Large Deviation

Contents

1 Introduction

2 Review of Convex Analysis

2.1 Convex Functions

2.2 Conjugate Functions

Conjugacy transform asscociates any function f, with a convex function f^*, called the conjugate of f. The conjugate function describes the function f in terms of affines functions that are majorized by f or in other words in terms of the envelop which hyperplanes which support the function. Consider an extended real-valued function $f : \mathbb{R}^n \to [-\infty, \infty]$. The conjugate function of f is the function $f^* : \mathbb{R}^n \to [-\infty, \infty]$ defines by

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \left\{ x^T y - f(x) \right\} \qquad y \in \mathbb{R}^n \tag{1}$$

Essentially, if $\mathcal{R}$ is the set of all functions $f : \mathbb{R} \to [-\infty, \infty]$ and $\mathcal{C}$ is the set of all convex functions, then conjugacy transform is a mapping from $\mathcal{R}$ to $\mathcal{C}$. If $\mathcal{C}'$ is the set of all closed proper convex functions, then this mapping is a bijection from $\mathcal{C}'$ to $\mathcal{C}$. In other words the transformation is symmetric, in the sense, f can be recovered by taking cojugate of the conjugate of f.

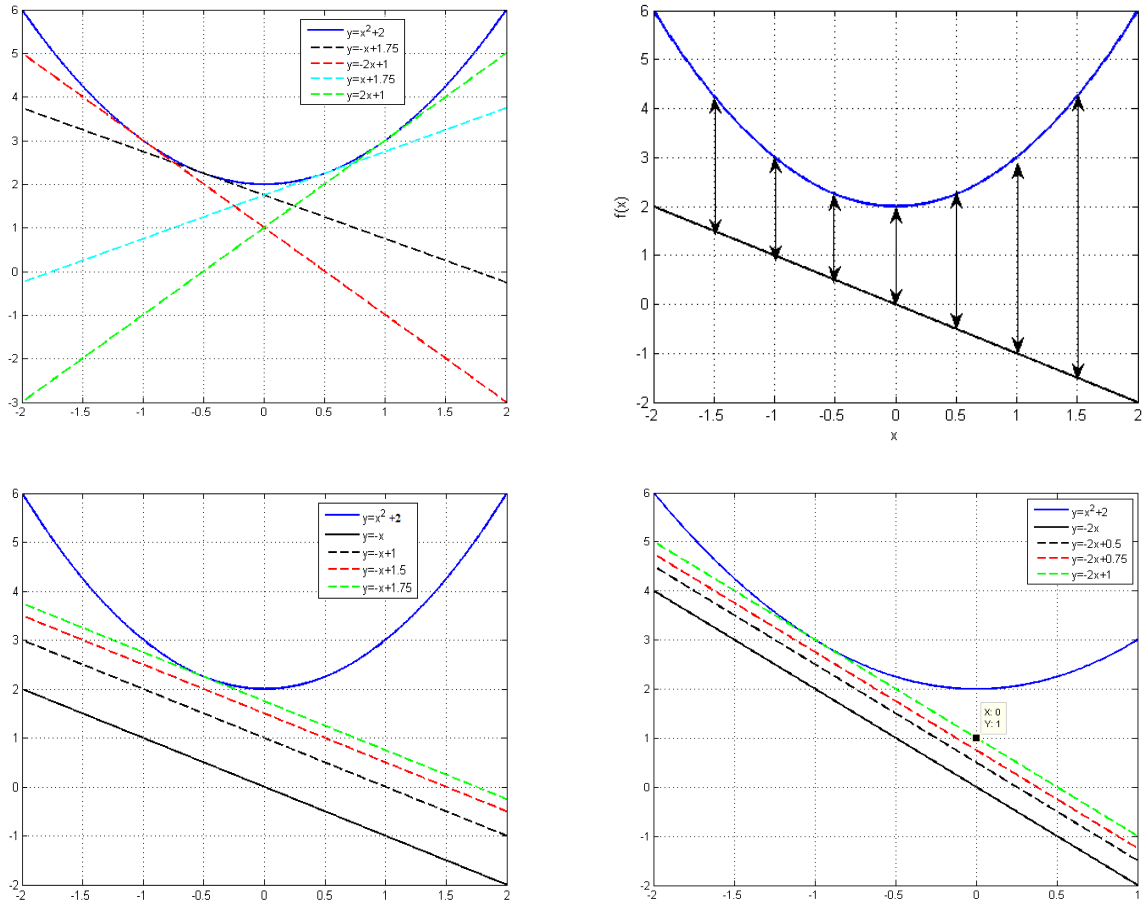
We know that a function is defined by specifying the value of the function for each value of x belonging to the domain of f, that is, by providing the locus of the path or the curve or the surface of the function. Alternatively, we can specify the function as the envelop of the hyperplanes or tangents. These hyperplanes or tangent are in turn specified by the their normals or slopes (or intercepts).
Let us first consider an simple example to understand what does equation (1) signifiy.
Example: Consider the following function $f : \mathbb{R} \to \infty$

$$f(x) = x^2 + 2 \tag{2}$$

As shown in the below figure, we can specify the function either by specifying the value of y for each value of x or we can specify the function by the giving the intercept of the tangents for each value of the slope.
In order to obtain the envelop of the the tangents, for each slope μ, we need to obtain the function $g^*_\mu(x) = \mu x + c^*_\mu$ which majorizes all the functions $g_\mu(x) = \mu x + c_\mu$ and is such that $g^*_\mu(x)$ is majozised f. Using this slope and intercept c^*_μ we could get a catalog of slopes versus intercepts for all the slopes and reconstruct the function, or atleast the convex hull of the function, which is nothing but the envelop of the tangents with slopes μ and intercept c^*_μ.

In the above figure (c) we see that for $\mu = -1$, $g_{-1}(x) = -x + c$, we see that $g^*_{-1} = -x + 1.75$. Similarly, for $\mu = -2$, $g_{-2}(x) = -2x + c$, we see that $g^*_{-2} = -2x + 1$. Thus the function f can be alternatively described using this information and we could form the preliminary definition of the conjugate function as follows:

$$f^*(\mu) = c^*_\mu \tag{3}$$



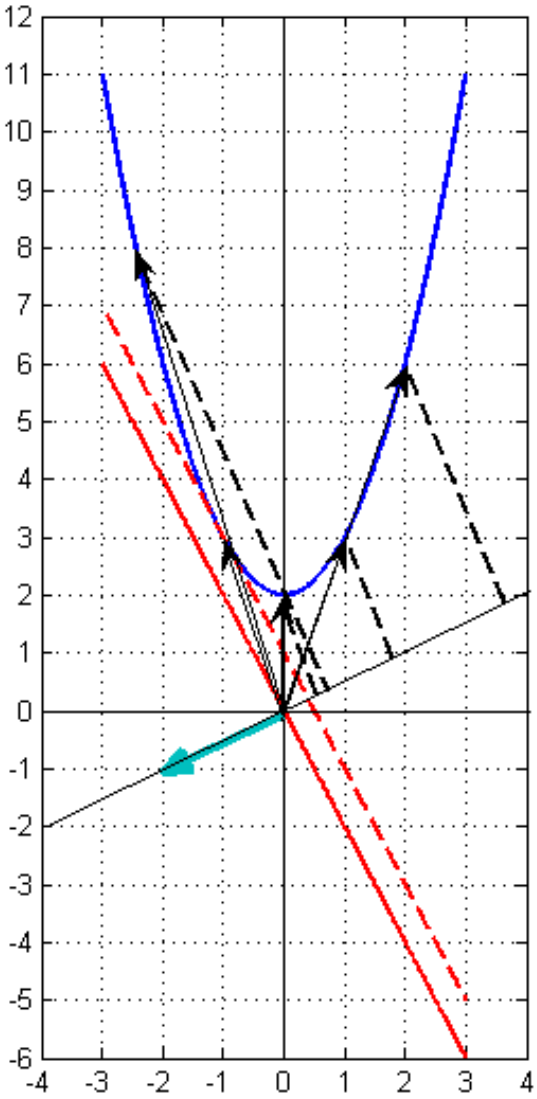
From the above definition of f^*, it is clear that f^* describes the function f in terms of the slopes and intercept. Rather f^* is like a catalog of slopes versus intercepts of the tangents which envelop the function f. This definition would suffice if we are considering the function in one dimension. When we go to higher dimensions, we take the envelop of hyperplanes which are the analog of tangents in the higher dimension and since hyperplanes are specified by normal vectors rather than their slopes, the definition of f^* has to be changed from slope versus intercept to normal versus intercept. Let us consider the above example again, but instead of the slope $\mu = 1$, we now consider the unit normal vector $\left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]$, which essentially represents the same tangent. But now let us make a slight modification. Instead of specifying the unit vector of the normal, let us specify the normal vector as $[\boldsymbol{\nu}, 1]$, that is, normalize the last co-ordinate of the normal vector to be 1. In this case, for slope $\mu = -1$, $\boldsymbol{\nu} = 1$ and $\mu = -2$, we need $\begin{bmatrix}\boldsymbol{\nu} & 1\end{bmatrix}\begin{bmatrix}-1\\2\end{bmatrix} = 0$, so we get $\boldsymbol{\nu} = 2$. Likewise for higher dimensions, $\boldsymbol{\nu}$ will be of appropriate dimensions. Now we can use $\boldsymbol{\nu}$ and intercept to specify $f^*(\boldsymbol{\nu})$. But we know that hyperplanes are defined as the set of all points in $\mathbb{R}^n$ such that

$$\{\mathbf{x} \in \mathbb{R}^n | \mathbf{a}^T\mathbf{x} = b\} \tag{4}$$

where $\mathbf{a}$ is the normal vector to the hyperplane. That means, there are two parameters which

specify the hyperplane, one is the normal vector $\mathbf{a}$ and the other is the value b, which is should be the value of the projection of the vectors on the normal vector. When we calculated $\boldsymbol{\nu}$, we got the normal of the hyperplane, now we need to find out where should the hyperplane be placed, that is, what should be the value b, such that the hyperplace just touches the function and thus all the hyperplanes together form the envelop of the function which is majorized by

f.



In the adjoining figure, the blue line represents the function $f(x) = x^2+2$. The red line represents the hyperplane which passes through 0. The cyan coloured arrow represents the direction of the normal to the hyperplane. We take the projection of the vectors $[x, f(x)]$ on the normal vector. The vector which gives the maximum projection, say $[x^* f(x^*)]$, is the vector which will touch the hyperplane or on otherwords it will be the vector which would belong to the hyperplane if the hyperplane passing through 0 is shifted by that value b^* (which is given by $b^* = \begin{bmatrix} x^* & f(x^*) \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix}$) in the direction $b^*\mathbf{a}$.

$$b^* = \sup_x \left\{ \begin{bmatrix} x & f(x) \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} \right\} \quad (5)$$

In other words, we can catalog for each value of $\boldsymbol{\nu}$, the value of b^*. In the above case $\boldsymbol{\nu} = 2$. Therefore, more generally we have,

$$f^*(\boldsymbol{\nu}) = \sup_x \left\{ \begin{bmatrix} \mathbf{x}^T & f(x) \end{bmatrix} \begin{bmatrix} \boldsymbol{\nu} \\ -1 \end{bmatrix} \right\} \quad (6)$$

From the above example we understand how one arrives at equation (1).

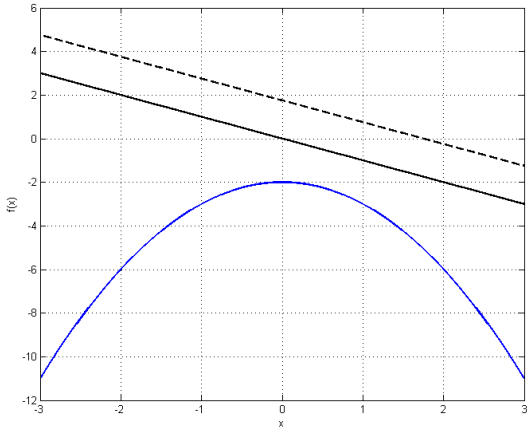
2.2.1 Discussion

We now discuss some of the points which probably might probably help further clear why the conjugate function is of the form of as shown in equation (1).

- Why did we take the normal vector of the form $\begin{bmatrix}\boldsymbol{\nu} & -1\end{bmatrix}$ and not $\begin{bmatrix}\boldsymbol{\nu} & 1\end{bmatrix}$?
  It is just a convention. We could take $\begin{bmatrix}\boldsymbol{\nu} & 1\end{bmatrix}$ instead as well. Then we would have infimum instead of supremum to get the vector at which the hyperplane touches the $[xf(x)]$.

$$f^*(\boldsymbol{\nu}) = \inf_x \left\{ \begin{bmatrix}\mathbf{x}^T & f(x)\end{bmatrix} \begin{bmatrix}\boldsymbol{\nu} \\ 1\end{bmatrix} \right\} \tag{7}$$

- Suppose we have a function $f(x) = -x^2 - 2$, then what would be its conjugate?



In this case, $f^*(\boldsymbol{\nu}) = -\infty$, for all $\boldsymbol{\nu}$, because from the fundamental notion of conjugate function, we were looking at the those hyperplanes which are majorized by the function f. In this case all the hyperplane need to have $b = -\infty$.

2.2.2 Legendre-Fenchel transform

3 Review of Random variables

3.1 Moment Generating function

3.2 Cumulant and cumulant generating function

4 Large Deviation Principle

4.1 Inequalities

4.1.1 Markov Inequality

Lemma (Markov Inequality): Consider a probability space $(\Omega, \mathcal{F}, P)$. Consider a non-negative random variable X defined on it. Let P_X be the induced measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$.

Markov inequality states that for any $c > 0$

$$P_X(\{X : X \geq c\}) \leq \frac{\mathbb{E}[X]}{c} \tag{8}$$

Proof:

$$c\mathbb{I}_{\{X \geq c\}} \leq X \tag{9}$$

Taking expectation on both sides we get,

$$\begin{aligned} \int_x c\mathbb{I}_{\{X \geq c\}} dP_X(x) &\leq \mathbb{E}[X] && (10)\\ cP_X(\{X : X \geq c\}) &\leq \mathbb{E}[X] && (11)\\ P_X(\{X : X \geq c\}) &\leq \frac{\mathbb{E}[X]}{c} && (12) \end{aligned}$$

4.1.2 Chebyshev's Inequality

Lemma (Chebyshev's Inequality): Chebychev inequality states that if X is a random variable with finite mean μ and variance σ^2, then for any $d > 0$,

$$P_X\big(\{X : |X - \mu| \geq d\}\big) \leq \frac{\sigma^2}{d^2} \tag{13}$$

Proof:

$$\begin{aligned} P_X\big(\{X : |X - \mu| \geq d\}\big) &= P_X\big(\{X : |X - \mu|^2 \geq d^2\}\big) && (14)\\ &\leq \frac{\mathbb{E}[|X - \mu|^2]}{d^2} && (15)\\ &= \frac{\sigma^2}{d^2} && (16) \end{aligned}$$

4.2 Weak Law of Large Numbers

We shall now consider the weak law of large numbers which is a direct consequence of Chebyshev's inequality. Consider the sum of N identically distributed statisitically independent random variables $\{X_i\}$, each with mean μ_x and variance σ_x^2. Let the sum of N such random variables be S_N.

$$S_N = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{17}$$

The mean and variance of S_N is given by

$$\begin{aligned} \mu_s &= \mathbb{E}[S_N] = \mu_x && (18)\\ \sigma_s^2 &= \mathbb{E}[(S_N - \mu_s)^2] && (19)\\ &= \mathbb{E}\left[\Big(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu_s\Big)^2\right] && (20)\\ & && (21) \end{aligned}$$

$$= \frac{1}{N^2}\mathbb{E}\left[\Big(\sum_{i=1}^{N} X_i - N\mu_s\Big)^2\right] \tag{22}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\Big(X_i - \mu_s\Big)^2\right] \tag{23}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sigma_x^2 \tag{24}$$

$$= \frac{\sigma_x^2}{N} \tag{25}$$

Let $Y_N = S_N - \mu_x$. Therefore,

$$\mu_y = 0 \tag{26}$$

$$\sigma_y^2 = \sigma_s^2 = \frac{\sigma_x^2}{N} \tag{27}$$

From Chebyshev's inequality we have,

$$P_{Y_N}(|y_N| \geq \epsilon) \leq \frac{\sigma_y^2}{\epsilon^2} \tag{28}$$

$$= \frac{\sigma_x^2}{N\epsilon^2} \tag{29}$$

$$\tag{30}$$

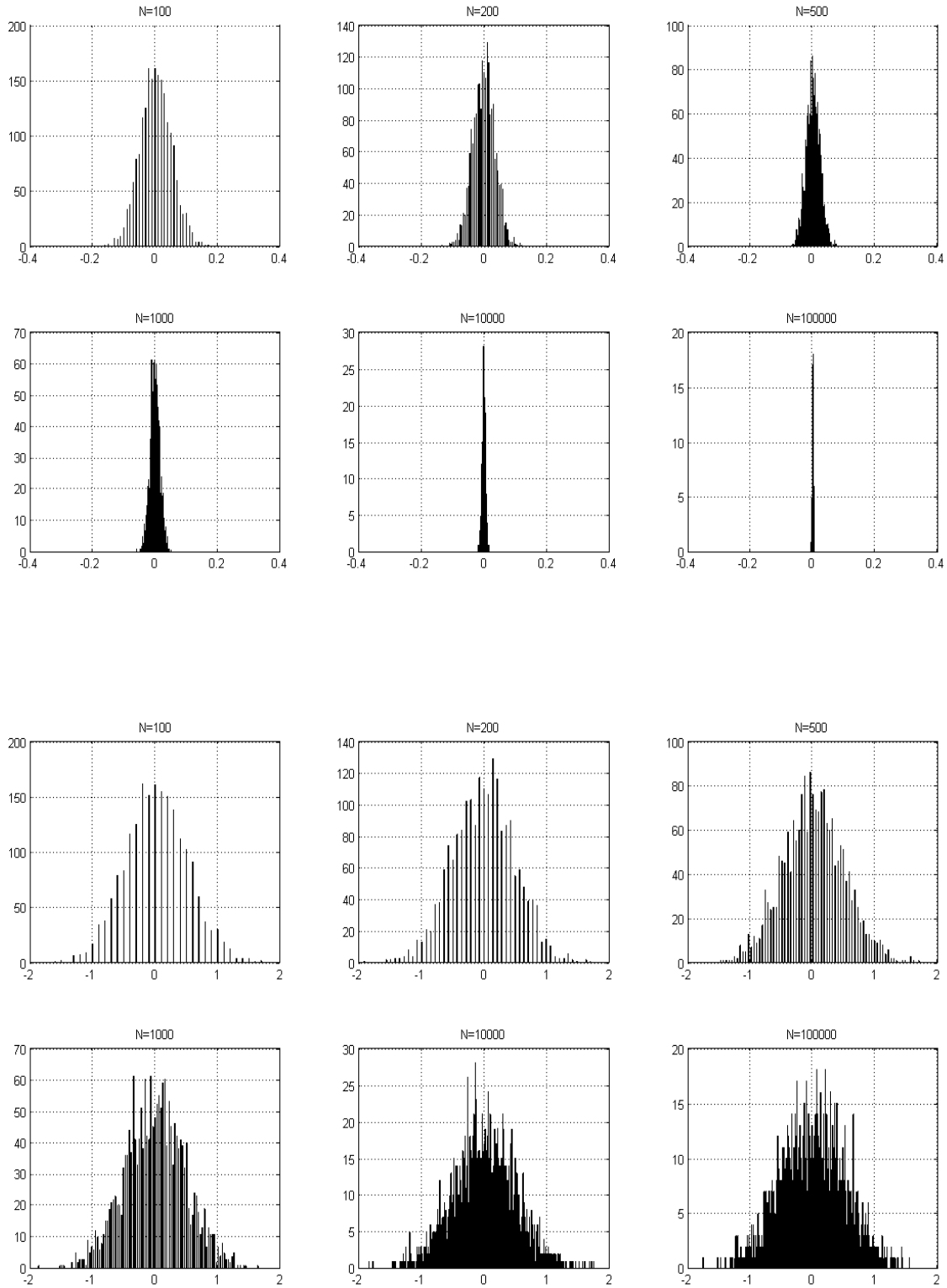
In other words,

$$P(|S_N - \mu_X| \geq \epsilon) \leq \frac{\sigma_x^2}{N\epsilon^2} \tag{31}$$

$$\lim_{N\to\infty} P(|S_N - \mu_X| \geq \epsilon) = 0 \tag{32}$$

The above equation in the statement of weak law of large numbers.

4.3 Central Limt Theorem



Proposition: (Central Limit Theorem) Suppose that $X_1, X_2, \cdots$ are i.i.d, each with mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Then the normalized sum

$$\frac{S_n - \mu}{\sqrt{n}} \tag{33}$$

converges in distribution to the $\mathcal{N}(0, \sigma^2)$ distribution as $n \to \infty$.

4.4 Chernoff Bound

4.5 Cramer's Theorem

In this section we try to bound the tail probablities, essentially bound the probability of rare events.

4.5.1 Cramer's Theorem

5 Vardhan's Integral

References

[1] Bruce Hajek, "Notes for ECE 534, An Exploration of Random Processes for Engineers", July 2011.