Disclaimer

- This is my presentation of the final project from the course Convex and Nonsmooth Optimization taught by Dr. Michael L. Overton, Silver Professor of Computer Science and Mathematics, at Courant Institute, NYU.
- This presentation has materials directedly cited from [Dru17] and class slides from Dr. Vandenberghe and Dr. Ang, for which all rights are reserved by their respective owners. The content is presented for educational purposes only and I, Rex Liu, do not claim any ownership over the material.

The Proximal Point Method

Class Project, Convex and Nonsmooth Optimization

Rex Liu, cl5682@nyu.edu

May 9, 2024

Introduction

To begin with[1], we want to minimize a closed[2], proper[3], convex, and possibly non-smooth function f, where the gradient descent does not apply.
In class we have seen the subgradient method. There is another way considering the *proximal operator*

$$\mathrm{prox}_f(v) := \arg\min_x \left(f(x) + \frac{1}{2}\|x - v\|^2 \right)$$
$$\mathrm{prox}_{\lambda f}(v) := \arg\min_x \left(f(x) + \frac{1}{2\lambda}\|x - v\|^2 \right)$$

called the *proximal point method*.

[1] In fact, f can be nonconvex in many cases, and weakly convex is sufficient.
[2] The epigraph of f is a closed set (and iff f is lower semicontinuous)
[3] $f : X \to \overline{\mathbb{R}}$ where $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$; in other words, f never attains the value $-\infty$ and its effective domain is nonempty

Setup

Given an iterate x_t, the method defines x_{t+1} to be any minimizer of the proximal subproblem

$$\arg\min_x \left(f(x) + \frac{1}{2\lambda}\|x - x_t\|^2 \right)$$

for an appropriately chosen parameter $\lambda > 0$. That is,

$$\text{choose } x_{t+1} \in \text{prox}_{\lambda f}(x_t)$$

The addition of the quadratic penalty term $\frac{1}{2\lambda}\|x - v\|^2$ often regularizes the subproblems and makes them well-conditioned. It can have larger strong convexity parameter thereby guaranteeing a unique solution for each subproblem regardless of the smoothness of f, facilitating faster numerical methods. ([Dru17])

Convergence Proof

By definition, consider

$$\mathbf{x}_{k+1} = \operatorname{prox}_f(\mathbf{x}_k) = \arg\min_{\mathbf{u}} \left(f(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}_k\|_2^2 \right)$$

From the subgradient first-order optimality condition:

$$\mathbf{0} \in \partial f(\mathbf{x}_{k+1}) + \mathbf{x}_{k+1} - \mathbf{x}_k \implies -(\mathbf{x}_{k+1} - \mathbf{x}_k) \in \partial f(\mathbf{x}_{k+1})$$

Since we assume for simplicity that f is convex:

$$\begin{aligned}
f(\mathbf{z}) &\geq f(\mathbf{x}_{k+1}) + \mathbf{q}^\top(\mathbf{z} - \mathbf{x}_{k+1}), \quad \forall \mathbf{q} \in \partial f(\mathbf{x}_{k+1}) \\
\implies f(\mathbf{z}) &\geq f(\mathbf{x}_{k+1}) - (\mathbf{x}_{k+1} - \mathbf{x}_k)^\top(\mathbf{z} - \mathbf{x}_{k+1}) \\
\implies f(\mathbf{x}_{k+1}) &\leq f(\mathbf{z}) + (\mathbf{x}_{k+1} - \mathbf{x}_k)^\top(\mathbf{z} - \mathbf{x}_{k+1}) \\
&= f(\mathbf{z}) - (\mathbf{x}_k - \mathbf{x}_{k+1})^\top(\mathbf{z} - \mathbf{x}_{k+1})
\end{aligned}$$

Convergence proof, conti.

Choosing $\mathbf{z} = \mathbf{x}^*$, where $\mathbf{x}^*$ is the optimal point:

$$\begin{aligned}
f\left(\mathbf{x}_{k+1}\right) - f^* &\leq -\left(\mathbf{x}_k - \mathbf{x}_{k+1}\right)^\top \left(\mathbf{x}^* - \mathbf{x}_{k+1}\right) \\
&\leq -\left(\mathbf{x}_k - \mathbf{x}_{k+1}\right)^\top \left(\mathbf{x}^* - \mathbf{x}_{k+1}\right) + \frac{1}{2}\left\|\mathbf{x}_k - \mathbf{x}_{k+1}\right\|_2^2 \\
&= \frac{1}{2}\left(\left\|\mathbf{x}_k - \mathbf{x}_{k+1} - \left(\mathbf{x}^* - \mathbf{x}_{k+1}\right)\right\|_2^2 - \left\|\mathbf{x}^* - \mathbf{x}_{k+1}\right\|_2^2\right) \\
&= \frac{1}{2}\left(\left\|\mathbf{x}_k - \mathbf{x}^*\right\|_2^2 - \left\|\mathbf{x}^* - \mathbf{x}_{k+1}\right\|_2^2\right)
\end{aligned}$$

Convergence proof, concl.

Summing from $k = 0$ to k:

$$\sum_{i=0}^{k}(f(\mathbf{x}_i) - f^*) \leq \frac{1}{2}(\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^* - \mathbf{x}_{k+1}\|_2^2) \leq \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

Since $f(\mathbf{x}_k)$ is non-increasing:

$$\sum_{i=0}^{k}(f(\mathbf{x}_k) - f^*) \leq \sum_{i=0}^{k}(f(\mathbf{x}_i) - f^*) \leq \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

Therefore, we have:

$$f(\mathbf{x}_k) - f^* \leq \frac{1}{2k}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

This result shows that the function values converge at a rate proportional to $1/k$.

A simple example

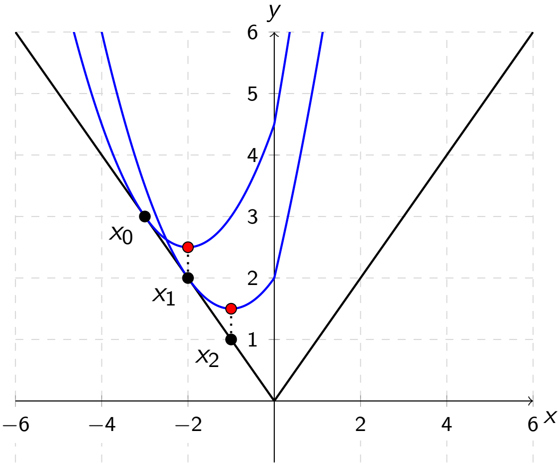
Suppose we have $f(x) = |x|$ and we want to find $\min_x f(x)$.
Let $x_0 = -3$ and suppose $\lambda = 1$.
Then

$$\mathrm{prox}_f(x_0) = f(x) + \frac{1}{2\lambda}\|x_0 - x\|^2 = |x| + \frac{(x_0 - x)^2}{2}$$

$$\implies x_1 = \arg\min_x \mathrm{prox}_f(x_0) = -2,$$

and we have $f(x_1) = 2$. Repeatedly, we have $x_2 = -1$, and so on...

A simple example, cont.



Issues

- The subproblem still requires invoking an iterative solver.
- In general, if f is already difficult to minimize, adding a quadratic makes it even more difficult to minimize. Only in some special cases, solving the prox is easier than minimizing f directly.
- Therefore historically it has not found many applications until recently.

Contemporary applications

In the past few years, this viewpoint has undergone a major revision. In a variety of circumstances, the proximal point method with a judicious choice of the control parameter λ and an appropriate iterative method for the subproblems can lead to practical and theoretically sound numerical methods. ([Dru17])

- Applications are needed: including machine learning / signal processing ([DG18]), portfolio optimization ([SLLC23]), etc.
- Improvements are given: for example, [LMH15] introduces a "catalyst" approach that solves a sequence of well-chosen auxiliary problems that incorporate a quadratic regularization term.

Now we take a closer look at the proximally guided subgradient method [DG18].

Stochastic approximation

Consider the problem of minimizing the expectation:[4]

$$\min F(x) = \mathbb{E}_{\zeta \sim \mathbb{P}} f(x, \zeta).$$

Here, ζ is a random variable following an fixed but unknown distribution $\mathbb{P}$, $x \in \mathcal{X} \subset \mathbb{R}^d$ is closed and convex, f is a known loss function, and the only access to F is by sampling ζ.

When the problem is convex, the stochastic subgradient method has strong theoretical guarantees and is often the method of choice.

The problem is well-studied (rates of convergence are given) when $f(\cdot, \zeta)$ is convex using stochastic (sub)gradient:

$$\begin{aligned}&\text{Sample } z_t \sim \mathbb{P}\\&\text{Set } x_{t+1} = x_t - \alpha_t \nabla_x f(x_t, z_t)\end{aligned}$$

[4] For simplicity of the exposition, the minimization problem is unconstrained. Simple constraints can be accommodated using a projection operation.

The proximally guided subgradient method

Now suppose f is nonsmooth and nonconvex. [DG18] shows how to use the proximal point method to guide the subgradient iterates in this broader setting, with rigorous guarantees.
Assume that the function $x \mapsto f(x, \zeta)$ is ρ-weakly convex[5] and L-Lipschitz for each ζ.
[DG18] proposed the scheme outlined in the following algorithm (PGSG)[6]:

Data: $x_0 \in \mathbb{R}^d$, $\{j_t\} \subset \mathbb{N}$, $\{\alpha_j\} \subset \mathbb{R}_0$
for $t = 0, \ldots, T$ **do**
  Set $y_0 = x_t$
  for $j = 0, \ldots, j_t - 2$ **do**
    Sample ζ and choose
    $v_j \in \partial \left(f(\cdot, \zeta) + \rho \left\| \cdot - x_t \right\|^2 \right) (y_j)$
    $y_{j+1} = y_j - \alpha_j v_j$
  end for
  $x_{t+1} = \frac{1}{j_t} \sum_{j=0}^{j_t - 1} y_j$
end for

---

[5] Here ρ can be understood as $1/\lambda$ in the previous notion.
[6] Here is a Python implementation by the authors.

The rate of convergence of PGSG

The method proceeds by applying a proximal point method with each subproblem approximately solved by a stochastic subgradient method.
It is proved that, by setting $j_t = t + \lceil 648 \log(648) \rceil$ and $\alpha_j = \frac{2}{\rho(j+49)}$ in the PGSG algorithm, the scheme will generate an iterate x satisfying $\mathbb{E}[\|\nabla F(x)\|^2] \leq \varepsilon$ after at most

$$O\left(\frac{\rho^2 \left(F\left(x_0\right) - \inf F\right)^2}{\varepsilon^2} + \frac{L^4 \log^4\left(\varepsilon^{-1}\right)}{\varepsilon^2}\right)$$

subgradient evaluations. This rate agrees with analogous guarantees for stochastic gradient methods for smooth nonconvex functions.

References

Damek Davis and Benjamin Grimmer.
Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems, 2018.
arXiv:1707.03505.

Dmitriy Drusvyatskiy.
The proximal point method revisited, 2017.
arXiv:1712.06038.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui.
A universal catalyst for first-order optimization, 2015.
arXiv:1506.02186.

Hong Seng Sim, Wendy Shin Yie Ling, Wah June Leong, and Chuei Yee Chen.
Proximal linearized method for sparse equity portfolio optimization with minimum transaction cost.
Journal of Inequalities and Applications, 2023(1):152, 2023.
doi:10.1186/s13660-023-03055-4.