9 Smoothing

In this lecture we are interested in the problem of minimizing a nonsmooth convex function $f(x)$ over $\mathbb{R}^n$. The subgradient method is quite slow, in theory and practice, and requires $1/\epsilon^2$ subgradient calls to reach accuracy ϵ. We will see that if we have more information about the structure of f, then one can reach accuracy ϵ in just $1/\epsilon$ iterations. The main idea will be to *approximate* the function $f(x)$ by a smooth one $f_\mu(x)$, and to apply the fast gradient method to $f_\mu(x)$. This idea is based on [Nes05]. For this to be possible, we are going to assume that we have access to the "internal structure" of f, in particular to a dual formulation of f as a pointwise maximum of linear functions.

Smoothing via the conjugate function Assume f is given as $f(x) = h^*(Ax+b)$ where h is closed and convex defined on a *compact domain* D, i.e.,

$$f(x) = \sup_{y \in D} y^T(Ax+b) - h(y).$$

An example is $f(x) = \|Ax+b\|_1 = h^*(Ax+b)$ where h is the indicator function of the ℓ_∞ ball; indeed $\|Ax+b\|_1 = \max_{\|y\|_\infty \leq 1} y^T(Ax+b)$. Let d be a 1-*strongly convex* function defined on D. We further assume that $d \geq 0$ on D. Then $(h+\mu d)$ is μ-strongly convex, and as such f_μ defined by

$$f_\mu(x) = (h+\mu d)^*(Ax+b)$$

is smooth with a Lipschitz continuous gradient. We know from the previous lecture that

$$\nabla (h+\mu d)^*(z) = \operatorname*{argmax}_{v \in D} \left\{ z^T v - h(v) - \mu d(v) \right\}.$$

The gradient of f_μ is thus given by

$$\nabla f_\mu(x) = A^T \nabla (h+\mu d)^*(Ax+b) = A^T \operatorname*{argmax}_{v \in D} \left\{ (Ax+b)^T v - h(v) - \mu d(v) \right\}.$$

Since $h+\mu d$ is μ-strongly convex, it follows that $(h+\mu d)^*$ is $1/\mu$-smooth. As a consequence ∇f_μ is $(\|A\|^2/\mu)$-Lipschitz where $\|A\|$ is the operator norm of A defined as

$$\|A\| = \max_{\|x\|_2 = 1} \|Ax\|_2.$$

How good is the approximation quality of f_μ to f?

- Since $d \geq 0$ it follows that $h + \mu d \geq h$ and thus $(h+\mu d)^* \leq h^*$ which implies that $f_\mu \leq f$.
- Let $R = \max_{x \in D} d(x)$. Then $h + \mu d \leq h + \mu R$ which implies that $(h+\mu d)^* \geq h^* - \mu R$.

At the end we get that $f(x) - \mu R \leq f_\mu(x) \leq f(x)$.

This means that if we minimize $f_\mu(x)$ with accuracy ϵ_μ, then this gives a solution to $\min_{x \in \mathbb{R}^n} f(x)$ with accuracy $\epsilon = \epsilon_\mu - \mu R$, i.e.,

$$f_\mu(x) - f_\mu^* \leq \epsilon_\mu \quad \Rightarrow \quad f(x) - f^* \leq \epsilon_\mu + \mu R.$$

So to get accuracy ϵ on f we need to set $\epsilon_\mu = \epsilon - \mu R$.

Using the fast gradient method, we can get accuracy ϵ_μ on f_μ with a number of iterations proportional to $\sqrt{\frac{L_\mu}{\epsilon_\mu}}$ where $L_\mu = \|A\|^2/\mu$ is the smoothness parameter of f_μ. If we set $\mu = \epsilon/(2R)$ so that $\epsilon_\mu = \epsilon - \mu R = \epsilon/2$, we can reach accuracy ϵ on f^* in a number of iterations proportional to

$$\frac{\|A\|\sqrt{R}}{\epsilon}.$$

Compare this with the subgradient method which requires $1/\epsilon^2$ iterations.

Examples

- Consider the problem of minimizing $f(x) = \|Ax + b\|_1$ where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Note that $f(x) = h^*(Ax + b)$ where h is the indicator function of the unit ℓ_∞ ball. In fact we have

  $$f(x) = \max_{\|y\|_\infty \le 1} y^T(Ax + b).$$

  Choose $d(y) = \|y\|_2^2/2$, which is 1-strongly convex and nonnegative, and let

  $$f_\mu(x) = (h + \mu d)^*(x) = \max_{\|y\|_\infty \le 1} \left\{ y^T(Ax + b) - \mu d(y) \right\}.$$

  The maximization problem on the right-hand side is separable in y. It is not difficult to verify that

  $$\max_{|y_i|\le 1} y_i t - \mu y_i^2/2 = \begin{cases} t^2/2\mu & \text{if } |t| \le \mu \\ |t| - \mu/2 & \text{if } |t| \ge \mu. \end{cases}$$

  Call $\phi_\mu(t)$ the function on the right-hand side, called the *Huber penalty function*. Then we have

  $$f_\mu(x) = \sum_{i=1}^m \phi_\mu((Ax + b)_i).$$

- We consider the same problem $f(x) = \|Ax + b\|_1$ but we use a different strongly convex function d. We use $d(y) = \sum_{i=1}^m \delta(y_i)$ where $\delta(y_i) = 1 - \sqrt{1 - y_i^2}$. One can check that $\delta''(y) = 1/\sqrt{1 - y^2} \ge 1$ and so d is 1-strongly convex. With this choice of d we have

  $$f_\mu(x) = \max_{\|y\|_\infty \le 1} \left\{ y^T(Ax + b) - \mu \sum_{i=1}^m \delta(y_i) \right\}.$$

  Again this problem is separable and we have to solve

  $$\max_{|y_i|\le 1} y_i t - \mu + \mu\sqrt{1 - y_i^2}$$

  which one can show is equal to $\psi_\mu(t) = \sqrt{t^2 + \mu^2} - \mu$. Thus in this case the function f is approximated by

  $$f_\mu(x) = \sum_{i=1}^m \psi_\mu((Ax + b)_i).$$

References

[Nes05] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005. 1