1 Nonsmooth Convex Optimization

We are interested in constrained minimization of convex, possibly nondifferentiable, $f : \mathbb{R}^n \to \mathbb{R}$

$$\text{minimize}_{x \in \mathcal{C}} \; f(x)$$

given first order oracle. $\mathcal{C}$ is a simple closed convex set.

1.1 Projected Subgradient Method

Subgradient method iteratively updates as follows

$$x^{k+1} = \mathcal{P}_{\mathcal{C}} \left(x^k - \alpha_k g^k \right)$$

where $g^k \in \partial f(x^k)$ is *any* subgradient of f and that $\mathcal{P}_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|^2$. First order optimality condition is $\langle g(x), x - x^* \rangle \geq 0$ for any $x \in \mathcal{C}$, which is impossible to test for nontrivial function f. Therefore, using $\left\|g^k\right\| \leq \epsilon$ is not informative and subgradient method does not really have a stopping criterion.

1.1.1 Connection to Mirror Descent

Each update involves solving a subproblem of the form

$$\begin{aligned} x^{k+1} &= \arg\min_{x \in \mathcal{C}} \left\| x^k - \alpha_k g^k - x \right\|_2^2 \\ &= \arg\min_{x \in \mathcal{C}} \left\{ \left\| x - x^k \right\|_2^2 + 2\alpha_k \left\langle x, \nabla f(x^k) \right\rangle + \left(\alpha_k \nabla f(x^k) \right)^2 \right\} \\ &= \arg\min_{x \in \mathcal{C}} \left\{ \left\langle x, \nabla f(x^k) \right\rangle + \frac{1}{\alpha_k} D^{\omega}(x, x^k) \right\} \end{aligned} \tag{1}$$

where $D^{\omega}(x, y) = \frac{1}{2} \|x - y\|_2^2$ is the Bregman divergence induced by $\omega(x) = \frac{1}{2} \|x\|_2^2$. In effect, projected subgradient method is mirror descent on $\mathcal{C}$ endowed with ℓ-2 norm.

1.1.2 Convergence

Given bounded subgradient $\left\|g^k\right\| \leq G$ and bounded domain $\left\|x^0 - x^*\right\| \leq R$, subgradient method is in a sense optimal as it achieves the lower bound $\mathcal{O}(\frac{1}{\epsilon^2})$ for this problem class. The derivation as follows

$$\begin{aligned} \left\| x^{k+1} - x^* \right\|_2^2 &= \left\| \mathcal{P}_{\mathcal{C}} \left(x^k - \alpha_k g^k \right) - \mathcal{P}_{\mathcal{C}}(x^*) \right\| && \text{(Try to bound a single update)} \\ &\leq \left\| x^k - \alpha_k g^k - x^* \right\|_2^2 && (\mathcal{P}_{\mathcal{C}} \text{ nonexpansive}) \\ &= \left\| x^k - x^* \right\|_2^2 - 2\alpha_k \left\langle g^k, x^k - x^* \right\rangle + \alpha_k^2 \left\| g^k \right\|_2^2 \\ &\leq \left\| x^k - x^* \right\|_2^2 - 2\alpha_k \left(f(x^k) - f(x^*) \right) + \alpha_k^2 \left\| g^k \right\|_2^2 \\ \left\| x^{k+1} - x^* \right\|_2^2 &\leq \left\| x^1 - x^* \right\|_2^2 - 2 \sum_{i=1}^{k} \alpha_i \left(f(x^i) - f(x^*) \right) + \sum_{i=1}^{k} \alpha_i^2 \left\| g^i \right\|_2^2 && \text{(Telescope)} \end{aligned}$$

Then rearrange, and bound

$$2\sum_{i=1}^{k}\left(f(x^i)-f(x^*)\right)\leq R^2+G^2\sum_{i=1}^{k}\alpha_i^2 \quad \Rightarrow \quad \min_{i\in[k]} f(x^i)-f(x^*)\leq \frac{R^2+G^2\sum_{i=1}^{k}\alpha_i^2}{2\sum_{i=1}^{k}\alpha_i}$$

We note that $\min_{i\in[k]} f(x^i)-f(x^*)\to 0$ if stepsize is square summable but not summable, i.e. $\sum_i \alpha_i^2<\infty$ and $\sum_i \alpha_i=\infty$. The choice of stepsize $\alpha_k=\frac{R}{\sqrt{k+1}}$ yield $\min_{i\in[k]} f(x^i)-f(x^*)=\mathcal{O}(\frac{1}{\epsilon^2})$. (3.2.3 in [2])

1.1.3 Solving Support Vector Machine w/ Subgradient Method

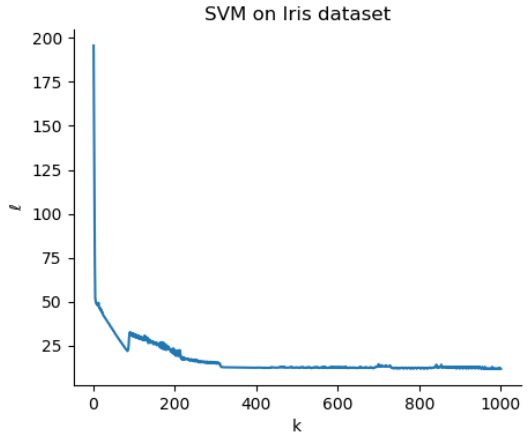
We are given data $\mathcal{D}=\{(x_i,y_i)\mid x_i\in\mathbb{R}^n\ y_i\in\{\pm 1\}\}$, support vector machine is supervised learning model that tries to find $w\in\mathbb{R}^n$ and $b\in\mathbb{R}$ such that the empirical risk and regularizer on w is minimized

$$\text{minimize}_{w,b}\quad \frac{1}{2}\|w\|_2^2+\lambda\sum_{i=1}^{m}\max\left[0,1-y_i(w^Tx_i+b)\right]\quad (:=f(w,b))$$

Support vector machines can be solved using subgradient method. We first find a subgradient of f

$$g_w^k=w^k-\lambda\sum_{i\in[m]:y_i(w^Tx_i+b)<1} y_ix_i$$
$$g_b=-\lambda\sum_{i\in[m]:y_i(w^Tx_i+b)<1} y_i$$

where we haved picked $0\in\partial(\max 0,1-y_i(w^Tx_i+b))$ when $y_i(w^Tx_i+b)=1$, the only case where the *max term* is non-differentiable. When tested on the Iris dataset, subgradient method worked!



1.2 Mirror Descent

Mirror descent has updates for the form given initialization $x^1=\arg\min_{x\in\mathcal{C}}\omega(x)$

$$x^{k+1}=\mathsf{prox}_{x^k}\left(\alpha_k g^k\right)\qquad \text{where}\qquad \mathsf{prox}_x(\xi):=\arg\min_{x\in\mathcal{C}}\left\{D_\omega(x,y)+\langle\xi,x\rangle\right\}\qquad (2)$$

where $\mathsf{prox}_x(\cdot)$ is called Bregman prox mapping and can be interpreted as extension of the standard prox operator regularized by Bregman divergence instead of ℓ-2 norm. By (1), we see mirror descent as a generalization of projected subgradient methods to domain other than $\mathbb{R}^n$. An alternative update can be derived by considering the optimality condition of (2) $0 \in \alpha_k g^k + \nabla\omega(x^{k+1}) - \nabla\omega(x^k) + \mathcal{N}_{\mathcal{C}}(x^{k+1})$ where $\mathcal{N}_C(\cdot)$ is normal cone of $\mathcal{C}$, or equivalently, $x^{k+1} \in (\nabla\omega + \mathcal{N}_{\mathcal{C}})^{-1}(\nabla\omega(x^k) - \alpha_k g^k)$. The update can be written as

$$x^{k+1} = \nabla\omega^* \left(\nabla\omega(x^k) - \alpha_k g^k\right) \tag{3}$$

where $w^*(y) = \sup_{z\in\mathcal{C}} [\langle z, y\rangle - \omega(z)]$

1.2.1 Bregman Divergence

Definition. ***(Bregman Divergence)*** *Let $\mathcal{C}$ be closed convex set. Let the distance generating function $h : \mathcal{C} \to \mathbb{R}$ be continuously differentiable and 1-strongly convex w.r.t. norm $\|\cdot\|$. Then the Bregman divergence is*

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y\rangle$$

which has the interpretation of the error in first order Taylor expansion of h around y evaluated at x. For example, $h(x) = \frac{1}{2}\|x\|_2^2$ is strongly convex w.r.t. ℓ-2 norm over $\mathcal{C} \subset \mathbb{R}^n$ with $D_h(x, y) = \frac{1}{2}\|x - y\|_2^2$; the negative entropy function $h(x) = \sum_i x_i \log x_i$ is strongly convex w.r.t. ℓ-1 norm over the probability simplex $\mathcal{C} \subset \triangle^n = \left\{x \in \mathbb{R}^n_+ \mid \sum_i x_i = 1\right\}$ with $D_h(x, y) = [\sum_i x_i \log(x_i/y_i) - \sum_i x_i + \sum_i y_i := D(x\|y)] \geq \frac{1}{2}\|x - y\|_1^2$ or KL divergence (inequality straight from Pinsker's inequality)

Some important properties of Bregman divergence include (reference)

$$D_h(x, y) \geq \frac{1}{2}\|x - y\|^2 \tag{4}$$

$$\langle \nabla h(x) - \nabla h(y), x - z\rangle = D_h(x, y) + D_h(z, x) - D_h(z, y) \tag{5}$$

$$\nabla_x D_h(x, y) = \nabla h(x) - \nabla h(y) \tag{6}$$

1.2.2 Convergence

Assume bounded gradient norm $\left\|g^k\right\|_* \leq G$ and an upper bound on divergence to z^1, i.e. $\Omega := \max_{u\in\mathcal{C}} D_\omega(u, z^1)$. The key to convergence of mirror descent is the MD lemma

$$\alpha_k \left\langle g^k, x^k - u\right\rangle \leq \frac{\alpha^2}{2}\left\|g^k\right\|_*^2 + D_\omega(u, x^k) - D_\omega(u, x^{k+1}) \qquad \forall\, u \in \mathcal{C} \tag{7}$$

Proof. Copied from [3]

$$\begin{aligned}
\alpha_k \left\langle g^k, x^k - u \right\rangle &= \left\langle \alpha_k g^k, x^k - x^{k+1} \right\rangle + \left\langle \alpha g^k, x^{k+1} - u \right\rangle \\
&\leq \left\langle \alpha_k g^k, x^k - x^{k+1} \right\rangle + \left\langle -\nabla D_\omega(x^{k+1}, x^k), x^{k+1} - u \right\rangle \\
&\text{(by 1st order optimality condition: } \left\langle \nabla D_\omega(x^{k+1}, x^k) + \alpha_k g^k, u - x^{k+1} \right\rangle \geq 0 \text{ for all } u \in \mathcal{C}) \\
&= \left\langle \alpha_k g^k, x^k - x^{k+1} \right\rangle + D_\omega(u, x^k) - D_\omega(u, x^{k+1}) - D_\omega(x^{k+1}, x^k) \\
&\qquad\qquad \text{(By (5,6))} \\
&\leq \left\langle \alpha_k g^k, x^k - x^{k+1} \right\rangle - \frac{1}{2} \left\| x^k - x^{k+1} \right\|^2 + D_\omega(u, x^k) - D_\omega(u, x^{k+1}) \\
&\qquad\qquad \text{(By (4))} \\
&\leq \frac{\alpha_k^2}{2} \left\| g^k \right\|_*^2 + D_\omega(u, x^k) - D_\omega(u, x^{k+1}) \quad \text{(completed square is nonzero)}
\end{aligned}$$

□

Using the MD lemma, we can derive an upper bound on the regret $\left\langle g^k, x^k - x^* \right\rangle$ by telescope

$$\sum_{k=1}^T \alpha_k \left\langle g^k, x^k - x^* \right\rangle \leq \sum_{k=1}^T \frac{\alpha_k^2}{2} \left\| g^k \right\|_*^2 + D_\omega(x^*, x^1) - D_\omega(x^*, x^{k+1}) \leq \frac{G^2}{2} \sum_{k=1}^T \alpha_k^2 + \Omega$$

By convexity of f and Jensen's inequality and let $\overline{x} = \sum_{k=1} \gamma_k x^k$ where $\gamma_k = \alpha_k / \sum_k \alpha_k$,

$$\sum_{k=1}^T \alpha_k \left\langle g^k, x^k - x^* \right\rangle \geq \sum_{k=1}^T \alpha_k \left(f(x^k) - f(x^*) \right) \geq \left(\sum_{k=1}^T \alpha_k \right) \left(f\left(\overline{x}\right) - f(x^*) \right)$$

Then we have a bound on the average iterate $\overline{x}$

$$f(\overline{x}) - f(x^*) \leq \frac{\frac{G^2}{2} \sum_{k=1}^T \alpha_k^2 + \Omega}{\sum_{k=1}^T \alpha_k}$$

Note the bound is really similar to bound we get for last iterate in convergence analysis for projected subgradient methods, it turns out the complexity is also the same, The choice of constant stepsize $\alpha_k = \frac{\sqrt{2\Omega}}{G\sqrt{T}}$ has $\sum \alpha_k = \frac{\sqrt{2\Omega T}}{G}$ and $\sum_k \alpha_k^2 = \frac{2\Omega}{G^2}$, then

$$f(\overline{x}) - f(x^*) \leq \frac{G(\Omega + \frac{G^2}{2} \frac{2\Omega}{G^2})}{\sqrt{2\Omega T}} = \frac{2\Omega G}{\sqrt{2\Omega T}} = G\sqrt{\frac{2\Omega}{T}}$$

which implies $\mathcal{O}(\frac{1}{\epsilon^2})$ complexity