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

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

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 space 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

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

Then rearrange, and bound

$$2\sum_{t=1}^{k}\big(f(x^t)-f(x^*)\big) \le R^2 + G^2\sum_{t=1}^{k}\alpha_t^2 \quad \Rightarrow \quad \min_{t\in[k]} f(x^t)-f(x^*) \le \frac{R^2+G^2\sum_{t=1}^{k}\alpha_t^2}{2\sum_{t=1}^{k}\alpha_t}$$

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

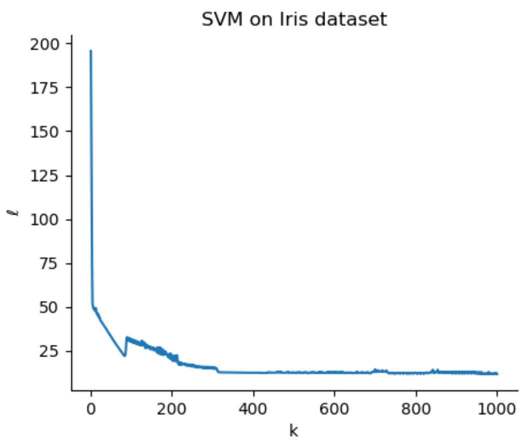
$$\text{minimize}_{w,b} \quad \frac{1}{2}\|w\|_2^2 + \lambda\sum_{i=1}^{m}\max\left[0, 1-y_i(w^Tx_i+b)\right] \quad (:= f(w,b))$$

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

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

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

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