1 Warm-up: Gaussian Mean Estimation

Suppose we have i.i.d. random variables $x_1, x_2, \ldots, x_n \sim \mathcal{N}(\theta_*, \sigma^2 I)$ where $\theta_* \in \mathbb{R}^d$ is unknown and σ^2 is known. Our goal is to estimate θ_* with an estimator $\hat{\theta}$ such that, $d(\hat{\theta}, \theta_*) < \epsilon$ for some small ϵ, where $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is some metric measuring the distance between $\hat{\theta}$ and θ_*. It is understood that $\hat{\theta}$ is a random variable whereas θ_* is deterministic.

There are many approaches that we can take to tackle this estimation problem. For example, we can use

- Sample mean estimator: $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$;
- Maximum Likelihood Estimator (exercise: in fact it reduces to sample mean)
- Maximum A posteriori Probability under some prior on θ_*
- ...

Let's take a look at the sample mean estimator as given by $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$, and find its performance.

Since x_i's are i.i.d. Gaussian random vectors, their linear combination is also Gaussian. One way to see this is by using the moment generating function (MGF) for Gaussian random vectors.

Lemma 1. *Given $Z_1 \sim \mathcal{N}(0, \Sigma_1), Z_2 \sim \mathcal{N}(0, \Sigma_2)$ independent random vectors, we have*

$$Z_1 + Z_2 \sim \mathcal{N}(0, \Sigma_1 + \Sigma_2).$$

Proof. Recall the definition of the MGF of a random variable X as $m_X(t) = \mathbb{E}[e^{\frac{1}{2}\langle t, X\rangle}]$. We have

$$m_{Z_1+Z_2}(t) = \mathbb{E}[e^{\frac{1}{2}\langle t, Z_1+Z_2\rangle}] = \mathbb{E}[e^{\frac{1}{2}\langle t, Z_1\rangle} e^{\frac{1}{2}\langle t, Z_2\rangle}] = \mathbb{E}[e^{\frac{1}{2}\langle t, Z_1\rangle}]\mathbb{E}[e^{\frac{1}{2}\langle t, Z_2\rangle}] = m_{Z_1}(t) m_{Z_2}(t)$$

by the independence of Z_1 and Z_2. Using the fact that the MGF for a Gaussian random variable $Z \sim \mathcal{N}(0, \Sigma)$ is $m_Z(t) = e^{\frac{1}{2}\langle t, \Sigma t\rangle}$, we have

$$m_{Z_1+Z_2}(t) = m_{Z_1}(t) m_{Z_2}(t) = e^{\frac{1}{2}\langle t, \Sigma_1 t\rangle} e^{\frac{1}{2}\langle t, \Sigma_2 t\rangle} = e^{\frac{1}{2}\langle t, (\Sigma_1+\Sigma_2) t\rangle}$$

which is the MGF of $\mathcal{N}(0, \Sigma_1 + \Sigma_2)$. Therefore, $Z_1 + Z_2 \sim \mathcal{N}(0, \Sigma_1 + \Sigma_2)$.

$\square$

If we look at the difference between the sample mean estimator and the true mean, we have

$$\hat{\theta} - \theta_* = \frac{1}{n} \sum_{i=1}^n (x_i - \theta_*)$$

Since each $x_i - \theta_* \sim \mathcal{N}(0, \sigma^2 I)$, applying Lemma 1 iteratively, we obtain

$$\hat{\theta} - \theta_* \sim \mathcal{N}\left(0, \frac{\sigma^2}{n} I\right). \tag{1.1}$$

Definition 2. *We define the notion of loss and risk as follows.*

- ***Loss*** *measures the distance. We will denote it by $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$. For example, squared L_2-norm can be a loss $\ell(\theta, \theta') = \|\theta - \theta'\|^2$.*

- ***Risk*** *is the expected loss (so it is a population quantity). Risk between an estimator and true parameter*

$$R(\hat{\theta}, \theta_*) = \mathbb{E}[\ell(\hat{\theta}, \theta_*)].$$

Here, the expectation is over $\hat{\theta}$.

Next, let's choose the loss function as the squared L_2-norm, i.e., $\ell(\theta, \theta') = \|\theta - \theta'\|^2$. Then the risk function is given as $R(\hat{\theta}, \theta_*) = \mathbb{E}[\ell(\hat{\theta}, \theta_*)] = \mathbb{E}[\|\hat{\theta} - \theta_*\|^2]$. For the sample mean estimator, we have

$$\ell(\hat{\theta}, \theta_*) = \|\hat{\theta} - \theta_*\|^2, \quad \text{and} \quad R(\hat{\theta}, \mu) = \mathbb{E}[\|\hat{\theta} - \theta_*\|^2] = \frac{\sigma^2 d}{n}, \tag{1.2}$$

where in the last step we used (1.1). Note that the risk $R(\hat{\theta}, \theta_*)$ increases with dimension d and decreases with the number of samples n. This dependence structure is commonly observed for most loss minimization problems. This intuitively means that estimation is harder in higher dimensions, but gets better with more observations.

Remark. The loss $\ell(\hat{\theta}, \theta_*) \sim \chi^2_d$ where χ^2_d denotes the chi-square distribution.

One concern about this estimator is that $\mathbb{E}[||\hat{\theta}||^2] = ||\theta_*||^2 + \frac{\sigma^2 d}{n} > ||\theta_*||^2$. This means that the second moment of our estimator is always significantly larger than that of the true parameter we are estimating. To resolve this, we can simply multiply $\hat{\theta}$ by a factor $(1 - \eta)$ to *shrink* it. This type of estimators called *shrinkage estimator*. In what follows, we show that MLE can be beaten.

1.1 SURE: Stein's Unbiased Risk Estimator

Lemma 3 (Stein's Lemma)**.** *Suppose $x \sim \mathcal{N}(\mu, \sigma^2 I)$, and $g : \mathbb{R}^d \to \mathbb{R}^d$ is weakly differentiable. Then*

$$\mathbb{E}[\langle x - \mu, g(x) \rangle] = \sigma^2 \mathbb{E}[\mathrm{Tr}(\nabla g(x))].$$

Remark. We are not giving a definition of *weak differentiablity*, but hereby we will assume g is differentiable which is a stronger assumption.

Proof. Let $\phi(x)$ denote the distribution of an isotropic Gaussian random vector. We can write

$$\mathbb{E}[\langle x - \mu, g(x) \rangle] = \int_{-\infty}^{\infty} \langle x - \mu, g(x) \rangle \phi(\tfrac{x-\mu}{\sigma}) \, \mathrm{d}x.$$

Using the fact that

$$\mathrm{d}\phi(\tfrac{x-\mu}{\sigma}) = -\frac{x - \mu}{\sigma^2} \phi(\tfrac{x-\mu}{\sigma}) \, \mathrm{d}x$$

and integration by parts, we have

$$\begin{aligned}\int_{-\infty}^{\infty} \langle x - \mu, g(x) \rangle \phi(\tfrac{x-\mu}{\sigma}) \, \mathrm{d}x &= -\sigma^2 \int_{-\infty}^{\infty} \langle \mathrm{d}\phi(\tfrac{x-\mu}{\sigma}), g(x) \rangle \\ &= \sigma^2 \int_{-\infty}^{\infty} \phi(\tfrac{x-\mu}{\sigma}) \, \mathrm{Tr}(\nabla g(x)) \, \mathrm{d}x = \sigma^2 \mathbb{E}[\mathrm{Tr}(\nabla g(x))].\end{aligned}$$

□

Remark. The above results is also referred to as Stein's identity, and has remarkable applications ranging from probability theory (non-asymptotic CLTs) to machine learning (Stein's variational gradient descent) and optimization (Newton-Stein method, Scaled Least Squares).

In the following we will consider the risk of estimators of a particular form and show that MLE can be beaten in terms of risk. Let $\hat{\theta}^s$ be an estimator of the form

$$\hat{\theta}^s = \hat{\theta} + g(\hat{\theta}), \tag{1.3}$$

where $\hat{\theta}$ is the sample mean and $g : \mathbb{R}^d \to \mathbb{R}^d$ is any differentiable function. Then

$$\begin{aligned} R(\hat{\theta}^s, \theta_*) =& \mathbb{E}[\|\hat{\theta}^s - \theta_*\|^2] = \mathbb{E}[\|\hat{\theta} + g(\hat{\theta}) - \theta_*\|^2], \\ =& \mathbb{E}[\|\hat{\theta} - \theta_*\|^2] + \mathbb{E}[\|g(\hat{\theta})\|^2] + 2\mathbb{E}[\langle \hat{\theta} - \theta_*, g(\hat{\theta})\rangle], \\ =& \frac{\sigma^2 d}{n} + \mathbb{E}[\|g(\hat{\theta})\|^2] + \frac{2\sigma^2}{n}\mathbb{E}[\mathrm{Tr}(\nabla g(\hat{\theta}))], \end{aligned} \tag{1.4}$$

where in the last step, we applied Stein's Lemma 3 on the last term.

This leads to the definition of the Stein's Unbiased Risk Estimator:

Definition 4 (SURE: Stein's Unbiased Risk Estimator)**.** *For an estimator of the form $\hat{\theta}^s = \hat{\theta} + g(\hat{\theta})$, we have the following unbiased estimator of the risk,*

$$SURE(\hat{\theta}) = \frac{\sigma^2 d}{n} + ||g(\hat{\theta})||^2 + \frac{2\sigma^2}{n}\,\mathrm{Tr}(\nabla g(\hat{\theta})).$$

The fact that $SURE(\hat{\mu})$ is an unbiased estimator for $R(\hat{\mu}^s, \mu)$ follows from (1.4). In other words, any estimator of the form (1.3), has the risk $\mathbb{E}[SURE(\hat{\mu})]$. Also note that the first term on the right hand side is the risk of MLE.

In the following, we will specify the function g in (1.3).

1.2 James-Stein Estimator

Definition 5 (James-Stein Estimator)**.** *Define the estimator*

$$\hat{\theta}^{js} = \left(1 - \frac{d-2}{||\hat{\theta}||^2}\frac{\sigma^2}{n}\right)\hat{\theta}.$$

The above estimator is of the form (1.3) with

$$g(x) = -\frac{\sigma^2}{n}\frac{d-2}{||x||^2}x, \quad \text{and} \quad \nabla g(x) = -\frac{\sigma^2}{n}\frac{d-2}{||x||^2}I + 2(d-2)\frac{\sigma^2}{n}\frac{xx^T}{||x||^4}.$$

This gives

$$\|g(x)\|^2 = \frac{\sigma^4}{n^2}\frac{(d-2)^2}{\|x\|^2} \quad \text{and} \quad \mathrm{Tr}(\nabla g(x)) = \frac{-d(d-2) + 2(d-2)}{||x||^2}\frac{\sigma^2}{n} = -\frac{(d-2)^2}{||x||^2}\frac{\sigma^2}{n}.$$

Therefore the risk of the James-Stein estimator is given as

$$\begin{aligned} R(\hat{\theta}^{js}, \theta_*) =& \frac{\sigma^2 d}{n} + \frac{\sigma^4}{n^2}\mathbb{E}\left[\frac{(d-2)^2}{\|\hat{\theta}\|^2}\right] - 2\frac{\sigma^4}{n^2}\mathbb{E}\left[\frac{(d-2)^2}{||\hat{\theta}||^2}\right] = \frac{\sigma^2 d}{n} - \frac{\sigma^4}{n^2}\mathbb{E}\left[\frac{(d-2)^2}{||\hat{\theta}||^2}\right] \\ <& R(\hat{\theta}, \theta_*), \end{aligned}$$

where the last step follows from $R(\hat{\theta}, \theta_*) = \sigma^2 d/n$ as derived in (1.2) Therefore, the James-Stein estimator is a strictly better estimator than the sample mean estimator based on the measure of the risk function. Note that this result holds for $d > 2$.

If we go one step further by applying Jensen's inequality ($x \to 1/x$ is convex for $x > 0$), we obtain

$$\mathbb{E}\left[\frac{1}{\|\hat{\theta}\|^2}\right] \geq \frac{1}{\mathbb{E}[\|\hat{\theta}\|^2]} = \frac{1}{\|\theta_*\|^2 + \frac{\sigma^2 d}{n}}.$$

Using this in the last step above, our bound for the risk of James-Stein estimator yields

$$\begin{aligned} R(\hat{\theta}^{js}, \theta_*) =& \frac{\sigma^2 d}{n} - \frac{\sigma^4}{n^2}\mathbb{E}[\frac{(d-2)^2}{\|\hat{\theta}\|^2}], \\ \leq& \frac{\sigma^2 d}{n} - \frac{\sigma^4}{n^2}\frac{(d-2)^2}{\|\theta_*\|^2 + \frac{\sigma^2 d}{n}}. \end{aligned}$$

Remark. A more careful treatment yields the following bound

$$R(\hat{\theta}^{js}, \theta_*) \leq \frac{\sigma^2 d}{n} - \frac{\sigma^4}{n^2}\frac{(d-2)^2}{\|\theta_*\|^2 + \frac{\sigma^2 (d-2)}{n}}.$$

- James-Stein is one the most significant advances in statistics.
- It shows that MLE can be beaten (inadmissable) for $d > 2$.
- This phenomenon is also known as Stein's paradox.