EECS 598-005: Theoretical Foundations of Machine Learning **Fall 2015**

Lecture 3: Concentration Inqualities

Lecturer: Jacob Abernethy *Scribes: Yuchen Jiang, Editors: Yuan Zhuang*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

3.1 Fenchel duality review

Definition 3.1. *Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, its Fenchel dual (conjugate) function $f^* : \mathbb{R}^n \to \mathbb{R}$ is defined by*

$$f^*(\theta) = \sup_{\mathbf{x} \in \mathrm{dom}(f)} \{\mathbf{x}^T\theta - f(\mathbf{x})\}$$

Fact 3.2. *The function $f^*(\theta)$ is convex in θ. (Inside the supremum is a linear function in θ, and the supremum of convex functions is convex).*

Fact 3.3 (Fenchel-Young inequality)**.** *The inequality $f(\mathbf{x}) + f^*(\theta) \geq \mathbf{x}^T\theta$ holds for all $\mathbf{x} \in \mathrm{dom}(f)$ and $\theta \in \mathrm{dom}(f^*)$.*

Proof: For all $\mathbf{x} \in \mathrm{dom}(f)$ and $\theta \in \mathrm{dom}(f^*)$, we have

$$\begin{aligned} f(\mathbf{x}) + f^*(\theta) &= f(\mathbf{x}) + \sup_{\mathbf{x}' \in \mathrm{dom}(f)} \{\mathbf{x}'^T\theta - f(\mathbf{x}')\} \\ &\geq f(\mathbf{x}) + \mathbf{x}^T\theta - f(\mathbf{x}) = \mathbf{x}^T\theta \end{aligned}$$

■

Examples

- Compute the dual of $f(x) = \frac{1}{p}|x|^p$ $(p > 1)$.

  Let $\Phi(x) = x\theta - \frac{1}{p}|x|^p$ so $f^*(\theta) = \sup_x \Phi(x)$. Setting $\nabla_x \Phi(x) = 0$, we get $\theta = |x|^{p-1}\,\mathrm{sgn}(x)$[1]. Noting that x and θ must have the same sign, we conclude $x = |\theta|^{\frac{1}{p-1}} \cdot \mathrm{sgn}(\theta)$. Therefore,

  $$\begin{aligned} f^*(\theta) &= |\theta|^{\frac{1}{p-1}} \cdot \mathrm{sgn}(\theta) \cdot \theta - \frac{1}{p}|\theta|^{\frac{p}{p-1}} \\ &= \left(1 - \frac{1}{p}\right)|\theta|^{\frac{p}{p-1}} \\ &= \frac{1}{q}|\theta|^q, \end{aligned}$$

  where $1/p + 1/q = 1$. In this case, Fenchel-Young inequality reduces to Young's inequality.

- The Fenchel dual of the function $f(\mathbf{x}) = \frac{1}{p}\|x\|_p^p$ is $f^*(\theta) = \frac{1}{q}\|\theta\|_q^q$, where p and q are nonnegative real numbers satisfying $1/p + 1/q = 1$.

[1] $\mathrm{sgn}(x) = \begin{cases} 1 & \text{if } x > 0; \\ 0 & \text{if } x = 0; \\ -1 & \text{if } x < 0. \end{cases}$

3.2 Random variables and deviation bounds

3.2.1 A few concepts about random variables.

- A *random variable* (r.v.) is a measurable function $X : \Omega \to \mathbb{R}$, where Ω is a *sample space*.
- For any *measurable set* $A \subseteq \mathbb{R}$, we have $\mathbb{P}(X \in A)$, the probability that X belongs to A.
- Every random variable X on $\mathbb{R}$ has a *cumulative distribution function (CDF)*: $F(t) = \mathbb{P}(X \leq t)$.
- If F is differentiable, we have *probability density function (PDF)*: $f(t) = F'(t)$.
- Given a PDF f we have $\mathbb{P}(a \leq X \leq b) = \int_a^b f(t)dt$.
- Two r.v.s X and Y are *independent* if for all $A, B \subseteq \mathbb{R}$, $\mathbb{P}(X \in A \text{ and } Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$
- X and Y are independent implies $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.
- $Var(X) := \mathbb{E}[(X - E[X])^2]$.

Fact 3.4. *1. If X, Y are independent, $Var(X + Y) = Var(X) + Var(Y)$.*

2. $Var(\alpha X) = \alpha^2 Var(X)$.

Therefore if $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables (i.i.d. r.v.s), we have $Var(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{Var(X_1)}{n}$.

3.2.2 Deviation bounds

Theorem 3.5 (Markov's inequality)**.** *Lex X be a r.v. taking only non-negative values. Then for any $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof: Note that $X \geq t\mathbf{1}_{\{X \geq t\}}$ [2]. Therefore, $\mathbb{E}[X] \geq \mathbb{E}[t\mathbf{1}_{\{X \geq t\}}] = t\,\mathbb{P}(X \geq t)$. ■

Theorem 3.6 (Chebyshev's inequality)**.** *Let X be a r.v. with $\mathbb{E}[X] = \mu$, $Var(X) = \sigma^2$, then*

$$\mathbb{P}(|X - \mu| \geq t\sigma) \leq \frac{1}{t^2}.$$

Proof: Let $Y = (X - \mathbb{E}[X])^2$. Using Markov's inequality,

$$\begin{aligned}\mathbb{P}(|X - \mu| \geq t\sigma) &= \mathbb{P}((X - \mu)^2 \geq t^2\sigma^2)) \\ &= \mathbb{P}(Y \geq t^2\,\mathbb{E}[Y]) \\ &\leq \frac{1}{t^2}.\end{aligned}$$

■

The Chebyshev deviation bound is much too weak for a specfic case, namely sums/averages of independent random variables. The Central Limit Theorem tells us that the average of n independent r.v.'s, when scaled appropriately, looks gaussian and hence the tail of the distribution decays very quickly.

[2] For any set A, $\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A; \\ 0 & \text{if } x \notin A \end{cases}$

Theorem 3.7 (Central Limit Theorem (CLT))**.** *Let $X_1, \ldots, X_n$ be i.i.d. r.v.s, with $\mathbb{E}[X_1] = \mu$, $Var(X_1) = \sigma^2$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then,*

$$\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} N(0,1),$$

that is,

$$\mathbb{P}\left(\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma} \leq t\right) \to \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

as $n \to \infty$.

By applying Chebyshev's inequality to the random variable $\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma}$, we have

$$\mathbb{P}\left(\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma} \geq t\right) \leq \frac{1}{t^2}.$$

Chebyshev's inequality is tight, in the sense that there exist r.v.s for which Chebyshev's inequality is tight. However, the decay rate given by Chebyshev's for sums of i.i.d. r.v.s is slower than we might expect given the central limit theorem. That is, it is slower than the decay rate for the tail probability of standard normal distribution. In fact, we can get a nice and simple expression to control the tail probabilities of sums of independent random variables (under certain restrictions) by using so-called *Chernoff Bounds*, one of the most popular of which is due to *Hoeffding*.

Theorem 3.8 (Hoeffding's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent r.v.s, taking values in $[0,1]$. $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\mu = \mathbb{E}(\bar{X}_n)$, then*

$$\mathbb{P}(\bar{X}_n - \mu \geq t) \leq \exp(-2nt^2).$$

Lemma 3.9 (Hoeffding's lemma)**.** *Assume X takes values in $[a,b]$, $\mathbb{E}[X] = 0$ then, for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[e^{\lambda x}] \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right).$$

Using Hoeffding's lemma and Markov's inequality, we are able to prove Theorem 3.8. We will finish the proof of this in the next lecture, but here is a sketch.

$$\begin{aligned}
\mathbb{P}(\bar{X}_n > t) &= \mathbb{P}(e^{\lambda \bar{X}_n} > e^{\lambda t}) \\
&\leq \frac{\mathbb{E}[e^{\lambda \bar{X}_n}]}{e^{\lambda t}} \\
&= \frac{1}{e^{\lambda t}} \left(\mathbb{E}[e^{\frac{\lambda X_1}{n}}]\right)^n \\
&= \exp\left(\frac{\lambda^2}{8n} - \lambda t\right),
\end{aligned}$$

where the first inequality follows from Markov's inequality, the next lines follows from independence and the final line follows from Hoeffding's lemma. The right hand side attains its minimum of $\exp(-2nt^2)$ when $\lambda = 4nt$.