1 Outline

In this lecture, we study

- the subgradient method,
- gradient descent for smooth functions,
- Nesterov's acceleration,
- composite minimization.

2 Subgradient Method

2.1 Subgradients and Subdifferentials

The first-order characterization of convex functions states that a differentiable function f is convex if and only if $\mathrm{dom}(f)$ is convex and $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ for all $x, y \in \mathrm{dom}(f)$. For a function that is not necessarily differentiable, we can define the notion of *subgradients* as well as *subdifferentials*.

Definition 5.1. Given a convex function $f : \mathbb{R}^d \to \mathbb{R}$ and a fixed point $x \in \mathrm{dom}(f)$, the *subdifferential* of f at x is defined as

$$\partial f(x) = \left\{ g : \ f(y) \geq f(x) + g^\top (y - x) \ \ \forall y \in \mathrm{dom}(f) \right\}.$$

Here, any $g \in \partial f(x)$ is called a *subgradient* of f at x.

Conversely, the subdifferential is the set of subgradients. If function f is differentiable at x, then we have $\partial f(x) = \{\nabla f(x)\}$, and therefore, the subdifferential reduces to the gradient. In contrast, a non-differentiable function may have more than one subgradient. Moreover, note that for any subgradient g at x, $f(x) + g^\top (y - x)$ provies a lower approximation of the function f.

Recall that for a differentiable univariate function f, the gradient of f at some point x is the slope of the line tangent to f at x. We have a similar geometric intuition for subgradients. Consider the the absolute value function $f(x) = |x|$ over $x \in \mathbb{R}$, which is not differentiable at $x = 0$.

As depicted in Figure 5.1, there are multiple lines that are below $f(x) = |x|$ and go through $x = 0$. In fact, the subdifferential of f can be computed as follows.

$$\partial f(x) = \begin{cases} \{-1\} = \{\mathrm{sign}(x)\}, & \text{for } x < 0 \\ [-1, 1], & \text{for } x = 0 \\ \{+1\} = \{\mathrm{sign}(x)\}, & \text{for } x > 0 \end{cases} = \begin{cases} \{\mathrm{sign}(x)\}, & \text{for } x \neq 0 \\ [-1, 1], & \text{for } x = 0. \end{cases}$$

Let us consider a few more examples.

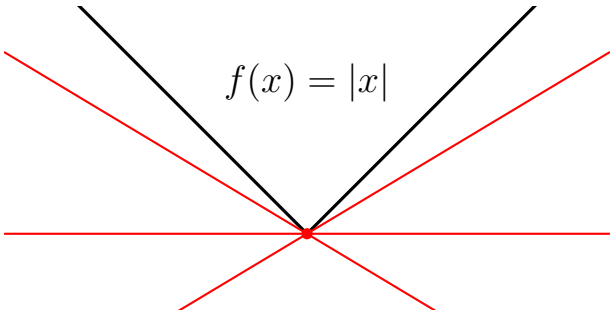

Figure 5.1: Subgradients of $f(x) = |x|$ at $x = 0$

Example 5.2. Let $f(x) = \|x\|_1 : \mathbb{R}^d \to \mathbb{R}$. Then the subdifferential of f at any point $x = (x_1, \ldots, x_d)^\top$ is the set of vectors $g = (g_1, \ldots, g_d)^\top$ such that for each $i \in [d]$,

$$g_i = \begin{cases} \text{sign}(x_i), & \text{if } x_i \neq 0 \\ [-1, 1], & \text{if } x_i = 0. \end{cases}.$$

Example 5.3. Let $f_1, \ldots, f_k$ be convex functions, and let f is defined as the pointwise maximum of $f_1, \ldots, f_k$. Given a point x, if $f(x) = f_i(x)$ for some $i \in [k]$, then any subgradient of f_i is a subgradient of f.

Example 5.4. Given a convex set $C \subseteq \mathbb{R}^d$, the indicator function $I_C(x)$ at a point $x \in \mathbb{R}^d$ is defined as

$$I_C(x) = \begin{cases} 0, & \text{if } x \in C \\ +\infty, & \text{if } x \notin C \end{cases}.$$

For a point $x \in C$, what is the subdifferential of the indicator function at x? Note that

$$\partial I_C(x) = \left\{ g \in \mathbb{R}^d : \ 0 \geq 0 + g^\top (y - x) \ \ \forall y \in C \right\} = N_C(x).$$

Therefore, the subdifferential is precisely the normal cone of C at x.

2.2 Optimality Conditions for Convex Minimization

Now we consider the convex minimization problem with a general convex objective function that is not necessarily differentiable.

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array} \quad = \quad \begin{array}{ll} \text{minimize} & f(x) + I_C(x) \\ \text{subject to} & x \in \mathbb{R}^d. \end{array}$$

The left is the constrained version, and the right formulation shows its unconstrained version with the indicator function. We discussed optimality conditions for convex minimization problems with a differentiable objective. In this section, we state and prove optimality conditions for the general case, in which the objective can be non-differentiable.

Theorem 5.5. *For a convex optimization problem* $\min_{x \in C} f(x)$, $x^* \in C$ *is an optimal solution if and only if there exists* $s \in \partial f(x^*)$ *such that*

$$s^\top (x - x^*) \geq 0 \quad \textit{for all } x \in C.$$

An immediate corollary of Theorem 5.5 is the following optimality condition for unconstrained problems.

Corollary 5.6. *For a convex optimization problem* $\min_{x\in\mathbb{R}^d} f(x)$, $x^* \in \mathbb{R}^d$ *is an optimal solution if and only if* $0 \in \partial f(x^*)$.

Corollary 5.6 can be applied to the unconstrained formulation of constrained convex minimization. We argued that for the differentiable case, the optimality condition is $-\nabla f(x^*) \in N_C(x^*)$ and $0 \in \{\nabla f(x^*)\} + N_C(x^*)$. As $\partial I_C(x)$ is equivalent to the normal cone $N_C(x)$, we obtain the following as a corollary of Corollary 5.6.

Corollary 5.7. *For a convex optimization problem* $\min_{x\in C} f(x)$, $x^* \in C$ *is an optimal solution if and only if*

$$0 \in \partial f(x^*) + N_C(x^*).$$

In this section, we will prove Theorem 5.5 which states the optimality condition for convex minimization. A tool that we need is the separating hyperplane theorem, which is an important result in convex analysis on its own. We state the separating hyperplane theorem without proof.

Theorem 5.8 (Separating hyperplane theorem)**.** *Let* $C, D \subseteq \mathbb{R}^d$ *be disjoint convex sets, i.e.,* $C \cap D = \emptyset$*, then there exists* $a \in \mathbb{R}^d \setminus \{0\}$ *and* $b \in \mathbb{R}$ *such that*

$$a^\top x \geq b \quad \text{for all } x \in C \quad \text{and} \quad a^\top x \leq b \quad \text{for all } x \in D$$

Let us prove Theorem 5.5 using Theorem 5.8.

Proof of Theorem 5.5. ($\Leftarrow$) Assume that there exists $s \in \partial f(x^*)$ such that $s^\top(x - x^*) \geq 0$ holds for all $x \in C$. Then it follows from the definition of subgradients that

$$f(x) - f(x^*) \geq s^\top(x - x^*) \geq 0 \quad \text{for all } x \in C.$$

This implies that $f(x) \geq f(x^*)$ for all $x \in C$, so x^* is optimal.

($\Rightarrow$) Let us consider the following two sets.

$$C = \{(x - x^*, t) : \ f(x) - f(x^*) \leq t\},$$
$$D = \{(x - x^*, t) : \ x \in C, \ t < 0\}.$$

Since $f(x) - f(x^*) \geq 0$ for any $x \in C$, these two sets are disjoint. Then by Theorem 5.8, there exists $a \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $c \in \mathbb{R}$ such that $(a, b) \neq (0, 0)$ and

$$a^\top(x - x^*) + bt \geq c, \quad \forall x \in \mathbb{R}^d, \ f(x) - f(x^*) \leq t \tag{5.1}$$
$$a^\top(x - x^*) + bt \leq c, \quad \forall x \in C, \ t < 0. \tag{5.2}$$

In (5.2), t can be arbitrarily small, so $b \geq 0$. Suppose that $b = 0$, in which case (5.1) becomes

$$a^\top(x - x^*)0 \geq c, \quad \forall x \in \mathbb{R}^d, \ f(x) - f(x^*) \leq t.$$

Here, $x - x^*$ can be $\lambda \cdot a$ where λ is an arbitrarily small number. This implies that $a = 0$. However, this contradicts the condition that $(a, b) \neq (0, 0)$. Therefore, $b > 0$. Then, without loss of generality, we may assume that $b = 1$. Then taking $x = x^*$ and $t = 0$ in (5.1), we obtain $0 \geq c$. Moreover,

taking $x = x^*$ and a number that is arbitrarily close to 0 for t, it follows that $0 \leq c$. Hence, $c = 0$. Then (5.1) and (5.2) become

$$a^\top (x - x^*) + t \geq 0, \quad \forall x \in \mathbb{R}^d,\ f(x) - f(x^*) \leq t \tag{5.3}$$

$$a^\top (x - x^*) + t \leq 0, \quad \forall x \in C,\ t < 0. \tag{5.4}$$

Here, we take $t = f(x) - f(x^*)$ in (5.3). Then (5.3) becomes

$$f(x) \geq f(x^*) - a^\top (x - x^*),$$

which implies that $-a \in \partial f(x^*)$. Moreover, we take a number that is arbitrarily close to 0 for t in (5.4). Then it becomes $a^\top (x - x^*) \leq 0$, which is equivalent to $-a^\top (x - x^*) \geq 0$. Hence, $-a$ is the desired vector. □

2.3 Convergence of Subgradient Method

We discussed the gradient descent method for minimizing a differentiable convex function. For non-differentiable convex functions, we can consider subgradients and use the subgradient method described as follows.

Algorithm 1 Subgradient method

Initialize $x_1 \in \mathrm{dom}(f)$.
for $t = 1, \ldots, T$ **do**
  Obtain a subgradient $g_t \in \partial f(x_t)$.
  $x_{t+1} = x_t - \eta_t g_t$ for a step size $\eta_t > 0$.
end for

We can show that the subgradient method given by Algorithm 1 converges if the subgradients of f are bounded. Recall that for the differentiable case, the ℓ_2 norm of f's gradient is bounded if and only if f is Lipschitz continuous.

Theorem 5.9. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function such that $\|g\|_2 \leq L$ for any $g \in \partial f(x)$ for every $x \in \mathbb{R}^d$. Let $\{x_t : t = 1, \ldots, T\}$ be the sequence of iterates generated by the subgradient method with step size*

$$\eta_t = \frac{1}{\sqrt{t}}$$

for each t. Then

$$f\left(\frac{\sum_{t=1}^T \eta_t x_t}{\sum_{t=1}^T \eta_t}\right) - f(x^*) \leq \frac{\|x_1 - x^*\|_2^2}{\sqrt{T}} + \frac{L^2(1 + \log T)}{2\sqrt{T}}$$

where x^* is an optimal solution to $\min_{x \in \mathbb{R}^d} f(x)$.

Remark 5.10. In fact, one may deduce a slightly better convergence rate of $O(1/\sqrt{T})$ by choosing a constant step size $\eta_t = 1/\sqrt{T}$. However, this requires knowing the total number of iterations T in advance. Another way to obtain a convergence rate of $O(1/\sqrt{T})$ is to use the so-called *doubling trick*, which we will not discuss here.

3 Gradient Descent for Smooth Functions

Remember that gradient descent for a Lipschitz continuous function has a conergence rate of $O(1/\sqrt{T})$ and uses a constant step size of order $O(1/\sqrt{T})$. In this section, we consider **smooth** convex functions for which gradient descent achieves a faster convergence rate.

We say that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth with respect to the ℓ_2 norm for some $\beta > 0$ if

$$\|\nabla f(x) - \nabla f(y)||_2 \leq \beta \|x - y\|_2$$

holds for any $x, y \in \mathbb{R}^d$. Smooth functions have the **self-tuning** property! By the optimality condition (for unconstrained problmes), we have $\nabla f(x^*) = 0$ for any optimal solution x^*. Then the smoothness assumption implies that the gradient gets close to 0 as we approach an optimal solution. This is in contrast to a non-differentiable function, e.g., $f(x) = |x|$ over $\mathbb{R}$.

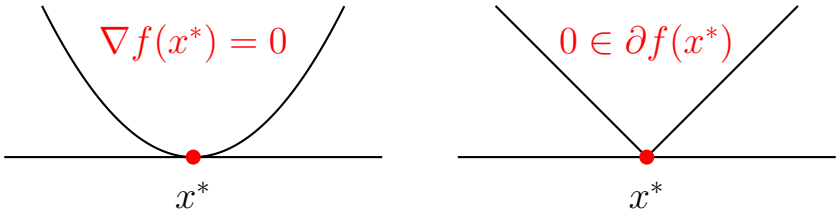

Figure 5.2: Smooth functions vs non-smooth functions

Recall that the gradient descent method for Lipschitz continuous functions requires a constant but small step size $O(1/\sqrt{T})$ where T is the total number of iterations. This is partly because the subgradient does not get smaller even when we get realy close to an optimal solution. In contrast, for smooth functions, we can take large step sizes, because the gradient gets reduced as we converge to an optimal solution. This is referred to as the self-tuning property.

Next we prove convergence for smooth functions. The first thing we observe is that a gradient step for a smooth function always leads to a strict improvement. To explain this, take a differentiable and β-smooth function $f : \mathbb{R}^d \to \mathbb{R}$.

Lemma 5.11. *If f is β-smooth in the ℓ_2 norm, then*

$$\left| f(y) - f(x) - \nabla f(x)^\top (y - x) \right| \leq \frac{\beta}{2} \|y - x\|^2.$$

Proof. By the fundamental theorem of calculus and the Cauchy-Schwarz inequality, we obtain the following.

$$\begin{aligned}
\left| f(y) - f(x) - \nabla f(x)^\top (y - x) \right| &= \left| \int_0^1 (y - x)^\top \left(\nabla f(x + t(y - x)) - \nabla f(x) \right) dt \right| \\
&\leq \int_0^1 \|y - x\|_2 \left\| \nabla f(x + t(y - x)) - \nabla f(x) \right\|_2 dt \\
&\leq \int_0^1 \beta t \|y - x\|_2^2 dt \\
&= \frac{\beta}{2} \|y - x\|_2^2
\end{aligned}$$

where the equality is due to the fundamental theorem of calculus, the first inequality is by the Cauchy-Schwarz inequality, and the second inequality is from the β-smoothness of f. □

Consider a gradient step, given by

$$x_{t+1} = x_t - \eta_t \nabla f(x_t).$$

Note that if $\eta_t \leq 1/\beta$,

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{\beta}{2}\|x_{t+1} - x_t\|_2^2 \\ &= f(x_t) + \left(-\eta_t + \frac{\eta_t^2 \beta}{2}\right) \|\nabla f(x_t)\|_2^2 \\ &\leq f(x_t) - \frac{\eta_t}{2}\|\nabla f(x_t)\|_2^2 \end{aligned}$$

where the first inequality follows from the β-smoothness of f and the second inequality is because $\eta_t \leq 1/\beta$. Therefore, when $\eta_t \leq 1/\beta$, we obtain

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta_t}{2}\|\nabla f(x_t)\|_2^2,$$

which implies that $f(x_{t+1})$ is strictly better than $f(x_t)$ when x_t is not an optimal solution. Based on this observation, we can prove the following convergence result for smooth functions.

Theorem 5.12. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be β-smooth in the ℓ_2-norm and convex, and let $\{x_t : t = 1, \ldots, T+1\}$ be the sequence of iterates generated by gradient descent with a constant step size*

$$\eta_t = \eta \leq \frac{1}{\beta}$$

for each t. Then

$$f(x_{T+1}) - f(x^*) \leq \frac{\|x_1 - x^*\|_2^2}{2\eta T}$$

where x^* is an optimal solution to $\min_{x \in \mathbb{R}^d} f(x)$.

Proof. Note that

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2 \\ &\leq f(x^*) - \nabla f(x_t)^\top (x^* - x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2 \\ &= f(x^*) + \frac{1}{2\eta}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) \end{aligned}$$

where the second inequality is because $f(x_t) + \nabla f(x_t)^\top (x - x_t)$ is a lower bound on f and the equality follows because $x_{t+1} = x_t - \eta \nabla f(x_t)$. This implies that

$$f(x_{t+1}) - f(x^*) \leq \frac{1}{2\eta}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right),$$

summing which over $t = 1, \ldots, T$ and dividing the resulting one by T, we obtain

$$\frac{1}{T}\sum_{t=1}^{T} f(x_{t+1}) - f(x^*) \leq \frac{1}{2\eta T}\left(\|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2\right) \leq \frac{1}{2\eta T}\|x_1 - x^*\|_2^2.$$

Recall that each gradient step for smooth functions leads to an improvement, i.e., $f(x_{t+1}) \leq f(x_t)$. Therefore,

$$f(x_{T+1}) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T} f(x_{t+1}) - f(x^*) \leq \frac{1}{2\eta T}\|x_1 - x^*\|_2^2,$$

as required. □

The important takeaway is that we may take a constant step size, which does not depend on the number of iterations T. This is due to the self-tuning property of smooth functions. Although we do not shrink the step size, the change between the current iterate x_t and the next iterate x_{t+1} gets reduced as we approach an optimal solution.

As discussed before, the term $\|x_1 - x^*\|_2$ and the smoothness parameter β are all fixed constants. Hence, the convergence rate is $O(1/T)$. Therefore, after $T = O(1/\epsilon)$ iterations, we have

$$f(x_{T+1}) - f(x^*) \leq \epsilon.$$

Note that the convergence results for smooth functions improves over $O(1/\sqrt{T})$ and $O(1/\epsilon^2)$ for the subgradient method.

4 Accelerated Gradient Method

For smooth functions, there is an algorithm that achieves a faster convergence rate than the vanilla gradient descent. The algorithm is due to Nesterov [Nes83, Nes04], and it is referred to as Nesterov's accelerated gradient descent. The algorithm is proven to achieve the optimal asymptotic convergence rate. Let us describe the algorithm and explain how it achieves a better convergence rate.

The main idea behind Nesterov's acceleration is to use "momentum", so the algorithm is often called gradient descent with momentum. Recall that gradient descent for a β-smooth function follows the update rule of

$$x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$$

from a given point x_t. The idea of momentum is to incorporate the direction $x_t - x_{t-1}$ that we took when moving from x_{t-1} to x_t to obtain the next iterate x_{t+1}. Then x_{t+1} is determined by not only the previous iterate x_t but also x_{t-1} which is the one before x_t. Figure 5.3 illustrates how the idea of momentum applies. Instead of applying the gradient descent update to x_t, we move a

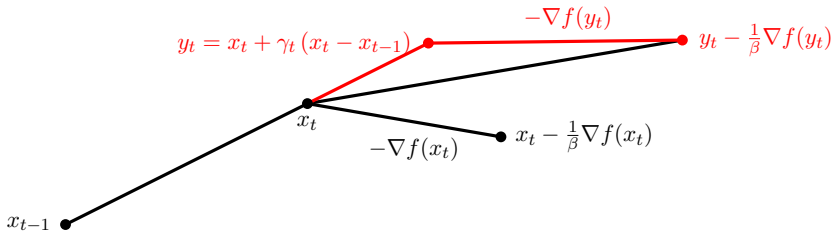

Figure 5.3: Illustration of gradient descent with momentum

bit further from x_t along the momentum direction that we took from x_{t-1} to x_t. Let $\gamma_t > 0$ be a weight, and

$$y_t = x_t + \gamma_t(x_t - x_{t-1}).$$

Then we apply the gradient descent update on y_t to obtain the next point x_{t+1}, as follows.

$$x_{t+1} = y_t - \frac{1}{\beta}\nabla f(y_t).$$

Algorithm 2 summarizes Nesterov's accelerated gradient descent that we just explained. The following shows a convergence result of the accelerated gradient descent method for smooth functions.

Algorithm 2 Nesterov's accelerated gradient descent

Initialize $x_1 \in \mathrm{dom}(f)$.
Set $x_0 = x_1$.
for $t = 1, \ldots, T$ **do**
  $y_t = x_t + \gamma_t(x_t - x_{t-1})$ for some $\gamma_t > 0$.
  $x_{t+1} = y_t - \frac{1}{\beta}\nabla f(y_t)$.
end for
Return x_{T+1}.

Theorem 5.13. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a β-smooth convex function in the ℓ_2 norm. We set*

$$\gamma_t = \frac{t-2}{t+1}.$$

Then

$$f(x_{T+1}) - f(x^*) \le \frac{2\beta\|x_1 - x^*\|_2^2}{(T+1)^2}$$

where x^* is an optimal solution to $\min_{x\in\mathbb{R}^d} f(x)$.

We will prove a more general result in the next lecture, which implies this theorem.

5 Composite Convex Minimization

We consider the following composite convex optimization problem.

$$\min_{x\in\mathbb{R}^d} \quad f(x) = g(x) + h(x)$$

where we assume that g is a smooth convex function and h is convex. For example, we may consider the following problems.

Example 5.14. Constrained smooth convex optimization: Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set. We consider the constrained smooth convex optimization problem

$$\min_{x\in\mathcal{X}} \quad g(x) \quad = \quad \min_{x\in\mathbb{R}^d} \quad g(x) + I_{\mathcal{X}}(x).$$

We can rewrite the above problem as a composite convex optimization problem by taking $h(x) = I_{\mathcal{X}}(x)$, the indicator function of $\mathcal{X}$.

Example 5.15. LASSO: Given $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, and $\lambda > 0$, we consider the least absolute shrinkage and selection operator, or simply the LASSO problem

$$\min_{x\in\mathbb{R}^d} \quad \underbrace{\frac{1}{n}\|Ax - b\|_2^2}_{\text{smooth convex function } g(x)} + \underbrace{\lambda\|x\|_1}_{\text{convex function } h(x)}.$$

Here, h is not necessarily smooth, so we cannot apply gradient descent directly. However, we can use the so-called **proximal gradient** method. The key tool is the **proximal operator** associated with h. Given a convex function $h : \mathbb{R}^d \to \mathbb{R}$ and $\eta > 0$, the proximal operator of h with parameter η is defined as

$$\mathrm{prox}_{\eta h}(x) = \operatorname*{argmin}_{u\in\mathbb{R}^d} \left\{ \eta h(u) + \frac{1}{2}\|x - u\|_2^2 \right\}.$$

For the indicator function $h(x) = I_{\mathcal{X}}(x)$, the associated prox operator is equivalent to the projection operator onto $\mathcal{X}$, as

$$\text{prox}_{\eta I_{\mathcal{X}}(\cdot)}(x) = \underset{u \in \mathbb{R}^d}{\text{argmin}} \left\{ \eta I_{\mathcal{X}}(u) + \frac{1}{2}\|x-u\|_2^2 \right\} = \underset{u \in \mathcal{X}}{\text{argmin}} \left\{ \frac{1}{2}\|x-u\|_2^2 \right\} = \text{proj}_{\mathcal{X}}(x).$$

For LASSO, we will prove in the next lecture that the prox operator of the ℓ_1-norm $h(x) = \lambda\|x\|_1$ is given by

$$\text{prox}_{\eta\lambda\|\cdot\|_1}(x) = \left(\underbrace{\max\{0, |x_i| - \eta\lambda\}}_{\text{shirinkage operator}} \cdot \text{sign}(x_i) \right)_{i \in [d]}$$

The proximal gradient algorithm applies to this composite problem proceeds with the following update rule.

$$x_{t+1} = \text{prox}_{\eta h}(x_t - \eta \nabla g(x_t)).$$

When f is smooth with parameter β, we set the step size $\eta = 1/\beta$ is in the smooth convex minimization setting. Proximal Gradient Descent applied to LASSO is referred to as Iterative

Algorithm 3 Proximal Gradient Descent

Initialize $x_1 \in \mathbb{R}^d$.
for $t = 1, \ldots, T$ **do**
  Update $x_{t+1} = \text{prox}_{\eta h}(x_t - (1/\beta)\nabla g(x_t))$ where β is the smoothness parameter of g.
end for
Return x_T.

Shrinkage-Thresholding Algorithm (ISTA). We can prove the following convergence result for Proximal Gradient Descent.

Theorem 5.16. *Let $f = g + h$ where g is a β-smooth convex function in the ℓ_2 norm and h is convex. Then x_{T+1} returned by Proximal Gradient Descent (Algorithm 3) satisfies*

$$f(x_{T+1}) - f(x^*) \leq \frac{\beta\|x_1 - x^*\|_2^2}{2}.$$

Note that this recovers the convergence result for gradient descent for smooth functions by considering the case $h = 0$.

References

[Nes83] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983. 4

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, Norwell, 2004. 4