2.1 If $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$, then

$$\begin{aligned}\|t_k - \hat{y}\|^2 &= (1 - \hat{y}_k)^2 + \sum_{i \neq k} \hat{y}_i^2 \\ &= 1 - 2\hat{y}_k + \|\hat{y}\|^2\end{aligned}$$

So the k that minimizes this will maximize $\hat{y}_k$.

2.2 The Bayes classifier picks the k which maximizes $\Pr(G_k|X = x)$. In this case, the boundary consists of points for which $\Pr(\text{orange}|X = x) = \Pr(\text{blue}|X = x) = \frac{1}{2}$. Letting $f_o(x), f_b(x)$ be the density functions for the blue and orange groups, respectively, we find

$$\Pr(\text{orange}|X = x) = \frac{f_o(x)}{f_b(x) + f_o(x)},$$

and so the decision boundary consists of points satisfying $f_o(x) = f_b(x)$.

2.3 Let $f(x)$ be the density function of the distance of a single point selected uniformly from the unit p-ball and $F(x)$ the cdf. If $x_1, \dots, x_N$ are iid points, let $x_{(c)}$ denote the closest point to the origin. Then

$$\begin{aligned}\Pr(d(0, x_{(c)}) \leq r) &= 1 - \Pr(d(0, x_i) \geq r \ \forall i) \\ &= 1 - \prod_i \Pr(d(0, x_i) \geq r) \\ &= 1 - \prod_i (1 - F(r)) \\ &= 1 - (1 - F(r))^N\end{aligned}$$

The median of the closest distance is the r for which this is one half, so solving gives

$$F(r_{\text{med}}) = 1 - \left(\frac{1}{2}\right)^{\frac{1}{N}}.$$

The density function expressed for radial coordinates has the form $\frac{1}{\text{Vol}(D^p)} r^{p-1} g(\theta_1, \dots, \theta_{p-1})$, where the angles parametrize the sphere S^{p-1}. So we find that

$$\begin{aligned}F(r) &= \frac{1}{\text{Vol}(D^p)} \int_0^r \int_{S^{p-1}} \rho^{p-1} g(\overline{\theta}) d\rho d\overline{\theta} \\ &= \frac{1}{\text{Vol}(D^p)} \left(\int_0^r \rho^{p-1} d\rho\right) \left(\int_{S^{p-1}} Cg(\overline{\theta}) d\overline{\theta}\right) \\ &= \frac{\text{Vol}(S^{p-1})}{p\text{Vol}(D^p)} r^p\end{aligned}$$

But the form of the volume form shows

$$\mathrm{Vol}(D^p) = \mathrm{Vol}(S^{p-1}) \int_0^1 \rho^{p-1} d\rho = \frac{1}{p}\mathrm{Vol}(S^{p-1}),$$

so

$$F(r) = r^p$$

and hence

$$r_{\mathrm{med}} = \left(1 - \left(\frac{1}{2}\right)^{\frac{1}{N}}\right)^{\frac{1}{p}}.$$

2.4 The density function has the form $Ce^{-\frac{1}{2}\|x\|^2}$, which is clearly invariant under orthonormal basis changes. Hence the projections of points sampled from this distribution onto any unit vector must be $\mathcal{N}(0,1)$. The expected square distance of these z_i is just the variance of $\mathcal{N}(0,1)$, which is 1. The expected square distance of x_0 from the origin is p, as mentioned in the problem.

2.5.a The first equality is the definition. For the second, we expand

$$(y_0 - \hat{y}_0)^2 = y_0^2 - 2y_0\hat{y}_0 + \hat{y}_0^2.$$

Taking $E_{\mathcal{T}}$ and noting that y_0 has no dependence on $\mathcal{T}$, we arrive at

$$\begin{aligned} & y_0^2 - 2y_0 E_{\mathcal{T}}[\hat{y}_0] + E_{\mathcal{T}}[\hat{y}_0^2] \\ = & y_0^2 - 2y_0 E_{\mathcal{T}}[\hat{y}_0] + E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}[\hat{y}_0]]^2 + (E_{\mathcal{T}}[\hat{y}_0])^2 \end{aligned}$$

Now take $E_{y_0|x_0}$ to obtain

$$\begin{aligned} & \mathrm{Var}(y_0|x_0) + (x_0^T\beta)^2 - 2x_0^T\beta E_{\mathcal{T}}[\hat{y}_0] + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + (E_{\mathcal{T}}[\hat{y}_0])^2 \\ = & \mathrm{Var}(y_0|x_0) + E_{\mathcal{T}}[x_0^T\beta - \hat{y}_0]^2 + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0), \end{aligned}$$

which is the second and third lines.

To obtain the fourth line, work term-by-term: For the first, use the fact that ε is $N(0,\sigma^2)$ to write $\mathrm{Var}(y_0|x_0) = \mathrm{Var}(x_0 + \varepsilon|x_0) = \sigma^2$. The fact that least squares in unbiased means the second term is 0. This leaves the variance

$\mathrm{Var}_{\mathcal{T}}(\hat{y}_0)$. Since there is no bias, the definition gives that

$$\begin{aligned}
\mathrm{Var}_{\mathcal{T}}(\hat{y}_0) &= \mathrm{Var}_{\mathcal{T}}(\sum_i l_i(x_0)\varepsilon_i) \\
&= \sum_{i,j} \mathrm{Cov}_{\mathcal{T}}(l_i(x_0)\varepsilon_i, l_j(x_0)\varepsilon_j) \\
&= \sum_{i,j} E_{\mathcal{T}}[\varepsilon_i\varepsilon_j] E_{\mathcal{T}}[l_i(x_0)l_j(x_0)] \\
&= \sigma^2 \sum_i E_{\mathcal{T}}[l_i(x_0)^2] \\
&= \sigma^2 E_{\mathcal{T}}\big[\sum_i l_i(x_0)^2\big]
\end{aligned}$$

Now, since $l_i(x_0)$ is the ith row of $X(X^TX)^{-1}x_0$, we find that

$$\sum_i l_i(x_0)^2 = x_0^T(X^TX)^{-T}X^TX(X^TX)^{-1}x_0 = x_0^T(XX^T)^{-1}x_0$$

which gives the final line.

2.5.b Two facts: If A is any $n \times n$ matrix and b is a column vector, we find that

$$b^TAb = \sum_{i,j} A_{ij}b_jb_j.$$

Also that if A, B are symmetric, then

$$\sum_{ij} A_{ij}B_{ij} = \sum_i \sum_j A_{ij}B_{ji} = \sum_i (AB)_{ii} = \mathrm{trace}(AB).$$

So (assuming $E[X] = 0$)

$$\begin{aligned}
E_{x_0}[x_0^T \mathrm{Cov}(X)^{-1}x_0] &= \sum_{i,j} \mathrm{Cov}(X)^{-1}_{ij} E_{x_0}[(x_0)_i(x_0)_j] \\
&= \sum_{i,j} \mathrm{Cov}(X)^{-1}_{ij} \mathrm{Cov}(x_0)_{ij} \\
&= \mathrm{trace}(\mathrm{Cov}(X)^{-1}\mathrm{Cov}(x_0)) \\
&= \mathrm{trace}(I) \\
&= p.
\end{aligned}$$

Plugging this into the equation gives the desired expression.

2.6 Suppose $a_1, \ldots, a_k$ is some sample of values. Then if μ is the average of these and $\mathrm{Var}(a)$ is the sample variance

$$\sum_i (a_i - c)^2 = k\,\mathrm{Var}(a) + k(\mu - c)^2$$

Let $\pi_X \mathcal{T}$ denote the set of x values in the training sample. Let $\mathcal{T}_x$ denote the set of y values of training points with x-coordinate x and μ_x be the average of those points. (View $\mathcal{T}_x$ as a multiset if necessary.) Then

$$\begin{aligned}
RSS(\theta) &= \sum_i (y_i - f_\theta(x_i))^2 \\
&= \sum_{x \in \pi_X \mathcal{T}} \sum_{y \in \mathcal{T}_x} (y - f_\theta(x))^2 \\
&= \sum_{x \in \pi_X \mathcal{T}} |\mathcal{T}_x|\,\mathrm{Var}(\mathcal{T}_x) + |\mathcal{T}_x|(\mu_x - f_\theta(x))^2 \\
&= \left(\sum_{x \in \pi_X \mathcal{T}} |\mathcal{T}_x|\,\mathrm{Var}(\mathcal{T}_x) \right) + \mathrm{WRSS}(\theta)
\end{aligned}$$

where WRSS is an RSS expression with one sample point for each x value, μ_x, and each x is weighted by $|\mathcal{T}_x|$. Since the first term in the sum has no dependence on θ, minimizing WRSS will minimize RSS.

2.7.a For k-nearest neighbor regression, we take $l_i(x_0; \mathcal{X})$ to be 1 if y_i is one of the k-nearest neighbors of x_0 and 0 otherwise. For linear regression, the prediction is

$$x_0^T \hat{\beta} = x_0^T (X^T X)^{-1} X^T y$$

and so we take $l_i(x_0; \mathcal{X})$ to be the ith component of $x_0^T (X^T X)^{-1} X^T$.

2.7.b Let $E_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) = \mu_{\mathcal{Y}|\mathcal{X}}(x_0)$.

$$\begin{aligned}
&(f(x_0) - \hat{f}(x_0))^2 \\
=&(y_0 + \varepsilon_0 - \hat{f}(x_0))^2 \\
=&(y_0 - \mu_{\mathcal{Y}|\mathcal{X}}(x_0) + \mu_{\mathcal{Y}|\mathcal{X}}(x_0) - \hat{f}(x_0) + \varepsilon_0)^2 \\
=&(y_0 - \mu_{\mathcal{Y}|\mathcal{X}}(x_0))^2 + (\mu_{\mathcal{Y}|\mathcal{X}}(x_0) - \hat{f}(x_0))^2 + \varepsilon_0^2 + \text{ cross terms}
\end{aligned}$$

Now apply $E_{\mathcal{Y}|\mathcal{X}}$ to this expression. The cross terms will vanish by definition of $\mu_{\mathcal{Y}|\mathcal{X}}(x_0)$ and the fact that ε_0 is independent from (x_0, y_0) with mean 0. The other terms become

$$\mathrm{Bias}^2_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) + \mathrm{Var}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) + \sigma^2$$

2.7.c The argument above holds again, this time using $\mu(x_0) = E_{\mathcal{Y},\mathcal{X}}(x_0)$.

2.7.d The conditional bias term is

$$y_0 - E_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) = y_0 - \sum_i l_i(x_0;\mathcal{X}) E_{\mathcal{Y}|\mathcal{X}}(y_i) = y_0 - \sum_i l_i(x_0;\mathcal{X}) f(x_i),$$

while the unconditional bias term is

$$\begin{aligned}
&y_0 - \sum_i E_{\mathcal{Y},\mathcal{X}}(l_i(x_0;\mathcal{X})(f(x_i)+\varepsilon_i)\\
&= y_0 - \sum_i E_{\mathcal{Y},\mathcal{X}}(l_i(x_0;\mathcal{X}) f(x_i))\\
&= y_0 - \sum_i E_{\mathcal{X}}(l_i(x_0;\mathcal{X}) f(x_i))
\end{aligned}$$

so

$$\mathrm{Bias}_{\mathcal{X},\mathcal{Y}}(\hat{f}(x_0)) = E_{\mathcal{X}}(\mathrm{Bias}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))).$$

For the variances, we calculate the conditional variance as

$$\begin{aligned}
&\mathrm{Var}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))\\
&= E_{\mathcal{Y}|\mathcal{X}}(\mu_{\mathcal{Y}|\mathcal{X}}(x_0) - \hat{f}(x_0))^2\\
&= E_{\mathcal{Y}|\mathcal{X}}\left(\sum_i l_i(x_0;\mathcal{X}) f(x_i) - \sum_i l_i(x_0;\mathcal{X})(f(x_i)+\varepsilon_i)\right)^2\\
&= E_{\mathcal{Y}|\mathcal{X}}\left(\sum_i l_i(x_0;\mathcal{X})\varepsilon_i\right)^2\\
&= \sigma^2 \sum_i l_i(x_0;\mathcal{X})^2.
\end{aligned}$$

For the unconditional variance, start with

$$\begin{aligned}
&\left(E_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0)) - \hat{f}(x_0)\right)^2\\
&= \left(\sum_i E_{\mathcal{X}}(l_i(x_0;\mathcal{X}) f(x_i)) - \sum_i l_i(x_0;\mathcal{X})(f(x_i)+\varepsilon_i)\right)^2\\
&= \left(\sum_i E_{\mathcal{X}}(l_i(x_0;\mathcal{X}) f(x_i)) - l_i(x_0;\mathcal{X}) f(x_i)\right)^2 + \left(\sum_i l_i(x_0;\mathcal{X})\varepsilon_i\right)^2\\
&\quad + 2\left(\sum_i E_{\mathcal{X}}(l_i(x_0;\mathcal{X}) f(x_i)) - l_i(x_0;\mathcal{X}) f(x_i)\right)\left(\sum_i l_i(x_0;\mathcal{X})\varepsilon_i\right)
\end{aligned}$$

To compute the variance of $\hat{f}(x_0)$, we take $E_{\mathcal{Y},\mathcal{X}}$ of this expression. We calculate each term separately:

$$E_{\mathcal{Y},\mathcal{X}}\left(\sum_i E_{\mathcal{X}}(l_i(x_0;\mathcal{X})f(x_i)) - l_i(x_0;\mathcal{X})f(x_i)\right)^2 = \mathrm{Var}_{\mathcal{X}}\left(\sum_i l_i(x_0;\mathcal{X})f(x_i)\right)$$

$$E_{\mathcal{Y},\mathcal{X}}\left(\sum_i l_i(x_0;\mathcal{X})\varepsilon_i\right)^2 = E_{\mathcal{X}}\,\mathrm{Var}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))$$

We split the cross term up into two pieces:

$$\begin{aligned}
&E_{\mathcal{Y},\mathcal{X}}\left(\sum_{ij} E_{\mathcal{X}}[l_i(x_0;\mathcal{X})f(x_i)]\ l_j(x_0;\mathcal{X})\varepsilon_j\right)\\
&= \sum_{ij} E_{\mathcal{X}}[l_i(x_0;\mathcal{X})f(x_i)]E_{\mathcal{Y},\mathcal{X}}[l_j(x_0;\mathcal{X})\varepsilon_j]\\
&= \sum_{ij} E_{\mathcal{X}}[l_i(x_0;\mathcal{X})f(x)i)]\ E_{\mathcal{Y},\mathcal{X}}[l_j(x_0;\mathcal{X})]\ E_{\mathcal{Y},\mathcal{X}}[\varepsilon_j]\\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
&E_{\mathcal{Y},\mathcal{X}}\left(\sum_{ij} l_i(x_0;\mathcal{X})f(x_i)l_j(x_0;\mathcal{X})\varepsilon_j\right)\\
&= 0,
\end{aligned}$$

by an argument similar to above using that the ε_j are independent of all other data and have mean 0. Hence

$$\mathrm{Var}_{\mathcal{Y},\mathcal{X}}(\hat{f}(x_0)) = E_{\mathcal{X}}\,\mathrm{Var}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) + \mathrm{Var}_{\mathcal{X}}\left(\sum_i l_i(x_0;\mathcal{X})f(x_i)\right).$$

2.9 The expected value of the quantity $\varepsilon(\hat{\beta}) := (\tilde{x}_j - \tilde{y}_j\hat{\beta})^2$ does not depend on j, since the distributions of $(x_1, y_1), \ldots, (x_N, y_N), (\tilde{x}_j, \tilde{y}_j)$ are all the same. So $R_{te}(\hat{\beta}) = E[\varepsilon(\hat{\beta})]$. We may then assume $N = M$. Let $\widehat{\beta^*}$ denote the regression coefficients obtained by using $\{(\tilde{x}_j, \tilde{y}_j)\}_{j=1}^N$. Then, by its minimality property,

$$\begin{aligned}
\varepsilon(\hat{\beta}) &= \frac{1}{N}\sum_{j=1}^{n}(\tilde{y}_j - \tilde{x}_j\hat{\beta})^2\\
&\geq \frac{1}{N}\sum_{j=1}^{n}(\tilde{y}_j - \tilde{x}_j\widehat{\beta^*})^2
\end{aligned}$$

But if we take the expectation over all that is random and observe that $\{(x_i, y_i)\}_{i=1}^N$ has the same distribution as $\{(\tilde{x}_j, \tilde{y}_j\}_{j=1}^N$, the inequality becomes

$$R_{te}(\hat{\beta}) \geq E\left[\frac{1}{N}\sum_{j=1}^{n}(\tilde{y}_j - \tilde{x}_j\widehat{\beta^*})^2\right] = E[R_{tr}(\beta)]$$