Useful Theorems and Results in Statistics, Optimization, Probability, and More

He Li
Stern School of Business, New York University
hli@stern.nyu.edu

1. Optimization

1.1 Convex Optimization

A function f is called α-*strongly convex* if for any $\boldsymbol{x}, \boldsymbol{y}$ in its domain,

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^\top (\boldsymbol{x} - \boldsymbol{y}) \geq \alpha \|\boldsymbol{x} - \boldsymbol{y}\|_2^2. \tag{1.1}$$

An equivalent condition is the following:

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \leq \nabla f(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{y}) - \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2. \tag{1.2}$$

A continuously differentiable function f is called β-*smooth* if the gradient of f is β-Lipschitz continuous, that is

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_* \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|, \ \ \forall \boldsymbol{x}, \boldsymbol{y}, \tag{1.3}$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. Usually we will use 2-norm for all.

Lemma 1.1. If a function f is β-smooth, then we have

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^\top (\boldsymbol{x} - \boldsymbol{y}) \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \tag{1.4}$$

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \geq \nabla f(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{y}) + \frac{\beta}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2. \tag{1.5}$$

Proof. We only prove Inequality (1.5). We define $g(t) = f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))$, by (1.4)

$$g'(t) - g'(0) = (\nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}))^\top (\boldsymbol{y} - \boldsymbol{x}) \leq t\beta \|\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

Therefore,

$$\begin{aligned} f(y) = g(1) = g(0) + \int_0^1 g'(t)\,\mathrm{d}t & \\ \leq g(0) + g'(0) + \frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 & \\ = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2. & \end{aligned}$$

$\square$

Remark 1.2. Notice that the technique $g(t) = f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))$ can also be used to prove Inequality (1.2).

Theorem 1.3 ((*co-coercivity*) Lemma 3.11 in Bubeck (2014))**.** Let f be β-smooth and α-strongly convex. Then for all $\boldsymbol{x}, \boldsymbol{y}$, we have

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^\top (\boldsymbol{x} - \boldsymbol{y}) \geq \frac{\alpha\beta}{\alpha+\beta}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{1}{\alpha+\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2. \tag{1.6}$$

Remark 1.4. Notice that this theorem can be used in convergence analysis of gradient method,

$$\begin{aligned}
\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|_2^2 &= \|\boldsymbol{w}_t - \eta\nabla f(\boldsymbol{w}_t) - \boldsymbol{w}^*\|_2^2 \\
&= \|\boldsymbol{w}_t - \boldsymbol{w}^*\|_2^2 - 2\eta\left(\nabla f(\boldsymbol{w}_t) - \nabla f(\boldsymbol{w}^*)\right)^\top (\boldsymbol{w}_t - \boldsymbol{w}^*) + \eta^2 \|\eta\nabla f(\boldsymbol{w}_t) - \eta\nabla f(\boldsymbol{w}^*)\|_2^2 \\
&\leq \left(1 - \frac{2\eta\alpha\beta}{\alpha+\beta}\right)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \left(\eta^2 - \frac{2\eta}{\alpha+\beta}\right)\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2.
\end{aligned}$$

Theorem 1.5 (Lemma 3.1 in Srebro et al. (2010))**.** If a non-negative function $f : \mathcal{W} \mapsto \mathbb{R}_+$ is β-smooth w.r.t. some norm $\|\cdot\|$ for some input space $\mathcal{W}$, then we have

$$\|\nabla f(\boldsymbol{w})\|_*^2 \leq 4\beta f(\boldsymbol{w}), \tag{1.7}$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

2. Concentration

Theorem 2.1 (Juditsky and Nemirovski (2008))**.** Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be independent copies of a zero-mean random vector $\boldsymbol{x}$. Then we have

$$\mathbb{E}\left\|\frac{1}{2}\sum_{i=1}^n \boldsymbol{x}_i\right\|_2^2 \leq \frac{1}{n}\mathbb{E}\|\boldsymbol{x}\|_2^2. \tag{2.1}$$

Remark 2.2 (Symmetrization Technique, Equation (4.17) in Wainwright (2019))**.** When our sequence

$$f(X_1, \ldots, X_n) \triangleq \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right|$$

is a bounded difference sequence with $\frac{c}{n}$, using McDiarmid's inequality, we have

$$f(X_1, \ldots, X_n) \leq \mathbb{E}f(X_1, \ldots, X_n) + C\sqrt{\frac{\log(1/\delta)}{n}},$$

with probability at least $1-\delta$. Now consider to bound the $\mathbb{E}f(X_1,\ldots,X_n)$ term using symmetrization argument:

$$\begin{aligned}
\mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}f(X)\right|\right] &= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \left(f(X_i)-\mathbb{E}_{Y_i}f(Y_i)\right)\right|\right] \\
&= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}_Y\frac{1}{n}\sum_{i=1}^n \left(f(X_i)-f(Y_i)\right)\right|\right] \\
&\le \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \left(f(X_i)-f(Y_i)\right)\right|\right],
\end{aligned}$$

and let $(\varepsilon_1,\ldots,\varepsilon_n)$ be an i.i.d. sequence of Rademacher variables, independent of X and Y. Hence we have,

$$\begin{aligned}
\mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \left(f(X_i)-f(Y_i)\right)\right|\right] &= \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(f(X_i)-f(Y_i)\right)\right|\right] \\
&\le 2\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] := 2\mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

Here $\mathcal{R}_n(\mathcal{F})$ is the Rademacher complexity of class $\mathcal{F}$ and can be bounded by Dudley's integral entropy (Wainwright, 2019, Theorem 5.22) bound and further bounded by its VC dimension (Wainwright, 2019, Equation (5.48)).

3. Useful Inequalities

$$\sqrt{1-x}\le 1-\frac{x}{2}. \tag{3.1}$$

References

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pages 2199–2207, 2010.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.