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 $\mathbb{E}[X], \mathsf{Var}(X)$ and $X \geqslant 0$:

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

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: $\mathbb{P}(|X| \geqslant t) \leqslant 2e^{-t^2/(2C^2)}$
- 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 empirical 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).

Estimation

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 $(\hat{\sigma}^2 - \sigma^2) = (\bar{\sigma}^2 - \sigma^2) + (\bar{\mu} - \mu)^2$

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 $f(\mathbf{w}_1) \leqslant f(\mathbf{w}_2) + \nabla f(\mathbf{w}_2)^\top(\mathbf{w}_1 - \mathbf{w}_2) + \frac{L}{2}\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2$

Gradient Descent

Assume f is once differentiable, $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta \nabla f(\mathbf{w}_t)$.

Gradient Descent Convergence

If f if $L-$smooth, gradient descent converges $\propto \frac{1}{t}$:

$$f(\mathbf{w}') \leqslant f(\mathbf{w}) + \nabla f(\mathbf{w})^\top(\mathbf{w}' - \mathbf{w}) + \frac{L}{2}\|\mathbf{w}' - \mathbf{w}\|_2^2$$

Substituting $\mathbf{w} = \mathbf{w}_t$ and $\mathbf{w}' = \mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)$,

$$f(\mathbf{w}_{t+1}) \leqslant f(\mathbf{w}_t) - \left(\eta_t - \frac{\eta_t^2 L}{2}\right)\|\nabla f(\mathbf{w}_t)\|_2^2$$

Choose $0 < \eta_t < \frac{2}{L}$, $f(\mathbf{w}_{t+1}) - f(\mathbf{w}_t) \leqslant \frac{-\eta}{2}\|\nabla f(\mathbf{w}_t)\|_2^2$

Telescopic sum and rearrangement:

$$\sum_{i=0}^t \|\nabla f(\mathbf{w}_i)\|_2^2 \leqslant \frac{2}{\eta}(f(\mathbf{w}_0) - f(\mathbf{w}_{t+1}))$$

$$\leqslant \frac{2}{\eta}(f(\mathbf{w}_0) - f_{\text{min}})$$

$$\min_i \|\nabla f(\mathbf{w}_i)\|_2^2 \leqslant \frac{2(f(\mathbf{w}_0) - f_{\text{min}})}{\eta(t+1)}$$

if we want gradient $\leqslant \epsilon$, set $t \geqslant \left\lceil \frac{\frac{2}{\eta}(f(\mathbf{w}_0) - f_\star)}{\epsilon^2} \right\rceil$

Reducing the Optimality Gap

If f is also convex, the error converges $\propto \frac{1}{t}$:

$$f(\mathbf{w}_*) \geqslant f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top(\mathbf{w}_* - \mathbf{w}_t).$$

Expanding the square:

$$\frac{1}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2 = \frac{1}{2}\|\mathbf{w}_t - \mathbf{w}_*\|_2^2$$

$$-\eta \nabla f(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}_*) + \frac{\eta^2}{2}\|\nabla f(\mathbf{w}_t)\|_2^2$$

Convexity implies

$$\eta \nabla f(\mathbf{w}_t)^\top(\mathbf{w}_* - \mathbf{w}_t) \leqslant \eta(f(\mathbf{w}_*) - f(\mathbf{w}_t))$$

and $L-$smoothness implies (proven prior)

$$\frac{\eta^2}{2}\|\nabla f(\mathbf{w}_t)\|_2^2 \leqslant \eta(f(\mathbf{w}_t) - f(\mathbf{w}_{t+1}))$$

put altogether

$$f(\mathbf{w}_{t+1}) - f(\mathbf{w}_*) \leqslant \frac{1}{2\eta}(\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2)$$

Summing telescopically and using how $f(\mathbf{w_t})$ decreases:

$$f(\mathbf{w}_{T+1}) - f(\mathbf{w}_*) \leqslant \frac{1}{2\eta(T+1)}\|\mathbf{w}_0 - \mathbf{w}_*\|_2^2$$

if we want LHS $\leqslant \epsilon$, set $t = \left\lceil \frac{1}{2\eta\epsilon}\|\mathbf{w}_0 - \mathbf{w}_*\|_2^2 \right\rceil$

Stochastic Gradient Descent

Let $f(\mathbf{w}) = \mathbb{E}_{x,y \sim \mathcal{D}}[\ell(\mathbf{w}; \mathbf{x}, \mathbf{y})]$ and the grad $\nabla_{\mathbf{w}} = \mathbb{E}_{x,y \sim \mathcal{D}}[\nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{x}, \mathbf{y})]$. Define $\mathbf{g_w} = \nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{x}, \mathbf{y})$. Then, $\nabla_{\mathbf{w}} = \mathbb{E}[\mathbf{g_w}]$. Finally, force $\mathbb{E}[\|\mathbf{g_w}\|_2^2] \leqslant G^2, \forall \mathbf{w}$:

$$\mathbb{E}_t[\frac{1}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2] = \mathbb{E}_t\left[\frac{1}{2}\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \eta \mathbf{g}_t^\top(\mathbf{w}_t - \mathbf{w}_*) + \frac{\eta^2}{2}\|\mathbf{g}_t\|_2^2\right]$$

$$= \frac{1}{2}\mathbb{E}_t[\|\mathbf{w}_t - \mathbf{w}_*\|_2^2] - \eta \mathbb{E}_t[\mathbf{g}_t^\top(\mathbf{w}_t - \mathbf{w}_*)] + \frac{\eta^2}{2}\mathbb{E}_t[\|\mathbf{g}_t\|_2^2]$$

$$\leq \frac{1}{2}\mathbb{E}_t[\|\mathbf{w}_t - \mathbf{w}_*\|_2^2] - \eta \nabla_{\mathbf{w}} f(\mathbf{w})(\mathbf{w}_t - \mathbf{w}_*) + \frac{\eta^2}{2}G^2$$

$$\eta \nabla_{\mathbf{w}} f(\mathbf{w})(\mathbf{w}_t - \mathbf{w}_*) \leq \frac{1}{2}\mathbb{E}_t[\|\mathbf{w}_t - \mathbf{w}_*\|_2^2] - \frac{1}{2}\mathbb{E}_t[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2] + \frac{\eta^2}{2}G^2$$

$$\eta(f(\mathbf{w}_t) - f(\mathbf{w}_*)) \leq (\text{convexity of } f)$$

$$\eta \mathbb{E}[(f(\mathbf{w}_t) - f(\mathbf{w}_*))] \leq \frac{1}{2}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|_2^2] - \frac{1}{2}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2] + \frac{\eta^2}{2}G^2$$

Summing telescopically and divide both sides by $\eta(T+1)$:

$$\frac{1}{T+1}\sum_{t=0}^T \mathbb{E}[f(\mathbf{w}_t) - f(\mathbf{w}_*)] \leqslant \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|_2^2}{2\eta(T+1)} + \frac{\eta G^2}{2}$$

Let $\overline{\mathbf{w}}_T = \frac{1}{T+1}\sum_{t=0}^T \mathbf{w}_t$ and $\eta = \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|_2}{G\sqrt{T+1}}$, we pick $T = \lceil \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|_2^2 + G^2}{2\epsilon^2} \rceil$ if we want error $\leqslant \epsilon$

PAC Learning

Define error: $\text{err}_{\mathcal{D}}(h) := \mathbb{P}_{\mathbf{x} \sim \mathcal{D}}(c(\mathbf{x}) \neq h(\mathbf{x}))$.

Learning Finite Classes

For finite $\mathcal{C}$, the probability error $\geqslant \epsilon$ yet h is chosen from n data points is $(1 - \epsilon)^n$. Suppose this for some h is $|\mathcal{C}|(1 - \epsilon)^n \leqslant \delta$ gives $n = \lceil \frac{\log(|\mathcal{C}|) + \log(1/\delta)}{\epsilon} \rceil$.

Bounding Empirical Error

Define $\hat{\text{err}}_{\mathcal{D}}(h) := \sum_{i=1}^n \mathbb{I}\{c(\mathbf{x}) \neq h(\mathbf{x})\}$. Hoeffding's bound give, for an h, $\mathbb{P}(|\hat{\text{err}}(h) - \text{err}(h)| \geqslant \epsilon) \leqslant 2e^{-2n\epsilon^2}$. Union bound for all classes gives $2|\mathcal{C}|e^{-2n\epsilon^2} \leqslant \delta$ which gives $n = \left\lceil \frac{1}{2\epsilon^2}\log\left(\frac{2|\mathcal{C}|}{\delta}\right) \right\rceil$.

Empirical Error Trick

Given $\mathbb{P}(\forall h, |\hat{\text{err}}(h) - \text{err}(h)| \leqslant \epsilon/2) \geqslant 1 - \delta$. Let $\hat{h}$ optimizes $\hat{\text{err}}$ and h_* optimizes $\text{err}(h)$. This means $\text{err}(\hat{h}) \leqslant \hat{\text{err}}(\hat{h}) + \epsilon/2 \leqslant \hat{\text{err}}(h_*) + \epsilon/2 \leqslant \text{err}(h_*) + \epsilon$

Perceptron Algorithm

Assuming $y_i = \pm 1$ and $\|\mathbf{x}_i\|_2^2, \|\mathbf{w}_*\|_2^2 \leqslant 1$,

$$\exists \mathbf{w}_*, y_i = \text{sign}(\mathbf{w}_*^\top \mathbf{x}) \text{ and } |\mathbf{w}_*^\top \mathbf{x}| \geqslant \gamma$$

The perceptron algorithm starts with $\mathbf{w}_0 = 0$ and updates $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i x_i$ for a misclassified i. First, $\mathbf{w}_*^\top \mathbf{w}_{t+1} = \mathbf{w}_*^\top(\mathbf{w}_t + y_i \mathbf{x}_i) = \mathbf{w}_*^\top \mathbf{w}_t + y_i \mathbf{w}_*^\top \mathbf{x}_i \geqslant \mathbf{w}_*^\top \mathbf{w}_t + \gamma$ so $\mathbf{w}_*^\top \mathbf{w}_t \geqslant t\gamma$. Moreover, $\|\mathbf{w}_{t+1}\|_2^2 = \|\mathbf{w}_t + y_i \mathbf{x}\|_2^2 = \|\mathbf{w}_t\|_2^2 + 2y_i \mathbf{w}_t^\top \mathbf{x}_i + y_i^2\|\mathbf{x}_i\|_2^2$. Since $y_i \mathbf{w}_t^\top \mathbf{x}_i \leqslant 0$, $\|\mathbf{w}_{t+1}\|_2^2 \leqslant \|\mathbf{w}_t\|_2^2 + 1$ so $\|\mathbf{w}_t\| \leqslant \sqrt{t}$. Put together, $t \leqslant \frac{1}{\gamma^2}$ is when the model converges.

Linear Regression

Suppose $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^{n \times 1}$, there exists $\mathbf{w}_* \in \mathbb{R}^{d \times 1}$ such that $\mathbf{y} = \mathbf{X}\mathbf{w}_* + \boldsymbol{\eta}$ for $\eta \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

MLE Solution (OLS)

The MLE gives the training objective to be $\ell(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$. The gradient is $2(\mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y})$ and hessian $2\mathbf{X}^\top \mathbf{X}$, giving $\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$.

Sample Complexity

$$\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2 = \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{X}\mathbf{w}_*\|_2^2$$

$$- 2\boldsymbol{\eta}^\top \mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}_*) + \|\boldsymbol{\eta}\|_2^2$$

$$\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2 \leqslant \|\boldsymbol{\eta}\|_2^2 \text{ because of MLE}$$

$$\|\mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}_*)\|_2^2 \leqslant 2\boldsymbol{\eta}^\top \mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}_*)$$

$$\|\mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}_*)\|_2 \leqslant \frac{2\boldsymbol{\eta}^\top \mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}_*)}{\|\mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}_*)\|_2}$$

Let $r = \text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}^\top \mathbf{X})$, then there exists orthonormal bases $\boldsymbol{\Phi} \in \mathbb{R}^{n \times r}$ such that $\boldsymbol{\Phi}^\top \boldsymbol{\Phi} = \mathbf{I}_r$ and $\mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}_*) = \boldsymbol{\Phi}\mathbf{z}$ for some $\mathbf{z} \in \mathbb{R}^r$. Cauchy-Schwartz:

$$\frac{\boldsymbol{\eta}^\top \mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}_*)}{\|\mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}_*)\|_2} = \frac{\boldsymbol{\eta}^\top \boldsymbol{\Phi}\mathbf{z}}{\|\boldsymbol{\Phi}\mathbf{z}\|_2} = \frac{\boldsymbol{\eta}_r^\top \mathbf{z}}{\|\mathbf{z}\|_2} \leqslant \|\boldsymbol{\eta}_r\|_2$$

Since $\boldsymbol{\eta}_r$ is normal, $\mathbb{E}[\|\boldsymbol{\eta}_r\|] = \sigma\sqrt{r}$, implying $\|\mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}_*)\|_2^2 \leqslant 4\|\boldsymbol{\eta}_r\|_2^2$ and $\mathbb{E}[\|\mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}_*)\|_2^2] \leqslant 4\sigma^2 r$. Define $V := \frac{1}{n}\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{X}\mathbf{w}_*\|_2^2 - 4\frac{\sigma^2 r}{n} \leqslant \frac{4}{n}\|\boldsymbol{\eta}_r\|_2^2 - 4\frac{\sigma^2 r}{n}$, then

$$\mathbb{E}\left[e^{\lambda V}\right] \leqslant \left(\mathbb{E}\left[e^{\frac{4\lambda}{n}\left(\eta_{r,1}^2/\sigma^2 - 1\right)}\right]\right)^r$$

$\eta_{r,1}^2/\sigma^2$ is standard normal. We use the MGF for $z^2 - 1$ if $z \in \mathcal{N}(0,1)$ as follows: $\mathbb{E}[e^{\lambda(z^2-1)}] \leqslant e^{2\lambda}$, set $\lambda = \frac{4\lambda\sigma^2}{n}$, $\mathbb{E}[e^{\lambda V}] \leqslant e^{\frac{32r\lambda^2\sigma^4}{n^2}}$.

Agnostic Linear Regression Model

Suppose $\mathbf{w}_*$ leads to the smallest true error $\mathsf{OPT} := \mathsf{err}(\mathbf{w}_*)$. We first bound $\hat{\mathsf{err}}(\mathbf{w}) \leqslant 4$ because of normalization constraint. For a given $\mathbf{w}$, Hoeffding's give

$$\mathbb{P}(|\hat{\mathsf{err}}(\mathbf{w}) - \mathsf{err}(\mathbf{w})| \geqslant t) \leqslant 2\exp\left(\frac{-2nt^2}{4}\right)$$

Quantization define a grid $\mathcal{H}_{\epsilon'}$ of size ϵ', $|\mathcal{H}_{\epsilon'}| = (\frac{2}{\epsilon'})^d$.

$$\mathbb{P}(\exists \mathbf{w} \in \mathcal{H}_{\epsilon'} : |\hat{\mathsf{err}}(\mathbf{w}) - \mathsf{err}(\mathbf{w})| \geqslant t) \leqslant 2(\frac{2}{\epsilon'})^d e^{(\frac{-2nt^2}{4})}$$

Set $t = \epsilon$ and RHS $\leqslant \delta$ gives $n = O(\frac{d\ln(1/\epsilon') + \ln(1/\delta)}{\epsilon^2})$. So $\mathsf{err}(\hat{\mathbf{v}}) \leqslant \mathsf{err}(\mathbf{v}_*) + \epsilon/2$ when $\mathbf{v}$ nearest to $\mathbf{w}$ in $\mathcal{H}_{\epsilon'}$.

$$\begin{aligned}
\mathsf{err}(\mathbf{v}) &= \mathbb{E}_{(\mathbf{x},y)}\left[(y - \mathbf{v}^\top\mathbf{x})^2\right] \\
&= \mathbb{E}_{(\mathbf{x},y)}\left[(y - \mathbf{w}^\top\mathbf{x} + \mathbf{w}^\top\mathbf{x} - \mathbf{v}^\top\mathbf{x})^2\right] \\
&= \mathbb{E}_{(\mathbf{x},y)}\left[(y - \mathbf{w}^\top\mathbf{x})^2\right] + \mathbb{E}_{(\mathbf{x},y)}\left[(\mathbf{w}^\top\mathbf{x} - \mathbf{v}^\top\mathbf{x})^2\right] \\
&\quad + 2\mathbb{E}_{(\mathbf{x},y)}\left[(y - \mathbf{w}^\top\mathbf{x})(\mathbf{w}^\top\mathbf{x} - \mathbf{v}^\top\mathbf{x})\right] \\
&\leqslant \mathsf{err}(\mathbf{w}) + \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\mathbf{w}^\top\mathbf{x} - \mathbf{v}^\top\mathbf{x})^2\right] \\
&\quad + 2\sqrt{\mathsf{err}(\mathbf{w}) \cdot \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\mathbf{w}^\top\mathbf{x} - \mathbf{v}^\top\mathbf{x})^2\right]} \\
&= \mathsf{err}(\mathbf{w}) + \mathsf{err}_q + 2\sqrt{\mathsf{err}(\mathbf{w}) \cdot \mathsf{err}_q} \\
&\leqslant \mathsf{err}(\mathbf{w}) + \mathsf{err}_q + 4\sqrt{\mathsf{err}_q}
\end{aligned}$$

Therefore, we want $\mathsf{err}_q + 4\sqrt{\mathsf{err}_q} \leqslant \epsilon/4$:

$$\begin{aligned}
\mathsf{err}_q &= \mathbb{E}_{(\mathbf{x},y)}\left[(\mathbf{w}^\top\mathbf{x} - \mathbf{v}^\top\mathbf{x})^2\right] \\
&= \mathbb{E}_{(\mathbf{x},y)}\left[\|\mathbf{w} - \mathbf{v}\|_2^2\|\mathbf{x}\|_2^2\right] \\
&\leqslant \|\mathbf{w} - \mathbf{v}\|_2^2 \cdot \mathbb{E}_{(\mathbf{x},y)}\left[\|\mathbf{x}\|_2^2\right] \\
&\leqslant \left(\frac{\epsilon}{2}\right)^2 \cdot \|\mathbf{1}_d\|_2^2 = \frac{(\epsilon)^2 d}{4}
\end{aligned}$$

As such, $\epsilon' \in O(d/\sqrt{\epsilon})$, $n = O(\frac{d\ln(d/\epsilon) + \ln(1/\delta)}{\epsilon^2})$

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 $\hat{p}_i = \frac{\sum_{j=1}^n \mathbb{I}(X_j = i)}{n}$.

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

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,

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

$$\mathbb{P}(|d_{\mathsf{TV}}(\hat{\mathbf{p}}, \mathbf{p})| \geqslant \epsilon) \leqslant 2^k 2e^{-2\epsilon^2 n}$$

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

Hypothesis Testing

H_0: null hypothesis (nothing happens) H_1: alternative. Define $\alpha = \mathbb{P}(\text{reject } H_0 | H_0)$ (significance) and $1 - \beta = \mathbb{P}(\text{reject } H_1 | H_0)$ (power). $p-$value: $p = \mathbb{P}(\text{event} | H_0 \text{ is accepted})$. $p \leqslant \alpha$ means we reject H_0.

Unknown Null Hypothesis

Given two unknown distr p_* (reference) and q_* (test). $q_1, q_2, \ldots, q_n \sim q_*$ Let $s = \frac{1}{n}\sum_{i=1}^n d_{\mathsf{TV}}(p_*, q_i)$ and $\bar{s} = \mathbb{E}_{q_*}[s]$. Hoeffding gives

$$p = \mathbb{P}(|s - \bar{s}| \geqslant |s_{\mathsf{obs}} - \bar{s}|) \leqslant 2e^{-2n|s_{\mathsf{obs}} - \bar{s}|^2}$$

if accept $H_0 := \{p_* = q_*\}$, sample $p_1, p_2, \ldots, p_N \sim p_*$, define $\hat{s} = \frac{1}{n}\sum_{i=1}^n d_{\mathsf{TV}}(p_*, p_i)$: $\mathbb{P}(|\hat{s} - \bar{s}| \geqslant \epsilon) \leqslant 2e^{-2N\epsilon^2}$

$$\begin{aligned}
p &\leqslant \mathbb{P}(|s - \bar{s}| \geqslant |s_{\mathsf{obs}} - \hat{s}| - |\hat{s} - \bar{s}|) \\
&\leqslant \mathbb{P}(|s - \bar{s}| \geqslant |s_{\mathsf{obs}} - \hat{s}| - |\hat{s} - \bar{s}| \,|\, |\hat{s} - \bar{s}| \leqslant \epsilon) \\
&\quad + \mathbb{P}(|\hat{s} - \bar{s}| \geqslant \epsilon) \text{ (total probability)} \\
&\leqslant \mathbb{P}(|s - \bar{s}| \geqslant |s_{\mathsf{obs}} - \hat{s}| - \epsilon) + \mathbb{P}(|\hat{s} - \bar{s}| \geqslant \epsilon) \\
&\leqslant 2e^{2n(|s_{\mathsf{obs}} - \bar{s}| - \epsilon)^2} + 2e^{2N\epsilon^2}
\end{aligned}$$

choose $\epsilon = \frac{|s_{\mathsf{obs}} - \bar{s}|}{2}$ and $n = N$: $p \leqslant 4e^{\frac{n|s_{\mathsf{obs}} - \bar{s}|^2}{2}}$

Multiple Null Hypotheses

Given $H_0^{(i)}, i = 1, 2, \ldots, K$, suppose each has significance α. Define FWER $:= \mathbb{P}(\geqslant 1 \text{ false rejections}) \leqslant \sum_{i=1}^K \mathbb{P}(\text{reject } H_0^{(i)} | H_0^{(i)}) = \alpha K$, we set $\alpha \to \alpha/K$

Improving Success Probability

Testing algorithm $\mathcal{A}$ passes with prob $1 - q$, $q \in [0, 0.5]$, using n samples (depends on k and ϵ). Suppose $\mathcal{A}_i \sim \mathsf{Ber}(q)$ and $\mathcal{A}' = \mathbb{I}\{\sum_{i=1}^m \mathcal{A}_i \geqslant \frac{m}{2}\}$.

$$\begin{aligned}
\mathbb{P}(\mathcal{A}' = 0) &= \mathbb{P}(\sum_{i=1}^m \mathcal{A}_i \geqslant \frac{m}{2}) \\
&= \mathbb{P}(\frac{1}{m}\sum_{i=1}^m \mathcal{A}_i - q \geqslant \frac{1}{2} - q) \\
&\leqslant e^{-2m(\frac{1}{2} - q)^2} = e^{-\frac{m}{2}(1 - 2q)^2}
\end{aligned}$$

Estimation Uniformity Testing

Let $\mathbf{u}_k = (1/k, 1/k. \ldots, 1/k)$. Let our estimator be $\hat{\mathbf{p}}$ such that $d_{\mathsf{TV}}(\hat{\mathbf{p}}, \mathbf{p}) \leqslant \epsilon/3$,

- If $\mathbf{p} = \mathbf{u}_k$, $d_{\mathsf{TV}}(\hat{\mathbf{p}}, \mathbf{u}_k) \leqslant \epsilon/3$
- If $d_{\mathsf{TV}}(\mathbf{p}, \mathbf{u}_k) \geqslant \epsilon$, d_{TV} is a metric so $d_{\mathsf{TV}}(\hat{\mathbf{p}}, \mathbf{u}_k) \geqslant d_{\mathsf{TV}}(\mathbf{p}, \mathbf{u}_k) - d_{\mathsf{TV}}(\mathbf{p}, \hat{\mathbf{p}}) \geqslant 2\epsilon/3$. $\epsilon/2$ could be used.

Sample complexity: $n \geqslant \frac{k\log(2) + \log(2\delta)}{2\epsilon^2}$

Collision Uniformity Testing

Let $C = \sum_{i<j} \mathbb{I}\{X_i = X_j\}$ be the number of pairwise collisions among n samples from $\mathbf{p}$. $\mathbb{E}[C] = \binom{n}{2}\|\mathbf{p}\|_2^2$.

- If $\mathbf{p} = \mathbf{u}_k$, $\mathbb{E}[C] = \binom{n}{2} \cdot \frac{1}{k}$
- $d_{\mathsf{TV}}(\mathbf{p}, \mathbf{u}_k) \geqslant \epsilon$, $\|\mathbf{p}\|_2^2 > \frac{1+\epsilon^2}{k} \Rightarrow \mathbb{E}[C] > \binom{n}{2} \cdot \frac{1+\epsilon^2}{k}$
- Set $T = \binom{n}{2} \cdot \frac{1+\epsilon^2/2}{k}$, accept if $C < T$, reject if $C \geqslant T$

Claim: $\mathsf{Var}(C) \leqslant n^2\|\mathbf{p}\|_2^2 + n^3\|\mathbf{p}\|_3^3$, $\|\mathbf{p}\|_3^3 = \sum_{i=1}^k p_i^3$ Observe that because $n \geqslant \sqrt{k}$, $\|\mathbf{p}\|_2 \geqslant \frac{1}{\sqrt{k}}$, and $\|\mathbf{p}\|_3 \leqslant \|\mathbf{p}\|_2$, we can further bound the variance by $\mathsf{Var}[C] \leqslant 2n^3\|\mathbf{p}\|_2^3$.

Using $\mathbb{P}(|C - \mathbb{E}[C]| \geqslant 2\sigma) \leqslant \frac{1}{4}$ and $\sigma = \sqrt{\mathsf{Var}[C]} \leqslant \sqrt{2n^3}\|\mathbf{p}\|_2^{3/2}$ we prove

- **Case 1 (Uniform):** $\mathbf{p} = \mathbf{u}_k$

$$\mathbb{E}[C] = \binom{n}{2} \cdot \frac{1}{k},$$

$$\sigma \leqslant \sqrt{2n^3} \cdot \frac{1}{k^{3/2}}$$

We want $C < T = \binom{n}{2} \cdot \frac{1+\epsilon^2/2}{k}$, so:

$$\mathbb{E}[C] + 2\sigma < T \quad \Rightarrow \quad n = O\left(\frac{\sqrt{k}}{\epsilon^2}\right)$$

- **Case 2 (Far):** $d_{\mathsf{TV}}(\mathbf{p}, \mathbf{u}_k) \geqslant \epsilon/2$

$$\|\mathbf{p}\|_2^2 \geqslant \frac{1 + \epsilon^2 + \alpha}{k},$$

$$\mathbb{E}[C] \geqslant \binom{n}{2} \cdot \frac{1 + \epsilon^2 + \alpha}{k}$$

We want $C \geqslant T$, so:

$$\mathbb{E}[C] - 2\sigma \geqslant T \quad \Rightarrow \quad n = O\left(\frac{\sqrt{k}}{\epsilon^2}\right)$$

Conclusion: $n = \Theta\left(\frac{\sqrt{k}}{\epsilon^2}\right)$ suffice to separate error $\leqslant 1/4$.

ℓ_2 Identity Testing

Goal: Given i.i.d. samples from unknown $\mathbf{q}$ and explicit $\mathbf{p}$, test:

- H_0: $\mathbf{p} = \mathbf{q}$ vs. H_1: $\|\mathbf{p} - \mathbf{q}\|_1 \geqslant \epsilon$

Key idea: Use a test statistic that estimates $\|\mathbf{p} - \mathbf{q}\|_2^2$.

Tester: Let $X_i \sim \mathsf{Pois}(nq_i)$, define

$$Z = \sum_{i=1}^k (X_i - np_i)^2 - X_i$$

Then:

$$\mathbb{E}[Z] = n^2\|\mathbf{p} - \mathbf{q}\|_2^2 \quad \text{(bias-corrected estimator)}$$

Theorem: With $n = \Theta\left(\frac{\sqrt{k}}{\epsilon^2}\right)$ samples, this tester distinguishes H_0 from H_1 with constant probability.

Why it works:

- Under H_0, $\mathbb{E}[Z] = 0$
- Under H_1, $\|\mathbf{p} - \mathbf{q}\|_1 \geqslant \epsilon \Rightarrow \|\cdot\|_2^2 \geqslant \epsilon^2/k$
- So $\mathbb{E}[Z] \geqslant n^2 \cdot \epsilon^2/k$
- With variance $\mathsf{Var}(Z) = O(n^2)$, Chebyshev implies signal is detectable with $n = \Theta(\sqrt{k}/\epsilon^2)$