DSCC/CSC 435 & ECE 412 Optimization for Machine Learning Lecture 9

Dual Methods

Lecturer: Jiaming Liang *October 10, 2024*

1 Dual proximal method

Consider the problem

$$\min\{f(x) + h(Ax) : x \in \mathbb{R}^n\}$$

where $A \in \mathbb{R}^{m\times n}$ and

- h is closed and convex;
- f is closed and μ-strongly convex;
- there exist $\hat{x} \in \mathrm{ri}\,(\mathrm{dom}\, f)$ and $\hat{z} \in \mathrm{ri}\,(\mathrm{dom}\, h)$ such that $A\hat{x} = \hat{z}$.

Strong duality holds in this case.

1.1 Dual problem

Consider an equivalent problem

$$\begin{aligned}&\min_{x,z\in\mathbb{R}^n} f(x) + h(z)\\ &\text{s.t. } Ax - z = 0.\end{aligned}$$

We define the Lagrangian as

$$L(x, z; y) = f(x) + h(z) - y^\top(Ax - z), \tag{1}$$

and the dual function is

$$\begin{aligned}d(y) &= \inf_{x,z} L(x, z; y)\\ &= \inf_x \left\{f(x) - y^\top Ax\right\} + \inf_z \left\{h(z) + y^\top z\right\}\\ &= -\sup_x \left\{(A^\top y)^\top x - f(x)\right\} - \sup_z \left\{(-y)^\top z - h(z)\right\}\\ &= -f^*(A^\top y) - h^*(-y),\end{aligned}$$

where f^* and h^* denote the conjugates of f and h, respectively. Thus, the dual problem is

$$\max_{y\in\mathbb{R}^n} d(y).$$

We consider the dual problem in its minimization form

$$\min_{y\in\mathbb{R}^m} F(y) + H(y) \tag{2}$$

where

$$F(y) = f^*(A^\top y), \quad H(y) = h^*(-y).$$

Example. Use Lagrangian to find the dual of the following linear programming

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & Ax = b \\ & Cx \le d. \end{array}$$

Associating dual variables $\lambda \ge 0$ and ν, the Lagrangian is

$$\begin{aligned} L(x,\lambda,\nu) &= c^\top x + \lambda^\top (Cx - d) + \nu^\top (Ax - b) \\ &= \left(c^\top + \lambda^\top C + \nu^\top A\right) x - d\lambda^\top - \nu^\top b, \end{aligned}$$

which is an affine function of x. It follows that the dual function is given by

$$d(\lambda,\nu) = \inf_x L(x,\lambda,\nu) = \begin{cases} -\lambda^\top d - \nu^\top b, & c + C^\top\lambda + A^\top\nu = 0 \\ -\infty, & \text{otherwise.} \end{cases}$$

The dual problem is

$$\begin{array}{ll} \text{maximize} & d(\lambda,\nu) \\ \text{subject to} & \lambda \ge 0. \end{array}$$

After making the implicit constraints explicit, we obtain

$$\begin{array}{ll} \text{maximize}_{\lambda,\nu} & -\lambda^\top d - \nu^\top b \\ \text{subject to} & c + C^\top\lambda + A^\top\nu = 0 \\ & \lambda \ge 0. \end{array}$$

Now, we have introduced the dual problem. Let us revisit the proximal gradient method studied in Lecture 6. Indeed, the method is a primal-dual method and provides a primal-dual convergence guarantee.

Theorem 1. *Consider the problem* $\min\{\phi(x) = f(x) + h(x) : x \in \mathbb{R}^n\}$ *in Lecture 5 and apply the proximal gradient mehtod, then we have the primal-dual gap bounded as follows,*

$$\hat{\phi}(\bar{x}_{k+1}) + f^*(\bar{y}_k) + \hat{h}^*(-\bar{y}_k) \le \frac{2d_0^2}{\lambda k},$$

where the notation is clear from the proof and skipped in the statement.

Proof. Recall from the proof of Theorem 5 in Lecture 6 that for every $x \in \operatorname{dom} h$,

$$\ell_f(x; x_k) + h(x) + \frac{1}{2\lambda}\|x - x_k\|^2 \geq \phi(x_{k+1}) + \frac{1}{2\lambda}\|x - x_{k+1}\|^2. \tag{3}$$

Taking $x = x_0^*$, which is the closest point in the solution set to x_0, and using the convexity of f, we have

$$\phi_* + \frac{1}{2\lambda}\|x_0^* - x_k\|^2 \geq \phi(x_{k+1}) + \frac{1}{2\lambda}\|x_0^* - x_{k+1}\|^2,$$

so

$$\|x_{k+1} - x_0^*\| \leq \|x_k - x_0^*\| \leq \|x_0 - x_0^*\| := d_0.$$

It follows from the triangle inequality that

$$\|x_k - x_0\| \leq \|x_k - x_0^*\| + \|x_0 - x_0^*\| \leq 2d_0.$$

We also have for every $x \in \mathbb{R}^n$,

$$\ell_f(x; x_k) \leq f(x), \quad y_k := \nabla f(x_k) = \nabla \ell_f(x; x_k).$$

Then, using Theorem 2 of Lecture 5, we have

$$\ell_f(x; x_k) = -(\ell_f(\cdot; x_k))^*(y_k) + \langle y_k, x\rangle \leq -f^*(y_k) + \langle y_k, x\rangle.$$

It thus follows from (3) that

$$\phi(x_{k+1}) + f^*(y_k) + \langle -y_k, x\rangle - h(x) \leq \frac{1}{2\lambda}\|x - x_k\|^2 - \frac{1}{2\lambda}\|x - x_{k+1}\|^2.$$

Averaging over the iterations and using convexity, we have

$$\phi(\bar{x}_{k+1}) + f^*(\bar{y}_k) + \langle -\bar{y}_k, x\rangle - h(x) \leq \frac{1}{2\lambda k}\|x - x_0\|^2.$$

Maximizing x over $B(x_0, 2d_0)$, we have

$$\phi(\bar{x}_{k+1}) + f^*(\bar{y}_k) + \max_{x \in B(x_0, 2d_0)}\{\langle -\bar{y}_k, x\rangle - h(x)\} \leq \frac{1}{2\lambda k}\max_{x \in B(x_0, 2d_0)}\|x - x_0\|^2 = \frac{2d_0^2}{\lambda k}.$$

Since

$$\max_{x \in B(x_0, 2d_0)}\{\langle -\bar{y}_k, x\rangle - h(x)\} = \max_{x \in \mathbb{R}^n}\{\langle -\bar{y}_k, x\rangle - \hat{h}(x)\} = \hat{h}^*(-\bar{y}_k)$$

where $\hat{h}(x) = h(x) + I_B(x)$,

$$\phi(\bar{x}_{k+1}) + f^*(\bar{y}_k) + \hat{h}^*(-\bar{y}_k) \leq \frac{2d_0^2}{\lambda k}.$$

Since $\bar{x}_{k+1} \in B(x_0, 2d_0)$,

$$\hat{\phi}(\bar{x}_{k+1}) + f^*(\bar{y}_k) + \hat{h}^*(-\bar{y}_k) \leq \frac{2d_0^2}{\lambda k}.$$

Note that $x_* \in B(x_0, 2d_0)$, to solve the problem $\min\{\phi(x) : x \in \mathbb{R}^n\}$, it suffices to solve

$$\min_{x \in \mathbb{R}^n}\{\hat{\phi}(x) = f(x) + \hat{h}(x)\}.$$

□

1.2 Dual proximal method

Let us go back to problem (2) and examine its properties.

Lemma 1. *We have F is convex and L_F-smooth where $L_F = \|A\|^2/\mu$ and H is closed and convex.*

Proof. Since f is μ-strongly convex, by conjugacy, we know f^* is $(1/\mu)$-smooth. Thus, for any $y_1, y_2 \in \mathbb{R}^m$, we have

$$\begin{aligned}\|\nabla F(y_1) - \nabla F(y_2)\| &= \|A\nabla f^*(A^\top y_1) - A\nabla f^*(A^\top y_2)\| \\ &\le \|A\|\|\nabla f^*(A^\top y_1) - \nabla f^*(A^\top y_2)\| \\ &\le \frac{\|A\|}{\mu}\|A^\top(y_1 - y_2)\| \\ &\le \frac{\|A\|^2}{\mu}\|y_1 - y_2\|.\end{aligned}$$

By conjugacy and the fact that convexity preserves under composition of a convex function and a linear mapping, we know both F and H are convex. □

Since the dual problem (2) is the sum of a convex smooth function $F(y)$ and a convex composite function $H(y)$, which has a proximal mapping because of the assumption on h and the Moreau decomposition theorem. This is exactly the setting for the proximal gradient method, we thus apply the method to (2).

Algorithm 1 Dual proximal method

Input: Initial point $y_0 \in \mathbb{R}^m$
for $k \ge 0$ **do**
  Compute $y_{k+1} = \mathrm{prox}_{\lambda H}(y_k - \lambda\nabla F(y_k))$.
end for

Since F is convex and L_F-smooth and H is closed and convex, invoking Theorem 5 of Lecture 6, we obtain the convergence rate of the dual sequence.

Theorem 2. *Choose $\lambda \in (0, 1/L_F]$. Then, Algorithm 1 generates a sequence of points $\{y_k\}$ satisfying*

$$d^* - d(y_k) \le \frac{\|y_0 - y^*\|^2}{2\lambda k}, \quad \forall k \ge 1.$$

Lemma 2. *The dual iteration $y_{k+1} = \mathrm{prox}_{\lambda H}(y_k - \lambda\nabla F(y_k))$ can be equivalently rewritten as*

$$x_{k+1} = \mathrm{argmax}_{x\in\mathbb{R}^n}\{\langle x, A^\top y_k\rangle - f(x)\}, \tag{4}$$

$$y_{k+1} = y_k - \lambda A x_{k+1} + \lambda\,\mathrm{prox}_{\frac{1}{\lambda}h}\left(Ax_{k+1} - \frac{1}{\lambda}y_k\right). \tag{5}$$

Proof. Note that the dual proximal update can be written as

$$y_{k+1} = \min_{y\in\mathbb{R}^m} \left\{ \ell_F(y; y_k) + H(y) + \frac{1}{2\lambda}\|y - y_k\|^2 \right\}.$$

Its optimality condition is

$$0 \in \nabla F(y_k) + \partial H(y_{k+1}) + \frac{y_{k+1} - y_k}{\lambda}. \tag{6}$$

It follows from Proposition 1 of Lecture 5 and (4) that

$$\nabla F(y_k) = A\nabla f^*(A^\top y_k) = A\,\mathrm{argmax}_x\{\langle A^\top y_k, x\rangle - f(x)\} \overset{(4)}{=} Ax_{k+1}.$$

Define

$$z_{k+1} = \frac{y_{k+1} - y_k}{\lambda} + \nabla F(y_k) = \frac{y_{k+1} - y_k}{\lambda} + Ax_{k+1}. \tag{7}$$

Then, it follows from the optimality condition (6) that

$$-z_{k+1} \in \partial H(y_{k+1}) = -\partial h^*(-y_{k+1}).$$

Using Theorem 2 of Lecture 5, we have

$$\partial h(z_{k+1}) \ni -y_{k+1}.$$

Hence,

$$0 \in y_{k+1} + \partial h(z_{k+1}).$$

Equivalently, by (7), we have

$$0 \in \partial h(z_{k+1}) + y_k + \lambda z_{k+1} - \lambda A x_{k+1}.$$

It is interesting to see that the above inclusion is also the optimality condition of

$$z_{k+1} = \mathrm{argmin}_{z\in\mathbb{R}^m} \left\{ h(z) + \langle z, y_k\rangle + \frac{\lambda}{2}\|z - Ax_{k+1}\|^2 \right\}.$$

Using Definition 1 of Lecture 6, we have

$$z_{k+1} = \mathrm{prox}_{\frac{1}{\lambda}h}\left(Ax_{k+1} - \frac{1}{\lambda}y_k \right).$$

Finally, it follows from (7) and the above formula for z_{k+1} that (5) holds. □

Note: The proof of Lemma 2 can be simplified using the extended Moreau decomposition. This alternative proof is left as a homework problem.

Using Lemma 2, we can rewrite Algorithm 1 in its primal form.

Algorithm 2 Dual proximal method (primal form)

Input: Initial point $y_0 \in \mathbb{R}^m$

for $k \geq 0$ **do**

Compute $x_{k+1} = \operatorname{argmax}_{x \in \mathbb{R}^n} \{\langle x, A^\top y_k \rangle - f(x)\}$.

Compute $y_{k+1} = y_k - \lambda A x_{k+1} + \lambda \operatorname{prox}_{\frac{1}{\lambda}h}\left(A x_{k+1} - \frac{1}{\lambda} y_k\right)$.

end for

It is clear from the proof of Lemma 2 that the dual proximal method has another presentation in the alternating minimization form using the z sequence.

Algorithm 3 Dual proximal method (alternating minimization form)

Input: Initial point $y_0 \in \mathbb{R}^m$

for $k \geq 0$ **do**

Compute $x_{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} \{f(x) - \langle x, A^\top y_k \rangle\}$.

Compute $z_{k+1} = \operatorname{argmin}_{z \in \mathbb{R}^m} \left\{h(z) + \langle z, y_k \rangle + \frac{\lambda}{2}\|z - A x_{k+1}\|^2\right\}$.

Compute $y_{k+1} = y_k - \lambda A x_{k+1} + \lambda z_{k+1}$.

end for

In fact, Algorithm 3 can be understood from the augmented Lagrangian perspective. Recall the Lagrange function $L(x, z; y)$ is defined in (1). We define the augmented Lagrange function as follows

$$L_\lambda(x, z; y) := L(x, z; y) + \frac{\lambda}{2}\|Ax - z\|^2 = f(x) + h(z) - y^\top (Ax - z) + \frac{\lambda}{2}\|Ax - z\|^2.$$

Then, we rewrite Algorithm 3 as

$$\begin{aligned}
x_{k+1} &= \operatorname{argmin}_{x \in \mathbb{R}^n} L(x, z_k; y_k), \\
z_{k+1} &= \operatorname{argmin}_{z \in \mathbb{R}^m} L_\lambda(x_{k+1}, z; y_k), \\
y_{k+1} &= y_k - \lambda(A x_{k+1} - z_{k+1}) = y_k + \lambda \nabla_y L(x_{k+1}, z_{k+1}; y).
\end{aligned}$$

With the understanding that the dual proximal method is the proximal gradient method applied to the dual problem in mind, we also note that the dual ascent method is the subgradient method applied to the dual problem, and the augmented Lagrangian method is the proximal point method applied to the dual problem.

2 Duality between Frank-Wolfe and mirror descent

We present a fascinating connection between Frank-Wolfe and mirror descent, that is, Frank-Wolfe applied to the dual problem is equivalent to mirror descent applied to the primal problem. We consider the following primal and dual problems.

Primal

$$\min_{x\in\mathbb{R}^n}\{\phi(x) := f(Ax) + h(x)\}$$

and dual

$$\max_{y\in C}\{\psi(y) := -h^*(-A^\top y) - f^*(y)\}.$$

Assume $f : \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuous everywhere and $h : \mathbb{R}^n \to (-\infty, \infty]$ is μ-strongly convex, and $A \in \mathbb{R}^{m\times n}$. This implies that $\operatorname{dom} f^*$ is bounded. Define

$$R = \max_{y_1,y_2\in\operatorname{dom} f^*} \|A^\top(y_1 - y_2)\|_* = \operatorname{diam}(A^\top \operatorname{dom} f^*). \tag{8}$$

Applying Frank-Wolf from Lecture 8 to the dual problem, we have the following dual Frank-Wolfe method.

Algorithm 4 Frank-Wolfe method for dual problem

Input: Initial point $y_0 \in \operatorname{dom} f^*$

for $k \geq 0$ **do**

Step 1. Compute $x_k = \operatorname{argmin}_{x\in\mathbb{R}^n}\{\langle x, A^\top y_k\rangle + h(x)\} = \nabla(h^*)(-A^\top y_k)$.

Step 2. Compute $\bar{y}_k \in \operatorname{Argmax}_{y\in C}\{\langle y, Ax_k\rangle - f^*(y)\} = \partial f(Ax_k)$.

Step 3. Choose $t_k \in [0, 1]$ and set $y_{k+1} = (1 - t_k)y_k + t_k\bar{y}_k$.

end for

Applying Theorem 1 of Lecture 8 directly gives the following convergence result for the dual problem.

Theorem 3. *For every $k \geq 1$, we have*

$$\psi^* - \psi(y_k) \leq \frac{2R^2}{\mu(k+1)}.$$

2.1 Mirror descent

Consider the primal problem

$$\min_{x\in\mathbb{R}^n}\{\phi(x) := f(Ax) + h(x)\},$$

we present the following special mirror descent method for the primal problem.

Algorithm 5 Mirror descent for primal problem

Input: Given $y_0 \in \operatorname{dom} f^*$, set initial point $x_0 = \nabla(h^*)(-A^\top y_0)$ and $h'(x_0) = -A^\top y_0$.

for $k \geq 0$ **do**

Step 1. Choose $t_k \in [0,1]$ and compute $x_{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^n} \left\{ \ell_\phi(x; x_k) + \frac{1}{t_k} D_h(x, x_k) \right\}$.

Step 2. Set $h'(x_{k+1}) = (1-t_k)h'(x_k) - t_k A^\top f'(Ax_k)$.

end for

Note that we linearize the whole primal function ϕ and use the μ-strongly convex function h as the distance generating function.

The following theorem show that the dual Frank-Wolfe method is equivalent to the above mirror descent method.

Theorem 4. *If both Algorithms 4 and 5 use the same subgradient oracle of f, i.e., $\bar{y}_k = f'(Ax_k)$ where $f'(Ax_k)$ is the one used in Step 1 of Algorithm 5, then given the same initial point $y_0 \in \operatorname{dom} f^*$, both algorithms generate same iterates $\{x_k\}$.*

Proof. It follows from Step 1 of Algorithm 5 that

$$0 \in t_k \left(A^\top f'(Ax_k) + h'(x_k) \right) + \partial h(x_{k+1}) - h'(x_k),$$

and hence that

$$0 \in -(1-t_k)h'(x_k) + t_k A^\top f'(Ax_k) + \partial h(x_{k+1}).$$

This is equivalent to

$$\partial h(x_{k+1}) \ni (1-t_k)h'(x_k) - t_k A^\top f'(Ax_k).$$

Using Theorem 3 of Lecture 3, we have

$$x_{k+1} \in \partial h^* \left((1-t_k)h'(x_k) - t_k A^\top f'(Ax_k) \right).$$

Since h is strongly convex, we know h^* is smooth and $\partial h^* = \nabla h^*$. This means x_{k+1} is unique

$$x_{k+1} = \nabla h^* \left((1-t_k)h'(x_k) - t_k A^\top f'(Ax_k) \right). \tag{9}$$

Next, we consider Algorithm 4 and prove that $-A^\top y_k$ from Algorithm 4 is equal to $h'(x_k)$ from Algorithm 5, i.e.,

$$-A^\top y_k = h'(x_k). \tag{10}$$

We prove this relation by induction. It clearly holds for $k = 0$ in view of the input of Algorithm 5. Suppose (10) holds for some $k \geq 0$. Then, it follows from Step 3 of Algorithm 4 and the assumption that $\bar{y}_k = f'(Ax_k)$ that

$$-A^\top y_{k+1} = -(1-t_k)A^\top y_k - t_k A^\top \bar{y}_k = (1-t_k)h'(x_k) - t_k A^\top f'(Ax_k) = h'(x_{k+1}),$$

where the last identity is due to Step 2 of Algorithm 5. Hence, we prove (10).

Now, using Step 2 of Algorithm 5 and (10), we conclude that (9) is equivalent to

$$x_{k+1} = \nabla h^* \left(-A^\top y_{k+1} \right),$$

which agrees with Step 1 of Algorithm 4. Therefore, we finally prove that dual Frank-Wolfe and mirror descent are equivalent. □

Convergence of Algorithm 5 is left as a homework problem.

3 Dual averaging

Recall from Lecture 5 that an iteration of the mirror descent reads as

$$x_{k+1} = \operatorname{argmin}_{x \in Q} \left\{ f(x_k) + \langle f'(x_k), x - x_k \rangle + \frac{1}{a_k} D_w(x, x_k) \right\}, \tag{11}$$

or $x_{k+1} = \operatorname{proj}_Q(y_{k+1})$ and y_{k+1} satisfies

$$\nabla w(y_{k+1}) = \nabla w(x_k) - a_k f'(x_k).$$

Nesterov's dual averaging method is a lazy version of mirror descent, where $x_{k+1} = \operatorname{proj}_Q(y_{k+1})$ and y_{k+1} satisfies

$$\nabla w(y_{k+1}) = \nabla w(y_k) - a_k f'(x_k), \quad \nabla w(y_0) = 0.$$

That is, different form mirror descent, which goes back and forth between primal and dual spaces, dual averaging simply averages the dual variables (i.e., gradients), and takes the inverse mirror map as in mirror descent only if asked for a point in the primal (i.e., y_{k+1}). Clearly,

$$\nabla w(y_{k+1}) = -\sum_{i=0}^{k} a_i f'(x_i).$$

Hence, dual averaging can be equivalently written as

$$x_{k+1} = \operatorname{argmin}_{x \in Q} \left\{ A_{k+1} \hat{\Gamma}_{k+1}(x) + w(x) \right\}, \tag{12}$$

where

$$\hat{\Gamma}_{k+1} = \frac{A_k}{A_{k+1}} \hat{\Gamma}_k + \frac{a_k}{A_{k+1}} \ell_f(x; x_k), \quad A_{k+1} = A_k + a_k, \quad \hat{\Gamma}_0 \equiv 0, \quad A_k = 0.$$

If $a_k = \lambda$ constanst stepsize, then dual averaging is

$$x_{k+1} = \operatorname{argmin}_{x \in Q} \left\{ \sum_{i=0}^{k} \ell_f(x; x_i) + \frac{1}{\lambda} w(x) \right\}.$$

To setup the stage, we consider the composite optimization problem $\min\{\phi(x) = f(x) + h(x)\}$ with the same assumptions on f and w as in Lecture 5 for mirror descent. We assume h is closed and convex.

Algorithm 6 Dual averaging

Input: Initial point $x_0 \in \operatorname{dom} h$, set $A_0 = 0$ and $\Gamma_0 \equiv 0$

for $k \geq 0$ **do**

Step 1. Choose a_k and compute $A_{k+1} = A_k + a_k$.

Step 2. Compute

$$x_{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} \left\{ A_{k+1}\Gamma_{k+1}(x) + w(x) \right\}, \tag{13}$$

where

$$\Gamma_{k+1} = \frac{A_k}{A_{k+1}}\Gamma_k + \frac{a_k}{A_{k+1}}\gamma_k, \quad \gamma_k = \ell_f(\cdot; x_k) + h. \tag{14}$$

end for

Lemma 3. *For every $k \geq 1$, we have*

$$A_k\phi(\bar{x}_k) - \frac{2M^2}{\rho}\sum_{i=1}^{k-1} a_i^2 + w_* \leq \min_{x \in \mathbb{R}^n} \left\{ A_k\Gamma_k(x) + w(x) \right\},$$

where $w_* = \min_{x \in \mathbb{R}^n} w(x)$, $\bar{x}_1 = x_1$, and for every $k \geq 1$,

$$\bar{x}_{k+1} = \frac{A_k}{A_{k+1}}\bar{x}_k + \frac{a_k}{A_{k+1}}x_{k+1}.$$

Proof. Proof by induction. Since $A_0 = 0$, the case $k = 1$ is trivial. Assume the claim is true for some $k \geq 1$. It follow form (13) and (14) that

$$\begin{aligned} A_{k+1}\Gamma_{k+1}(x_{k+1}) + w(x_{k+1}) &= A_k\Gamma_k(x_{k+1}) + a_k\gamma_k(x_{k+1}) + w(x_{k+1}) \\ &\geq A_k\Gamma_k(x_k) + w(x_k) + D_w(x_{k+1}, x_k) + a_k\gamma_k(x_{k+1}), \end{aligned}$$

where we also use the fact that w is "1-strongly convex" in D_w in the inequality. Using the induction hypothesis, (13), and the assumption that w is ρ-strongly convex in $\|\cdot\|_2$, we have

$$\begin{aligned} A_{k+1}\Gamma_{k+1}(x_{k+1}) + w(x_{k+1}) &\geq A_k\phi(\bar{x}_k) - \frac{2M^2}{\rho}\sum_{i=1}^{k-1} a_i^2 + w_* + \frac{\rho}{2}\|x_{k+1} - x_k\|^2 + a_k\gamma_k(x_{k+1}) \\ &= A_k\phi(\bar{x}_k) - \frac{2M^2}{\rho}\sum_{i=1}^{k-1} a_i^2 + w_* + a_k\left[\gamma_k(x_{k+1}) + \frac{\rho}{2a_k}\|x_{k+1} - x_k\|^2\right]. \end{aligned}$$

Since f is M-Lipschitz continuous, we know

$$\phi(x_{k+1}) - \gamma_k(x_{k+1}) = f(x_{k+1}) - \ell_f(x_{k+1}; x_k) \leq 2M\|x_{k+1} - x_k\|$$

and

$$\gamma_k(x_{k+1}) + \frac{\rho}{2a_k}\|x_{k+1} - x_k\|^2 \geq \phi(x_{k+1}) - 2M\|x_{k+1} - x_k\| + \frac{\rho}{2a_k}\|x_{k+1} - x_k\|^2 \geq \phi(x_{k+1}) - \frac{2a_k M^2}{\rho}.$$

Therefore, using the definition of $\bar{x}_{k+1}$ and the convexity of ϕ, we conclude that

$$\begin{aligned} A_{k+1}\Gamma_{k+1}(x_{k+1}) + w(x_{k+1}) &\geq A_k\phi(\bar{x}_k) - \frac{2M^2}{\rho}\sum_{i=1}^{k-1} a_i^2 + w_* + a_k\phi(x_{k+1}) - \frac{2a_k^2 M^2}{\rho} \\ &\geq A_{k+1}\phi(\bar{x}_{k+1}) - \frac{2M^2}{\rho}\sum_{i=1}^{k} a_i^2 + w_*, \end{aligned}$$

and hence that the claim for $k+1$ is proved. □

Theorem 5. *For every $k \geq 1$, we have*

$$\phi(\bar{x}_k) - \phi_* \leq \frac{w(x_*) - w_*}{2A_k} + \frac{2M^2}{\rho}\frac{\sum_{i=1}^{k-1} a_i^2}{A_k}.$$

Proof. Using Lemma 3, we have for every $x \in \mathbb{R}^n$,

$$A_k\phi(\bar{x}_k) - \frac{2M^2}{\rho}\sum_{i=1}^{k-1} a_i^2 + w_* \leq \min_{x\in\mathbb{R}^n}\{A_k\Gamma_k(x) + w(x)\} \leq A_k\Gamma_k(x) + w(x).$$

Taking $x = x_*$ gives

$$A_k\phi(\bar{x}_k) - \frac{2M^2}{\rho}\sum_{i=1}^{k-1} a_i^2 + w_* \leq A_k\Gamma_k(x_*) + w(x_*) \leq A_k\phi_* + w(x_*)$$

Therefore, the theorem is proved. □

If we take constant stepsize

$$a_k = \lambda, \quad A_k = k\lambda,$$

then

$$\phi(y_k) - \phi_* \leq \frac{w(x_*) - w_*}{2\lambda k} + \frac{2\lambda M^2}{\rho}.$$