

Lecture 18: Stein's phenomenon and shrinkage estimation

Lecturer: Yiqiao Zhong[Report an issue](#)*Last update: October 19, 2026*

These notes support STAT 709 at the University of Wisconsin–Madison. Use them alongside the cited textbooks; corrections and questions are welcome.

Optional part. This lecture belongs to optional Part VI and is not examined. Its high, mid and low labels organize a suggested reading path within the optional material; they do not add exam requirements. See the Part VI overview.

High priority *Overview*

An estimator can be best among unbiased rules and still be improved by allowing bias. We begin with the risk of fixed **shrinkage**, then use Stein's identity to verify the James–Stein improvement in dimension at least three. Tweedie's formula explains shrinkage through a posterior mean. Superharmonic constructions and diffusion models are optional connections built on that identity.

Recap from earlier lectures

From Lectures 1, 5 and 7: earlier ideas used here

Reminder only: this block introduces no new exam requirements.

- **Bias and variance.** Squared-error risk balances squared bias and variance. Used in the fixed-shrinkage comparison.
- **Gaussian transformations.** Rescaling a Gaussian vector rescales its mean and covariance. Used to reduce to unit noise.
- **Conditional means.** The conditional mean minimizes squared prediction error. Used for the Bayes estimator.

Learning goals

- Compare fixed shrinkage with the data-dependent James–Stein rule.
- Follow the Stein identity calculation and explain why dimension three matters.
- Derive a posterior mean from a Gaussian marginal density.
- Explore the optional connection between denoising and diffusion models.

1 Motivation**High priority** *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

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

Each coordinate of $\bar{X}$ is the UMVUE, and in this Normal model the sample mean also attains the 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

Mid priority *A high-dimensional perspective*

An initial observation was made 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

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

The mean and variance make this heuristic precise:

$$\mathbb{E} \|X\|^2 = \|\mu\|^2 + p, \quad \text{Var}(\|X\|^2) = 2p + 4\|\mu\|^2.$$

The cross covariance vanishes because centered Gaussian third moments are zero. Chebyshev's inequality now gives $\|X\|^2 / (\|\mu\|^2 + p) \rightarrow_p 1$ as $p \rightarrow \infty$, even if μ varies with p . In particular, 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 a biased estimator $\hat{\mu}$.



1.2 Bias-variance tradeoff perspective

High priority *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 . To verify minimaxity, put a $\text{Beta}(\sqrt{n}/2, \sqrt{n}/2)$ prior on p : its posterior mean is exactly T_0 . Any estimator’s maximum risk is at least its average risk under this prior, and that average risk is minimized by T_0 . Since T_0 has constant risk, it attains this lower bound. This uses the posterior-mean principle from Lecture 5. 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

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

For each fixed μ and p , a small amount of shrinkage reduces risk:

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

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

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

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: if p is sufficiently large,

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

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 $p = 3$.

Bias can lower risk. For $X \sim N(\mu, I_p)$, constant shrinkage aX has risk $a^2p + (1-a)^2\|\mu\|^2$. Reduced variance can outweigh squared bias, but a fixed a below one cannot dominate X over all means.

High priority *Practice: When does fixed shrinkage beat the data?*

Exercise prompt. For $T_a(X) = aX$, $0 < a < 1$, find the set of means for which its squared-error risk is smaller than p .

See Exercise 1 in Section 7 for the full statement and complete solution.

2 Shrinkage estimators

High priority *Shrinkage estimators*

For $p \geq 3$, the 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 a fixed, data-independent $c \in \mathbb{R}^p$,

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

We interpret this result from the signal vs. noise perspective. If we believe signal strength relative to the center, $\|\mu - c\|^2$, is large as evidenced by large $\|X - c\|^2$, then we shrink less, to avoid pulling a strong signal too far toward the center. 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 either of the two shrinkage estimators in Equation 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**: another estimator has no larger risk at any mean and smaller risk at some mean (here, at every mean). The comparison concerns total expected squared error, not every observed sample or every coordinate separately. 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 $\mathbb{E}\|\nabla g(Z)\| < \infty$, $\mathbb{E}|Z_i g(Z)| < \infty$ for every i , and vanishing boundary terms (or a valid cutoff approximation), 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

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

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

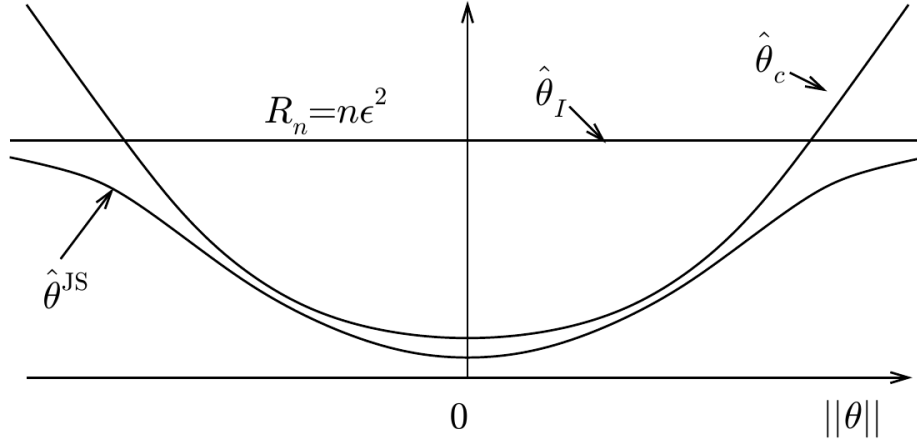


Figure 1: Expected squared-error risk: fixed linear shrinkage eventually exceeds the raw-data risk, whereas James–Stein stays below it. In the source notation, $n = p$, $\epsilon = 1$ and $\theta = \mu$; the picture uses shrinkage toward zero. Figure: Johnstone, Figure 2.1. The comparison is in expectation, not at every observation.

Applying the scalar identity to each component gives its vector-valued version. If $h : \mathbb{R}^p \rightarrow \mathbb{R}^p$ satisfies the same differentiability, integrability and boundary conditions componentwise, then (see Exercise 5)

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

where $\nabla \cdot h(z) = \sum_j \partial_j h_j(z)$ is its divergence. Below, g denotes a scalar shrinkage factor and h a vector correction.

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

Proof. Let us consider estimators of the form $X - g(X)X$ where $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is differentiable with the required integrability. 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 (Equation 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 Equations 4–5 we set $a = 0$ and $b = r(p - 2)$, which leads to the expression in our target equation (2). $\square$

Why the singularity is integrable. For $p > 2$, the Gaussian density is bounded near c , and the radial volume element is proportional to $\rho^{p-1}d\rho$. Thus ρ^{-2} is locally integrable because $\int_0^1 \rho^{p-3}d\rho < \infty$; outside the unit ball it is bounded. Apply the smooth risk calculation first with $a > 0$ and then dominated convergence as $a \downarrow 0$. Values assigned at $X = c$ do not affect risk because this event has probability zero.

The positive part also improves risk. Write $b = r(p-2) > 0$ and translate c to zero. The ordinary rule equals $t(R)X$, where $R = \|X\|$ and $t(R) = 1 - b/R^2$. It differs from the positive part only when $R^2 < b$. Given R , the conditional direction has density proportional to $e^{\mu^\top X}$ on the sphere. Pairing opposite directions shows $\mathbb{E}[\mu^\top X \mid R] \geq 0$: their paired contribution is proportional to $a \sinh(a) \geq 0$. On $R^2 < b$, the conditional loss of $t(R)X$ minus that of zero is

$$t(R)^2 R^2 - 2t(R) \mathbb{E}[\mu^\top X \mid R] > 0.$$

This event has positive probability. Consequently the positive-part rule strictly improves the ordinary rule, completing the dominance claim.

Stein risk identity. For a sufficiently regular vector function h and $X \sim N(\mu, I_p)$, the risk of $X + h(X)$ is $p + \mathbb{E}[2 \operatorname{div} h(X) + \|h(X)\|^2]$. The derivative and integrability conditions matter at the shrinkage singularity.

High priority *Practice: Translate the shrinkage center*

Exercise prompt. Explain rigorously why the James–Stein risk calculation may assume $c = 0$ for a fixed center c .

See Exercise 2 in Section 7 for the full statement and complete solution.

3 Tweedie's formula

3.1 Bayesian perspective

Mid priority *Bayesian perspective*

Suppose that μ has a prior distribution $\mu \sim \pi$ where π is a density on $\mathbb{R}^p$ with $\mathbb{E}_\pi \|\mu\|^2 < \infty$. The observation now has conditional density $X \mid \mu \sim \mathcal{N}(\mu, I_p)$. The marginal density of X is

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

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 . Tonelli's theorem applies to the nonnegative squared loss, giving

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

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

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

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

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

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

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

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

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

The gradient is with respect to the observation x , not the unknown parameter as in a likelihood score. Differentiation under the integral is valid because Gaussian derivatives times a polynomial are bounded as functions of $x - \mu$. Thus the Bayes estimator under prior π is

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

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

Instructor's Note *Follow the slope of the noisy density*

Tweedie's formula starts at the observation and adds a correction pointing toward where the noisy data distribution becomes denser. For a centered Gaussian prior that correction points toward zero. A more structured prior produces a direction that depends on the local shape of the density.

Posterior means and marginal scores. In Gaussian noise, the posterior mean equals the observation plus the noise variance times the gradient of the log marginal density. This score is a data-variable derivative, distinct from a parameter score in likelihood theory.

Tweedie's formula also applies to a discrete prior: replace its integrals by weighted sums. In the symmetric two-point example below, posterior odds are e^{4x} , so $\mathbb{E}[\mu | X = x] = 2 \tanh(2x)$. This illustrates why a structured prior need not shrink toward a single center.

Mid priority *Practice: Tweedie with a Gaussian prior*

Exercise prompt. If $\mu \sim N(0, \tau^2 I_p)$ and $X | \mu \sim N(\mu, I_p)$, use Tweedie to find the posterior mean. See Exercise 3 in Section 7 for the full statement and complete solution.

3.2 Shrinkage from superharmonic functions

Low priority *Shrinkage from superharmonic functions*

Tweedie's formula suggests considering estimators of the form $X + \nabla \log f(X)$ and placing assumptions directly on a positive function f . In this construction, f need not be a probability density or the marginal density from a prior; the risk calculation stands on its own.

For a twice differentiable 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 differentiable function f is superharmonic iff $\Delta f(x) \leq 0$ for every $x \in \mathbb{R}^p$.

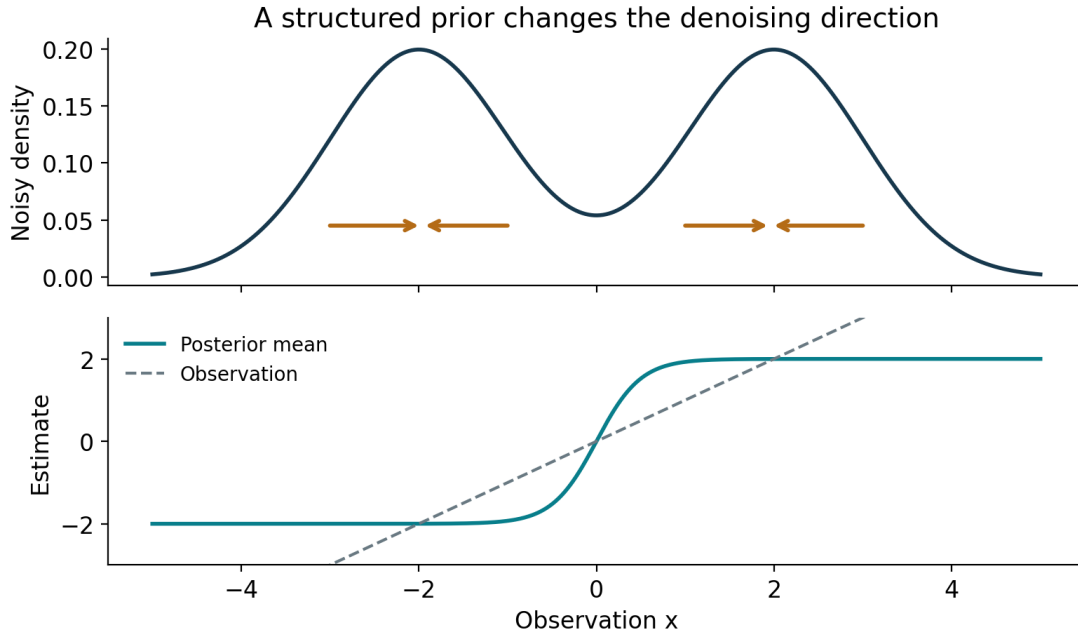


Figure 2: For a prior assigning probability $1/2$ to each of -2 and 2 , and unit Gaussian noise, the posterior mean is $2 \tanh(2x)$. The score correction can move an observation away from zero toward either mode.

Theorem 3. Let $f \in C^2(\mathbb{R}^p)$ take positive values and suppose $\sqrt{f}$ 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

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

Exercise 7 supplies the full risk calculation.

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 for $x \neq 0$

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

Thus, as long as $m \in (0, p-2]$, $\sqrt{f}$ is superharmonic away from zero¹. For this function, using the regularization $\|x\|^2 \mapsto a + \|x\|^2$, followed by $a \downarrow 0$, justifies the risk formula when $p \geq 3$, because $\mathbb{E} \|X\|^{-2} < \infty$. It recovers the JS estimator when $m = (p-2)/2$.

¹In particular, for $m = p-2$ it is harmonic away from the origin.

4 A modern connection: diffusion model

Low priority *A modern connection: diffusion model*

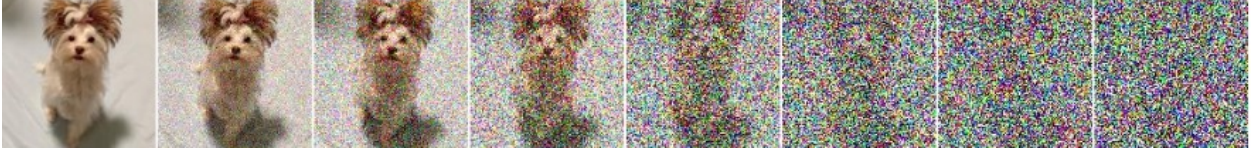


Figure 3: Increasing Gaussian noise progressively hides the image from left to right. A generative model learns to move from noise back toward data; this row illustrates the forward corruption process, not generated samples. Image: Yang Song (2021), unchanged.

How time flies! Now we have tech companies running gigantic models. Figure 3 shows the idea of generating new images using a diffusion model. Let x_0 represent a randomly sampled image (vectorized pixel values).

- **Forward process.** At each fixed time, with prescribed marginal scales (α_t) and (β_t) , progressively adding noise to an image x_0 :

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

- **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,

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

where $z_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$. These step parameters determine the marginal scales α_t, β_t above; they are not interchangeable with them. The first schedule requires nondecreasing σ_t^2 ; in the second schedule take $0 < \sigma_t^2 < 1$ so the following inverse scaling is defined. 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 estimating it from data gives $s_\theta(x_t, t)$. The following update is a schematic small-step denoising sampler, not an exact finite-step conditional distribution. In general the reverse transition is not Gaussian with covariance $\sigma_t^2 I$; a chosen diffusion discretization determines its noise scale. It uses the score as follows: starting with $\bar{X}_T \sim p_T$, approximated by a simple Gaussian when the terminal noise dominates the signal,

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

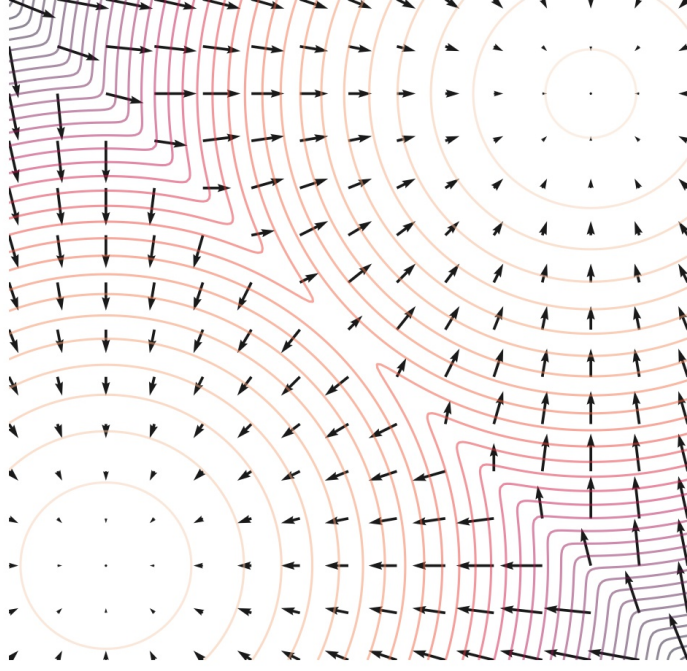


Figure 4: The score $\nabla \log p(x)$ points toward locally increasing density; contours and arrows show a two-Gaussian mixture. The denoiser uses an estimated version of this vector field at each noise level. Figure: Yang Song (2021), unchanged. The score here differentiates with respect to the observation x , not a model parameter.

The score estimate $s_\theta(x_t, t)$ is obtained from a neural network. The ideal objective compares it with $\nabla \log p_t$, but that target is unknown. For the known Gaussian forward marginal $X_t = \alpha_t X_0 + \beta_t Z$, $\beta_t > 0$, a usable denoising objective is

$$\min_{\theta} \mathbb{E} \left[w(t) \left\| s_\theta(X_t, t) + \frac{X_t - \alpha_t X_0}{\beta_t^2} \right\|^2 \right].$$

The target is the conditional score of $X_t | X_0$. Conditioning on X_t gives its mean $\nabla \log p_t(X_t)$; hence least-squares projection shows that this objective differs from weighted marginal-score error only by a constant independent of the network. Sample a training image, a positive noise time and Gaussian noise, then optimize this observable loss. The weighting $w(t) > 0$ and time range must make the expectations finite. See [2] for a specified sampler and schedule.

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 $X_0 \sim p_0$ and $(W_t)_{t \geq 0}$ is the Wiener process (Brownian motion).² Then,

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

Here t runs from T down to 0, so dt is negative, and $\bar{W}$ is Brownian noise in the reverse-time parameterization. Start with $\bar{X}_T \sim p_T$. Under the smoothness, nondegeneracy and integrability hypotheses of the time-reversal

²Brownian motion is a continuous Gaussian process with independent increments: over a time interval of length h , its increment has mean zero and covariance hI . The notation dW_t describes this random forcing on an infinitesimal time scale.

theorem, this reverse process has marginals p_t at the corresponding times. Equivalently, with increasing time $s = T - t$, the drift is $-f(x, T - s) + g(T - s)^2 \nabla \log p_{T-s}(x)$. The sign convention and time direction are part of the formula [1, 2].

Low priority *Practice: Tweedie with general Gaussian noise*

Exercise prompt. If $Y = aU + \sigma Z$ with $a > 0$, $\sigma > 0$, $\mathbb{E} \|U\|^2 < \infty$, and standard Gaussian Z independent of U , derive $\mathbb{E}[U \mid Y = y]$ from the density p of Y .

See Exercise 4 in Section 7 for the full statement and complete solution.

5 Further reading

Further reading *Further reading*

For an accessible account of Stein risk calculations, see Tibshirani's Stein notes (with Larry Wasserman). For the posterior-mean identity, see Efron (2011), Equation (1.4). The primary references are [3, 4]. The references for diffusion models are [1, 2].

Key ideas

- Fixed shrinkage trades variance against bias; its best amount depends on the unknown mean.
- James–Stein uses the observed norm to adapt the shrinkage and improves total risk when the dimension is at least three.
- Tweedie's formula links the posterior mean to the gradient of a noisy marginal density.
- The optional diffusion connection uses such scores at several noise levels.

6 Detailed reference

High priority *Bias can lower risk*

For $X \sim N(\mu, I_p)$, constant shrinkage aX has risk $a^2 p + (1 - a)^2 \|\mu\|^2$. Reduced variance can outweigh squared bias, but a fixed a below one cannot dominate X over all means. Return to the earlier explanation.

High priority *Stein risk identity*

For a sufficiently regular vector function h and $X \sim N(\mu, I_p)$, the risk of $X + h(X)$ is $p + \mathbb{E}[2 \operatorname{div} h(X) + \|h(X)\|^2]$. The derivative and integrability conditions matter at the shrinkage singularity. Return to the earlier explanation.

Mid priority *Posterior means and marginal scores*

In Gaussian noise, the posterior mean equals the observation plus the noise variance times the gradient of the log marginal density. This score is a data-variable derivative, distinct from a parameter score in likelihood theory. Return to the earlier explanation.

Looking ahead

Lecture 19 asks how to improve shrinkage when the signal is sparse in a useful coordinate system. Thresholding adapts coordinate by coordinate rather than using one radial factor. Continue within optional Part VI.

7 Exercises and complete solutions

High priority *When does fixed shrinkage beat the data?*

Exercise 1 (When does fixed shrinkage beat the data?).

For $T_a(X) = aX$, $0 < a < 1$, find the set of means for which its squared-error risk is smaller than p .

Solution.

The risk is $(1 - a)^2 \|\mu\|^2 + a^2 p$. It is less than p exactly when $(1 - a)^2 \|\mu\|^2 < (1 - a^2)p$, or $\|\mu\|^2 < p(1 + a)/(1 - a)$. No fixed $a < 1$ improves every mean; this motivates a data-dependent shrinkage factor. $\square$

High priority *Translate the shrinkage center*

Exercise 2 (Translate the shrinkage center).

Explain rigorously why the James–Stein risk calculation may assume $c = 0$ for a fixed center c .

Solution.

Set $Y = X - c \sim N(\mu - c, I_p)$. The centered rule has form $c + \delta(Y)$, and its error is $\delta(Y) - (\mu - c)$. Thus its risk at μ equals the zero-center risk at $\mu - c$. A bound valid for every mean transfers to every fixed center. $\square$

Mid priority *Tweedie with a Gaussian prior*

Exercise 3 (Tweedie with a Gaussian prior).

If $\mu \sim N(0, \tau^2 I_p)$ and $X | \mu \sim N(\mu, I_p)$, use Tweedie to find the posterior mean.

Solution.

The marginal law is $N(0, (1 + \tau^2)I_p)$, so $\nabla \log f_X(x) = -x/(1 + \tau^2)$. Adding this to x gives $\tau^2 x/(1 + \tau^2)$. At $\tau^2 = 0$ the limiting prior is a point mass at zero and the estimate is zero. $\square$

Low priority *Tweedie with general Gaussian noise*

Exercise 4 (Tweedie with general Gaussian noise).

If $Y = aU + \sigma Z$ with $a > 0$, $\sigma > 0$, $\mathbb{E}\|U\|^2 < \infty$, and standard Gaussian Z independent of U , derive $\mathbb{E}[U | Y = y]$ from the density p of Y .

Solution.

Differentiating the Gaussian mixture gives $\nabla p(y) = \sigma^{-2} p(y) \{a \mathbb{E}[U | Y = y] - y\}$. Since the Gaussian convolution has positive density, divide by p and solve: $\mathbb{E}[U | Y = y] = [y + \sigma^2 \nabla \log p(y)]/a$. Differentiation is justified by Gaussian smoothing, with the relevant posterior moments finite. $\square$

High priority *From coordinate Stein identities to divergence*

Exercise 5 (From coordinate Stein identities to divergence).

Prove $\mathbb{E}[Z^\top h(Z)] = \mathbb{E}[\text{div } h(Z)]$ for a vector field satisfying the coordinate Stein assumptions.

Solution.

Apply the scalar identity to h_j in coordinate j : $\mathbb{E}[Z_j h_j(Z)] = \mathbb{E}[\partial_j h_j(Z)]$. Sum the p identities; the finite sum and the assumed integrability permit taking expectations termwise. This yields the displayed divergence identity. □

High priority *Risk at the shrinkage center*

Exercise 6 (Risk at the shrinkage center).

For the ordinary James–Stein rule with $r = 1$ and $\mu = c$, compute the risk exactly for $p > 2$.

Solution.

Here $Q = \|X - c\|^2 \sim \chi_p^2$. Its Gamma density gives $\mathbb{E}Q^{-1} = \Gamma(p/2 - 1)/[2\Gamma(p/2)] = 1/(p - 2)$. The risk formula becomes $p - (p - 2)^2/(p - 2) = 2$, compared with the unshrunk risk p . □

Low priority *The superharmonic risk identity*

Exercise 7 (The superharmonic risk identity).

Using Stein, prove $R_{X + \nabla \log f}(\mu) = p + 4\mathbb{E}[\Delta\sqrt{f}(X)/\sqrt{f}(X)]$ under the stated assumptions.

Solution.

For $h = \nabla \log f$, expand the risk: $p + \mathbb{E}\|h\|^2 + 2\mathbb{E}\text{div } h$. The integrand is $\|\nabla f\|^2/f^2 + 2\Delta f/f - 2\|\nabla f\|^2/f^2 = 2\Delta f/f - \|\nabla f\|^2/f^2$. Differentiating $\sqrt{f}$ gives four times $\Delta\sqrt{f}/\sqrt{f}$ equal to this expression. Superharmonicity makes it nonpositive. □

Low priority *Why constant Normal risk is minimax*

Exercise 8 (Why constant Normal risk is minimax).

Use priors $\mu \sim N(0, \tau^2 I_p)$ to show that the minimax risk for estimating a unit-noise Normal mean is p .

Solution.

For every rule, its supremum risk is at least its Bayes risk under any prior. The Gaussian posterior covariance is $\tau^2/(1 + \tau^2)I_p$, so the optimal Bayes risk is $p\tau^2/(1 + \tau^2)$. Letting $\tau^2 \rightarrow \infty$ yields the lower bound p . The rule X has constant risk p , so it attains the bound. □

What changed

October 19, 2026: Corrected notation, wording and cross-references after the course-wide errata review; refreshed the linked HTML and PDF editions.

October 16, 2026: Clarified risk reduction versus error at an individual observation, justified the Bernoulli minimax example, and distinguished a Tweedie-form estimator from a genuine posterior mean. Added optional-part guidance, a backward recap and concise reading summaries. October 17, 2026: Published in optional Part VI with lecture links, student downloads and issue reporting; this material is not examined.

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.