Problem 1)

a)

$\mathrm{E}[x_\alpha - z_\alpha] = \mathrm{E}[x_\alpha] - \mathrm{E}[z_\alpha] = 0$

$\mathrm{VAR}[x_\alpha - z_\alpha] = \mathrm{VAR}[x_\alpha] + \mathrm{VAR}[z_\alpha] = 2$

This is the Gaussian distribution

b)

$$\mu_D = E\left[\sum_{\alpha=1}^{d}(x_\alpha - z_\alpha)^2\right] = \sum_{\alpha=1}^{d} E[(x_\alpha - z_\alpha)^2]$$

$E[(x_\alpha - z_\alpha)^2] = \mathrm{VAR}[x_\alpha - z_\alpha] + \mathrm{E}[x_\alpha - z_\alpha]^2 = 2$

$\therefore \mu_D = 2d$

c)

$$\sigma_D^2 = VAR\left[\sum_{\alpha=1}^{d}(x_\alpha - z_\alpha)^2\right] = \sum_{\alpha=1}^{d} VAR[(x_\alpha - z_\alpha)^2]$$

$VAR[(x_\alpha - z_\alpha)^2] = E[(x_\alpha - z_\alpha)^4] - (E[(x_\alpha - z_\alpha)^2])^2$

$E[(x_\alpha - z_\alpha)^4] = 12$

$E[(x_\alpha - z_\alpha)^2] = 2$

$\therefore\ VAR[(x_\alpha - z_\alpha)^2] = 8$

$\sigma_D^2 = 8d$

d)

$$\frac{4\sigma_D}{\mu_D} = \frac{4\sqrt{2}}{\sqrt{d}} \to 0$$

Problem 2

(a) The data set is:

$(1, 1)$ ——— ①

$(3, -1)$ ——— ②

$(4, 1)$ ——— ③

$(5, -1)$ ——— ④

The labels are

$[1, -1, 1, -1]$

When we execute perceptron algorithm on data set in order $(1, 4, 3, 2)$; then it converges FAST in 1 iteration.

When we execute in order $(1, 3, 4, 2)$, it converges SLOWER in 3 iterations.

(b) Given vectors $x_1, x_2, x_3, \ldots x_n$

(i) When perceptron converges after a single pass, regardless of order, the fastest convergence is achieved after presenting '1' vector, and slowest and presenting 'n' vectors.
Thus, max difference is $(n-1)$

(ii) If the perceptron might take more than one pass, depending on the order, then the fastest convergence is achieved after presenting '1' vector, and slowest after presenting the max times on every vector till convergence.
We know $k_i \leq \frac{w^T x}{x^T x}$ for vector x_i.
$\therefore$ Max slowest convergence is at $\sum_{i=1}^{n} k_i$ for all vectors x_i.

(c) Yes. If ~~Given~~ give a data set, we find a hyperplane that has the biggest γ_{margin}. Then this hyperplane should be better than others. Because $k \le \frac{1}{\gamma_{margin}^2}$ if we get the maximum γ_{margin}. Then We get the minimum k.

(d) From the previous questions, we understand that the ordering of the vector set affectr the speed of convergence of the perceptron.
We can thus utilize this by randomizing the vector set in every iteration of the perceptron training, as it effectively is averaging multiple hyperplanes to find the one that is converging faster.
From the previous question, we understand that this averaged hyperplane would have a big γ_{margin} and since, $k \le \frac{1}{\gamma^2_{margin}}$, the perceptron converges fast.

Problem 3)

a)

$$\mathrm{P(D|\lambda)} = \prod_{i=1}^{n} \lambda e^{-\lambda X_i}$$

$$\mathrm{logP(D|\lambda)} = \sum_{i=1}^{n} (\log \lambda - \lambda X_i) = nlog\mathrm{\lambda} - \mathrm{\lambda} \sum_{i=1}^{n} X_i$$

$$\operatorname*{argmax}_{\lambda} \mathrm{logP(D|\lambda)}$$

$$\frac{\partial \log P(D|\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0$$

$$\mathrm{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i}$$

b)

$$\mathrm{P(D|p)} = \prod_{i=1}^{n} p(1-p)^{X_i}$$

$$\mathrm{logP(D|p)} = \sum_{i=1}^{n} \log p + X_i \log(1-p) = nlog\mathrm{p} + \log(1-\mathrm{p}) \sum_{i=1}^{n} X_i$$

$$\frac{\partial \log P(D|p)}{\partial p} = \frac{n}{p} - \frac{1}{1-p} \sum_{i=1}^{n} X_i = 0$$

$$\frac{n}{p} = \frac{1}{1-p} \sum_{i=1}^{n} X_i$$

$$\frac{1-p}{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\mathrm{p} = \frac{n}{\sum_{i=1}^{n} X_i + n}$$

a is the inverse of mean, b is the inverse of mean+1

This is because the exponential distribution is the convergence of geometric distribution

c)

$$\mathrm{P}(\mathrm{D}|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

$$\log\mathrm{P}(\mathrm{D}|\mu, \sigma^2) = \sum_{i=1}^{n} \left(-\frac{(X_i-\mu)^2}{2\sigma^2} - \log\left(\sigma\sqrt{2\pi}\right)\right)$$

$$= -\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2} - \mathrm{n}\log(\sigma\sqrt{2\pi})$$

$$\frac{\partial logP}{\partial \mu} = \frac{2\sum_{i=1}^{n}(X_i-\mu)}{2\sigma^2} = 0$$

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\frac{\partial logP}{\partial \sigma} = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^3} - \frac{n}{\sigma} = \frac{\sum_{i=1}^{n}(X_i-\mu)^2 - n\sigma^2}{\sigma^3} = 0$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}$$