Lecture 26

26.1 Contracting-Point Acceleration

We discuss a conceptual acceleration scheme that can be used to potentially accelerate *any* optimization algorithm, including sophisticated ones (e.g., stochastic methods, such as coordinate descent or methods with variance reduction, as well as second-order algorithms).

While direct acceleration (i.e., acceleration developed for a specific method) is usually preferable from a practical standpoint, the conceptual scheme that we will discuss is useful for quickly identifying the expected rate of convergence one should aim for, while remaining remarkably simple.

26.1.1 Problem Setup

We consider the minimization of a convex function $f : Q \to \mathbb{R}$ defined on an open convex set $Q \subseteq \mathbb{R}^n$:

$$\min_{x \in Q} f(x), \tag{26.1}$$

and we assume that a minimizer $x^\star$ exists.

Thus far, we do not assume any additional conditions on the objective, such as smoothness, although such conditions are often crucial for acceleration. Recall that the subgradient method is optimal for the black-box minimization of non-differentiable Lipschitz convex functions; therefore, acceleration is not possible for every combination of algorithm and problem class.

To solve (26.1), we fix a differentiable convex *regularizer* $d : Q \to \mathbb{R}$ and define the associated *Bregman divergence*:

$$\beta_d(x; y) \;\; := \;\; d(y) - d(x) - \langle \nabla d(x), y - x \rangle \;\; \geq \;\; 0,$$

which serves as a measure of the distance from y to x.

We use the following main fact about the Bregman divergence (see Lemma 16.2.3 in Lecture 16). For any convex $g : Q \to \mathbb{R}$, consider the solution to the regularized subproblem for a fixed center $v \in Q$:

$$v^+ \;\; := \;\; \arg\min_{y \in Q} \Big\{ g(y) + \beta_d(v; y) \Big\}$$

Then, it holds:

$$g(y) + \beta_d(v; y) \;\; \geq \;\; g(v^+) + \beta_d(v; v^+) + \beta_d(v^+; y), \qquad y \in Q. \tag{26.2}$$

Inequality (26.2) improves upon the trivial bound that holds by the definition of the minimum, due tp additional non-negative term $\beta_d(v^+; y) \geq 0$, which is very useful for the analysis.

26.1.2 Accelerated Scheme

In the accelerated method, we construct two sequences of points, both starting from some initialization $x_0 = v_0 \in Q$:

- An auxiliary sequence of *prox centers* $\{v_k\}_{k \geq 0}$,
- A sequence of *main iterates* $\{x_k\}_{k \geq 0}$.

We also have an increasing sequence of controlling parameters $\{A_k\}_{k \geq 0}$, starting from $A_0 = 0$. We denote the partial differences by:

$$a_{k+1} \;:=\; A_{k+1} - A_k \;>\; 0 \qquad \Leftrightarrow \qquad A_k \;=\; \sum_{i=1}^{k} a_i,$$

and the *contracting coefficients*:

$$\gamma_k \;:=\; \tfrac{a_{k+1}}{A_{k+1}} \;\in\; (0, 1].$$

Then, our goal is to ensure the following inequality, for any $k \geq 0$:

$$\beta_d(x_0; x) + A_k f(x) \;\geq\; \beta_d(v_k; x) + A_k f(x_k), \qquad x \in Q. \tag{26.3}$$

Note that plugging $x := x^\star$ into (26.3) and rearranging the terms, we obtain the following convergence rate:

$$f(x_k) - f^\star \;\leq\; \tfrac{\beta_d(x_0; x^\star)}{A_k}, \qquad k \geq 1, \tag{26.4}$$

and, therefore, we are interested to increase $A_k \to +\infty$ as fast as possible.

It is easy to check that inequality (26.3) holds for $k = 0$ due to our choices: $A_0 = 0$ and $x_0 = v_0$. Now, we assume that it holds for a current iteration $k \geq 0$ and see how we can propagate this inequality for the next iteration. We have,

$$\begin{aligned} \beta_d(x_0; x) + A_{k+1} f(x) \;&=\; \beta_d(x_0; x) + A_k f(x) + a_{k+1} f(x) \\ &\overset{(26.3)}{\geq}\; \beta_d(v_k; x) + A_k f(x_k) + a_{k+1} f(x) \\ &\geq\; \beta_d(v_k; x) + A_{k+1} f\big(\gamma_k x + (1 - \gamma_k) x_k\big), \end{aligned} \tag{26.5}$$

where in the last inequality we used convexity of f. Let us denote by v_{k+1} the minimum of the right-hand side of (26.5):

$$v_{k+1} \;:=\; \arg\min_{x \in Q} \Big\{ A_{k+1} f\big(\gamma_k x + (1 - \gamma_k) x_k\big) + \beta_d(v_k; x) \Big\}. \tag{26.6}$$

Applying the main Bregman divergence inequality (26.2), we obtain:

$$\begin{aligned} \beta_d(x_0; x) + A_{k+1} f(x) \;&\overset{(26.5)}{\geq}\; \beta_d(v_k; x) + A_{k+1} f\big(\gamma_k x + (1 - \gamma_k) x_k\big) \\ &\overset{(26.2)}{\geq}\; \beta_d(v_k; v_{k+1}) + A_{k+1} f\big(\gamma_k v_{k+1} + (1 - \gamma_k) x_k\big) + \beta_d(v_{k+1}; x) \\ &\geq\; \beta_d(v_{k+1}; x) + A_{k+1} f(x_{k+1}), \end{aligned}$$

where in the last inequality we dropped[1] the non-negative term $\beta_d(v_k; v_{k+1}) \geq 0$, and set the next main iterate as

$$x_{k+1} \;:=\; \gamma v_{k+1} + (1-\gamma_k) x_k.$$

Therefore, we established (26.3) for the next iteration, and thus proved it by induction for all $k \geq 0$.

26.1.3 Algorithm

We can write down these iterations in algorithmic form.

Algorithm 26.1: *Contracting-Point Scheme for Acceleration.*

Initialization: $x_0 \in \mathbb{R}^n$. Choose regularizer $d(\cdot)$. Set $v_0 = x_0$ and $A_0 = 0$. Fix $K \geq 1$.

For $k = 0 \dots K-1$ **iterate:**

1. Choose a new coefficient $a_{k+1} > 0$. Set $A_{k+1} := A_k + a_{k+1}$ and $\gamma_k := \frac{a_{k+1}}{A_{k+1}}$
2. Form the contracted objective with Bregman regularization:

$$h_k(x) \;:=\; A_{k+1} f\big(\gamma_k x + (1-\gamma_k) x_k\big) + \beta_d(v_k; x)$$

3. Compute

$$v_{k+1} \;\approx\; \arg\min_{x \in Q} h_k(x)$$

4. Set a new point from the triangle rule: $x_{k+1} := \gamma_k v_{k+1} + (1-\gamma_k) x_k$

Return x_K

From our previous reasoning, we obtain the following convergence result.

Theorem 26.1.1. *Let v_{k+1} be the exact minimizer of $h_k(\cdot)$. Then, we have*

$$f(x_k) - f^\star \;\leq\; \tfrac{\beta_d(x_0; x^\star)}{A_k}, \qquad k \geq 1. \tag{26.7}$$

A similar convergence rate can be established when v_{k+1} is an approximate minimizer of $h_k(\cdot)$ with sufficient accuracy.

Note that the classic fast gradient method can be viewed as an instance of this scheme, where we use the Euclidean prox function $d(x) = \frac{1}{2}\|x\|_2^2$ and, in Step 3, additionally linearize the contracted objective $f(\gamma_k x + (1-\gamma_k)x_k)$ around the point v_k:

$$\begin{aligned} h_k(x) \;&=\; A_{k+1} f\big(\gamma_k x + (1-\gamma_k) x_k\big) + \tfrac{1}{2}\|x - v_k\|^2 \\ &\approx\; A_{k+1}\Big[f(y_k) + \gamma_k \langle \nabla f(y_k), x - v_k\rangle\Big] + \tfrac{1}{2}\|x - v_k\|^2, \end{aligned} \tag{26.8}$$

where $y_k := \gamma_k v_k + (1-\gamma_k) x_k$. In the fast gradient method, we then set v_{k+1} to be the minimizer of the right-hand side of (26.8).

[1] To obtain the fastest possible rate, it is actually better to keep all the terms, which we omit here for simplicity.

26.2 Example: Gradient Methods

Let us consider an example of using the general contracting-point scheme to accelerate the basic gradient method.

The first crucial choice is to fix the regularizer $d(x)$. The simplest one is the square of the Euclidean norm:

$$d(x) \;:=\; \tfrac{1}{2}\|x\|_2^2,$$

which makes the Bregman divergence to be:

$$\beta_d(x;y) \;=\; \tfrac{1}{2}\|y-x\|_2^2.$$

At step 3 of the algorithm, we need to solve the following subproblem:

$$\min_{x\in Q}\Big\{\; h_k(x) \;:=\; g_k(x) + \tfrac{1}{2}\|x-v_k\|_2^2 \;\Big\}, \tag{26.9}$$

where

$$g_k(x) \;:=\; A_{k+1} f\big(\gamma_k x + (1-\gamma_k)x_k\big).$$

it the *contracted* objective, which gives the name to the whole scheme. Notice that

$$\nabla h_k(x) \;=\; a_{k+1}\nabla f\big(\gamma_k x + (1-\gamma_k)x_k\big) + (x-v_k),$$

$$\nabla^2 h_k(x) \;=\; \tfrac{a_{k+1}^2}{A_{k+1}}\nabla^2 f\big(\gamma_k x + (1-\gamma_k)x_k\big) + I.$$

Now, assume that f has the Lipschitz continuous gradient with constant L_f, with respect to the Euclidean norm. Then,

$$I \;\preceq\; \nabla^2 h_k(x) \;\preceq\; \Big(\tfrac{a_{k+1}^2}{A_{k+1}}L_f + 1\Big)I,$$

and we conclude that $h_k(\cdot)$ is *strongly convex* with parameter $\mu_k := 1$ and it has the Lipschitz gradient with parameter $L_k := \frac{a_{k+1}^2}{A_{k+1}}L_f + 1$.

We know that the basic gradient method, as applied to (26.9), will then exhibit a linear rate of convergence, and the main complexity factor will be the *condition number*:

$$\tfrac{L_k}{\mu_k} \;=\; \tfrac{a_{k+1}^2}{A_{k+1}}L_f + 1. \tag{26.10}$$

Therefore, by choosing a_{k+1} in a smart manner we can ensure that the condition number (26.10) is an absolute constant. For example, we can find a_{k+1} from the quadratic equation:

$$\boxed{\tfrac{a_{k+1}^2}{A_{k+1}} \;=\; \tfrac{1}{L_f},} \tag{26.11}$$

which makes $\frac{L_k}{\mu_k} = 2$. Note that equation (26.11) is exactly the one we used in deriving the rate of the fast gradient method (Lecture 8), which leads to the following rate of growth of the controlling coefficients:

$$A_k \;\geq\; \tfrac{k^2}{4L_f}. \tag{26.12}$$

We conclude that the resulting contracting-point scheme will have the accelerated optimal rate:

$$f(x_k) - f^\star \;\overset{(26.7),(26.12)}{\leq}\; \tfrac{2L_f\|x_0-x^\star\|_2^2}{k^2}, \qquad k \geq 1, \tag{26.13}$$

while each subproblem in Step 3 can be solved in $\tilde{O}(1)$ iterations of the gradient method, where $\tilde{O}(\cdot)$ notation hides logarithmic factors.

Compared to direct acceleration, the fast gradient method performs exactly one gradient step per iteration, achieving the same optimal rate (26.13). An extra logarithmic factor seems to be a reasonable price to pay for the generality. Utilizing the same reasoning, we can obtain acceleration for second-order methods.