Convex Analysis and Convex Optimization

Alessandro Rudi, Pierre Gaillard

March 5, 2021

1 Constrained optimization problems

Let $f : D \mapsto \mathbb{R}^d$ convex and $C \subset D$ convex. We consider the constrained minimization problem

$$\inf_{x \in C} f(x) \,.$$

C is the constraint set. It is often defined as the intersection of sets of the form $\{h_i(x) = 0\}$ and $\{g_j(x) \leqslant 0\}$.

Example 1.1. *The minimization of a linear function over a compact $A \subset \mathbb{R}^d$. Let $A \subset \mathbb{R}^d$ compact (non-necessarily convex) and $a \in \mathbb{R}^d \neq 0$ then we can reformulate the non-convex minimization on A as a constrained convex optimization problem on* $\operatorname{Conv}(A)$

$$\min_{x \in A} \left\{ a^\top x \right\} = \min_{x \in \operatorname{Conv}(A)} \left\{ a^\top x \right\}$$

Lagrangian duality A useful notion to solve constrained optimization problems is Lagrangian duality. Assume that we are interested in the following constrained optimization problem:

$$\min_{x \in D} f(x) \quad \text{such that} \quad \left\{ \begin{array}{ll} h_i(x) = 0 & \text{for } i = 1, \dots, m \\ g_j(x) \leqslant 0 & \text{for } j = 1, \dots, r \end{array} \right. . \tag{P}$$

We denote by $D^* \subseteq D$ the set of points that satisfy the constraints. Remark that equality constraints $h_i(x) = 0$ can be rewritten as inequalities

$$h_i(x) \leqslant 0 \qquad \text{and} \qquad -h_i(x) \leqslant 0 \,.$$

Contrary to unconstrained optimization problems, canceling the gradient does not necessarily provide a solution for constrained optimization problems. The basic idea of Lagrangian duality is to take the constraint D^* into account in the minimization problem by augmenting the objective function with a weighted sum of the constraint functions.

Definition 1.1 (Lagrangian)**.** *The Lagrangian associated to the optimization problem* (P) *is the function $\mathcal{L} : D \times \mathbb{R}^m \times \mathbb{R}_+^r$ defined by:*

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \lambda^\top h(x) + \mu^\top g(x) \,.$$

Definition 1.2 (Primal function)**.** *We define the primal function $\bar{f} : D \to \mathbb{R} \cup \{+\infty\}$ associated to* (P) *by, for all $x \in D$*

$$\bar{f}(x) = \sup_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}_+^r} \mathcal{L}(x, \lambda, \mu) = \left\{ \begin{array}{ll} f(x) & \textit{if } x \in D^* \\ +\infty & \textit{otherwise} \end{array} \right. .$$

With these definitions, we remark that the optimization problem (P) can be re-written by using the primal function without the constraints

$$\begin{aligned}\inf_{x\in D^*} f(x) &= \inf_{x\in D} \bar{f}(x)\\ &= \inf_{x\in D} \sup_{\lambda\in\mathbb{R}^m,\mu\in\mathbb{R}^r_+} \mathcal{L}(x,\lambda,\mu)\,. \end{aligned} \tag{Primal problem}$$

This optimization problem is thus called the *Primal problem.*

The *Dual problem* is obtained by exchanging inf and sup in the primal problem.

$$\sup_{\lambda\in\mathbb{R}^m,\mu\in\mathbb{R}^r_+} f^*(\lambda,\mu) = \sup_{\lambda\in\mathbb{R}^m,\mu\in\mathbb{R}^r_+} \inf_{x\in D} \mathcal{L}(x,\lambda,\mu)\,, \tag{Dual problem}$$

where $f^* : (\lambda,\mu) \mapsto \inf_{x\in D} \mathcal{L}(x,\lambda,\mu)$ is the dual function. If f is convex this function is concave. Remark that the dual of the dual is the primal.

The denote by $D^* = \{x \in D : \bar{f}(x) < \infty\}$ the admissibility domain of the primal. Similarly we denote by $C^* = \{(\lambda,\mu) \in \mathbb{R}^m \times \mathbb{R}^r_+ : f^*(\lambda,\mu) > -\infty\}$ the admissibility domain of the dual. If there is no solution to the optimization problem (P) then $D^* = \emptyset$. If the problem is unbounded then $C^* = \emptyset$.

Link between the primal and the dual problems If they are not necessarily identical the primal and the dual problems have strong relationship. For any (λ,μ), $f^*(\lambda,\mu)$ provides a lower bound on the solution of (P). The dual problem finds the best lower bound.

Proposition 1.1 (Weak duality principle)**.** *We have the inequality*

$$d^* := \sup_{\lambda\in\mathbb{R}^m,\mu\in\mathbb{R}^r_+} \inf_{x\in D} \mathcal{L}(x,\lambda,\mu) \leqslant \inf_{x\in D} \sup_{\lambda\in\mathbb{R}^m,\mu\in\mathbb{R}^r_+} \mathcal{L}(x,\lambda,\mu) := p^*\,.$$

Therefore, the solution of the dual problem is always smaller than the solution of the primal. A good mnemonic to remember this inequality is “the largest dwarf is always smaller than the smallest giant”.

Definition 1.3 (Dual gap and strong duality)**.** *The dual gap of the optimization problem is the difference between the primal and dual solutions: $p^* - d^* \geqslant 0$. We say that there is* strong duality *if $p^* = d^*$.*

If the duality gap is non-zero, the solutions of the primal and the dual problems are not really related. But when there is no gap, we say that there is strong duality. The two problems are equivalent (they share the same solutions). In this case, the existence of the solutions are related with the existence of saddle points of the Lagrangian. It is worth emphasizing that strong duality does not always holds.

When is there strong duality? Sometimes the dual problem is easier to solve than the primal problem. It is then useful to know if there is strong duality.

Definition 1.4 (Slater's condition)**.** *There exists a point $x_0 \in D$ strictly feasible*

$$\exists x_0 \in D \quad \textit{such that} \quad \begin{cases} \forall 1 \leqslant i \leqslant m & h_i(x_0) = 0 \\ \forall 1 \leqslant j \leqslant r & g_j(x_0) < 0 \end{cases}\,.$$

Theorem 1.2 (Strong duality)**.** *If the optimization problem* (P) *is convex: i.e.,*

- *f and D are convex,*
- *the equality constraint functions h_i are affine*
- *the inequality constraint functions g_j are convex*

and if Slater's condition holds than there is strong duality ($p^* = d^*$).

In this case, we can solve the dual problem to find a solution of the primal problem.

Example 1.2. *Let us compute the dual of the following linear programing problem over the set $D = \mathbb{R}^d_+$*

$$\min_{x \geqslant 0:\ Ax=b} c^\top x \,,$$

where A is a $m \times d$ matrix and $b \in \mathbb{R}^m$. The constraints can be written as $Ax - b = 0$. We can thus define the Lagrangian $\mathcal{L} : (x, \lambda) \in \mathbb{R}^d_+ \times \mathbb{R}' \to c^\top x + \lambda^\top(b - Ax)$ and re-write the primal problem with the Lagrangian

$$\begin{aligned} \min_{x \geqslant 0:\ Ax=b} c^\top x &= \min_{x \geqslant 0} \sup_{\lambda \in \mathbb{R}^m} \left\{ c^\top x + \lambda^\top (b - Ax) \right\} \\ &= \min_{x \geqslant 0} \sup_{\lambda \in \mathbb{R}^m} \left\{ b^\top \lambda + x^\top (c - A^\top \lambda) \right\} . \end{aligned}$$

By Slaters condition (the problem is convex since the objective function is convex and the equality constraints are affine), there is strong duality. We can thus swap the min *and the* sup*, we get*

$$\begin{aligned} \min_{x \geqslant 0,\ Ax=b} c^\top x &= \sup_{\lambda \in \mathbb{R}^m} \min_{x \geqslant 0} \left\{ b^\top \lambda + x^\top (c - A^\top \lambda) \right\} \\ &= \sup_{\lambda \in \mathbb{R}^m : A^\top \lambda \leqslant c} \left\{ b^\top \lambda \right\} . \end{aligned}$$

The latter is the dual formulation of the problem.

Optimality condition Now, we see conditions that play the same role as canceling the gradients for unconstrained optimization problems. These conditions will be useful to find equations to compute analytically the solution of (P).

Assume that the functions f, h_i and g_j are all differentiable. Let x^* and (λ^*, μ^*) be any primal and dual solutions and assume that there is strong duality (no duality gap). Then, we have

1. By definition x^* minimizes $\mathcal{L}(x, \lambda^*, \mu^*)$ over x. Therefore, its gradient must be canceled at x^*, i.e.,

$$\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla h(x^*) + \sum_{j=1}^r \mu_j \nabla g_j(x^*) = 0 \tag{KKT1}$$

2. Since $x^* \in D^*$ and $(\lambda^*, \mu^*) \in C^*$ are feasible we have

$$\begin{aligned} h_i(x^*) = 0 &\quad \forall 1 \leqslant i \leqslant m \\ g_j(x^*) \leqslant 0 &\quad \forall 1 \leqslant j \leqslant r \\ \mu_j^* \geqslant 0 &\quad \forall 1 \leqslant j \leqslant r \,. \end{aligned} \tag{KKT2}$$

3. The *complementary condition* holds

$$\mu_j^* g_j(x^*) = 0 \quad \forall 1 \leqslant j \leqslant r \,. \tag{KKT3}$$

Otherwise, we can improve μ^* by setting $\mu_j^* = 0$ since $g_j(x^*) \leqslant 0$ and (λ^*, μ^*) maximizes $\mathcal{L}(x^*, \lambda, \mu) = f(x^*) + \sum_i \lambda_i h_i(x^*) + \sum_j \mu_j g_j(x^*)$.

These conditions (KKT1-3) are called the Karush-Kuhn-Tucker (KKT) conditions. When the primal problem is convex (see Thm. 1.2) these conditions are also sufficient.

Theorem 1.3. *If there is strong duality then*

$$\left.\begin{array}{r}x^* \text{ is a solution of the primal problem} \\ (\lambda^*, \mu^*) \text{ is a solution of the dual problem}\end{array}\right\} \Leftrightarrow \text{(KKT) conditions are satisfied.}$$

The KKT conditions play an important role in optimization. In some cases, it is possible to solve them analytically. Many optimization methods are conceived for solving the KKT conditions.

2 Optimization algorithms for unconstrained convex optimization

In this section we will see two widely used optimization algorithms for the problem of unconstrained optimization. *Gradient descent* and *Stochastic Gradient descent.*

Since the goal of minimizing a function $f : \mathbb{R}^d \to \mathbb{R}$ ($d \in \mathbb{N}$) is to find the point x for which the function has minimum value, the fundamental idea behind gradient descent, consists in starting from a given $x_0 \in \mathbb{R}^d$ and finding the next point following a descent direction iteratively. In particular we will consider f to be a convex function.

2.1 Gradient descent

Note that when f is differentiable, the gradient of f, denoted by $\nabla f(x)$ determines the direction of maximum increase of the function in a suitable neighborood of x (and so $-\nabla f$ determines the direction of maximum decrease of the function). The gradient descent algorithm then reads as follows

$$x_{t+1} = x_t - \gamma_t \nabla f(x_t), \quad \forall t = 1, \dots, T$$

where x_0 is denoted starting point and is given (e.g. $x_0 = 0$), and γ_t is denoted as *step-size* and is small enough such that $-\gamma_t \nabla f(x_t)$ is still a decrease direction in the neighborhood of x_t.

The choice of γ_t is crucial for the optimization algorithm, indeed a γ_t that is too big, makes the algorithm unstable and possibly diverging, since it follows the direction $-\nabla f(x_t)$ outside of the region where it is a descent direction. On the other hand if γ_t is too small, the chosen direction is a descent direction, but each step is very short leading to a larger number of steps required to arrive to the minimum solution (with a big impact on the total computational complexity).

In the next theorem we show that there exists a step-size that guarantees the convergence of the solution of the gradient descent algorithm to the minimizer of f and we characterize how fast gradient descent achieves it.

Convergence of Gradient Descent for strongly convex functions. In this paragraph we assume that f is strongly convex with gradients that are L-Lipschitz continuous.

Definition 2.1 (L-Lipschitz continuous gradients)**.** *Let $L > 0$. f has L-Lipschitz continuous gradients if for all $x, y \in \mathbb{R}^d$ the following holds*

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|,$$

When f is L-Lipschitz, the following holds

Lemma 2.1. *Let $L > 0$ and $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function with L-Lipschitz continuous gradients, then*

$$f(y) \leqslant f(x) + \nabla f(x)^\top (y - x) + L\|y - x\|^2. \tag{1}$$

Proof. By the characterization of differentiable convex functions we have seen in the previous class we have

$$f(x) - f(y) \geqslant \nabla f(y)^\top (x - y),$$

from which, by reordering the terms

$$f(y) \leqslant f(x) + \nabla f(y)^\top (y - x).$$

By adding and subtracting $\nabla f(x)^\top (y - x)$, we have

$$f(y) \leqslant f(x) + \nabla f(x)^\top (y - x) + (\nabla f(y) - \nabla f(x))^\top (y - x).$$

Finally, by Cauchy-Schwarz and L-Lipschitzianity of the gradient, we bound the third term of the right hand side of the equation above as

$$(\nabla f(y) - \nabla f(x))^\top (y - x) \leqslant \|\nabla f(y) - \nabla f(x)\|\|y - x\| \leqslant L\|x - y\|^2.$$

□

Remark 2.1. By using a different argument it is possible to prove the tighter result

$$f(y) \leqslant f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2. \tag{2}$$

We are ready to see that gradient descent satisfies $f(x_t) < f(x_{t-1})$ when $\gamma \in (0, 1/L)$, as proven in the following lemma

Lemma 2.2 (Gradient descent is a descent algorithm with $\gamma_t \in (0, 1/L)$)**.** *Let f be convex and with L-Lipschitz gradients, let $x_0 \in \mathbb{R}^d$ and $\gamma_t > 0$, then*

$$f(x_t) \leqslant f(x_{t-1}) - \gamma_t(1 - L\gamma_t)\|\nabla f(x_{t-1})\|^2, \quad \forall\ t \in \mathbb{N}. \tag{3}$$

In particular if $\gamma_t \in (0, 1/L)$ we have that $\gamma_t(1 - L\gamma_t) > 0$ and so for all $t \in \mathbb{N}$,

$$f(x_t) < f(x_{t-1})$$

whenever x_{t-1} is not a global optimum, otherwise $x_t = x_{t-1}$ and $f(x_t) = f(x_{t-1})$.

Proof. Choose $\gamma \in (0, 1/L)$, by applying the lemma above, we have

$$f(x_t) \leqslant f(x_{t-1}) + \nabla f(x_{t-1})(x_t - x_{t-1}) + L\|x_{t-1} - x_t\|^2,$$

however $x_t - x_{t-1} = -\nabla f(x_{t-1})$, then

$$f(x_t) \leqslant f(x_{t-1}) - \gamma(1 - L\gamma)\|\nabla f(x_{t-1})\|^2.$$

To conclude, since f is convex and differentiable, note that x is a global optimum iff $\nabla f(x) = 0$. By choosing $\gamma_t \in (0, 1/L)$ we ensure that $\gamma(1 - L\gamma) > 0$, so if x_{t-1} is not a global optimum, then $\gamma_t(1 - L\gamma_t)\|\nabla f(x_{t-1})\|^2 > 0$, that implies $f(x_t) < f(x_{t-1})$. If x_{t-1} is a global optimum, then $\nabla f(x_{t-1}) = 0$, so

$$x_t = x_{t-1} - \gamma_t \nabla f(x_{t-1}) = x_{t-1}.$$

□

To prove the convergence rate of gradient descent in the case of strongly convex functions, we recall from previous class, that, when f is μ-strongly convex, we have

$$f(y) \geqslant f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|x-y\|^2. \tag{4}$$

We are going to use this property in the next theorem

Theorem 2.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a μ-strongly convex function with L-Lipschitz continuous gradients. Let $\gamma \in (0, 1/(2L))$ and choose a constant step-size $\gamma_t = \gamma$ for $t \in \mathbb{N}$. Denote by x^* the global optimum of f, let $x_0 \in \mathbb{R}^d$ and $T \in \mathbb{N}$, then*

$$\|x_T - x^*\|^2 \leqslant (1-\gamma\mu)^T \|x_0 - x^*\|^2.$$

Proof. We have

$$\begin{aligned}\|x_t - x^*\|^2 &= \|x_{t-1} - \gamma\nabla f(x_{t-1}) - x^*\|^2 && (5)\\ &= \|x_{t-1} - x^*\|^2 - 2\gamma \nabla f(x_{t-1})^\top (x_{t-1} - x^*) + \gamma^2 \|\nabla f(x_{t-1})\|^2. && (6)\end{aligned}$$

However by strong convexity of f, reordering Eq. (4) with $x = x_{t-1}$ and $y = x^*$, we have

$$-\nabla f(x_{t-1})^\top (x_{t-1} - x^*) \leqslant f(x_*) - f(x_{t-1}) - \frac{\mu}{2}\|x_{t-1} - x^*\|^2. \tag{7}$$

By substituting the upper bound for $-\nabla f(x_{t-1})^\top (x_{t-1} - x^*)$ in Eq. (6), we have

$$\begin{aligned}\|x_t - x^*\|^2 &\leqslant \|x_{t-1} - x^*\|^2 - 2\gamma(f(x_{t-1}) - f(x_*)) - \gamma\mu\|x_t - x^*\|^2 + \gamma^2\|\nabla f(x_{t-1})\|^2\\ &= (1-\gamma\mu)\|x_{t-1} - x^*\|^2 - 2\gamma(f(x_{t-1}) - f(x_*)) + \gamma^2\|\nabla f(x_{t-1})\|^2.\end{aligned}$$

By using the fact that Gradient descent is a descent algorithm, see Eq. (3), we have

$$f(x^*) \leqslant f(x_t) \leqslant f(x_{t-1}) - \gamma(1 - L\gamma)\|\nabla f(x_{t-1})\|^2,$$

from which

$$\|\nabla f(x_{t-1})\|^2 \leqslant \frac{1}{\gamma(1-\gamma L)}(f(x_{t-1}) - f(x^*)). \tag{8}$$

By applying this result above, and noting that $(1-\gamma L)^{-1} \leqslant 2$ when $\gamma \leqslant 1/(2L)$ we have

$$\|x_t - x^*\|^2 \leqslant (1-\gamma\mu)\|x_{t-1} - x^*\|^2 - \gamma\frac{2 - 2L\gamma - 1}{1 - L\gamma}(f(x_{t-1}) - f(x_*)).$$

Note that when $\gamma \in (0, 1/(2L))$ we have that $\gamma\frac{2-2L\gamma-1}{1-L\gamma} > 0$, then

$$\begin{aligned}\|x_t - x^*\|^2 &\leqslant (1-\gamma\mu)\|x_{t-1} - x^*\|^2 - \gamma\frac{2 - 2L\gamma - 1}{1 - L\gamma}(f(x_{t-1}) - f(x_*))\\ &\leqslant (1-\gamma\mu)\|x_{t-1} - x^*\|^2,\end{aligned}$$

since $f(x_{t-1}) \geqslant f(x^*)$ by the fact that x^* is a global optimum. To conclude, by unrolling the iteration we have

$$\|x_T - x^*\|^2 \leqslant (1-\gamma\mu)\|x_{T-1} - x^*\|^2 \leqslant \ldots \leqslant (1-\gamma\mu)^T\|x_0 - x^*\|^2.$$

□

Remark 2.2. It is possible to extend the result to $\gamma \in (0, 1/L)$, by using the tighter inequality (2) in Lemma 2 and the resulting tighter statement in the theorem above.

Corollary 2.4.

$$(1-\gamma\mu)^T\|x_0 - x^*\|^2 \leqslant e^{-\gamma\mu T}\|x_0 - x^*\|^2$$

By selecting $T = \frac{1}{\gamma\mu}\log(\|x_0 - x^*\|^2/\varepsilon)$, we have

$$\|x_T - x^*\|^2 \leqslant \varepsilon.$$

2.2 Stochastic Gradient Descent

In this section we deal with Stochastic Gradient Descent. This algorithm is useful to find the global minimum of convex functions of the form

$$f(x) = \mathbb{E}_\theta \; g(x, \theta), \quad \text{where} \quad \mathbb{E}_\theta \; g(x, \theta) = \int_\Omega g(x, \theta) dp(\theta),$$

p is a probability distribution over Ω and $g : \mathbb{R}^d \times \Omega \to \mathbb{R}$.

Example 2.1 (Empirical Risk Minimization)**.**

$$f(x) = \frac{1}{n} \sum_{i=1}^n L(x, \theta_i), \quad L(x, \theta_i) = \ell(x^\top z_i, y_i), \quad \theta_i = (z_i, y_i).$$

In this context we are not able to evaluate the integral above, but we assume that we are able to sample some θ_t from p and to compute the the gradient $\nabla_x g(x, \theta_i)$. The algorithm is as follows.

$$x_t = x_{t-1} - \gamma_t \nabla_x g(x, \theta_t), \quad \textit{where} \quad \theta_t \sim p,$$

where $\gamma_t > 0$ is a sequence of step-sizes, x_0 is given and θ_t is independently and identically distributed according to p. Note that, by linearity of the integral $\nabla f(x) = \mathbb{E}_\theta \nabla_x g(x, \theta)$, so

$$\begin{aligned} \mathbb{E}_{\theta_t} x_t &= x_{t-1} - \gamma_t \mathbb{E}_{\theta_t} \nabla_x g(x, \theta_t) \\ &= x_{t-1} - \gamma_t \nabla f(x_{t-1}). \end{aligned}$$

Then in expectation stochastic gradient descent seems to behave like gradient descent. Let's analyze this property more in detail with the following theorem. First we define the variance of the of the estimator of the gradient as

$$\sigma^2(x) = \mathbb{E}_\theta \|\nabla f(x) - \nabla_x g(x, \theta)\|^2.$$

Theorem 2.5. *Assuming that $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and with L-Lipschitz continuous gradients. Let $\gamma_t = \gamma$ for $t \in \mathbb{N}$, with $\gamma \in (0, 1/(2L))$. Assume that there exists σ^2 such that $\sigma^2(x) \leqslant \sigma^2$ for all $x \in \mathbb{R}^d$, we have that*

$$\mathbb{E}_{\theta_1, \dots, \theta_T} \|x_T - x^*\|^2 \leqslant (1 - \mu\gamma)^T \|x_0 - x^*\|^2 + \frac{\gamma}{\mu} \sigma^2.$$

Proof. Denote by ζ_t the random variable $\zeta_t = \nabla f(x_{t-1}) - \nabla_x g(x_{t-1}, \theta_t)$. Note that in particular, that

$$\mathbb{E}_{\theta_t} \zeta_t = 0,$$

since x_{t-1} does not depend on θ_t, moreover $\mathbb{E}_{\theta_t} \|\zeta_t\|^2 = \sigma^2(x_{t-1}) \leqslant \sigma^2$. Analogously to the proof for gradient descent we have

$$\begin{aligned} \mathbb{E}_{\theta_1, \dots, \theta_t} [\|x_t - x^*\|^2] &= \mathbb{E}_{\theta_1, \dots, \theta_t} \Big[\|x_{t-1} - x^* - \gamma_t \nabla f(x_{t-1}) + \gamma_t \zeta_t\|^2 \Big] \\ &= \mathbb{E}_{\theta_1, \dots, \theta_t} \Big[\|x_{t-1} - x^*\|^2 - 2\gamma_t \nabla f(x_{t-1})^\top (x_{t-1} - x^*) + \gamma_t^2 \|\nabla f(x_{t-1})\|^2 + \\ &\qquad + \; 2\gamma_t (x_{t-1} - x^* - \gamma_t \nabla f(x_{t-1}))^\top \zeta_t + \gamma_t^2 \|\zeta_t\|^2 \Big] \\ &= \mathbb{E}_{\theta_1, \dots, \theta_{t-1}} \Big[\|x_{t-1} - x^*\|^2 - 2\gamma_t \nabla f(x_{t-1})^\top (x_{t-1} - x^*) + \gamma_t^2 \|\nabla f(x_{t-1})\|^2 + \\ &\qquad + \; 2\gamma_t (x_{t-1} - x^* - \gamma_t \nabla f(x_{t-1}))^\top \mathbb{E}_{\theta_t} \zeta_t + \gamma_t^2 \mathbb{E}_{\theta_t} \|\zeta_t\|^2 \Big] \\ &= \mathbb{E}_{\theta_1, \dots, \theta_{t-1}} \Big[\|x_{t-1} - x^*\|^2 - 2\gamma_t \nabla f(x_{t-1})^\top (x_{t-1} - x^*) + \gamma_t^2 \|\nabla f(x_{t-1})\|^2 + \gamma_t^2 \mathbb{E}_{\theta_t} \|\zeta_t\|^2 \Big] \\ &\leqslant \mathbb{E}_{\theta_1, \dots, \theta_{t-1}} \Big[\|x_{t-1} - x^*\|^2 - 2\gamma_t \nabla f(x_{t-1})^\top (x_{t-1} - x^*) + \gamma_t^2 \|\nabla f(x_{t-1})\|^2 \Big] \; + \; \gamma_t^2 \sigma^2. \end{aligned}$$

Now by the same reasoning we did for the gradient descent algorithm, in particular bounding $-\nabla f(x_{t-1})^\top (x_{t-1} - x^*)$ with (7) and $\|\nabla f(x_{t-1})\|^2$ with (8), we obtain

$$\mathbb{E}_{\theta_1,\dots,\theta_{t-1}}\Big[\|x_{t-1}-x^*\|^2-2\gamma_t\nabla f(x_{t-1})^\top(x_{t-1}-x^*)+\gamma_t^2\|\nabla f(x_{t-1})\|^2\Big] \leqslant (1-\mu\gamma_t)\mathbb{E}_{\theta_1,\dots,\theta_{t-1}}\Big[\|x_{t-1}-x^*\|^2\Big],$$

when $\gamma_t \in (0, 1/(2L))$. Finally, by denoting with R_t the quantity

$$R_t := \mathbb{E}_{\theta_1,\dots,\theta_t}[\|x_t - x^*\|^2],$$

and $R_0 := \|x_0 - x^*\|^2$, we obtain a recursion that is

$$R_t = (1 - \mu\gamma_t)R_{t-1} + \gamma_t^2\sigma^2.$$

By selecting a fixed step-size $\gamma_t = \gamma$ and unfolding the recursion we obtain

$$R_t = (1 - \mu\gamma)^t R_0 + \gamma^2\sigma^2 \sum_{j=0}^{t-1}(1 - \mu\gamma)^j.$$

Now note that $\sum_{j=0}^{t-1}(1 - \mu\gamma)^j \leqslant \sum_{j=0}^{\infty}(1 - \mu\gamma)^j \leqslant 1/(\mu\gamma)$, then

$$R_T \leqslant (1 - \mu\gamma)^T\|x_0 - x^*\|^2 + \frac{\gamma}{\mu}\sigma^2.$$

□