- **Independence**: $f_{X,Y}(X,Y) = f_X(x)f_Y(y)$
- $\mathbb{E}[X] = \int_{-\infty}^{\infty} f_X(x)dx$
- $\forall X, Y, \mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- X, Y indep, $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$
- X, Y follows $\mathcal{N}$ and indep, so does $X + Y$

Concentration Inequalities

Markov's inequality

For RV X with finite $\mathbb{E}[X]$ and $X \geqslant 0$:

$$\boxed{\mathbb{P}(X \geqslant t) \leqslant \mathbb{E}[X]/t \quad t > 0}$$

Pf. Refer to the property

$$\int_0^\infty \mathbb{P}(X \geqslant x)dx = x\mathbb{P}(X \geqslant x)\Big|_0^\infty - \int_0^\infty x d\mathbb{P}(X \geqslant x)$$
$$= 0 + \int_0^\infty f_X(x)dx = \mathbb{E}[X]$$

From here, consider

$$\int_0^\infty \mathbb{P}(X \geqslant x)dx \geqslant \int_0^t \mathbb{P}(X \geqslant t)dx \geqslant t\mathbb{P}(X \geqslant t)$$

Chebyshev's inequality

For RV X with finite $\mathbb{E}[X], \text{Var}(X)$ and $X \geqslant 0$:

$$\boxed{\mathbb{P}(|X - \mathbb{E}[X]| \geqslant t) \leqslant \text{Var}(X)/t^2 \quad t > 0}$$

Pf. Apply Markov's with $Y = |X - \mathbb{E}[X]|^2$

Chernoff's bound

For RV X with finite MGF $= \mathbb{E}[e^{\lambda X}]$, for any $t \geqslant 0$:

$$\boxed{\mathbb{P}(X \geqslant t) = \mathbb{P}(e^{\lambda X} \geqslant e^{\lambda t}) \leqslant \mathbb{E}[e^{\lambda X - \lambda t}]}$$

$$\boxed{\mathbb{P}(X \leqslant -t) = \mathbb{P}(e^{-\lambda X} \geqslant e^{\lambda t}) \leqslant \mathbb{E}[e^{-\lambda X - \lambda t}]}$$

Pf. Apply Markov's with $Y = e^{\pm \lambda X}$ (notice $Y \geqslant 0$). One may select a λ that to maximize $\mathbb{E}[e^{\lambda X - \lambda t}]$.

Chernoff's bound corollaries

- C^2-**Subgaussian RV** $\mathbb{E}[X] = 0$ and $\mathbb{E}[e^{\lambda X}] \geqslant e^{\lambda^2 C^2/2}$, using Chernoff's bound: $\mathbb{P}(X \geqslant t) \leqslant e^{\lambda^2 C^2/2 - \lambda t}$. Setting $\lambda = \frac{t}{C^2}$ yields the tightest bound: $\mathbb{P}(|X| \geqslant t) \leqslant e^{-t^2/(2C^2)}$
- If $X \in [a, b]$, then $X - \mathbb{E}[X]$ is $\frac{(a-b)^2}{2}$ subgaussian.

Hoeffding's bound

For IID RV $\in [a, b]$ with mean μ and emperical mean $\hat{\mu}_n$,

$$\boxed{\mathbb{P}(|\hat{\mu}_n - \mu| \geqslant t) \leqslant 2e^{\frac{-2nt^2}{(b-a)^2}} \quad t > 0}$$

Pf. $\hat{\mu}_n - \mu$ is $\frac{(b-a)^2}{4n}$ subgaussian (use the corollary).

Geometry of Higher Dimension Gaussian

Gaussian PDF and MGF

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

$$\boxed{M_X(t) = \mathbb{E}[X] = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)}$$

For a Gaussian RV $X \sim \mathcal{N}(\mu, \sigma^2)$

Gaussian Annulus Theorem

Let $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ then for any $t \in (0, \sqrt{d})$,

$$\boxed{\mathbb{P}(\|\mathbf{z}\|_2 \in [\sqrt{d} - t, \sqrt{d} + t]) \geqslant 1 - 2e^{-t^2/8}}$$

Pf. Write $\|\mathbf{z}\|_2^2 = \sum_{i=1}^n z_i^2$ then

$$\mathbb{E}[e^{\lambda(\|\mathbf{z}\|_2^2 - d)}] = \left(e^{-\lambda} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z_i^2(1-2\lambda)}{2}} dz_i\right)^d$$

The term inside the integral is a $\mathcal{N}$ pdf with variance $\frac{1}{1-2\lambda}$. For $\lambda < \frac{1}{2}$, the expectation is $\frac{e^{-d\lambda}}{(1-2\lambda)^{d/2}}$, which is $\leqslant e^{2d\lambda^2}$ for $|\lambda| < \frac{1}{4}$. Hoeffing's on $\|\mathbf{z}\|_2^2 - d$ yields

$$\mathbb{P}(|\|\mathbf{z}\|_2^2 - d| \geqslant t\sqrt{d}) \leqslant e^{2d\lambda^2 - \lambda t\sqrt{d}} \leqslant 2e^{-t^2/8}$$

Finally, $|\|\mathbf{z}\|_2^2 - \sqrt{d}| \geqslant t$ implies $|\|\mathbf{z}\|_2^2 - d| \geqslant t\sqrt{d}$

Near Orthogonality

Lem1. Fix $\|\mathbf{x}\|_2 = \ell$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, then for any $t \geqslant 0$

$$\boxed{\mathbb{P}(|\mathbf{z}^\top \mathbf{x}| \geqslant t) \leqslant 2e^{-\frac{t^2}{2\ell^2}}}$$

Pf. $|\mathbf{z}^\top \mathbf{x}| \sim \mathcal{N}(0, \ell^2)$ hence $\mathbb{E}[e^{\lambda(\mathbf{z}^\top \mathbf{x})}] = e^{\frac{\lambda^2 l^2}{2}}$. Apply Hoeffding's bound on $\mathbf{z}^\top \mathbf{x}$.

Lem2. $\mathbf{z}_1, \mathbf{z}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ are indep, then for any $t \geqslant 0$

$$\boxed{\mathbb{P}(|\mathbf{z}_1^\top \mathbf{z}_2| \geqslant td) \leqslant 2e^{-\frac{t^2 d}{4}}}$$

Pf. The MGF of the product of two standard Gaussians is $\frac{1}{\sqrt{1-\lambda^2}}$. Therefore, $\mathbb{E}[e^{\lambda(\mathbf{z}_1^\top \mathbf{z}_2)}] = (1-\lambda^2)^{-\frac{d}{2}} \leqslant (e^{-2\lambda^2})^{-\frac{d}{2}} = e^{\lambda^2 d}$ for $\lambda \in (0, 0.8]$. Apply Hoeffding's bound on $\mathbf{z}_1^\top \mathbf{z}_2$

Lem3. Same conditions as **lem2**, then for any $t \in (0, 1]$

$$\boxed{\mathbb{P}\left(\left|\frac{\mathbf{z_1}^\top \mathbf{z_2}}{||\mathbf{z_1}||_2 ||\mathbf{z_2}||_2}\right| \geqslant t\right) \leqslant 4e^{-\frac{t^2 d}{16}}}$$

Pf. Define event $\mathcal{E} := \{||\mathbf{z_1}||_2 \geqslant \sqrt{d/2}$ and $||\mathbf{z_2}||_2 \geqslant \sqrt{d/2}\}$ then $\mathcal{E}^c := \{||\mathbf{z_1}||_2 \leqslant \sqrt{d/2}$ or $||\mathbf{z_2}||_2 \leqslant \sqrt{d/2}\}$

$$\mathbb{P}(\mathcal{E}^c) \leqslant \mathbb{P}(||\mathbf{z_1}||_2 \leqslant \sqrt{d/2}) + \mathbb{P}(||\mathbf{z_2}||_2 \leqslant \sqrt{d/2})$$
$$\leqslant 2\mathbb{P}(||\mathbf{z_1}||_2 \leqslant \sqrt{d} - \sqrt{d/2})$$
$$\leqslant 2e^{-(\sqrt{d/2})^2/8} \text{ (Annulus Theorem)} = 2e^{-d/16}$$

Define $\mathcal{F} = \left\{\left|\frac{\mathbf{z_1}^\top \mathbf{z_2}}{||\mathbf{z_1}||_2 ||\mathbf{z_2}||_2}\right| \geqslant t\right\}$: $\mathbb{P}(\mathcal{F}) = \mathbb{P}(\mathcal{F}|\mathcal{E})\mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{F}|\mathcal{E}^c)\mathbb{P}(\mathcal{E}^c) \leqslant \mathbb{P}(\mathcal{F}|\mathcal{E}) + \mathbb{P}(\mathcal{E}^c)$. $\mathcal{F}$ given $\mathcal{E}$ implies $|\mathbf{z_1}^\top \mathbf{z_2}| \geqslant t^2 d/2$ which is $\leqslant 2e^{-dt^2/16}$ (**lem2**). Therefore, $\mathbb{P}(\mathcal{F}) \leqslant 2e^{-dt^2/16} + 2e^{-d/16} \leqslant 4e^{-dt^2/16}$

Johnson-Lindenstrauss Lemma

Given d-dimensional dataset $\mathbf{v}_1, \ldots, \mathbf{v}_n$, for $k < d$ sample $\mathbf{u}_1, \ldots, \mathbf{u}_k$ IID from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Define

$$\boxed{\pi(\mathbf{v}) = \frac{1}{\sqrt{k}}[\mathbf{u_1}^\top \mathbf{v}, \mathbf{u_2}^\top \mathbf{v}, \ldots, \mathbf{u_k}^\top \mathbf{v}]}$$

Lem4. π preserves magnitude: fix $\delta > 0$ and $\epsilon > 0$, for $k \geqslant \frac{8}{\varepsilon^2} \ln \frac{2}{\delta}$, the probability holds

$$\boxed{\mathbb{P}(|\|\pi(\mathbf{v})\|_2 - \|\mathbf{v}\|_2| \geqslant \epsilon\|\mathbf{v}\|_2) \leqslant \delta}$$

Pf. Since $\mathbf{u}_i^\top \mathbf{v} \in \mathcal{N}(0, \|\mathbf{v}\|_2^2)$, $\frac{\sqrt{k}}{||\mathbf{v}||_2}\pi(\mathbf{v}) \sim \mathcal{N}(0, \mathbf{I}_k)$. We apply the Annulus theorem:

$$\mathbb{P}\left(\left|\frac{\sqrt{k}}{\|\mathbf{v}\|_2}\|\pi(\mathbf{v})\|_2 - \sqrt{k}\right| \geqslant t\right) \leqslant 2e^{-\frac{t^2}{8}}$$

$$\mathbb{P}\left(|\|\pi(\mathbf{v})\|_2 - \|\mathbf{v}\|_2| \geqslant \frac{t\|\mathbf{v}\|_2}{k}\right) \leqslant 2e^{-\frac{t^2}{8}}$$

Set $\epsilon = \frac{t}{\sqrt{k}}$ and $2e^{-\frac{\epsilon^2 k}{8}} \leqslant \delta$, $k \geqslant \frac{8}{\varepsilon^2} \ln \frac{2}{\delta}$.

Lem5. π preserves distance: fix $\delta > 0$ and $\epsilon > 0$, for $k \geqslant \frac{16}{\varepsilon^2} \ln \frac{n}{\delta}$, the probability holds ($\mathbf{v}_{i,j} = \mathbf{v}_i - \mathbf{v}_j$)

$$\boxed{\mathbb{P}(\forall i, j, |\|\pi(\mathbf{v}_{i,j})\|_2 - \|\mathbf{v}_{i,j}\|_2| \geqslant \epsilon\|\mathbf{v}_{i,j}\|_2) \leqslant \delta}$$

Pf. The probability of each condition is $\leqslant 2e^{-\frac{\epsilon^2 k}{8}}$. Union bound suggests the total probability is $\leqslant \binom{n}{2} 2e^{-\frac{\epsilon^2 k}{8}} \leqslant n^2 e^{-\frac{\epsilon^2 k}{8}}$. Setting this to $\leqslant \delta$ yields $k \geqslant \frac{16}{\varepsilon^2} \ln \frac{n}{\delta}$

Estimation

Maximum Likelihood Estimation (MLE)

Given data $x \sim X$, the MLE is computes the most likely parameter θ: $\hat{\theta} = \max_\theta f(\theta) = \max_\theta f_X(x|\theta)$.

Finding the MLE it is often easier to optimize the log-likelihood $g(\theta) = \log(f(\theta))$. The **1-dimensional derivative test** suggests that if 1) $g'(\theta) = 0$ and 2) $g''(\theta) > 0$, then θ is a local minima.

Convex $g''(x) \geqslant 0$ always, global minima exists

Concave $g''(x) \leqslant 0$ always, global maxima exists

Emperical Mean and Variances

Let $X \in [[a, b]$ be a random variable, $\bar{\mu}$ be the empirical mean, $\bar{\sigma}^2$ the empirical variance computed with μ, and $\hat{\sigma}^2$ with $\bar{\mu}$ respectively. One may bound $|\bar{\mu} - \mu|$ with the Hoeffding's bound on X_i, $|\bar{\sigma}^2 - \sigma|$ with the Hoeffding's bound on $Y_i = |X_i - \mu|^2$, and $|\hat{\sigma}^2 - \sigma^2|$ using the identity $(\hat{\sigma}^2 - \sigma^2) = (\bar{\sigma}^2 - \sigma^2) + (\bar{\mu} - \mu)^2$

Separating Gaussian Mixture

Given data $X_1 \sim \mathcal{N}(\mu_1, \mathbf{I}_d)$ and $X_2 \sim \mathcal{N}(\mu_2, \mathbf{I}_d)$ with unknown μ_1, μ_2, we wish to classify them correctly.

Proof outline:

- **Distance bound for same Gaussian:** If $\mathbf{x}_1, \mathbf{x}_2$ are from the same distribution, $\frac{\mathbf{x_1} - \mathbf{x_2}}{\sqrt{2}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, we use the Annulus Theorem to bound their pairwise distance:
$$\mathbb{P}(\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leqslant \sqrt{2}(\sqrt{d} + t)) \geqslant 1 - e^{-t^2/8}$$
- **Separation of different Gaussians:** If $\mathbf{x}_1 \sim \mathcal{N}(\mu_1, I)$ and $\mathbf{x}_2 \sim \mathcal{N}(\mu_2, I)$, define:
$$\mathbf{z} = \frac{\mathbf{x}_1 - \mathbf{x}_2 - (\mu_1 - \mu_2)}{\sqrt{2}} \sim \mathcal{N}(0, I).$$
Then, $\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 = \|\mathbf{z}\|_2^2 + \Delta^2 + 2(\mathbf{x}_1 - \mathbf{x}_2)^\top(\mu_1 - \mu_2)$. Using the Gaussian Annulus Theorem and concentration bounds on $(\mathbf{x}_1 - \mathbf{x}_2)^\top(\mu_1 - \mu_2)$, we conclude that:
$$\mathbb{P}(\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \geqslant \sqrt{2}(\sqrt{d} + t))$$
is high when $\|\mu_1 - \mu_2\|_2 = \Delta \gtrsim d^{1/4}\sqrt{\log(n/\delta)}$.
- **Union bound:** We apply Union bound on n points to show the property extends pairwise to any n points.

Algorithm: Pick a reference vector, assign all within $\sqrt{2}(\sqrt{d} + t)$ to one cluster, others to the second.

Total Variation Distance

Given observations from $\mathbf{p} = (p_1, \ldots, p_k)$, we measure the closesness of an estimate $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_k)$:

$$d_{\mathsf{TV}}(\hat{\mathbf{p}}, \mathbf{p}) = \max_{S \subseteq \{1,2,\ldots,k\}} \sum_{i \in S} (p_i - \hat{p}_i)$$

Break it down to positive and negative terms:

$$\sum_{i: p_i > \hat{p}_i} |p_i - \hat{p}_i| - \sum_{i: p_i < \hat{p}_i} |\hat{p}_i - p_i|$$

This value is maximal when $p_i > \hat{p}_i$. Since positive and negative sums are equal in magnitude,

$$\max_{S \subseteq \{1,2,\ldots,k\}} \sum_{i \in S} (p_i - \hat{p}_i) = \frac{1}{2} \sum_{i=1}^k |p_i - \hat{p}_i|$$

Estimator: $X_i : i = 1, 2, \ldots, n$, we assign a Bernoulli random variable for each class where $\hat{p}_i = \frac{\sum_{j=1}^n \mathbb{I}(X_j = i)}{n}$.

$$d_{\mathsf{TV}}(\hat{\mathbf{p}}, \mathbf{p}) = \max_{S \subseteq \{1,2,\ldots,k\}} \sum_{i \in S} (p_i - \hat{p}_i)$$
$$= \max_{S \subseteq \{1,2,\ldots,k\}} \sum_{i \in S} \frac{\mathbb{I}(X_j = i)}{n} - \sum_{i \in S} p_i$$

By linearity of expectation, $\mathbb{E}\left[\sum_{i \in S} \frac{\mathbb{I}(X_j = i)}{n}\right] = \sum_{i \in S} p_i$.

This means $\sum_{i \in S} \frac{\mathbb{I}(X_j = i)}{n}$ is a random variable with mean $\sum_{i \in S} p_i$ and range $[0, 1]$. Applying Hoeffding's bound,

$$\mathbb{P}\left(\left|\sum_{i \in S} \frac{\mathbb{I}(X_j = i)}{n} - \sum_{i \in S} p_i\right| \geqslant \epsilon\right) \leqslant 2e^{-2\epsilon^2 n}$$

$$\mathbb{P}(|d_{\mathsf{TV}}(\hat{\mathbf{p}}, \mathbf{p})| \leqslant \epsilon) \leqslant 2^k 2e^{-2\epsilon^2 n} \text{ (no. of } S\text{)}$$

2^k is the number of possible S. Hence, $2^k 2e^{-2\epsilon^2 n} \leqslant \delta$ yields the bound for $n \geqslant \frac{k\log(2) + \log(2\delta)}{2\epsilon^2}$,

Optimization

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ which we want to optimize wrt to its input $\mathbf{x}$: $\hat{\mathbf{x}} = \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$.

Derivative test

Gradient the first derivative $\nabla f(\mathbf{w})$ is defined by

$$\boxed{\begin{bmatrix} \frac{\partial f}{\partial w_1}(\mathbf{w}) & \frac{\partial f}{\partial w_2}(\mathbf{w}) & \cdots & \frac{\partial f}{\partial w_d}(\mathbf{w}) \end{bmatrix}^\top}$$

Hessian the second derivative $\nabla^2 f(\mathbf{w})$ is defined by

$$\boxed{\begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2}(\mathbf{w}) & \frac{\partial^2 f}{\partial w_1 \partial w_2}(\mathbf{w}) & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_d}(\mathbf{w}) \\ \frac{\partial^2 f}{\partial w_2 \partial w_1}(\mathbf{w}) & \frac{\partial^2 f}{\partial w_2^2}(\mathbf{w}) & \cdots & \frac{\partial^2 f}{\partial w_2 \partial w_d}(\mathbf{w}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial w_d \partial w_1}(\mathbf{w}) & \frac{\partial^2 f}{\partial w_d \partial w_2}(\mathbf{w}) & \cdots & \frac{\partial^2 f}{\partial w_d^2}(\mathbf{w}) \end{bmatrix}}$$

Local min: $\mathbf{x}^*$ s.t. for some neighborhood around $\mathbf{x}^*$, $f(\mathbf{x}) \geqslant f(\mathbf{x}^*)$ for all $\mathbf{x}$ from the neighborhood

Global min: $\mathbf{x}^*$ s.t. $\forall \mathbf{x} \in \mathbb{R}^d, f(\mathbf{x}) \geqslant f(\mathbf{x}^*)$

Stationary points: $\mathbf{x}$ s.t. $\nabla f(\mathbf{x}) = \mathbf{0}$. It can either be a local max, min, or saddle point.

Conditions for local min: On top of $\nabla f(\mathbf{x}) = \mathbf{0}$:

- (necessary) $\mathbf{v}^\top \nabla^2 f(\mathbf{x}) \mathbf{v} \geqslant 0$, $\forall \mathbf{v} \in \mathbb{R}^d$
- (sufficient) $\mathbf{v}^\top \nabla^2 f(\mathbf{x}) \mathbf{v} > 0$, $\forall \mathbf{v} \in \mathbb{R}^d - \{\mathbf{0}\}$

Alternatively, the entries of $\nabla^2 f(\mathbf{x})$ must be $\geqslant 0$

Convexity

f is **convex** if for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ and $\alpha \in (0, 1)$

$$\boxed{f(\alpha \mathbf{w}_1 + (1-\alpha)\mathbf{w}_2) \leqslant \alpha f(\mathbf{w}_1) + (1-\alpha) f(\mathbf{w}_2)}$$

If f is differentiable, then

- Once: $f(\mathbf{w}_1) \geqslant f(\mathbf{w}_2) + \nabla f(\mathbf{w}_2)^\top (\mathbf{w}_1 - \mathbf{w}_2)$
- Twice: $\nabla^2 f(\mathbf{x})$ is positive semi-definite everywhere.

Jensen's inequality if f is convex, for any distribution D

$$\boxed{f(\mathbb{E}_{\mathbf{w} \sim D}[\mathbf{w}]) \leqslant \mathbb{E}_{\mathbf{w} \sim D}[f(\mathbf{w})]}$$

Lipchitz smooth $\|\nabla f(\mathbf{w}_1) - \nabla f(\mathbf{w}_2)\|_2 \leqslant L \|\mathbf{w}_1 - \mathbf{w}_2\|_2$. All $L-$smooth function must satisfy $f(\mathbf{w}_1) \leqslant f(\mathbf{w}_2) + \nabla f(\mathbf{w}_2)^\top (\mathbf{w}_1 - \mathbf{w}_2) + \frac{L}{2} \|\mathbf{w}_1 - \mathbf{w}_2\|_2^2$