5.4 Outlook on nonsmooth optimization

- The (stochastic) gradient descent method can not be applied to compute (regularized) ERM-hypotheses based on the hinge loss.

SVM

- This requires an optimization procedure for nondifferentiable (but convex) objective functions – the subgradient method.

Definition 5.8:

For a function $F\colon \mathcal{W} \to \mathbb{R}$, $\mathcal{W} \subseteq \mathbb{R}^p$, and a $\mathbf{w} \in \mathcal{W}$ we call a vector $\mathbf{v} \in \mathbb{R}^p$ a subgradient of F at $\mathbf{w}$ if

$$F(\mathbf{u}) \geq F(\mathbf{w}) + \mathbf{v} \cdot (\mathbf{u} - \mathbf{w}) \qquad \forall \mathbf{u} \in \mathcal{W}.$$

The set of all such vectors $\mathbf{v}$ is denoted by $\partial F(\mathbf{w})$ and called subdifferential of F at $\mathbf{w}$.

$F(w)$ 凸函数

Illustration

- The condition on the subgradient $\mathbf{v}$

$$F(\mathbf{u}) \geq F(\mathbf{w}) + \mathbf{v} \cdot (\mathbf{u} - \mathbf{w})$$

mimics the property of the gradient for smooth convex functions F:

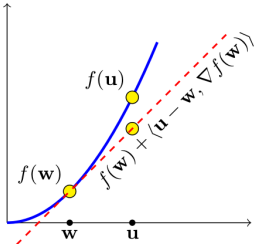
$$F(\mathbf{u}) \geq F(\mathbf{w}) + \nabla F(\mathbf{w})^\top (\mathbf{u} - \mathbf{w})$$

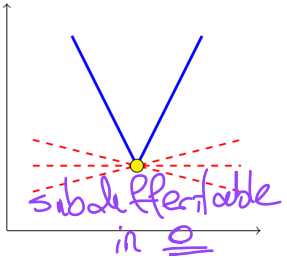
- Further interpretation: the affine approximation to F

$$\widetilde{F}(\mathbf{u}) := F(\mathbf{w}) + \mathbf{v} \cdot (\mathbf{u} - \mathbf{w}), \qquad \mathbf{u} \in \mathcal{W}$$

is a lower bound or undershoots F.

- In particular, Lipschitz continuous functions like $F(x) = |x|$ are subdifferentiable.





Source: "Understanding Machine Learning" (2014)

Subgradients for the Hinge loss

- Based on simple arithmetics we obtain for the Hinge loss

$$\ell_{\text{hinge}}(\mathbf{w};\mathbf{x},y) = \max\{0, 1 - y(\mathbf{w}\cdot\mathbf{x})\},$$

for fixed $\mathbf{x}$ and y, that

$$\partial\ell(\mathbf{w};\mathbf{x},y) \ni \mathbf{v} = \begin{cases} \mathbf{0}, & 1 - y(\mathbf{w}\cdot\mathbf{x}) \leq 0 \\ -y\,\mathbf{x}, & 1 - y(\mathbf{w}\cdot\mathbf{x}) > 0. \end{cases}$$

- For the (regularizred) empirical Hinge risk with $\lambda \geq 0$

$$F_s(\mathbf{w}) = \lambda\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{j=1}^{m} \ell_{\text{hinge}}(\mathbf{w};\mathbf{x}_i,y_i)$$

we obtain by linearity of the subdifferential that

$$2\lambda\mathbf{w} - \frac{1}{m}\sum_{i:\, y_i(\mathbf{w}\cdot\mathbf{x}_i)<1} y_i\,\mathbf{x}_i \quad \in \quad \partial F_S(\mathbf{w}).$$

Subgradient descent

- To minimize convex, non-differentiable but Lipschitz continuous $F\colon \mathbb{R}^p \to \mathbb{R}$ we can apply

Subgradient descent method

Given initial state $\mathbf{w}_0 \in \mathbb{R}^p$ compute for $k = 0, 1, 2, \ldots$

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k \mathbf{v}_k, \qquad \mathbf{v}_k \in \partial F(\mathbf{w}_k)$$

with suitable stepsizes $\eta_k > 0$.

- In general, the sequence of function values $F(\mathbf{w}_k)$ does **not** decrease monotonically. Therefore one often considers

$$F_k^{\text{best}} := \min_{j=0,\ldots,k} F(\mathbf{w}_j), \qquad \mathbf{w}_k^{\text{best}} := \operatorname*{argmin}_{\mathbf{w}_j \in \{\mathbf{w}_0,\ldots,\mathbf{w}_k\}} F(\mathbf{w}_j).$$

or the running mean

$$\overline{\mathbf{w}}_k := \frac{1}{1+k} \sum_{j=0}^{k} \mathbf{w}_j$$

Convergence analysis of subgradient descent

Some facts about properties of functions $F\colon \mathcal{W} \to \mathbb{R}$ and their relation to the subdifferential (without proof):

1. F is convex **if and only if** $\partial F(\mathbf{w}) \neq \emptyset$ for all $\mathbf{w} \in \mathcal{W}$. otherwise it might stuck (algo)
2. If F is differentiable in $\mathbf{w} \in \mathcal{W}$, then $\partial F(\mathbf{w}) = \{\nabla F(\mathbf{w})\}$.
3. A convex F is L-Lipschitz, i.e., $|F(\mathbf{w}) - F(\mathbf{v})| \leq L\|\mathbf{v} - \mathbf{w}\|$, **if and only if** for all $\mathbf{w} \in \mathcal{W}$ and $\mathbf{v} \in \partial F(\mathbf{w})$ we have $\|\mathbf{v}\| \leq L$.

By that we obtain as a first result:

Lemma 5.9:

Let $F\colon \mathbb{R}^p \to \mathbb{R}$ be L-Lipschitz and convex, then we have for subgradient descent:

$$F_k^{\text{best}} - F(\mathbf{w}^*) \leq \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + L^2 \sum_{j=0}^k \eta_j^2}{2 \sum_{j=0}^k \eta_j}$$

- The previous lemma yields a sufficient condition for convergence regarding the decay of the stepsizes η_k:

$$\sum_{k=0}^{\infty} \eta_k = +\infty, \qquad \sum_{k=0}^{\infty} \eta_k^2 < +\infty.$$

Hence, the stepsizes η_k should decay, but not too fast

- We will encounter this condition again for the stochastic gradient method.

Theorem 5.10:

Let $F\colon \mathbb{R}^p \to \mathbb{R}$ be convex and L-Lipschitz. Then we have for the subgradient method with stepsizes

$$\eta_j := \frac{\eta_0}{\sqrt{j+1}}, \qquad \eta_0 > 0,\ j = 0, 1, 2 \ldots,$$

that

$$F_k^{\text{best}} - F(\mathbf{w}^*) \in \mathcal{O}\left(\frac{1}{k^q}\right), \qquad q < \frac{1}{2}.$$

Stochastic subgradient descent

- The stochastic subgradient descent works analogously with updates

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k \mathbf{v}_k, \qquad \mathbf{v}_k \in \partial f_{i_k}(\mathbf{w}_k).$$

where i_k is drawn at random from $\mathrm{U}([m])$ and even $\mathbf{v}_k$ could be drawn randomly from the subdifferential $\partial f_{i_k}(\mathbf{w}_k)$.

- Stochastic subgradient descent can be applied, for instance, to soft (kernel) SVM for large data sets

$$F_s(\mathbf{w}) = \lambda \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{j=1}^{m} \ell_{\mathsf{hinge}}(\mathbf{w}; \mathbf{x}_i, y_i)$$

where then with i_k drawn from $\mathrm{U}([m])$

$$\mathbf{v}_k = \begin{cases} 2\lambda \mathbf{w}_k, & y_{i_k}(\mathbf{w}_k \cdot \mathbf{x}_{i_k}) \geq 1 \\ 2\lambda \mathbf{w}_k - y_{i_k} \mathbf{x}_{i_k}, & y_{i_k}(\mathbf{w}_k \cdot \mathbf{x}_{i_k}) < 1. \end{cases}$$

The convergence analysis can be done analogously to the deterministic version:

Lemma 5.11:

Are all $f_i\colon \mathbb{R}^p \to \mathbb{R}$, $i = 1, \dots, m$, L-Lipschitz and convex, the we have for the stochastic subgradient descent with initial vector $\mathbf{w}_0 \in \mathbb{R}^p$

$$\mathbb{E}\big[F_k^{\text{best}} - F(\mathbf{w}^*)\big] \leq \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + L^2 \sum_{j=0}^{k} \eta_j^2}{2 \sum_{j=0}^{k} \eta_j}$$

Theorem 5.12:

Let $F\colon \mathbb{R}^p \to \mathbb{R}$ be convex and L-Lipschitz. Then we have for the stochastic subgradient method with stepsizes

$$\eta_j := \frac{\eta_0}{\sqrt{j+1}}, \qquad \eta_0 > 0,\ j = 0, 1, 2 \dots,$$

that

$$\mathbb{E}\big[F_k^{\text{best}} - F(\mathbf{w}^*)\big] \in \mathcal{O}\left(\frac{1}{k^q}\right), \qquad q < \frac{1}{2}.$$