General Unconstrained Convex Optimization

1 Convex Optimization Problem

In this lecture, we consider the general unconstrained convex optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x).$$

There are four different possibilities:

- Strongly convex and smooth
- Strong convex and non-smooth
- Non-strongly convex and smooth
- Non-strongly convex and non-smooth.

In this lecture, we will explore the relationships among these methods. In particular, we show that if we have an algorithm that works for the strongly convex and smooth case, then it can also be applied to the other three situations.

2 Non-smooth and Strongly Convex Optimization

In this section, we consider the situation that $f(x)$ is non-smooth but G-Lipschitz, and λ strongly convex. We show that for subgradient method, better convergence rate can be obtained than the rate for the non-strongly convex case.

Example 1 *We consider the SVM formulation*

$$\min_w f(w) := \left[\frac{1}{n}\sum_{i=1}^n (1 - w^\top x_i y_i)_+ + \frac{\lambda}{2}\|w\|_2^2\right]$$

This is non-smooth. The function $f(w)$ is λ strongly convex. It is also Lipschitz within the region of interests:

$$\|\nabla f(w)\|_2 \le G = \frac{1}{n}\sum_{i=1}^n \|x_i\|_2 + \sqrt{2\lambda}.$$

2.1 Subgradient Method

The following algorithm has been considered in the last lecture.

Algorithm 1: Subgradient Descent Method

Input: $f(x)$, x_0, $\eta_1, \eta_2, \ldots$

Output: x_T

for $t = 1, \ldots, T$ **do**

  Let $x_t = x_{t-1} - \eta_t g_t$, where $g_t \in \partial f(x_{t-1})$ is a subgradient

Return: x_T

We showed that for non-smooth and non-strongly convex functions, it is possible to take a learning rate of $\eta_t = O(1/\sqrt{t})$ to obtain a convergence rate of $\tilde{O}(1/\sqrt{T})$, where $\tilde{O}(\cdot)$ may hide a logarithmic factor in T.

In the following, we show that for non-smooth but strongly convex function, subgradient descent method can achieve a faster convergence rate.

Theorem 1 *Assume $f(x)$ is λ-strongly convex, and G-Lipschitz. Let $\eta_t = 1/(\lambda t)$, then we have*

$$\frac{1}{T}\sum_{t=1}^{T} f(x_{t-1}) \leq \min_x f(x) + \frac{(\ln T + 1)G^2}{2\lambda T}.$$

Proof Given any x, we have

$$\begin{aligned}
\|x_t - x\|_2^2 =& \|(x_t - x_{t-1}) + (x_{t-1} - x)\|_2^2 \\
=& \|x_t - x_{t-1}\|_2^2 + 2(x_t - x_{t-1})^\top (x_{t-1} - x) + \|x_{t-1} - x\|_2^2 \\
=& \eta_t^2 \|g_t\|_2^2 - 2\eta_t g_t^\top (x_{t-1} - x) + \|x_{t-1} - x\|_2^2 \\
\leq& \|x_{t-1} - x\|_2^2 + 2\eta_t g_t^\top (x - x_{t-1}) + \eta_t^2 G^2 \\
\leq& \|x_{t-1} - x\|_2^2 + 2\eta_t \left[f(x) - f(x_{t-1}) - \frac{\lambda}{2}\|x - x_{t-1}\|_2^2 \right] + \eta_t^2 G^2.
\end{aligned}$$

The first inequality is due to the condition $\|g_t\|_2 \leq G$, and the second inequality is due to the definition of subgradient g_t at x_{t-1} and strong convexity. Dividing by η_t^{-1}, we obtain

$$\frac{1}{\eta_t}\|x_t - x\|_2^2 \leq \left(\frac{1}{\eta_t} - \lambda\right)\|x_{t-1} - x\|_2^2 + 2[f(x) - f(x_{t-1})] + \eta_t G^2.$$

This is equivalent to

$$\lambda t \|x_t - x\|_2^2 \leq \lambda(t-1)\|x_{t-1} - x\|_2^2 + 2[f(x) - f(x_{t-1})] + \frac{G^2}{\lambda t}.$$

By summing over $t = 1$ to T, we obtain

$$\lambda T \|x_T - x\|_2^2 \leq 2\sum_{t=1}^{T}[f(x) - f(x_{t-1})] + \sum_{t=1}^{T}\frac{G^2}{\lambda t}.$$

Since $\sum_{t=1}^{T} t^{-1} \leq \ln T + 1$, we obtain

$$\lambda \|x_T - x\|_2^2 \leq \frac{2}{T}\sum_{t=1}^{T}[f(x) - f(x_{t-1})] + \frac{(\ln T + 1)G^2}{\lambda T}.$$

This proves the desired bound. ■

Note that in the theorem, we use a smaller learning rate of $\eta_t = O(1/\lambda t)$, and achieves a faster convergence rate of $O(1/t)$. In other word, in order to achieve ϵ accuracy, we need

$$O(G^2/\lambda\epsilon)$$

number of iterations.

2.2 Smoothing with Nesterov's Acceleration

We may employ smoothing to the objective function and obtain a smoothed approximation. For a non-smooth but G Lipschitz function, we may smooth it and obtain an (L, ϵ)-smooth approximation that is $L = G^2/2\epsilon$ smooth. If the strong convexity comes from the regularizer, and we do not smooth the regularizer, then typically the resulting smoothed formulation is still λ-strongly convex. An example is apply smoothing directly to the SVM loss function, but not the regularizer. In general, after smoothing, we can assume that the strong convexity parameter is close to λ.

By applying the smooth and strong convex version of Nesterov's acceleration algorithm, we obtain a convergence to ϵ-suboptimality in

$$T = O(\sqrt{L/\lambda}\log(1/\epsilon)) = O(G/\sqrt{\lambda\epsilon}\log(1/\epsilon))$$

number of iterations.

3 Reduction to Smooth and Strongly Convex Solver

Assume that we have an optimization algorithm $\mathcal{A}$ for L-smooth and λ-strongly convex optimization, then we can use it to solve the optimization problem for the other three situations, as we will show in this section.

If we take $\mathcal{A}$ as the gradient descent method, then the number of iterations required to achieve ϵ-accuracy is

$$T = O((L/\lambda)\log(1/\epsilon)) = \tilde{O}(L/\lambda).$$

Learning rate is $\eta = O(1/L)$.

If we take $\mathcal{A}$ as the accelerated gradient descent method, then the number of iterations required to achieve ϵ-accuracy is

$$T = O(\sqrt{L/\lambda}\log(1/\epsilon)) = \tilde{O}(\sqrt{L/\lambda}).$$

We set $\alpha = O(1/L)$ and set $\beta = (1-\sqrt{\alpha\lambda})/(1+\sqrt{\alpha\lambda})$ or set adaptively.

Smooth and Non-Strongly Convex Problem

Assume $f(x)$ is L-smooth but not strongly convex (that is, we can take the strong convex parameter $\lambda = 0$). Then given an accuracy ϵ, we may use solver $\mathcal{A}$ to solve the following problem

$$\min_x \tilde{x}(x), \qquad \tilde{f}(x) = f(x) + \frac{\epsilon}{2}\|x - x_0\|_2^2.$$

This function is $L+\epsilon$-smooth and $\lambda=\epsilon$ strongly convex. This solves the problem up to $O(\epsilon)$ accuracy. The condition number is $(L+\epsilon)/\epsilon$. We can apply a smooth and strongly convex solver.

If we use gradient descent, then in order to achieve ϵ accuracy, we need

$$T=\tilde{O}(L/\epsilon)$$

iterations, which is consistent with the analysis of smooth and non-strongly convex optimization. The learning rate is $O(1/L)$ which is independent of T.

If we use accelerated gradient descent, then in order to achieve ϵ accuracy, we need

$$T=\tilde{O}(\sqrt{L/\epsilon}),$$

which is consistent with that of the general Nesterov's acceleration method for smooth and non-strongly convex functions. The learning rate is $O(1/L)$ which is independent of T.

Non-Smooth and Strongly Convex Problem

If $f(x)$ is non-smooth but G-Lipschitz, and λ-strongly convex, then one can smooth f using

$$\tilde{f}(x)=\min_z\left[f(z)+\frac{L}{2}\|x-z\|_2^2\right],$$

with $L=G^2/2\epsilon$. Then $\tilde{f}(x)$ is (L,ϵ)-smooth approximation, and at least $\lambda/(1+\lambda/L)$ strongly convex.

This leads to a condition number of $O(G^2/(\epsilon\lambda))$. We can apply a smooth and strongly convex solver.

If we use gradient descent, then in order to achieve ϵ accuracy, we need

$$T=\tilde{O}(G^2/(\epsilon\lambda))$$

iterations, which is consistent with the analysis of non-smooth and strongly convex optimization. The learning rate is $O(1/L)=O(\epsilon/G^2)=O(1/\lambda T)$.

If we use Nesterov's acceleration, then in order to achieve ϵ accuracy, we need

$$T=\tilde{O}(G/\sqrt{\epsilon\lambda}),$$

which is consistent with the analysis of non-smooth and strongly convex optimization. The learning rate is $O(1/L)=O(\epsilon/G^2)=O(1/\lambda T^2)$.

Non-Smooth and Non-Strongly Convex Problem

We can find $\tilde{f}$ such that

$$\tilde{f}(x)=\min_z\left[f(z)+\frac{L}{2}\|x-z\|_2^2\right]+\frac{\epsilon}{2}\|x-x_0\|_2^2.$$

This gives $L=(G^2/2\epsilon)+\epsilon$-smooth and ϵ-strongly convex.

This gives a condition number of $O(G^2/\epsilon^2)$. We can apply a smooth and strongly convex solver.

If we use gradient descent, then in order to achieve ϵ accuracy, we need

$$T = \tilde{O}(G^2/\epsilon^2)$$

iterations, which is consistent with the analysis of non-smooth and non-strongly convex optimization for subgradient method. The learning rate is $O(1/L) = O(\epsilon/G^2) = O(1/G\sqrt{T})$

If we use Nesterov's acceleration, then in order to achieve ϵ accuracy, we need

$$T = \tilde{O}(G/\epsilon),$$

which is consistent with the analysis of non-smooth and non-strongly convex optimization. The learning rate is $O(1/L) = O(\epsilon/G^2) = O(1/GT)$.

4 General Results for Convex Optimization

We summarize results for the different situations with the two algorithms. The smoothness parameter is L, the strong convexity parameter is λ, and the Lipschitz parameter is G.

Gradient Descent

In order to obtain ϵ accurate result, the following iteration complexity can be achieved.

	smooth	non-smooth
strongly-convex	$\tilde{O}(L/\lambda)$ gradient descent	$\tilde{O}(G^2/\lambda\epsilon)$ sub-gradient
non-strongly-convex	$\tilde{O}(L/\epsilon)$ gradient descent	$\tilde{O}(G^2/\epsilon^2)$ sub-gradient

Table 1: Optimization Complexity for Gradient Descent

Accelerated Gradient Descent

In order to obtain ϵ accurate result, the following iteration complexity can be achieved.

	smooth	non-smooth
strongly-convex	$\tilde{O}(\sqrt{L/\lambda})$ accelerated gradient	$\tilde{O}(G/\sqrt{\lambda\epsilon})$ accelerated gradient with smoothing
non-strongly-convex	$\tilde{O}(\sqrt{L/\epsilon})$ accelerated gradient	$\tilde{O}(G/\epsilon)$ accelerated gradient with smoothing

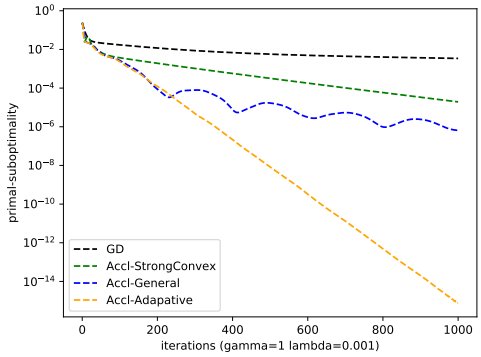
Table 2: Optimization Complexity for Accelerated Gradient Descent

5 Empirical Studies

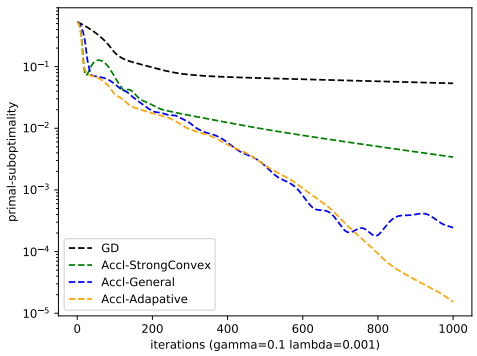
We study the effect of smoothing for gradient descent and accelerated gradient methods for SVM. We use a smoothing of the hinge loss for SVM, where the hinge loss $(1-z)_+$ is replaced by

$$\phi_\gamma(z) = \max_z \left[(1-z)_+ + \frac{1}{2\gamma}(x-z)^2 \right].$$

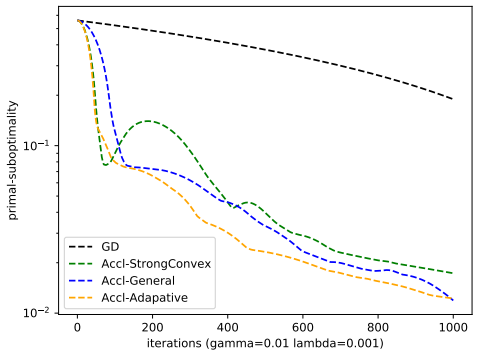
As we can see, even with a very small amount of smoothing or no smoothing, acceleration works well. However, the rate of convergence slows down with less smooth objective functions (that is, when γ is small). This is consistent with our theoretical results.



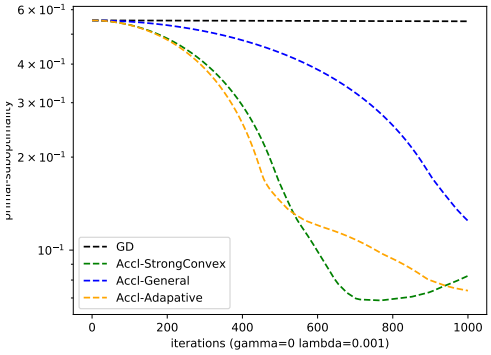
(a) $\gamma = 1$



(b) $\gamma = 0.1$



(c) $\gamma = 0.01$



(d) $\gamma = 0$

Figure 1: Convergence Comparisons with Different Smoothing Parameter