DSCC/CSC 435 & ECE 412 Optimization for Machine Learning — Lecture 13

Quasi-Newton Methods

Lecturer: Jiaming Liang — *November 21, 2024*

1 Proximal point method

We are interested in solving

$$\min\{\phi(x) := f(x) + h(x)\}$$

- h is closed and convex;
- f is closed and convex, $\operatorname{dom} h \subseteq \operatorname{dom} f$;
- the optimal set X_* is nonempty.

Algorithm 1 Proximal point method

```
  **Input:** Initial point $x_0 \in \operatorname{dom} h$ and constant stepsize $\lambda > 0$
  **for** $k \geq 0$ **do**
    Solve $x_{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^n}\{\phi(x) + \frac{1}{2\lambda}\|x - x_k\|^2\}$.
  **end for**
```

Theorem 1.

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

Proof. It follows from the optimality of x_{k+1} that for every $x \in \operatorname{dom} h$,

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

Taking $x = x_k$, we have

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

and hence this is a descent method. Taking $x = x_*$, we have

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

Rearranging the above inequality, we obtain

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

Summing the resulting inequality and using the descent property, we have

$$k[\phi(x_k)-\phi_*] \le \sum_{i=1}^{k}[\phi(x_i)-\phi_*] \le \frac{1}{2\lambda}\|x_0-x_*\|^2 - \frac{1}{2\lambda}\|x_k-x_*\|^2 \le \frac{1}{2\lambda}\|x_0-x_*\|^2.$$

Therefore, the conclusion of the theorem follows. □

2 Inexact proximal point framework

The proximal point method is more conceptual than practical. In practice, we usually design algorithms to approximate the solution x_{k+1} to the proximal subproblem. Algorithms solving the proximal subproblem approximately can be described and analyzed under the inexact proximal point (IPP) framework.

2.1 Algorithm

Algorithm 2 Inexact proximal point framework

```
  Input: Initial point $x_0 \in \operatorname{dom} h$ and scalar $\sigma \in (0,1]$
  for $k \ge 1$ do
    Step 1. Choose $\lambda_k > 0$ and $\delta_k \ge 0$.
    Step 2. Compute $(x_k, \tilde{x}_k, \varepsilon_k)$ such that
        $\frac{x_{k-1}-x_k}{\lambda_k} \in \partial_{\varepsilon_k}\phi(\tilde{x}_k),$   (1)
        $\|x_k-\tilde{x}_k\|^2 + 2\lambda_k\varepsilon_k \le \sigma\|\tilde{x}_k - x_{k-1}\|^2 + \delta_k.$   (2)
  end for
```

The inclusion (1) in the IPP framework means

$$\tilde{v}_k = \frac{\tilde{x}_k - x_k}{\lambda_k} \in \partial_{\varepsilon_k}\left(\phi(\cdot) + \frac{1}{2\lambda_k}\|\cdot - x_{k-1}\|^2\right)(\tilde{x}_k).$$

In contrast to the PPM, the above inclusion provides two relaxations $\tilde{v}_k$ and ε_k. If both $\tilde{v}_k = 0$ (i.e., $\tilde{x}_k = x_k$) and $\varepsilon_k = 0$, then

$$0 \in \partial\left(\phi(\cdot) + \frac{1}{2\lambda_k}\|\cdot - x_{k-1}\|^2\right)(x_k),$$

i.e., the proximal problem is solved exactly

$$x_k = \operatorname{argmin}_{x\in\mathbb{R}^n}\left\{\phi(x) + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2\right\}.$$

Moreover, the inequality (2) is automatically satisfied

$$\|x_k - \tilde{x}_k\|^2 + 2\lambda_k\varepsilon_k = \|\tilde{v}_k\|^2 + 2\lambda_k\varepsilon_k = 0 \le \sigma\|\tilde{x}_k - x_{k-1}\|^2 + \delta_k.$$

Hence, the IPP framework becomes the PPM.

2.2 Proximal gradient method as an example

In this subsection, we assume f is L-smooth, then we show that the proximal gradient (PG) method with stepsize $\lambda_k \le \sigma/L$ for some $\sigma \in (0,1]$ is an instance of the IPP framework. We begin with an iteration of the PG method

$$x_k = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\ell_f(x;x_{k-1}) + h(x) + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2\right\}. \tag{3}$$

The optimality condition is

$$\frac{x_{k-1} - x_k}{\lambda_k} \in \partial[\ell_f(\cdot;x_{k-1}) + h(\cdot)](x_k),$$

which means for every $x \in \operatorname{dom} h$,

$$\begin{aligned}\phi(x) &\ge \ell_f(x;x_{k-1}) + h(x)\\ &\ge \ell_f(x_k;x_{k-1}) + h(x_k) + \frac{1}{\lambda_k}\langle x_{k-1} - x_k, x - x_k\rangle\\ &= \phi(x_k) + \frac{1}{\lambda_k}\langle x_{k-1} - x_k, x - x_k\rangle - \varepsilon_k\end{aligned}$$

where the first inequality is due to the convexity of f and ε_k is defined as

$$\varepsilon_k := f(x_k) - \ell_f(x_k;x_{k-1}).$$

Hence, PG satisfies the inclusion (1) of IPP with

$$\tilde{x}_k = x_k, \quad \varepsilon_k = f(x_k) - \ell_f(x_k;x_{k-1}), \quad \delta_k = 0.$$

Moreover, it follows from the assumption that f is L-smooth that

$$\varepsilon_k = f(x_k) - \ell_f(x_k;x_{k-1}) \le \frac{L}{2}\|x_k - x_{k-1}\|^2 = \frac{L}{2}\|\tilde{x}_k - x_{k-1}\|^2 \le \frac{\sigma}{2\lambda_k}\|\tilde{x}_k - x_{k-1}\|^2.$$

Hence, the inequality (2) of IPP is also satisfied. Now, we have shown PG is an instance of IPP.

Next, let us show the convergence of PG using the general convergence guarantee of IPP. It follows from (3) and L-smoothness of f that

$$\begin{aligned}\phi(x_{k-1}) &\ge \ell_f(x_k;x_{k-1}) + h(x_k) + \frac{1}{\lambda_k}\|x_k - x_{k-1}\|^2\\ &\ge \phi(x_k) + \left(\frac{1}{\lambda_k} - \frac{L}{2}\right)\|x_k - x_{k-1}\|^2\\ &\ge \phi(x_k) + \frac{2-\sigma}{2\lambda_k}\|x_k - x_{k-1}\|^2 \ge \phi(x_k) + \frac{1}{2\lambda_k}\|x_k - x_{k-1}\|^2,\end{aligned}$$

where the second last inequality is due to $\lambda_k \le \sigma/L$ and the last inequality is due to $\sigma \le 1$.

Other examples

For simplicity, we consider the case $h \equiv 0$ and f is differentiable. The examples below could be extended to the versions with h.

(1) if f is μ-strongly convex, approximating f by $\ell_f(x; x_{k-1}) + \frac{\mu}{2}\|x - x_{k-1}\|^2$, then

$$\begin{aligned}\tilde{x}_k = x_k &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \ell_f(x; x_{k-1}) + \frac{\mu}{2}\|x - x_{k-1}\|^2 + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2 \right\} \\ &= x_{k-1} - \frac{\lambda_k}{1 + \lambda_k \mu} \nabla f(x_{k-1});\end{aligned}$$

(2) the extragradient method: approximating f by $\ell_f(x; x_{k-1})$ and $\ell_f(x; \tilde{x}_k)$, then

$$\begin{aligned}\tilde{x}_k &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \ell_f(x; x_{k-1}) + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2 \right\} = x_{k-1} - \lambda_k \nabla f(x_{k-1}), \\ x_k &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \ell_f(x; \tilde{x}_k) + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2 \right\} = x_{k-1} - \lambda_k \nabla f(\tilde{x}_k);\end{aligned}$$

(3) Newton's method: if f is twice differentiable, approximating f by

$$q_f(x; x_{k-1}) = \ell_f(x; x_{k-1}) + \frac{1}{2}\|x - x_{k-1}\|^2_{\nabla^2 f(x_{k-1})}$$

and removing the quadratic term form (3), then

$$x_k = \operatorname*{argmin}_{x \in \mathbb{R}^n} q_f(x; x_{k-1}) = x_{k-1} - \left[\nabla^2 f(x_{k-1})\right]^{-1} \nabla f(x_{k-1});$$

(4) regularized Newton's method: if f is twice differentiable, approximating f by $q_f(x; x_{k-1})$, then

$$\begin{aligned}x_k &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ q_f(x; x_{k-1}) + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2 \right\} \\ &= x_{k-1} - \left[\nabla^2 f(x_{k-1}) + \frac{1}{\lambda_k} I\right]^{-1} \nabla f(x_{k-1}) \\ &= x_{k-1} - \left[\lambda_k \nabla^2 f(x_{k-1}) + I\right]^{-1} \lambda_k \nabla f(x_{k-1});\end{aligned}$$

if we consider $\phi = f + h$, then the regularized Newton's method generalizes to the proximal Newton's method;

(5) Newton proximal extragradient: if f is twice differentiable, approximating f by $q_f(x; x_{k-1})$ and $\ell_f(x; \tilde{x}_k)$, then

$$\tilde{x}_k = \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ q_f(x; x_{k-1}) + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2 \right\} = x_{k-1} - \left[\lambda_k \nabla^2 f(x_{k-1}) + I\right]^{-1} \lambda_k \nabla f(x_{k-1}),$$

$$x_k = \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ \ell_f(x; \tilde{x}_k) + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2 \right\} = x_{k-1} - \lambda_k \nabla f(\tilde{x}_k);$$

if we consider $\phi = f + h$, then the regularized Newton's method generalizes to the proximal Newton's method;

(6) quasi Newton proximal extragradient: if f is twice differentiable and μ-strongly convex, approximating f by

$$\tilde{q}_f(x; x_{k-1}) = \ell_f(x; x_{k-1}) + \frac{1}{2}\|x - x_{k-1}\|_{B_k}^2$$

and $\ell_f(x; \tilde{x}_k) + \frac{\mu}{2}\|x - x_{k-1}\|^2$, then

$$\begin{aligned} \tilde{x}_k &= \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ \tilde{q}_f(x; x_{k-1}) + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2 \right\} \\ &= x_{k-1} - \left[\lambda_k B_k + I\right]^{-1} \lambda_k \nabla f(x_{k-1}), \\ x_k &= \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ \ell_f(x; \tilde{x}_k) + \frac{\mu}{2}\|x - x_{k-1}\|^2 + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2 \right\} \\ &= \frac{1}{1+\lambda_k\mu}[x_{k-1} - \lambda_k \nabla f(\tilde{x}_k)] + \frac{\lambda_k \mu}{1+\lambda_k\mu}. \end{aligned}$$

3 Quasi Newton proximal extragradient

Consider $\min_{x\in\mathbb{R}^n} f(x)$ where f is closed, convex, μ-strongly convex and L_1-smooth, and assume that $\nabla^2 f$ is L_2-Lipschitz continuous, i.e.,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L_2\|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

Algorithm 3 Quasi Newton proximal extragradient

```
Input: Initial point $x_0 \in \mathbb{R}^n$, scalars $\alpha_1 \in [0,1)$, $\alpha_2 \in (0,1)$ satisfying $\alpha_1 + \alpha_2 < 1$
for $k \geq 1$ do
  Step 1. Given $B_k$ and $x_{k-1}$, find the stepsize $\lambda_k$ and $\tilde{x}_k$ such that
      $\|\tilde{x}_k - x_{k-1} + \lambda_k \left(\nabla f(x_{k-1}) + B_k(\tilde{x}_k - x_{k-1})\right)\| \leq \alpha_1 \|\tilde{x}_k - x_{k-1}\|,$   (4)
      $\lambda_k \|\nabla f(\tilde{x}_k) - \nabla f(x_{k-1}) - B_k(\tilde{x}_k - x_{k-1})\| \leq \alpha_2 \|\tilde{x}_k - x_{k-1}\|;$   (5)
  Step 2. Set
      $x_k = \frac{1}{1+\lambda_k \mu}[x_{k-1} - \lambda_k \nabla f(\tilde{x}_k)] + \frac{\lambda_k \mu}{1+\lambda_k \mu}\tilde{x}_k;$   (6)
  Step 3. Update $B_{k+1}$ using a subroutine.
end for
```

Proposition 1. *Let $\{x_k\}$ be the iterates generated by Algorithm 3 then we have for every $k \geq 1$,*

$$\|x_k - x_*\|^2 \leq \frac{\|x_{k-1} - x_*\|^2}{1+\lambda_k \mu}. \tag{7}$$

Proof. To begin with, using the triangle inequality and conditions (4) and (5), we have

$$\begin{aligned}
&\|\tilde{x}_k - x_{k-1} + \lambda_k \nabla f(\tilde{x}_k)\| \\
\leq &\|\tilde{x}_k - x_{k-1} + \lambda_k \left(\nabla f(x_{k-1}) + B_k(\tilde{x}_k - x_{k-1})\right)\| + \lambda_k \|\nabla f(\tilde{x}_k) - \nabla f(x_{k-1}) - B_k(\tilde{x}_k - x_{k-1})\| \\
\leq &(\alpha_1 + \alpha_2)\|\tilde{x}_k - x_{k-1}\| = \alpha \|\tilde{x}_k - x_{k-1}\|.
\end{aligned} \tag{8}$$

It is easy to observe that

$$\langle x_{k-1} - \tilde{x}_k, \tilde{x}_k - x \rangle = \frac{1}{2}\|x_{k-1} - x\|^2 - \frac{1}{2}\|x_{k-1} - \tilde{x}_k\|^2 - \frac{1}{2}\|\tilde{x}_k - x\|^2. \tag{9}$$

Using the Cauchy-Schwarz inequality and (8) and (9), we have for every $x \in \mathbb{R}^n$,

$$\begin{aligned}
\lambda_k \langle \nabla f(\tilde{x}_k), \tilde{x}_k - x \rangle &= \langle \tilde{x}_k - x_{k-1} + \lambda_k \nabla f(\tilde{x}_k), \tilde{x}_k - x \rangle + \langle x_{k-1} - \tilde{x}_k, \tilde{x}_k - x \rangle \\
&\leq \|\tilde{x}_k - x_{k-1} + \lambda_k \nabla f(\tilde{x}_k)\| \|\tilde{x}_k - x\| + \langle x_{k-1} - \tilde{x}_k, \tilde{x}_k - x \rangle \\
&\leq \alpha \|\tilde{x}_k - x_{k-1}\| \|\tilde{x}_k - x\| + \frac{1}{2}\|x_{k-1} - x\|^2 - \frac{1}{2}\|x_{k-1} - \tilde{x}_k\|^2 - \frac{1}{2}\|\tilde{x}_k - x\|^2 \\
&\leq \frac{1}{2}\|x_{k-1} - x\|^2 - \frac{1-\alpha}{2}\|x_{k-1} - \tilde{x}_k\|^2 - \frac{1-\alpha}{2}\|\tilde{x}_k - x\|^2.
\end{aligned} \tag{10}$$

It follows from (6) that for every $x \in \mathbb{R}^n$,

$$\begin{aligned}
&\lambda_k \langle \nabla f(\tilde{x}_k), x_k - x \rangle = \langle x_{k-1} - x_k, x_k - x \rangle + \lambda_k \mu \langle \tilde{x}_k - x_k, x_k - x \rangle \\
=&\frac{1}{2}\|x_{k-1} - x\|^2 - \frac{1}{2}\|x_{k-1} - x_k\|^2 - \frac{1+\lambda_k \mu}{2}\|x_k - x\|^2 + \frac{\lambda_k \mu}{2}\|\tilde{x}_k - x\|^2 - \frac{\lambda_k \mu}{2}\|\tilde{x}_k - x_k\|^2.
\end{aligned} \tag{11}$$

Combining (10) with $x = x_k$ and (11) with $x = x_*$, we obtain

$$\begin{aligned}\lambda_k\langle\nabla f(\tilde{x}_k), \tilde{x}_k - x_*\rangle &= \lambda_k\langle\nabla f(\tilde{x}_k), \tilde{x}_k - x_k\rangle + \lambda_k\langle\nabla f(\tilde{x}_k), x_k - x_*\rangle \\ &\le -\frac{1-\alpha}{2}\|x_{k-1} - \tilde{x}_k\|^2 + \frac{1}{2}\|x_{k-1} - x_*\|^2 - \frac{1+\lambda_k\mu}{2}\|x_k - x_*\|^2 \\ &\quad + \frac{\lambda_k\mu}{2}\|\tilde{x}_k - x_*\|^2 - \left(\frac{1-\alpha}{2} + \frac{\lambda_k\mu}{2}\right)\|\tilde{x}_k - x_k\|^2. \end{aligned} \tag{12}$$

Since f is μ-strongly convex, it follows from Lemma 4 of Lecture 3 that

$$\langle\nabla f(\tilde{x}_k), \tilde{x}_k - x_*\rangle = \langle\nabla f(\tilde{x}_k) - \nabla f(x_*), \tilde{x}_k - x_*\rangle \ge \mu\|\tilde{x}_k - x_*\|^2.$$

Applying the above inequality to the left-hand side of (12), we have

$$\begin{aligned}\frac{1+\lambda_k\mu}{2}\|x_k - x_*\|^2 &\le \frac{1}{2}\|x_{k-1} - x_*\|^2 - \frac{1-\alpha}{2}\|x_{k-1} - \tilde{x}_k\|^2 - \left(\frac{1-\alpha}{2} + \frac{\lambda_k\mu}{2}\right)\|\tilde{x}_k - x_k\|^2 \\ &\le \frac{1}{2}\|x_{k-1} - x_*\|^2 - \frac{1-\alpha}{2}\|x_{k-1} - \tilde{x}_k\|^2.\end{aligned}$$

Hence, (7) immediately follows. □

Lemma 1. *For every $k \ge 1$, $\lambda_k \ge 1/(8L_1)$.*

Theorem 2. *Under mild conditions, we have for every $k \ge 1$*

(a) linear convergence

$$\frac{\|x_k - x_*\|}{\|x_{k-1} - x_*\|} \le \left(1 + \frac{\mu}{8L_1}\right)^{-1};$$

(b) superlinear convergence

$$\lim_{k\to\infty}\frac{\|x_k - x_*\|}{\|x_{k-1} - x_*\|} = 0;$$

moreover, for every $k \ge 1$

$$\frac{\|x_k - x_*\|}{\|x_0 - x_*\|} \le \left(1 + \frac{3}{16}\mu\sqrt{\frac{k}{L_1^2 + 36\|B_0 - \nabla^2 f(x_*)\|^2 + \left(27 + \frac{32L_1}{\mu}\right)L_2^2\|x_0 - x_*\|^2}}\right)^{-k}.$$

Proof. (a) This case immediately follows from Proposition 1 and Lemma 1.

(b) Noting that $x \mapsto \log(1 + x^{-1})$ is convex, using the Jensen's inequality and the Cauchy-Schwarz inequality, we have

$$\frac{\|x_k - x_*\|}{\|x_0 - x_*\|} \le \prod_{i=1}^k \frac{\|x_i - x_*\|}{\|x_{i-1} - x_*\|} \le \prod_{i=1}^k \frac{1}{1 + \lambda_i\mu} \le \left(1 + \mu\sqrt{\frac{k}{\sum_{i=1}^k 1/\lambda_i^2}}\right)^{-k}.$$

To obtain the superlinear convergence of QNPE, it suffices to bound $\sum_{i=1}^{k} 1/\lambda_i^2$. A technical result shows that

$$\sum_{i=1}^{k} \frac{1}{\lambda_i^2} \leq \frac{1}{\left(1-\beta^2\right)\sigma_0^2} + \frac{1}{\left(1-\beta^2\right)\alpha_2^2\beta^2} \sum_{i\in\mathcal{B}} \frac{\left\|y_i - B_i s_i\right\|^2}{\left\|s_i\right\|^2},$$

where $y_i = \nabla f(\tilde{x}_i) - \nabla f(x_{k-1})$ and $s_i = \tilde{x}_i - x_{i-1}$. Indeed, defining the loss function at iteration k

$$\ell_k(B) = \begin{cases} 0, & \text{if } k \notin B \\ \frac{\|y_k - Bs_k\|^2}{2\|s_k\|^2}, & \text{otherwise,} \end{cases}$$

then the process of finding B_{k+1} can be viewed as an online optimization problem. Hence, the subroutine step 3 can be any online optimization method, e.g., FTRL. The bound on $\sum_{i=1}^{k} 1/\lambda_i^2$ is controlled by the regret bound of FTRL. We skipped the details here and recommend the interested readers to refer to the paper “Online Learning Guided Curvature Approximation: A Quasi-Newton Method with Global Non-Asymptotic Superlinear Convergence”. □