1}\alpha_{k+1}}\xi'_{k+1},$$
$$x_{k+1} = x_k - \gamma_{k+1}\widetilde{\nabla} f(x_{k+1/2}) + \sqrt{2\gamma_{k+1}}\,\xi_{k+1},$$

where $\{\alpha_k\}$, ξ_k, ξ'_k, and $\widetilde{\nabla} f$ are the same as in (RMM), and the noise and bias are $U_{k+1} \coloneqq \widetilde{\nabla} f(x_{k+1/2}) - \nabla f(x_{k+1/2})$ and $b_{k+1} \coloneqq \nabla f(x_{k+1/2}) - \nabla f(x_k)$. We have

$$\begin{aligned}
\mathbb{E}[\|b_{k+1}\|^2 \mid \mathscr{F}_k] &= \mathbb{E}[\|\nabla f(x_{k+1/2}) - \nabla f(x_k)\|^2 \mid \mathscr{F}_k] \\
&\le L^2 \mathbb{E}[\|x_{k+1/2} - x_k\|^2 \mid \mathscr{F}_k] \\
&= L^2 \mathbb{E}[\|-\gamma_{k+1}\alpha_{k+1}\widetilde{\nabla} f(x_{k-\frac{1}{2}}) + \sqrt{2\gamma_{k+1}\alpha_{k+1}}\xi'_{k+1}\|^2 \mid \mathscr{F}_k] \\
&\le 2L^2\gamma_{k+1}^2 \mathbb{E}[\|\nabla f(x_{k-\frac{1}{2}})\| \mid \mathscr{F}_k] + 2L^2\gamma_{k+1}^2\sigma^2 + 4L^2 d\gamma_{k+1}.
\end{aligned}$$

Similar to the proof for Example 6, notice that $\|\nabla f(x_{k-\frac{1}{2}})\|^2 \le 2\|\nabla f(x_{k-\frac{1}{2}}) - \nabla f(x_k)\|^2 + 2\|\nabla f(x_k)\|^2$. As $\gamma_k \to 0$, one can assume that $2L^2\gamma_{k+1}^2 < \frac{1}{2}$, and we get

$$\mathbb{E}[\|b_{k+1}\|^2 \mid \mathscr{F}_k] \le 4L^2\gamma_{k+1}^2\|\nabla f(x_k)\|^2 + 4L^2\gamma_{k+1}^2\sigma^2 + 8L^2 d\gamma_{k+1} = \mathcal{O}(\gamma_{k+1}^2\|\nabla f(x_k)\|^2 + \gamma_{k+1}),$$

as desired. ■

§ Proof for Example 3. The iterates of stochastic Runge-Kutta Langevin algorithm is as follows:

$$\begin{aligned}
h_1 &= x_k + \sqrt{2\gamma_{k+1}}\Big[(1/2 + 1/\sqrt{6})\,\xi_{k+1} + \xi'_{k+1}/\sqrt{12}\Big] \\
h_2 &= x_k - \gamma_{k+1}\{\nabla f(x_k) + U'_{k+1}\} + \sqrt{2\gamma_{k+1}}\Big[(1/2 - 1/\sqrt{6})\,\xi_{k+1} + \xi'_{k+1}/\sqrt{12}\Big] \\
x_{k+1} &= x_k - \frac{\gamma_{k+1}}{2}(\nabla f(h_1) + \nabla f(h_2)) + \gamma_{k+1}U_{k+1} + \sqrt{2\gamma_{k+1}}\,\xi_{k+1},
\end{aligned}$$

where ξ_{k+1} and ξ'_{k+1} are independent standard Gaussian random variables independent of x_k, and U_{k+1} and U'_{k+1} are noise in the evaluation of f.

Observe that

$$b_{k+1} = \frac{1}{2}(\nabla f(h_1) - \nabla f(x_k)) + \frac{1}{2}(\nabla f(h_2) - \nabla f(x_k)).$$

We have

$$\mathbb{E}[\|\nabla f(h_1) - \nabla f(x_k)\|^2 \mid \mathscr{F}_k] \le 2L^2 d(1/4 + 1/6 + 1/12)\gamma_{k+1} = \mathcal{O}(\gamma_{k+1}),$$

and

$$\begin{aligned}
\mathbb{E}[\|\nabla f(h_2) - \nabla f(x_k)\|^2 \mid \mathscr{F}_k] &\le 2L^2\Big(\gamma_{k+1}^2\|\nabla f(x_k)\|^2 + 2\gamma_{k+1}^2\sigma^2 + 2d(1/4 - 1/6 + 1/12)\gamma_{k+1}\Big) \\
&= \mathcal{O}(\gamma_{k+1}^2\|\nabla f(x_k)\|^2 + \gamma_{k+1}).
\end{aligned}$$

We thus have

$$\begin{aligned}
\mathbb{E}[\|b_{k+1}\|^2 \mid \mathscr{F}_t] &\le \frac{1}{2}\mathbb{E}[\|\nabla f(h_1) - \nabla f(x_k)\|^2 \mid \mathscr{F}_k] + \frac{1}{2}\mathbb{E}[\|\nabla f(h_2) - \nabla f(x_k)\|^2 \mid \mathscr{F}_k] \\
&= \mathcal{O}(\gamma_{k+1}^2\|\nabla f(x_k)\|^2 + \gamma_{k+1}),
\end{aligned}$$

as desired. ■

§ Proof for Example 4. Suppose ϕ is a Legendre function [52] for $\mathbb{R}^d$, and consider the iterates

$$x_{k+1} = x_k - \gamma_{k+1}\nabla f(\nabla\phi^*(x_k)) + \sqrt{2\gamma_{k+1}}(\nabla^2\phi^*(x_k)^{-1})^{1/2}\,\xi_{k+1},$$

where ϕ^* is the *Fenchel dual* of ϕ, that is, $\phi^*(x) = \sup_{y\in\mathbb{R}^d}(\langle x, y\rangle - \phi(y))$. Also recall that [52]

$$\nabla\phi(\nabla\phi^*(x)) = x, \quad \nabla^2\phi^*(\nabla\phi(x))^{-1} = \nabla^2\phi(x), \quad \forall x \in \mathbb{R}^d.$$

Let $v = -\nabla f \circ \nabla\phi^*$ and $\sigma = (\nabla^2\phi^*)^{-1/2}$. First, we mention what our assumptions imply on f:

- The Lipschitzness of v corresponds to a similar condition in [31, A2]:

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

- The Lipschitzness of σ in Frobenius norm corresponds to *modified self-concordance* in [31, A1]:

$$\|\nabla^2\phi(x)^{1/2} - \nabla^2\phi(y)^{1/2}\|_F \leq L\|\nabla\phi(x) - \nabla\phi(y)\|.$$

- Boundedness of σ in Hilbert-Schmidt norm implies

$$\left\|\nabla^2\phi(x)^{-1/2}\right\|_F \leq C_\sigma.$$

- Dissipativity and weak-dissipativity of v corresponds to the conditions below, respectively:

$$\langle\nabla\phi(x), \nabla f(x)\rangle \geq \alpha\|\nabla\phi(x)\|^2 - \beta, \quad \langle\nabla\phi(x), \nabla f(x)\rangle \geq \alpha\|\nabla\phi(x)\|^{1+\kappa} - \beta.$$

If f and ϕ satisfy the conditions above, then the mirror Langevin algorithm Example 4 fits into the (LRM) scheme.

Remark. Note that this version of Mirror Langevin *cannot* handle the case where e^{-f} is supported on a compact domain; in that case, the Hessian of ϕ *has to* blow up near the boundary, and will not satisfy our boundedness assumption. The version of mirror Langevin we consider in this paper, though, can be thought as an adaptive conditioning method for densities supported on $\mathbb{R}^d$. This setting has also been studied in the literature, see [55].

§ Proof for Example 6. The iterates of (PLA) follow

$$x_{k+1} = x_k - \gamma_{k+1}\nabla f(x_{k+1}) + \sqrt{2\gamma_{k+1}}\,\xi_{k+1}. \tag{PLA}$$

We mentioned that the bias term is $b_{k+1} = \nabla f(x_{k+1}) - \nabla f(x_k)$. Now it remains to prove that it satisfies the conditions (5) and (11). We have

$$\begin{aligned}
\mathbb{E}[\|b_{k+1}\|^2 \mid \mathscr{F}_k] &= \mathbb{E}[\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \mid \mathscr{F}_k] \\
&\leq L^2\mathbb{E}[\|x_{k+1} - x_k\|^2 \mid \mathscr{F}_k] \\
&= L^2\mathbb{E}[\|-\gamma_{k+1}\nabla f(x_{k+1}) + \sqrt{2\gamma_{k+1}}\,\xi_{k+1}\|^2 \mid \mathscr{F}_k] \\
&\leq 2L^2\gamma_{k+1}^2\mathbb{E}[\|\nabla f(x_{k+1})\|^2 \mid \mathscr{F}_k] + 4L^2 d\gamma_{k+1}.
\end{aligned}$$

Now, notice that $\|\nabla f(x_{k+1})\|^2 \leq 2\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 + 2\|\nabla f(x_k)\|^2$. As $\gamma_k \to 0$, one can assume that $2L^2\gamma_{k+1}^2 < \frac{1}{2}$, and we get

$$\mathbb{E}[\|b_{k+1}\|^2 \mid \mathscr{F}_k] \leq \frac{1}{2}\mathbb{E}[\|b_{k+1}\|^2 \mid \mathscr{F}_k] + \|\nabla f(x_k)\|^2 + 4L^2 d\gamma_{k+1},$$

which implies

$$\mathbb{E}[\|b_{k+1}\|^2 \mid \mathscr{F}_k] \leq 4L^2\gamma_{k+1}^2\|\nabla f(x_k)\|^2 + 8L^2 d\gamma_{k+1} = O(\gamma_{k+1}^2\|\nabla f(x_k)\|^2 + \gamma_{k+1}),$$

as desired. ∎