

Lecture 18: Stein's phenomenon and shrinkage estimation

Lecturer: Yiqiao Zhong

Last update: 04:47, 2024-12-02

Disclaimer: materials in this pdf are used for STAT 709 "Mathematical Statistics" at the University of Wisconsin-Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

1 The Motivation

Suppose $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, I_p)$, where $\mu \in \mathbb{R}^p$ is the unknown parameter. The sample mean is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \frac{1}{n} I_p).$$

It is the UMVUE, and thus attains Cramér-Rao lower bound. **Q:** While we have a complete theory for unbiased estimation, what about biased estimation?

For simplicity, consider one observation

$$X \sim \mathcal{N}(\mu, I_p),$$

otherwise we can rescale $X = \sqrt{n}\bar{X}$, and $\sqrt{n}\mu$ replaces μ .

1.1 A high-dimensional perspective

An initial observation by Stein in 1956 [3]. Let

$$X = \mu + Z, \text{ where } Z \sim \mathcal{N}(0, I_p).$$

Then, when p is large, we have

$$\|X\|^2 = \|\mu\|^2 + \|Z\|^2 + 2\langle \mu, Z \rangle \approx \|\mu\|^2 + p,$$

where " $\approx$ " can be made rigorous using the CLT or concentration inequalities. It is clear that if $p \gg \|\mu\|^2$, then

$$\|X\| \gg \|\mu\| \quad \text{with high probability,}$$

since large variances are accumulated across p coordinates.

This leads to attempts to improve the natural unbiased estimator—shrinking the norm and obtaining biased estimator $\hat{\mu}$.

1.2 Bias-variance tradeoff perspective

From statistical decision theory, the risk function under the squared loss (MSE) of an estimator T is

$$R_T(\mu) = \mathbb{E}[\|T(X) - \mu\|^2].$$

where the expectation $\mathbb{E}$ is based on $\mathcal{N}(\mu, I_p)$.



A homework problem. Consider a loosely related problem. In the homework problem, we have observations $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, where the UMVUE is the sample mean $\bar{X}$ with

$$R_{\bar{X}}(p) = \frac{p(1-p)}{n}.$$

The risk is highest at $p = 1/2$. Another simple estimator is the constant estimate $1/2$, which has the lowest risk at $p = 1/2$ but highest at 0 and 1.

We can “even out” the risk by interpolating the two estimators: shrinking $\bar{X}$ towards the midpoint $\frac{1}{2}$, trading bias around endpoints for smaller variance around the midpoint, which arrives at the minimax risk estimator

$$T_0(X) = \frac{\sqrt{n}}{\sqrt{n}+1} \bar{X} + \frac{1}{2(\sqrt{n}+1)}.$$

We can calculate the risk.

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{\sqrt{n}}{\sqrt{n}+1} \bar{X} + \frac{1}{2(\sqrt{n}+1)} - p \right)^2 \right] \\ &= \underbrace{\frac{1}{(\sqrt{n}+1)^2} p(1-p)}_{\text{variance}} + \underbrace{\frac{(p-1/2)^2}{(\sqrt{n}+1)^2}}_{\text{bias}^2} \\ &= \frac{1}{4(\sqrt{n}+1)^2}. \end{aligned}$$

The risk is constant over p , and the estimator is minimax. Minimaxity will be studied in detail in STAT 710.

Beating the constant risk. In our Gaussian mean estimation, the natural unbiased estimator X already has constant risk over all $\mu \in \mathbb{R}^p$.

Consider an estimator of the form $\hat{\mu} = (1 - \varepsilon)X$ where $\varepsilon \in (0, 1)$. The risk function is

$$\mathbb{E}[\|\hat{\mu} - \mu\|^2] = \underbrace{\varepsilon^2 \|\mu\|^2}_{\text{Bias}^2} + \underbrace{(1 - \varepsilon)^2 p}_{\text{Variance}}.$$

Running a quick analysis, we find that $R_{\hat{\mu}} = (1 - 2\varepsilon)p + O(\varepsilon^2)$ as $\varepsilon \rightarrow 0$, so

$$R_{\hat{\mu}} = (1 - 2\varepsilon)p + O(\varepsilon^2).$$

Thus, shrinkage is advantageous if we can choose the correct ε . To determine the optimal ε , we solve

$$\min_{\varepsilon \in [0,1]} \{ \varepsilon \|\mu\|^2 + (1 - \varepsilon)^2 p \} \implies \varepsilon^* = \frac{p}{\|\mu\|^2 + p}.$$

Unfortunately, we don't know $\|\mu\|$ as it involves the unknown μ . Nevertheless, we only need to estimate the norm reasonably well instead of estimating the p -dimensional μ .

A heuristic is to use the high-dimensional intuition: The if p is sufficiently large,

$$\|\theta\|^2 + p \approx \|X\|^2.$$

so we can use $\|X\|^2$ to estimate $\|\mu\|^2 + p$. The James-Stein estimator optimizes the factor and achieves a smaller risk function in dimension as small as $p = 3$.

2 Shrinkage estimators

James-Stein estimator and its improvement are defined by

$$\hat{\mu}_{\text{JS}} = \left(1 - \frac{p-2}{\|X\|^2}\right) X, \quad \hat{\mu}_{\text{JS}+} = \left(1 - \frac{p-2}{\|X\|^2}\right)_+ X.$$

We can actually choose a nonzero center. More generally, given $c \in \mathbb{R}^p$,

$$\hat{\mu}_{\text{JS}} = \left(1 - \frac{r(p-2)}{\|X-c\|^2}\right) (X-c) + c, \quad \hat{\mu}_{\text{JS}+} = \left(1 - \frac{r(p-2)}{\|X-c\|^2}\right)_+ (X-c) + c. \quad (1)$$

We interpret this result from the signal vs. noise perspective. If we believe signal strength $\|\mu\|^2$ is large as evidenced by large $\|X\|^2$, then we need weaker shrinkage (less bias to balance smaller variance). On the other hand, if dimension p is large, then there is more accumulated noise, so we need stronger shrinkage.

Theorem 1. *If $\hat{\mu}$ is any of the two shrinkage estimators in Eqn. 1, then for any $r \in (0, 2)$, when $p \geq 3$,*

$$R_{\hat{\mu}}(\mu) < R_X(\mu), \quad \text{for every } \mu \in \mathbb{R}^p.$$

This inequality means that the natural mean estimator X is inadmissible. Moreover, for $\hat{\mu} = \hat{\mu}_{\text{JS}}$, we have

$$R_{\hat{\mu}}(\mu) = p - (2r - r^2)(p-2)^2 \mathbb{E}[\|X-c\|^{-2}]. \quad (2)$$

Stein's identity. First, we state one widely used proof component.

Theorem 2 (Stein's lemma). *If $g \in C^1(\mathbb{R}^p)$ with $\mathbb{E}[\|\nabla g(Z)\|] < \infty$, where $Z \sim \mathcal{N}(0, I_p)$, then*

$$\mathbb{E}[\nabla g(Z)] = \mathbb{E}[Zg(Z)].$$

Proof. For any $i \leq p$, we evaluate

$$\begin{aligned} \mathbb{E}[\partial_i g(Z)] &= \int \partial_i g(z) \frac{1}{\sqrt{2\pi}} e^{-\|z\|^2/2} dz = - \int g(z) \frac{1}{\sqrt{2\pi}} \partial_i e^{-\|z\|^2/2} dz \\ &= \int g(z) \frac{1}{\sqrt{2\pi}} z_i e^{-\|z\|^2/2} dz \\ &= \mathbb{E}[Z_i g(Z)] \end{aligned}$$

where we used integration by parts in the second equality. □

Stein's identity can be extended to multidimensional case. If $g \in C^1(\mathbb{R}^p; \mathbb{R}^p)$ is a map, then ([exercise](#))

$$\mathbb{E}[\nabla \cdot g(Z)] = \mathbb{E}[Z^\top g(Z)] \quad (3)$$

where $\nabla \cdot g(z) = \sum_{j=1}^p \partial_j g_j(z)$

Proof of Eqn. 2. Without loss of generality, we assume $c = 0$ (why?).

Proof. Let us consider estimators of the form $X - g(X)X$ where $g \in \mathbb{R}^1(\mathbb{R}^p)$. We introduce $Z = X - \mu \sim N(0, I_p)$. Now

$$\begin{aligned} R(\mu) &= \mathbb{E}[\|X - \mu - g(X)X\|^2] \\ &= \mathbb{E}[\|Z - g(X)X\|^2] \\ &= p + \mathbb{E}[\|X\|^2 g(X)^2] - 2 \underbrace{\mathbb{E}[Z^\top X g(X)]}_{\text{apply Stein's identity}} \\ &= p + \mathbb{E}[\|X\|^2 g(X)^2] - 2p\mathbb{E}[g(X)] - 2\mathbb{E}[X^\top \nabla g(X)]. \end{aligned}$$

where we applied Stein's identity (Eqn. 3) to the map $z \mapsto (z + \mu)g(z + \mu)$.

Now consider the form $g(X) = \frac{b}{a + \|X\|^2}$ where $a, b \geq 0$. We find

$$R(\mu) - p = b^2 \mathbb{E} \left[\frac{\|X\|^2}{(a + \|X\|^2)^2} \right] - 2pb \mathbb{E} \left[\frac{1}{a + \|X\|^2} \right] + 4b \mathbb{E} \left[\frac{\|X\|^2}{(a + \|X\|^2)^2} \right] \quad (4)$$

$$\leq (b^2 - 2pb + 4b) \mathbb{E} \left[\frac{1}{a + \|X\|^2} \right], \quad (5)$$

where we used

$$\mathbb{E} \left[\frac{\|X\|^2}{(a + \|X\|^2)^2} \right] \leq \mathbb{E} \left[\frac{1}{a + \|X\|^2} \right].$$

The equality holds if $a = 0$. In Eqn. 4 we set $a = 0$ and $b = r(p - 2)$, which leads to the expression in our target equation (2). $\square$

3 Tweedie's formula

3.1 Bayesian perspective

Suppose that μ has a prior distribution $\mu \sim \pi$ where π is a (Lebesgue) pdf on $\mathbb{R}^p$. Our observation becomes a conditional density $X|\mu \sim \mathcal{N}(\mu, I_p)$. The marginal density of X is

$$\begin{aligned} f_X(x) &= \int_{\mathbb{R}^p} f_{X|\mu}(x, \mu) d\mu = \int_{\mathbb{R}^p} f_{X|\mu}(x|\mu) \pi(\mu) d\mu \\ &= \frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} e^{-\|x - \mu\|^2/2} \pi(\mu) d\mu. \end{aligned}$$

Consider a map $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^p$ in order to minimize $\mathbb{E}[\|\mu - \phi(X)\|^2]$. This MSE is $\mathbb{E}_{\mu \sim \pi} R(\mu)$, i.e., mean risk function under the prior π . If the Fubini's condition is satisfied,

$$\mathbb{E}[\|\mu - \phi(X)\|^2] = \mathbb{E}_X \{ \mathbb{E}_{\mu|X} [\|\mu - \phi(X)\|^2] \}$$

Note that conditioning on $X = x$, $\phi(X)$ has a fixed value, so

$$\mathbb{E}_{\mu|X=x} [\|\mu - \phi(X)\|^2] \geq \mathbb{E}_{\mu|X=x} [\|\mu - \mathbb{E}_{\mu|X=x}(\mu)\|^2]$$

i.e., the minimizer of MSE must be the conditional mean of μ under the posterior distribution. In order for the equality to hold, we only need to choose

$$\phi(x) = \mathbb{E}_{\mu|X=x}(\mu).$$

It boils down to derive $\mathbb{E}_{\mu|X=x}(\mu)$. Observe

$$\mathbb{E}_{\mu|X=x}(\mu) = \int_{\mathbb{R}^p} \mu f_{\mu|X=x}(\mu|x) d\mu = \frac{\int_{\mathbb{R}^p} \mu f_{X=x|\mu}(x|\mu) \pi(\mu) d\mu}{\int_{\mathbb{R}^p} f_{X|\mu}(x|\mu) \pi(\mu) d\mu}$$

where we applied Bayes' formula to the conditional pdf $f_{\mu|X=x}(\mu|x)$. Notice the following equality for our case $X|\mu \sim \mathcal{N}(\mu, I_d)$.

$$\mathbb{E}_{\mu|X=x}(\mu - x) = \frac{\int_{\mathbb{R}^p} (\mu - x) e^{-\|x - \mu\|^2/2} \pi(\mu) d\mu}{\int_{\mathbb{R}^p} e^{-\|x - \mu\|^2/2} \pi(\mu) d\mu} = \frac{\nabla f_X(x)}{f_X(x)} = \nabla \log f_X(x).$$

Thus, we derived the Bayes estimator under prior π

$$\phi_\pi(X) = X + \nabla \log f_X(X). \quad (6)$$

Remark 1. From this perspective, shrinkage is a result of a prior on the unknown π .

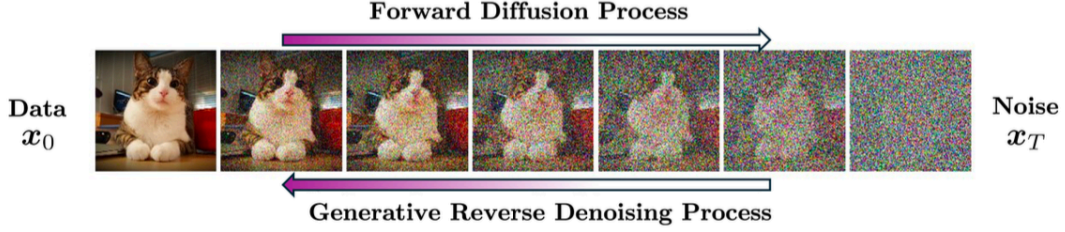


Figure 1: Diffusion model: generating a new image through progressive denoising.

3.2 Shrinkage under superharmonic function

We don't know the prior distribution, but we can consider estimators of the form $X + \nabla \log f(X)$ —known as Tweedie's formula, and directly make assumptions on f .

For a twice differential function $f \in C^2(\mathbb{R}^p)$, denote the Laplace operator by $\Delta f(x) = \sum_{j=1}^p \partial_j^2 f(x)$. We say that a twice differential function f is superharmonic iff $\Delta f(x) \leq 0$ for every $x \in \mathbb{R}^p$.

Theorem 3. *Let $f \in C^2(\mathbb{R}^p)$ taking positive values and $\sqrt{f}$ is superharmonic. Under the regularity conditions*

$$\mathbb{E} \left[\frac{1}{f(X)} \sum_{j=1}^p |\partial_j^2 f(X)| \right] < \infty, \quad \mathbb{E} [\|\nabla \log f(X)\|^2] < \infty,$$

we have

$$\mathbb{E} [\|X + \nabla \log f(X) - \mu\|^2] = p + 4 \mathbb{E} \left[\frac{\Delta \sqrt{f(X)}}{\sqrt{f(X)}} \right] \leq p$$

This theorem is a result of direct calculation (homework exercise).

Remark 2 (Connection to minimax). *The naive estimator X has constant risk and is a minimax estimator. So this theorem automatically verifies minimaxity for $X + \nabla \log f(X)$. Moreover, there is a wide class of estimators constructed from superharmonic functions that beat the naive estimator.*

If we choose $f(x) = \|x\|^{-2m}$, then we can check that f

$$\Delta \sqrt{f(x)} = m \cdot \frac{m+2-p}{\|x\|^{m+2}}.$$

Thus, as long as $m \in (0, p-2]$, the function is superharmonic¹ and satisfies the regularity conditions. It recovers the JS estimator when $m = (p-2)/2$.

4 A modern connection: diffusion model

How time flies! Now we have tech companies running gigantic models. Figure 2 shows the idea of generating new images using a diffusion model. Let x_0 represent a randomly sampled image (vectorized pixel values).

- **Forward process.** Given pre-specified scalar scales (α_t) and (β_t) , progressively adding noise to an image x_0 :

$$x_t = \alpha_t x_0 + \beta_t \varepsilon, \quad \text{where } \varepsilon \sim \mathcal{N}(0, I)$$

¹In particular, if $m = p-2$, it is a harmonic function.

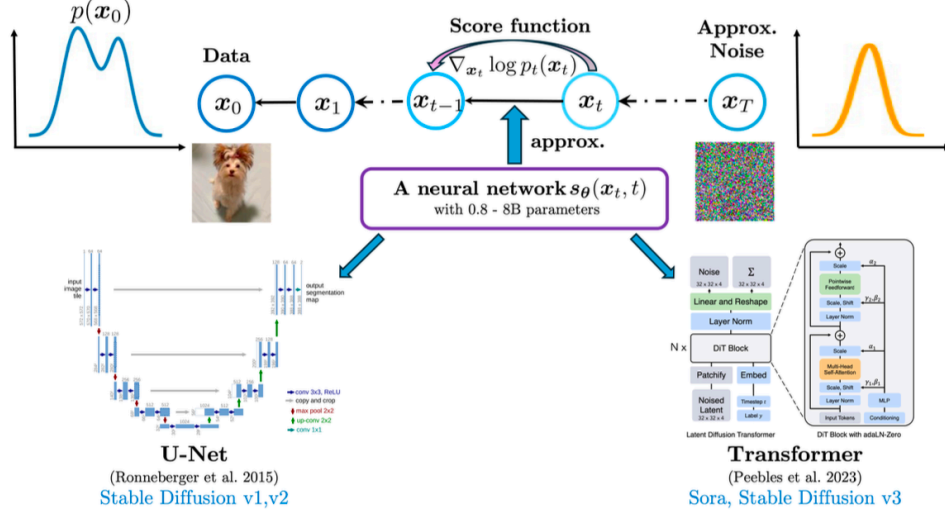


Figure 2: Score function estimation via neural networks (credit: Qing Qu).

- **Backward process.** Starting from a random noise ε , progressively denoising to generate an image x_0 .

The forward process is easy to implement: just adding independent Gaussian noise at each (discrete) time step, for example,

$$x_t = x_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2} z_{t-1}, \quad \text{or} \quad x_t = \sqrt{1 - \sigma_t^2} x_{t-1} + \sigma_t z_{t-1}$$

where $z_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$. The critical part is to denoise an image from x_t to x_{t-1} .

Score function. Suppose that at time t , the random vectors X_{t-1} and X_t are determined by

$$X_t = \sqrt{1 - \sigma_t^2} X_{t-1} + \sigma_t Z_t, \quad \text{where } Z_t \sim \mathcal{N}(0, I).$$

By Tweedie's formula, the posterior mean is given by

$$\mathbb{E}[X_{t-1} | X_t = x_t] = \frac{1}{\sqrt{1 - \sigma_t^2}} x_t + \frac{\sigma_t^2}{\sqrt{1 - \sigma_t^2}} \underbrace{\nabla \log p_t(x_t)}_{\text{score function}}$$

where p_t is the pdf of X_t . We don't know the score function exactly, but it can be estimated from data yields $s_\theta(x_t, t)$. The denoising step uses the score function as follows: starting from $\bar{X}_T = X_T$,

$$\bar{X}_{t-1} = \frac{1}{\sqrt{1 - \sigma_t^2}} \bar{X}_t + \frac{\sigma_t^2}{\sqrt{1 - \sigma_t^2}} s_\theta(\bar{X}_t, t) + \sigma_t Z_t, \quad \text{where } t = T, T-1, \dots, 1 \quad (7)$$

The score function estimate $s_\theta(x_t, t)$ is usually obtained by optimizing a neural network where θ contains all trainable parameters in the neural network:

$$\min_{\theta} \mathbb{E}_{t \sim \text{Unif}([0,1]), X_0 \sim p_0, X_t \sim p(x_t | X_0)} \left[\sigma_t \left\| \nabla \log p_t(X_t) - s_\theta(X_t, t) \right\|_2^2 \right]$$

In implementation, we sample an image X_0 from a training set, sample a time stamp t , and noised image X_t from the forward pass. Then we minimize this function by default (stochastic) gradient descent to find $s_\theta(x_t, t)$. Once this is done, we can sample a new image using Eqn. 7.

Stochastic differential equation*. The theoretical foundation for diffusion models can be found in [1]. Consider the stochastic differential equation (SDE):

$$\text{(forward pass)} \quad dX_t = f(X_t, t)dt + g(t)dW_t$$

where $X_t \sim p_0$ and $(W_t)_{t \geq 0}$ is the Wiener process (Brownian motion). Then,

$$\text{(backward pass)} \quad d\bar{X}_t = \left[f(\bar{X}_t, t) - [g(t)]^2 \nabla \log p_t(\bar{X}_t) \right] dt + g(t)d\bar{W}_t$$

where $\bar{X}_T \sim X_T$ and $(\bar{W}_t)_{t \geq 0}$ is another independent Wiener process. It is shown [1] that $(\bar{X}_t)_{t \in [0, T]}$ has the same distribution as $(X_t)_{t \in [0, T]}$.

5 Further reading

The main reference of this lecture note is [3, 4]. The references of diffusion models are [1, 2].

References

- [1] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- [2] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [3] Charles Stein et al. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1956.
- [4] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.