The Gaussian Distribution from Scratch

Karl Stratos

1 Definitions

Let $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}_{\succ 0}$ where we assert $\Sigma \succ 0$ (i.e., positive-definite) to avoid handling degenerate cases. We define the **Gaussian distribution** $\mathcal{N}(\mu, \Sigma) : \mathbb{R}^d \to [0, 1]$ as

$$\mathcal{N}(\mu, \Sigma)(x) := \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) > 0 \tag{1}$$

which integrates to 1 over $\mathbb{R}^d$ (Lemma E.4). The following statements about a random variable $X \in \mathbb{R}^d$ are equivalent (Lemma E.12):

1. $X \sim \mathcal{N}(\mu, \Sigma)$. That is, the probability of $X = x$ is $\mathcal{N}(\mu, \Sigma)(x)$ defined in (1).
2. The moment-generating function is $M_X(t) := \mathbf{E}[\exp(t^\top X)] = \exp(t^\top \mu + \frac{1}{2} t^\top \Sigma t)$ for all $t \in \mathbb{R}^d$.
3. $X = \mu + \Sigma^{1/2} Z$ where $Z \sim \mathcal{N}(0_d, I_{d \times d})$.
4. $a^\top X \sim \mathcal{N}(a^\top \mu, a^\top \Sigma a)$ for all nonzero $a \in \mathbb{R}^d$.[1]

If any holds, we say $X \in \mathbb{R}^d$ is **normally distributed** with parameters (μ, Σ). Note that 3 and 4 just reduce general normality to simpler forms of (1) (standard and univariate). These alternative definitions are useful in different contexts, for instance

- By 2, a point-mass distribution on $x \in \mathbb{R}^d$ is "normal" with parameters $(x, 0_{d \times d})$ since its MGF is $\mathbf{E}[\exp(t^\top X)] = \exp(t^\top x)$.
- 3 is exactly the reparameterization trick in GANs where we view X sampled from a Gaussian distribution as a differentiable perturbation of model parameters (μ, Σ).
- By 4, we can show that two random variables Y and Z are jointly normal (Section 2) by showing that any scalar projection of (Y, Z) using a nonzero vector is (univariate) normal.

Linear transformation. A critical property of the Gaussian distribution is that it is closed under linear transformation. Note that definitions 3 and 4 are consistent with this property. For any $A \in \mathbb{R}^{d' \times d}$ and $b \in \mathbb{R}^{d'}$ where A is full-rank with $d' \leq d$ (so that $A\Sigma A^\top \succ 0$), $X \sim \mathcal{N}(\mu, \Sigma)$ implies (Lemma C.2):

$$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top) \tag{2}$$

Sample mean and covariance. Another characteristic of the Gaussian distribution is that the sample mean and covariance are independent. For any iid $X_1 \dots X_N \sim$ **Unk** with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}_{\succ 0}$, unbiased estimators of the mean and covariance are given by

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i \qquad\qquad \bar{S}^2_N = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X}_N)(X_i - \bar{X}_N)^\top$$

It turns out that $\bar{X}_N$ and $\bar{S}^2_N$ are independent iff **Unk** is normal (Geary, 1936). In fact, if **Unk** is normal, then $\bar{X}_N \sim \mathcal{N}(\mu, (1/N)\Sigma)$ and, independently, $(N-1)\bar{S}^2_N \sim \mathcal{W}_d(N-1, \Sigma)$ where $\mathcal{W}_d$ is known as the Wishart distribution (proof).[2] If $d = 1$ and $\Sigma = \sigma^2 > 0$, this implies the better known form $(N-1)/\sigma^2 \bar{S}^2_N \sim \chi^2(N-1)$ where $\chi^2(k)$ is the chi-square distribution with k degrees of freedom.

[1] The mean and variance of $a^\top X$ are always $a^\top \mu$ and $a^\top \Sigma a$, so this is simply saying that $a^\top X$ has the distribution (1) with $d = 1$.

[2] Specifically, $\mathcal{W}_d(k, \Sigma)$ is the distribution over $(u_1 \dots u_k)^\top (u_1 \dots u_k) \in \mathbb{R}^{d \times d}$ where $u_1 \dots u_k \in \mathbb{R}^d$ are iid samples from $\mathcal{N}(0_d, \Sigma)$.

2 Joint Distribution

We say $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ are **jointly normally distributed** with parameters (μ, Σ) if the concatenation (X, Y) follows $\mathcal{N}(\mu, \Sigma)$. More explicitly,

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}\right)$$

where $\mu_X \in \mathbb{R}^d$, $\mu_Y \in \mathbb{R}^{d'}$, $\Sigma_X \in \mathbb{R}^{d\times d}_{\succ 0}$, $\Sigma_Y \in \mathbb{R}^{d'\times d'}_{\succ 0}$, $\Sigma_{XY} \in \mathbb{R}^{d\times d'}$, and $\Sigma_{YX} = \Sigma_{XY}^\top$.[3] A subtle fact is that X and Y can be individual normal but not jointly normal (Appendix F), so we must explicitly establish joint normality even for normal variables (e.g., by using 4). However, if X and Y are *independently* normal, then they are jointly normal since we can write

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_X & 0_{d\times d'} \\ 0_{d'\times d} & \Sigma_Y \end{bmatrix}\right)$$

On the other hand, if X, Y are jointly normal *and* uncorrelated, then they are independent. This follows from the form of conditional distribution (4):

$$\Sigma_{XY} = 0_{d\times d'} \quad \Rightarrow \quad \mathcal{N}(\mu, \Sigma)(y|x) = \mathcal{N}(\mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X), \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})(y) = \mathcal{N}(\mu_Y, \Sigma_Y)(y)$$

The following lemma is an application of this fact.

Lemma 2.1. Let $X \sim \mathcal{N}(\mu, \Sigma)$. For any $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{m\times d}$,

$$A\Sigma B^\top = 0_{n\times m} \qquad \Leftrightarrow \qquad AX \in \mathbb{R}^n \text{ and } BX \in \mathbb{R}^m \text{ are independent}$$

Proof. If A or B is zero then the statement is trivially true (a constant is independent by definition). Otherwise, for all nonzero $(u, v) \in \mathbb{R}^{n+m}$, $(u, v)^\top(AX, BX) = (u^\top A + v^\top B)X$ is normal by the closure under linear transformation (2). Thus (AX, BX) is normal by 4. Hence AX and BX are independent iff they are uncorrelated: $\mathbf{E}\left[A(X - \mu)(X - \mu)^\top B^\top\right] = A\Sigma B^\top = 0_{n\times m}$. □

As a reference we give a related theorem about the independence of quadratic forms under normal distributions attributed to Allen T. Craig. Despite its striking similarity to Lemma 2.1, it is difficult to prove and has a long and complicated history (Driscoll and Gundberg Jr, 1986).

Theorem 2.2 (Craig's theorem)**.** Let $X \sim \mathcal{N}(\mu, \Sigma)$. For any $A, B \in \mathbb{R}^{d\times d}$,

$$A\Sigma B = 0_{d\times d} \qquad \Leftrightarrow \qquad X^\top AX \in \mathbb{R} \text{ and } X^\top BX \in \mathbb{R} \text{ are independent}$$

A final remark: recall that uncorrelatedness generally does not imply independence, so we must show joint normality before claiming independence from uncorrelatedness. For instance, Appendix F gives $X, Y \in \mathbb{R}$ that are individually normal (but not jointly normal) and uncorrelated but not independent.

2.1 Linear Combinations

Let $A \in \mathbb{R}^{p\times d}$, $B \in \mathbb{R}^{p\times d'}$, and $b \in \mathbb{R}^p$ where A, B are full-rank with $p \le \min(d, d')$. If $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ are jointly normal with parameters (μ, Σ), we have from (2) that

$$AX + BY + b \sim \mathcal{N}(A\mu_X + B\mu_Y + b, A\Sigma_X A^\top + A\Sigma_{XY}B^\top + B\Sigma_{YX}A^\top + B\Sigma_Y B^\top) \tag{3}$$

In particular, if X and Y are independently normal, then their sum is normal:

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$$

Note that we need joint normality to guarantee the normality of a linear combination. In general a linear combination of normal variables may not be normal (e.g., (68)).

[3] $\Sigma_X, \Sigma_Y \succ 0$ since they are main-diagonal blocks of $\Sigma \succ 0$ (Lemma E.9) and $\Sigma_{XY} = \Sigma_{YX}^\top$ since Σ is symmetric.

2.2 Conditional Distribution

If $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ are jointly normal with parameters (μ, Σ), and if $\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ is invertible, then for all $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^{d'}$ (Lemma E.10):

$$\begin{aligned}\mathcal{N}(\mu, \Sigma)((x, y)) &= \mathcal{N}(\mu_X, \Sigma_X)(x) \\ &\quad \times \mathcal{N}(\mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X), \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})(y)\end{aligned} \tag{4}$$

Therefore the conditional distribution over $Y|X = x$ is also Gaussian. This is useful for applications like Kalman filtering.

3 Entropy

Let $\mu' \in \mathbb{R}^d$ and $\Sigma' \in \mathbb{R}^{d\times d}_{\succ 0}$ be parameters of an additional Gaussian distribution over $\mathbb{R}^d$. Then (Lemma E.6):

$$H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) = \frac{1}{2}(\mu' - \mu)^\top \Sigma^{-1}(\mu' - \mu) + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma'\right) + \frac{1}{2}\log((2\pi)^d \det(\Sigma))$$

It follows that

$$\begin{aligned}H(\mathcal{N}(\mu, \Sigma)) &= \frac{1}{2}\log\left((2\pi e)^d \det(\Sigma)\right) \\ D_{\mathrm{KL}}(\mathcal{N}(\mu', \Sigma')||\mathcal{N}(\mu, \Sigma)) &= \frac{1}{2}(\mu' - \mu)^\top \Sigma^{-1}(\mu' - \mu) + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma' - I_{d\times d}\right) + \frac{1}{2}\log\left(\frac{\det(\Sigma)}{\det(\Sigma')}\right)\end{aligned}$$

$\mathcal{N}(\mu, \Sigma)$ has the largest entropy among all distributions over $\mathbb{R}^d$ with mean μ and covariance Σ (Theorem B.1). This is mainly because it standardizes x inside the exponential function.

3.1 Mutual Information

Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ be jointly normal with parameters (μ, Σ). If $\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ is invertible, then (Lemma E.11):

$$\begin{aligned}H(Y|X) &= \frac{1}{2}\log\left((2\pi e)^{d'} \det(\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})\right) \\ I(X, Y) &= \frac{1}{2}\log\left(\frac{\det(\Sigma_X)\det(\Sigma_Y)}{\det(\Sigma)}\right)\end{aligned} \tag{5}$$

Note that $I(X, Y)$ is infinite if $Y = X$. By the noisy-channel coding theorem, mutual information is the capacity (highest information rate that can be achieved nearly error-free) of a communication channel between X and Y. Below we give some well-known models with controllable mutual information.

Additive white Gaussian noise channel. Let $X \sim \mathcal{N}(0, \sigma^2)$ and $Z \sim \mathcal{N}(0, \nu^2)$ independently, and define $Y = X + Z$. X and Y are jointly normal because $a_1X + a_2Y = (a_1 + a_2)X + a_2Z$ is a sum of independently normal variables and thus normal for all nonzero $a = (a_1, a_2)$ (definition 4). Since $\mathrm{Var}(Y) = \sigma^2 + \nu^2$ and $\mathrm{Cov}(X, Y) = \sigma^2$,

$$\begin{bmatrix}X \\ Y\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}\sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \nu^2\end{bmatrix}\right) \qquad \Rightarrow \qquad I(X, Y) = \frac{1}{2}\log\left(1 + \frac{\sigma^2}{\nu^2}\right)$$

Thus $I(X, X + Z)$ grows logarithmically in signal-to-noise ratio $\frac{\sigma^2}{\nu^2}$.

Correlated standard normal channel. Let $X, Y \in \mathbb{R}$ be jointly standard normal with correlation $\rho < 1$. One way to construct them is to let $X, Z \overset{iid}{\sim} \mathcal{N}(0, 1)$ and set $Y = \rho X + \sqrt{1 - \rho^2}Z$. Then

$$\begin{bmatrix}X \\ Y\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}\right) \qquad \Rightarrow \qquad I(X, Y) = -\frac{1}{2}\log\left(1 - \rho^2\right)$$

By taking the correlation $\rho \to 1$ we can arbitrarily increase $I(X, Y)$.

4 Central Limit Theorem

Let $\mathbf{Unk}(\mu, \sigma^2)$ denote an unknown distribution over $\mathbb{R}$ with mean μ and variance $\sigma^2 > 0$. It is often of interest to consider the sample average $\bar{X}_N$ defined as

$$X_1 \dots X_N \overset{\text{iid}}{\sim} \mathbf{Unk}(\mu, \sigma^2) \qquad\qquad \bar{X}_N := \frac{1}{N} \sum_{i=1}^{N} X_i$$

The average is itself random: every time we draw N iid samples from $\mathbf{Unk}(\mu, \sigma^2)$, we draw a single sample of $\bar{X}_N$. We can easily verify that $\mathbf{E}[\bar{X}_N] = \mu$ and $\text{Var}\left(\bar{X}_N\right) = \frac{\sigma^2}{N}$, which states that $\bar{X}_N$ concentrates around μ as $N \to \infty$ (this is called the "law of large numbers"). But what is the distribution of $\bar{X}_N$? The **central limit theorem** (CLT) states that $\bar{X}_N$ is asympotically normal. More precisely, as $N \to \infty$ we have

$$\sqrt{N}\left(\bar{X}_N - \mu\right) \overset{\text{approx.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{6}$$

or, using the closure under linear transformation,

$$\bar{X}_N \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right) \tag{7}$$

which is consistent with but not implied by the law of large numbers. CLT allows us to make probabilistic statements about sample averages regardless of the underlying distribution. For instance, if $X_1 \dots X_N$ are arbitrary iid samples with mean 42 and variance 7, then approximately $\bar{X}_N \sim \mathcal{N}(42, \frac{7}{N})$ so that we can calculate quantities like $\Pr(\bar{X}_N \leq 50)$ (e.g., by consulting a standard normal table).

A proof of CLT shows that the KL divergence between the distribution of $\sqrt{N}\left(\bar{X}_N - \mu\right)$ and $\mathcal{N}(0, \sigma^2)$ goes to zero as $N \to \infty$. It is nontrivial: we refer to Marsh (2013) for details. CLT generalizes naturally to multivariate. If $\mathbf{Unk}(\mu, \Sigma)$ is an unknown distribution over $\mathbb{R}^d$ with mean μ and covariance $\Sigma \succ 0$, then the average $\bar{X}_N$ of samples $X_1 \dots X_N \overset{\text{iid}}{\sim} \mathbf{Unk}(\mu, \Sigma)$ satisfies as $N \to \infty$:

$$\sqrt{N}\left(\bar{X}_N - \mu\right) \overset{\text{approx.}}{\sim} \mathcal{N}(0_d, \Sigma) \tag{8}$$

$$\bar{X}_N \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu, \frac{1}{N}\Sigma\right) \tag{9}$$

5 Exponential Family

The Gaussian distribution is an exponential family (Appendix D), with one parameterization (Lemma E.19)

$$\mathcal{N}(\mu, \Sigma)(x) = \underbrace{\frac{1}{(\sqrt{2\pi})^d}}_{h(x)} \exp\left(\underbrace{\begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\text{vec}\left(\Sigma^{-1}\right) \end{bmatrix}}_{\theta}^{\top} \underbrace{\begin{bmatrix} x \\ \text{vec}\left(xx^\top\right) \end{bmatrix}}_{\tau(x)} - \underbrace{\frac{1}{2}\left(\mu^\top \Sigma^{-1}\mu + \log\left(\det(\Sigma)\right)\right)}_{A_{h,\tau}(\theta)} \right) \tag{10}$$

where $h(x) \geq 0$ is the base measure, $\theta \in \mathbb{R}^{d(d+1)}$ is the natural parameter, $\tau(x) \in \mathbb{R}^{d(d+1)}$ is the sufficient statistic, and $A_{h,\tau}(\theta) = \log\left(\int_{x \in \mathbb{R}^d} h(x) \exp\left(\theta^\top \tau(x)\right) dx\right)$ is the log-partition function. Thus Gaussian distributions inherit the usual properties of an exponential family such as the concavity of the likelihood function and the availability of conjugate priors.

6 Exponential Tilting

The Gaussian distribution is closed under exponential tilting (Lemma E.22):

$$\Pr(X_t = x) \propto e^{t^\top x} \times \mathcal{N}(\mu, \Sigma)(x) \qquad \Rightarrow \qquad X_t \sim \mathcal{N}(\mu + \Sigma t, \Sigma) \tag{11}$$

7 Cumulant-Generating Function

The cumulant-generating function $\psi_X(t) := \log \mathbf{E}[e^{t^\top X}]$ of $X \sim \mathcal{N}(\mu, \Sigma)$ and the first two cumulants are

$$\begin{aligned}\psi_X(t) &= t^\top \mu + \frac{1}{2} t^\top \Sigma t \\ \nabla \psi_X(t) &= \mu + \Sigma t \\ \nabla^2 \psi_X(t) &= \Sigma\end{aligned}$$

The corresponding Legendre transform $\psi_X^*(t) := \sup_{\lambda \in \mathbb{R}^d} \lambda^\top t - \psi_X(\lambda)$ of ψ_X is (Lemma E.24)

$$\psi_X^*(t) = \frac{1}{2}(t-\mu)^\top \Sigma^{-1}(t-\mu) \tag{12}$$

8 Sub-Gaussian Distribution

A random scalar $S \in \mathbb{R}$ with $\mathbf{E}[S] = 0$ is **sub-Gaussian with variance factor** σ^2, denoted by $S \sim \mathcal{G}(\sigma^2)$, if

$$\psi_S(t) \le \psi_{Z \sim \mathcal{N}(0, \sigma^2)}(t) = \frac{\sigma^2 t^2}{2} \tag{13}$$

for all $t \in \mathbb{R}$. It is stable in the following sense:

1. $\mathrm{Var}(S) \le \sigma^2$ (Lemma E.25).
2. $-S \sim \mathcal{G}(\sigma^2)$. This can be seen by noting that $\psi_{-S}(t) = \psi_S(-t)$.
3. $\Pr(S \ge \epsilon) \le \exp(-\frac{\epsilon^2}{2\sigma^2})$ for all $\epsilon \ge 0$. Use Chernoff's inequality (E.15) with Lemma E.26 and (12).
4. If $S_1 \dots S_N$ are independent with $S_i \sim \mathcal{G}(\sigma_i^2)$, then $\sum_{i=1}^N S_i \sim \mathcal{G}(\sum_{i=1}^N \sigma_i^2)$.

Combining these properties, we have (Lemma E.27)

$$S_i \sim \mathcal{G}(\sigma_i^2) \text{ independently} \quad \Rightarrow \quad \Pr\left(\left|\frac{1}{N}\sum_{i=1}^N S_i\right| \ge \epsilon\right) \le 2\exp\left(-\frac{N^2\epsilon^2}{2\left(\sum_{i=1}^N \sigma_i^2\right)}\right) \tag{14}$$

An important class of sub-Gaussian variables is bounded scalars: if $X \in [a,b]$ then $X - \mathbf{E}[X] \sim \mathcal{G}(\frac{(b-a)^2}{4})$ (Hoeffding's lemma, E.23). This yields the following popular tail inequality.

Corollary 8.1 (Hoeffding's inequality)**.** If $X_1 \dots X_N \in [a,b]$ are iid with mean $\mu = \mathbf{E}[X_i] \in \mathbb{R}$,

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right| \ge \epsilon\right) \le 2\exp\left(-\frac{2N\epsilon^2}{(b-a)^2}\right) \tag{15}$$

Proof. By Hoeffding's lemma, $X_i - \mu \sim \mathcal{G}(\frac{(b-a)^2}{4})$. We get the statement by plugging $\sigma_i^2 = \frac{(b-a)^2}{4}$ in (14). □

9 TODO: High-Dimensional Behavior

10 TODO: Gaussian Process

References

Driscoll, M. F. and Gundberg Jr, W. R. (1986). A history of the development of craig's theorem. *The American Statistician*, **40**(1), 65–70.

Geary, R. (1936). The distribution of" student's" ratio for non-normal samples. *Supplement to the Journal of the Royal Statistical Society*, **3**(2), 178–184.

Marsh, C. (2013). Introduction to continuous entropy. *Department of Computer Science, Princeton University*.

A Integration for Dummies

A.1 Single-Variable

An **antiderivative** of $f : \mathbb{R} \to \mathbb{R}$ is a function $F : \mathbb{R} \to \mathbb{R}$ such that $F' = f$. If F is an antiderivative, then so is $F + C$ for any constant $C \in \mathbb{R}$. For instance, $(1/3)x^3 + 42$ is an antiderivative of x^2.

The (definite) **integral** of $f : \mathbb{R} \to \mathbb{R}$ over $[a, b]$ is a scalar $\int_a^b f(x)dx \in \mathbb{R}$ that represents the signed area of f on $[a, b]$. The quantity $f(x)dx$ is interpreted as the product of the function value and an infinitesimally small interval. There are different ways to formalize the area. The most common definition is the Riemannn integral which partitions $[a, b]$ into intervals $[i\delta, (i + 1)\delta]$ of width $\delta > 0$ and define

$$\int_a^b f(x)dx := \lim_{\delta \to 0} \sum_i f(x_i^\delta)\delta \tag{16}$$

where $x_i^\delta \in [i\delta, (i + 1)\delta]$. The finite sum $\sum_i f(x_i^\delta)\delta$ for a given width δ is called a **Riemann sum**. Thus an integral is simply the limiting value of a Riemann sum (if it exists it is unique). A more general definition is a Lebesque integral which partitions the range of f.

The **fundamental theorem of calculus** (FTC) allows us to evaluate integrals by antiderivatives: if F is any antiderivative of f, then

$$\int_a^b f(x)dx = F(x)\Big|_a^b := F(b) - F(a) \tag{17}$$

For instance, the signed area under x^2 over $[-1, 1]$ is $2/3$. Basic properties of integration include

$$\int_a^b \alpha f(x) + \beta g(x)dx = \alpha \int_a^b f(x)dx + \beta \int_a^b g(x)dx \qquad \text{(linearity)} \tag{18}$$

$$\int_a^b f(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f(u)du \qquad \text{(u-substitution)} \tag{19}$$

$$\int_a^b f(x)G(x)dx = F(x)G(x)\Big|_a^b - \int_a^b F(x)g(x)dx \qquad \text{(integration by parts)} \tag{20}$$

(Exercise: verify (19–20) using the chain rule and the product rule in differentiation.)

A.1.1 Substitution in practice

While (19) is the standard form of u-substitution, we often use it mechanically as follows. We wish to integrate f over the interval $a < b$. We view f as a (hopefully simpler) function of $u = g(x)$ where $g : \mathbb{R} \to \mathbb{R}$ is invertible and differentiable with nonzero derivative over (a, b). The infinitesimals are related as $du = g'(x)dx$ by the chain rule, or equivalently $dx = g'(g^{-1}(u))^{-1}du$. This yields a "plug-in" version of (19) where we substitute $g(x) = u$ and $dx = g'(g^{-1}(u))^{-1}du$,

$$\int_a^b f(g(x))dx = \int_{g(a)}^{g(b)} f(u)g'(g^{-1}(u))^{-1}du \tag{21}$$

For instance,

$$\begin{aligned}\int_0^{\sqrt{\frac{\pi}{2}}} 2x\cos\left(x^2\right)dx &= \int_0^{\frac{\pi}{2}} 2\sqrt{u}\cos(u)\left(\frac{1}{2\sqrt{u}}\right)du \\ &= \int_0^{\frac{\pi}{2}} \cos(u)\,du = \sin(u)\Big|_0^{\frac{\pi}{2}} = 1\end{aligned}$$

where $2x\cos\left(x^2\right) = 2\sqrt{u}\cos(u)$ with $u = g(x) = x^2$. Note that g is invertible on $(0, \sqrt{\frac{\pi}{2}})$ so that $x = \sqrt{u}$; it is also differentiable with nonzero derivative $g'(x) = 2x$. Writing $dx = (2\sqrt{u})^{-1}du$, we cancel terms and are finally able to use FTC (17).

Orientation of region. Observe that

$$1 = \int_0^1 1dx = \int_0^{-1} (-1)du = \int_{-1}^0 (+1)du$$

The first equality is by FTC. The second equality is by (21) with $f(x) = 1$ and $u = g(x) = -x$. The final equality is again by FTC, simply acknowledging that $(-x)|_0^{-1} = x|_{-1}^0 = 1$. More generally, when $g'(x)^{-1} < 0$ (i.e., u is moving in the opposite direction of x), we also change the "orientation of region" in integration (right-to-left instead of left-to-right). We can consider an alternative orientation-free formulation of u-substitution by always assuming integrating left-to-right. Let R denotes a region $a < b$, then

$$\int_R f(g(x))dx = \int_{g(R)} f(u) \left|g'(g^{-1}(u))^{-1}\right| du \tag{22}$$

where $g(R)$ is the output region of g when applied to R, integrated from a smaller value to a larger value. This formulation is useful because it generalizes to higher dimensions (24).

A.2 Multi-Variable

The integral of $f : \mathbb{R}^d \to \mathbb{R}$ over a region $R \subseteq \mathbb{R}^d$ is a scalar $\int_R f(x)dx \in \mathbb{R}$ that represents the signed hypervolume of f on R. Evaluation of such an integral is generally challenging because the region may take complicated forms (high-dimensional curves).

We can greatly simplify the problem by restricting the region to be a hypercube $R = [a, b]$ where $a, b \in \mathbb{R}^d$ specify a d-dimensional bounding box $[a_1, b_1] \times \cdots \times [a_d, b_d]$ (potentially all of $\mathbb{R}^d$). A central tool in this setting is **Fubini's thoerem**, which states that

$$\int_{[a,b]} f(x)dx = \int_{a_{\pi(d)}}^{b_{\pi(d)}} \left(\cdots \left(\int_{a_{\pi(1)}}^{b_{\pi(1)}} f(x_1 \dots x_d)dx_{\pi(1)} \right) \cdots \right) dx_{\pi(d)}$$

where π is any permutation of $\{1 \dots d\}$. Thus we can evaluate a multi-variable integral by iteratively evaluating a single-variable integral in any order.

Many properties of integration carry over (like linearity), but some need to be generalized. One important generalization is **multi-variable u-substitution**. Let $R \subseteq \mathbb{R}^d$ and $g : R \to \mathbb{R}^d$ such that $J_g(x) \in \mathbb{R}^{d \times d}$ (Jacobian of g) is nonzero for all $x \in R$. Then

$$\int_R f(g(x)) \left|\det(J_g(x))\right| dx = \int_{g(R)} f(u)du \tag{23}$$

Similar to the single-variable case, we often use substitution mechanically as follows. We integrate f over a region R by viewing it as a simpler function of $u = g(x)$ where $g : R \to \mathbb{R}^d$ is assumed to be invertible (i.e., $\det(J_g(x)) \neq 0$). The infinitesimals are related as $du = \left|\det(J_g(x))\right| dx$ or equivalently $dx = \left|\det(J_g(x))\right|^{-1} du$. This gives

$$\int_R f(g(x))dx = \int_{g(R)} f(u) \left|\det(J_g(g^{-1}(u)))\right|^{-1} du \tag{24}$$

where we "plug in" $g(x) = u$ and $dx = \left|\det(J_g(g^{-1}(u)))\right|^{-1} du$. This strictly generalizes (22).

A.2.1 Applications to probability

Let $X \in \mathbb{R}^d$ be a random vector with distribution p_X supported on $S \subseteq \mathbb{R}^d$ (i.e., $p_X(x) \geq 0$ and $\int_S p_X(x)dx = 1$). The probability that X lies in a region $R \subseteq S$ is

$$\Pr(X \in R) = \int_R p_X(x)dx$$

Let $t : S \to T$ be a smooth invertible function where $T \subseteq \mathbb{R}^d$. Define a new random vector $Y = t(X)$ supported on T. We claim that Y has the distribution

$$p_Y(y) = p_X(t^{-1}(y)) \left|\det(J_{t^{-1}}(y))\right| \qquad \forall y \in T \tag{25}$$

Equivalently,

$$p_Y(t(x)) = p_X(x)\,|\det(J_{t^{-1}}(t(x)))| \qquad \forall x \in S \tag{26}$$

Proof sketch. For any $R \subseteq T$,

$$\Pr(Y \in R) = \Pr(X \in t^{-1}(R)) = \int_{t^{-1}(R)} p_X(x)dx = \int_R p_X(t^{-1}(y))\,|\det(J_{t^{-1}}(y))|\,dy$$

where the last equality applies (23) with $g = t^{-1}$. This implies (25).

B Continuous Entropy and KL Divergence

We generalize results in Marsh (2013) to multivariate. The continuous/differential entropy of $X \in \mathbb{R}^d$ with density p_X supported on $S \subseteq \mathbb{R}^d$ is defined as[4]

$$H(X) := -\int_S p_X(x) \log p_X(x) dx \tag{27}$$

It is easily seen that entropy is additive for independent variables. That is, if $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ are independent then the entropy of $Z = (X, Y) \in \mathbb{R}^{d+d'}$ is $H(Z) = H(X) + H(Y)$.

- The uniform distribution $u_{[a,b]}(x) := \frac{1}{b-a}$ over $[a, b] \subset \mathbb{R}$ has entropy

$$H(X) = \int_a^b \frac{1}{b-a} \log(b-a) dx = \log(b-a) \tag{28}$$

- The Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ over $\mathbb{R}^d$ has entropy (Corollary E.7)

$$H(X) = \frac{1}{2} \log\left((2\pi e)^d \det(\Sigma)\right)$$

- The exponential distribution $e_\lambda(x) := \lambda \exp(-\lambda x)$ over $[0, \infty)$ with parameter $\lambda > 0$ has entropy (Lemma E.5)

$$H(X) = 1 - \log \lambda \tag{29}$$

Unfortunately, continuous entropy suffers from various shortcomings (reviewed in Section B.1), most notably negativity (e.g., (28) is negative if $b - a < 1$, (29) is negative if $\lambda > e$). On the other hand, let q_X be another density of X with support S. Define the continuous KL divergence (aka. relative entropy) between p_X and q_X as

$$D_{\mathrm{KL}}(p_X||q_X) := \int_S p_X(x) \log \frac{p_X(x)}{q_X(x)} dx \tag{30}$$

Continuous KL divergence is nonnegative:

$$\begin{aligned}
D_{\mathrm{KL}}(p_X||q_X) &= \mathop{\mathbf{E}}_{x\sim p_X}\left[\log \frac{p_X(x)}{q_X(x)}\right] \\
&= \mathop{\mathbf{E}}_{x\sim p_X}\left[-\log \frac{q_X(x)}{p_X(x)}\right] \\
&\geq -\log\left(\mathop{\mathbf{E}}_{x\sim p_X}\left[\frac{q_X(x)}{p_X(x)}\right]\right) && \text{(convexity of } -\log\text{)} \\
&= -\log\left(\int_S p_X(x)\frac{q_X(x)}{p_X(x)} dx\right) \\
&= -\log\left(\int_S q_X(x) dx\right) = 0
\end{aligned}$$

where $D_{\mathrm{KL}}(p_X||q_X) = 0$ iff $p_X = q_X$ almost everywhere. This has useful implications.

[4] We use the term "density" in this section to distingiush continuous vs discrete variables.

- The cross entropy between p_X and q_X upper bounds the entropy of p_X,

$$H(p_X, q_X) := H(p_X) + D_{\mathrm{KL}}(p_X || q_X) \geq H(p_X) \tag{31}$$

- Mutual information is nonnegative,

$$I(X, Y) := D_{\mathrm{KL}}(p_{XY} || p_X p_Y) \geq 0 \tag{32}$$

The cross entropy upper bound can be used to derive various maximum entropy densities.

Theorem B.1.

$$\mathcal{N}(\mu, \Sigma) \in \underset{p_X \colon \mathbf{E}[X]=\mu,\ \mathrm{Var}(X)=\Sigma}{\arg\max} H(p_X) \tag{33}$$

$$u_{[a,b]} \in \underset{p_X \colon \mathrm{Support}(p_X)=[a,b]}{\arg\max} H(p_X) \tag{34}$$

$$e_\lambda \in \underset{p_X \colon \mathrm{Support}(p_X)=\mathbb{R}^d_{\geq 0},\ \mathbf{E}[X]=\lambda^{-1}}{\arg\max} H(p_X) \tag{35}$$

where $u_{[a,b]}$ denotes the uniform distribution over $[a, b] \subset \mathbb{R}^d$ and e_λ denotes the product exponential density over $\mathbb{R}^d_{\geq 0}$ with $\lambda > 0_d$

Proof. (33): Let p_X with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \succ 0$. Then

$$\begin{aligned}
H(p_X, \mathcal{N}(\mu, \Sigma)) &= \int_{\mathbb{R}^d} p_X(x) \left(\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) + \frac{1}{2}\log((2\pi)^d \det(\Sigma)) \right) \\
&= \frac{1}{2} \mathop{\mathbf{E}}_{x \sim p_X} \left[(x-\mu)^\top \Sigma^{-1}(x-\mu) \right] + \frac{1}{2}\log((2\pi)^d \det(\Sigma)) \\
&= \frac{d}{2} + \frac{1}{2}\log((2\pi)^d \det(\Sigma)) \\
&= \frac{1}{2}\log((2\pi e)^d \det(\Sigma)) = H(\mathcal{N}(\mu, \Sigma)) \geq H(p_X)
\end{aligned}$$

(34): Assume $d = 1$. Given any p_X with support $[a, b]$ we have

$$H(p_X, u_X) = \int_a^b p_X(x) \log(b-a) dx = \log(b-a) = H(u_{[a,b]}) \geq H(p_X)$$

The statement holds for $d > 1$ since each dimension is independently optimized.

(35): Assume $d = 1$. Given any p_X with support $[0, \infty)$ and mean $\lambda^{-1} > 0$ we have

$$H(p_X, e_\lambda) = \int_0^\infty p_X(x)(\lambda x - \log\lambda) dx = \lambda \mathop{\mathbf{E}}_{x \sim p_X}[x] - \log\lambda = 1 - \log\lambda = H(e_\lambda) \geq H(p_X, e_\lambda)$$

The statement holds for $d > 1$ since each dimension is independently optimized. □

B.1 Shortcomings of Continuous Entropy

B.1.1 Inconsistency with Shannon entropy

The Shannon entropy of discrete $X \in \{x_1 \dots x_n\}$ with distribution p_X is

$$H(X) := -\sum_{i=1}^n p_X(x_i) \log p_X(x_i) \tag{36}$$

This definition was derived by Shannon as a solution that satisfies axioms of information (regarding monotonicity, non-negativity, zero information, and independence). (27) appears to be a natural continuous extension of (36) in the sense that both are $\mathbf{E}_{x \sim p_X}[-\log p_X(x)]$, but it fails to satisfy the axioms (e.g., it can be negative). One way to

better understand why is to show that (27) is inconsistent with (36) in the limit. Assume $d = 1$ and let p_X be a density supported on $[a, b]$. By definition

$$\int_a^b p_X(x)dx = \lim_{\delta \to 0} \sum_i p_X(x_i^\delta)\delta = 1 \tag{37}$$

where $\sum_i p_X(x_i^\delta)\delta$ is a finite Riemann sum of width $\delta > 0$. Thus we can cast the density p_X as an increasingly fine-grained discrete distribution with probabilities $p_X(x_i^\delta)\delta$ as $\delta \to 0$. Note that each value of $\delta > 0$ yields a discrete distribution with a well-defined Shannon entropy. This Shannon entropy, in the limit, is

$$\begin{aligned}
\lim_{\delta \to 0} \left(-\sum_i (p_X(x_i^\delta)\delta) \log(p_X(x_i^\delta)\delta) \right) &= -\lim_{\delta \to 0} \sum_i (p_X(x_i^\delta) \log p_X(x_i^\delta))\delta - \lim_{\delta \to 0} \sum_i p_X(x_i^\delta)\delta \log \delta \\
&= -\int_a^b p_X(x) \log p_X(x)dx - \lim_{\delta \to 0} \sum_i p_X(x_i^\delta)\delta \log \delta \\
&= H(X) - \left(\lim_{\delta \to 0} \sum_i p_X(x_i^\delta)\delta \right) \left(\lim_{\delta \to 0} \log \delta \right) \qquad (38) \\
&= H(X) + \infty \qquad (39)
\end{aligned}$$

where (38) follows from the generalized product rule of limits using (37).[5] So the limiting Shannon entropy diverges from the continuous entropy by an infinite offset.

B.1.2 Variability under change of coordinates

A good measure of information should not depend on the representation of samples from a distribution. For instance, let p_X be a distribution over finitely many circles, each of which can be specified by its radius or area. Clearly, the Shannon entropy of the circle is the same regardless of the representation. Now let p_X be a density over all circles. The continuous entropy of the circle under the radius representation is different from that under the area representation. A general statement that implies this result is given below.

Lemma B.2. Let $X \in \mathbb{R}^d$ with density p_X supported on S. For any invertible mapping t on S,

$$H(t(X)) = H(X) - \mathop{\mathbf{E}}_{x \sim p_X} [\log |\det(J_{t^{-1}}(t(x)))|]$$

Proof.

$$\begin{aligned}
H(t(X)) &= -\int_S p_X(x) \log p_X(t(x))dx \\
&= -\int_S p_X(x) \log p_X(x)dx - \int_S p_X(x) \log |\det(J_{t^{-1}}(t(x)))| \, dx \qquad \text{(by (26))} \\
&= H(X) - \mathop{\mathbf{E}}_{x \sim p_X} [\log |\det(J_{t^{-1}}(t(x)))|]
\end{aligned}$$

□

Corollary B.3. For any invertible $A \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^d$,

$$H(AX + b) = H(X) - \log \left|\det(A^{-1})\right| \tag{40}$$

Corollary B.4. For $\alpha > 0$,

$$H(\alpha X) = H(X) + d \log \alpha$$

[5] Assume $\lim_{x \to a} f(x) \neq 0$. If $g(x)$ does not oscillate around a,

$$\lim_{x \to a} f(x)g(x) = \lim_{x \to a} f(x) \lim_{y \to a} g(y)$$

If $g(x)$ oscillates around a, then so does $f(x)g(x)$.

Proof.

$$\begin{aligned} H(\alpha X) &= H(X) - \log\left|\det(\alpha^{-1} I_{d\times d})\right| && \text{(by (40))} \\ &= H(X) - \log\left|\alpha^{-d}\right| \\ &= H(X) - \log \alpha^{-d} && \text{(since } \alpha > 0\text{)} \\ &= H(X) + d\log\alpha \end{aligned}$$

□

Corollary B.4 states that we can vacuously increase the continuous entropy of $X \in \mathbb{R}^d$ to infinity by multiplying each value with a scalar α as we take $\alpha \to \infty$.

C Moment-Generating Function

Let $X \in \mathbb{R}^d$ denote a random vector with distribution p_X. The **moment-generating function (MGF)** of X is a real-valued positive mapping $M_X : \mathbb{R}^d \to (0, \infty)$ defined as

$$M_X(t) := \mathop{\mathbf{E}}_{x\sim p_X}\left[\exp\left(t^\top x\right)\right] \tag{41}$$

Not every distribution has a corresponding MGF (because (41) may diverge). But a classical result in probability theory is that an MGF uniquely determines a probability distribution. More formally, let $X, Y \in \mathbb{R}^d$ be random vectors with distributions p_X, p_Y with well-defined MGFs M_X, M_Y. Then $p_X = p_Y$ iff $M_X = M_Y$. Thus an MGF is an alternative characterization of a random variable.

What makes M_X special is obviously the exponential function. Since $e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$,

$$M_X(t) = 1 + t^\top \underbrace{\mathbf{E}\left[X\right]}_{\text{1st moment}} + \frac{1}{2} t^\top \underbrace{\mathbf{E}\left[XX^\top\right]}_{\text{2nd moment}} t + \cdots$$

so that $\frac{\partial^n M_X(t)}{\partial^n t}|_{t=0_d}$ is the n-th moment of p_X (hence the name).

Lemma C.1. Let $X \sim \mathcal{N}(\mu, \Sigma)$. Then

$$M_X(t) = \exp\left(t^\top \mu + \frac{1}{2} t^\top \Sigma t\right)$$

Proof. We use the same substitution in the proof of Lemma E.4. Let $\Sigma = U\Lambda U^\top$ denote an orthonormal eigendecomposition. Let $u = g(x)$ where $g(x) = \Lambda^{-1/2} U^\top (x - \mu)$, which implies $x = U\Lambda^{1/2} u + \mu$. Thus $|\det(J_g(x))| = \left|\det(\Lambda^{-1/2} U^\top)\right| = \det(\Lambda)^{-1/2}$, so we have the infinitesimal $dx = \sqrt{\det(\Lambda)} du$. Then

$$\begin{aligned} &\int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) \exp(t^\top x) dx \\ &= \int_{\mathbb{R}^d} \frac{\sqrt{\det(\Lambda)}}{(\sqrt{2\pi})^d \sqrt{\det(\Lambda)}} \exp\left(-\frac{1}{2} u^\top u\right) \exp(t^\top U\Lambda^{1/2} u + t^\top \mu) du \\ &= \exp(t^\top \mu) \int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{1}{2} u^\top u + t^\top U\Lambda^{1/2} u\right) du \\ &= \exp(t^\top \mu) \int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{1}{2}\left|\left|u - U\Lambda^{1/2} t\right|\right|^2 + \frac{1}{2} t^\top \Sigma t\right) du \\ &= \exp\left(t^\top \mu + \frac{1}{2} t^\top \Sigma t\right) \int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{1}{2}\left|\left|u - U\Lambda^{1/2} t\right|\right|^2\right) du \\ &= \exp\left(t^\top \mu + \frac{1}{2} t^\top \Sigma t\right) \end{aligned}$$

□

An interesting consequence of the Gaussian MGF in Lemma C.1 is that a point-mass density can be viewed as a degenerate Gaussian distribution with zero variance. That is, if $X \in \mathbb{R}^d$ takes value $a \in \mathbb{R}^d$ with probability 1, then $M_X(t) = \exp(a^\top t)$, which is equal to the Gaussian MGF with $\Sigma = 0_{d\times d}$.

One application of MGF is showing that a linear transformation of a Gaussian random variable is also Gaussian. Note that the MGF of a linear transformation of X is generally

$$M_{AX+b}(t) = \mathop{\mathbf{E}}_{x\sim p_X}\left[\exp\left(t^\top Ax\right)\exp\left(t^\top b\right)\right] = \exp\left(t^\top b\right) M_X(A^\top t) \tag{42}$$

Lemma C.2. Let $X \sim \mathcal{N}(\mu, \Sigma)$. Let $A \in \mathbb{R}^{d'\times d}$ and $b \in \mathbb{R}^{d'}$ where $d' \le d$ and A has full rank. Then $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top)$.

Proof. For any $t \in \mathbb{R}^{d'}$,

$$\begin{aligned} M_{AX+b}(t) &= \exp\left(t^\top b\right) M_X(A^\top t) && \text{(by (42))} \\ &= \exp\left(t^\top b\right)\exp\left(t^\top A\mu + \frac{1}{2}t^\top A\Sigma A^\top t\right) && \text{(by Lemma C.1)} \\ &= \exp\left(t^\top(A\mu + b) + \frac{1}{2}t^\top A\Sigma A^\top t\right) \end{aligned}$$

The last term is the MGF of a random variable with distribution $\mathcal{N}(A\mu + b, A\Sigma A^\top)$ where $A\Sigma A^\top \succ 0$. The statement follows from the one-to-one correspondence between MGFs and distributions. □

C.1 Cumulant-Generating Function

The log MGF $\psi_X(t) := \log \mathbf{E}[e^{t^\top X}]$ is called the **cumulant-generating function (CGF)** of X. We see that it is the (convex) log-partition function of t-tilted X_t distributed as (Appendix D)

$$p_{X_t}(x) = \frac{e^{t^\top x} p_X(x)}{\mathbf{E}[e^{t^\top X}]}$$

We call $\nabla^{(n)}\psi_X(t)$ the n-th **cumulant** of X. From (49–50), we have

$$\nabla\psi_X(t) = \mathbf{E}[X_t] \tag{43}$$
$$\nabla^2\psi_X(t) = \mathrm{Cov}(X_t) \tag{44}$$

In particular,

$$\nabla\psi_X(0_d) = \mathbf{E}[X] \tag{45}$$
$$\nabla^2\psi_X(0_d) = \mathrm{Cov}(X) \tag{46}$$

This fact is used in Hoeffding's lemma which bounds the CGF of a bounded scalar random variable by using Taylor's approximation of the CGF around 0 and then bounding the mean/variance of that variable (Lemma E.23).

Example. The CGF of $X \sim \mathcal{N}(\mu, \Sigma)$ is $\psi_X(t) = \mu^\top t + \frac{1}{2}t^\top \Sigma t$, so

$$\nabla\psi_X(t) = \mu + \Sigma t$$
$$\nabla^2\psi_X(t) = \Sigma$$

which is consistent with the fact that $X_t \sim \mathcal{N}(\mu + \Sigma t, \Sigma)$ (Lemma E.22).

D Exponential Family

D.1 Exponential Tilting

Given any "base" distribution p over $\mathbb{R}^d$, we can generate a set of distributions $q_{p,\tau,\theta}$ by

$$q_{p,\tau,\theta}(x) := \frac{e^{\theta^\top \tau(x)} p(x)}{\mathbf{E}_{x'\sim p}[e^{\theta^\top \tau(x')}]} \tag{47}$$

for any $\tau : \mathbb{R}^d \to \mathbb{R}^m$ and $\theta \in \mathbb{R}^m$ such that $\mathbf{E}_{x'\sim p}[e^{\theta^\top \tau(x')}]$ exists. Note that

- $q_{p,\tau,\theta}$ is nonnegative and sums/integrates to 1.
- $q_{p,\tau,\theta}$ has the same support as p.
- $q_{p,\tau,\theta}$ assigns a weight $e^{\theta^\top \tau(x)}$ on the probability of x, changing the tails of p.
- $q_{p,\tau,0_m} = p$.

This technique is called **exponential tilting** of p. We can rewrite (47) as

$$q_{p,\tau,\theta}(x) = p(x) \exp\left(\theta^\top \tau(x) - B_{p,\tau}(\theta)\right) \tag{48}$$

where the log-partition function $B_{p,\tau}(\theta) := \log \mathbf{E}_{x\sim p}[e^{\theta^\top \tau(x)}]$ normalizes $q_{p,\tau,\theta}$. We note several properties:

- $B_{p,\tau}(\theta)$ is convex (Lemma E.17).
- τ is a sufficient statistic for θ (Theorem E.16).
- Differentiating $B_{p,\tau}(\theta)$ generates the cumulants of $\tau(x)$ over $x \sim q_{p,\tau,\theta}$, for instance (Lemma E.18)

$$\nabla B_{p,\tau}(\theta) = \mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}[\tau(x)] \tag{49}$$

$$\nabla^2 B_{p,\tau}(\theta) = \mathop{\mathrm{Cov}}_{x\sim q_{p,\tau,\theta}}(\tau(x)) \tag{50}$$

  In particular, $\nabla B_{p,\tau}(0_m) = \mathbf{E}_{x\sim p}[\tau(x)]$ and $\nabla^2 B_{p,\tau}(0_m) = \mathrm{Cov}_{x\sim p}(\tau(x))$.
- Aside: (50) implies that $B_{p,\tau}(\theta)$ is convex since $\nabla^2 B_{p,\tau}(\theta) \succeq 0$.

Exponential tilting often preserves the distribution family. For instance, if $X \sim \mathcal{N}(\mu, \Sigma)$ and X_t is the t-tilted X with $t \in \mathbb{R}^d$ ($\tau(x) = x$), then $X_t \sim \mathcal{N}(\mu + \Sigma t, \Sigma)$ (Lemma E.22).

D.2 Unnormalized Form

More generally, we may consider any nonnegative function $h : \mathbb{R}^d \to (0, \infty)$ ("base measure") and define

$$q_{h,\tau,\theta}(x) = \frac{\exp\left(\theta^\top \tau(x)\right) h(x)}{\int_{x\in\mathbb{R}^d} \exp\left(\theta^\top \tau(x)\right) h(x) dx} \tag{51}$$

for any $\tau : \mathbb{R}^d \to \mathbb{R}^m$ and $\theta \in \mathbb{R}^m$ such that $\int_{x\in\mathbb{R}^d} \exp\left(\theta^\top \tau(x)\right) h(x) dx$ exists. We can rewrite (51) as

$$q_{h,\tau,\theta}(x) = h(x) \exp\left(\theta^\top \tau(x) - A_{h,\tau}(\theta)\right) \tag{52}$$

where $A_{h,\tau}(\theta) := \log\left(\int_{x\in\mathbb{R}^d} h(x) \exp\left(\theta^\top \tau(x)\right) dx\right)$ and τ is again a sufficient statistic for θ. Clearly, exponential tilting is a special case where the base measure is normalized. However, (51) is strictly more general since it allows for h such that $\int_x h(x)dx$ diverges. It is easy to check that the previous properties hold without a normalized base measure, specifically:

- Differentiating $A_{h,\tau}(\theta)$ generates the cumulants of $\tau(x)$ over $x \sim q_{h,\tau,\theta}$, in particular

$$\nabla A_{h,\tau}(\theta) = \mathop{\mathbf{E}}_{x\sim q_{h,\tau,\theta}}[\tau(x)] \tag{53}$$

$$\nabla^2 A_{h,\tau}(\theta) = \mathop{\mathrm{Cov}}_{x\sim q_{h,\tau,\theta}}(\tau(x)) \tag{54}$$

- (54) implies that $A_{h,\tau}(\theta)$ is convex.

A set of distributions that can be expressed in the form (52) is called an **exponential family**. $\theta \in \mathbb{R}^m$ is called its **natural parameter**. Note that there are many exponential families. For instance, the set of all normal distributions is one exponential family. The set of all categorical distributions is another exponential family.

D.2.1 Discussions

CGF. The CGF $\psi_{\tau(X)}(t) = \log \mathbf{E}[e^{t^\top \tau(X)}]$ of $\tau(x)$ takes the form (Lemma E.20):

$$\psi_{\tau(X)}(t) = A_{h,\tau}(\theta + t) - A_{h,\tau}(\theta) \tag{55}$$

where we see $\nabla^{(n)}\psi_{\tau(X)}(0_m) = \nabla^{(n)}A_{h,\tau}(\theta)$; this is consistent with the fact that in an exponential family, the log-partition function generates cumulants.

Conjugate prior. In Bayesian probability theory, a prior over the parameter of a distribution is called a **conjugate prior** if the implied posterior over the parameter conditioning on a sample from the distribution is in the same distribution family that the prior is in. For an exponential family, we can define a prior

$$\pi_{h,\tau}(\theta; \alpha, \beta) = \frac{1}{Z_{h,\tau}(\alpha, \beta)} \exp\left(\theta^\top \alpha - \beta A_{h,\tau}(\theta)\right) \tag{56}$$

for any "pseudo-counts" $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}$ such that $Z_{h,\tau}(\alpha, \beta) = \int_{\theta \in \mathbb{R}^m} \exp\left(\theta^\top \alpha - \beta A_{h,\tau}(\theta)\right) d\theta$ exists. Then the posterior over θ given $x \sim q_{h,\tau,\theta}$ is given by (Lemma E.21)

$$\kappa_{h,\tau}(\theta | x; \alpha, \beta) = \pi_{h,\tau}(\theta; \tau(x) + \alpha, 1 + \beta) \tag{57}$$

thus (56) is a conjugate prior.

Identifying an exponential family. To check if a set of distributions $\left\{p(x; \bar{\theta})\right\}_{\bar{\theta}}$ is an exponential family, it is sufficient to propose any $h(x) \geq 0$, a transformation of $\bar{\theta}$ into natural parameter form $\theta = g(\bar{\theta}) \in \mathbb{R}^m$ and x into sufficient statistic form $\tau(x) \in \mathbb{R}^m$, and *some* function $A_{h,\tau}(\theta)$, such that it can be written as (52):

$$p(x; \bar{\theta}) = q_{h,\tau,\theta}(x) = h(x) \exp\left(\theta^\top \tau(x) - A_{h,\tau}(\theta)\right)$$

In particular, we do not need to explicitly calculate $A_{h,\tau}(\theta) = \log\left(\int_{x \in \mathbb{R}^d} h(x) \exp\left(\theta^\top \tau(x)\right) dx\right)$ since the normalization of $p(x; \bar{\theta})$ enforces it (and guarantees its existence).

Non-unique parameterization. An exponential family has infinitely many equivalent parameterizations:

$$q_{h,\tau,\theta}(x) = q_{ah,u\odot\tau,\mathrm{inv}(u)\odot\theta}(x) \qquad \forall a \in \mathbb{R}\setminus\{0\},\ u \in (\mathbb{R}\setminus\{0\})^m$$

where $\odot$ is the elementwise multiplication and $\mathrm{inv}(u)$ is the elementwise inverse of vector u. It is often clear what a natural parameterization is (e.g., choose u that makes $\tau(x)$ as simple as possible).

Limitations. A dizzying array of distributions are exponential families, including the normal (Lemma E.19), categorical, exponential, geometric, Bernoulli, Poisson, beta, and many others. But there are certain properties that an exponential family cannot capture. First, the form

$$h(x) \exp\left(\theta^\top \tau(x) - A_{h,\tau}(\theta)\right)$$

implies that the support of this distribution cannot depend on the parameter θ. This rules out distributions like a uniform distribution on $[a, b] \subset \mathbb{R}$ whose support depends on the paramters a, b. Second, some distributions simply cannot be expressed using an inner product between the input and the parameter, for instance the Laplace distribution

$$\mathrm{Laplace}(\mu, b)(x) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

Third, an exponential family necessarily has a well-defined MGF by (55), so it rules out distributions without an MGF such as the Cauchy distribution.

E Lemmas

Lemma E.1 (Polar coordinates)**.** For any integrable $f : \mathbb{R} \to \mathbb{R}$,

$$\int_{\mathbb{R}^2} f(x^2+y^2)d(x,y) = 2\pi \int_0^\infty f(r^2) r dr$$

Proof. Let $R = [0,\infty) \times [0, 2\pi]$ and define $g : R \to \mathbb{R}^2$ by $g(r,\theta) = (r\cos\theta, r\sin\theta)$. Note that $r^2 = x^2 + y^2$ and $g(R) = \mathbb{R}^2$. The Jacobian of g at (r,θ) is

$$J_g(r,\theta) = \begin{bmatrix} \frac{\partial r\cos\theta}{\partial r} & \frac{\partial r\cos\theta}{\partial \theta} \\ \frac{\partial r\sin\theta}{\partial r} & \frac{\partial r\sin\theta}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}$$

Thus $|\det(J_g(r,\theta))| = \left|r(\cos^2\theta + \sin^2\theta)\right| = r$. Thus

$$\begin{aligned}
\int_{\mathbb{R}^2} f(x^2+y^2)d(x,y) &= \int_R f(g_1(r,\theta)^2 + g_2(r,\theta)^2)\, |J_g(r,\theta)|\, d(r,\theta) && \text{(by (23))} \\
&= \int_R f(r^2) r d(r,\theta) \\
&= \int_0^\infty \left(\int_0^{2\pi} \exp(-r^2) r d\theta \right) dr && \text{(Fubini)} \\
&= \int_0^\infty 2\pi \exp(-r^2) r dr && \text{(FTC)} \\
&= 2\pi \int_0^\infty \exp(-r^2) r dr && \text{(linearity)}
\end{aligned}$$

□

Lemma E.2 (Gaussian integral)**.**

$$\int_{-\infty}^{\infty} \exp\left(-x^2\right) dx = \sqrt{\pi} \tag{58}$$

Proof. A standard proof shows that $(\int_{-\infty}^{\infty} \exp\left(-x^2\right) dx)^2 = \pi$ as follows:

$$\begin{aligned}
\left(\int_{-\infty}^{\infty} \exp\left(-x^2\right) dx \right) \left(\int_{-\infty}^{\infty} \exp\left(-y^2\right) dy \right) &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \exp\left(-x^2\right) dx \right) \exp\left(-y^2\right) dy && \text{(linearity)} \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \exp\left(-x^2\right) \exp\left(-y^2\right) dx \right) dy && \text{(linearity)} \\
&= \int_{\mathbb{R}^2} \exp\left(-(x^2+y^2)\right) d(x,y) && \text{(Fubini)} \\
&= 2\pi \int_0^\infty \exp\left(-r^2\right) r dr && \text{(Lemma E.1)} \\
&= 2\pi \left(-\frac{1}{2} \exp(-r^2) \right) \Big|_0^\infty && \text{(FTC)} \\
&= 2\pi \left(0 + \frac{1}{2} \right) = \pi
\end{aligned}$$

□

Lemma E.3. For any $\mu \in \mathbb{R}$ and $\sigma^2 > 0$,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) dx = 1 \tag{59}$$

Proof. Let $u = \frac{x-\mu}{\sqrt{2}\sigma}$ which gives the infinitesimal $dx = \sqrt{2}\sigma du$. Then

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx &= \int_{-\infty}^{\infty} \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \exp\left(-u^2\right) du && \text{(by (21))} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-u^2\right) du \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(-u^2\right) du && \text{(linearity)} \\
&= 1 && \text{(Lemma E.2)}
\end{aligned}$$

□

Lemma E.4.

$$\int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) dx = 1$$

Proof. Let $\Sigma = U\Lambda U^\top$ denote an orthonormal eigendecomposition. Let $u = g(x)$ where $g(x) = \Lambda^{-1/2}U^\top(x-\mu)$. Thus $|\det(J_g(x))| = \left|\det(\Lambda^{-1/2}U^\top)\right| = \det(\Lambda)^{-1/2}$, so we have the infinitesimal $dx = \sqrt{\det(\Lambda)}du$. Then

$$\begin{aligned}
\int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi})^d \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) dx &= \int_{\mathbb{R}^d} \frac{\sqrt{\det(\Lambda)}}{(\sqrt{2\pi})^d \sqrt{\det(\Lambda)}} \exp\left(-\frac{1}{2}u^\top u\right) du \\
&= \int_{\mathbb{R}^d} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_i^2}{2}\right) du
\end{aligned}$$

By Fubini and linearity,

$$\begin{aligned}
\int_{\mathbb{R}^d} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_i^2}{2}\right) du &= \int_{-\infty}^{\infty} \left(\cdots \left(\int_{-\infty}^{\infty} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_i^2}{2}\right) du_1 \right) \cdots \right) du_d \\
&= \prod_{i=1}^{d} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u_i^2}{2}\right) du_i \\
&= \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \right)^d = 1
\end{aligned}$$

where the last step applies Lemma E.3 with $\mu = 0$ and $\sigma^2 = 1$. □

Lemma E.5. For any $\lambda > 0$, the exponential distribution $e_\lambda(x) := \lambda \exp(-\lambda x)$ over $[0, \infty)$ has entropy

$$H(X) = 1 - \log \lambda$$

Proof.

$$\begin{aligned}
H(X) &= -\int_0^\infty \lambda \exp(-\lambda x) \log(\lambda \exp(-\lambda x)) dx \\
&= -\log \lambda - \lambda \int_0^\infty \exp(-\lambda x)(-\lambda x) dx
\end{aligned}$$

We evaluate the last integral as follows. Let $u = g(x) = -\lambda x$, then $g'(x) = -\lambda$ so that $\left|g'(g^{-1}(u))^{-1}\right| = 1/\lambda$. Reorienting the region between $g(0) = 0$ and $g(\infty) = -\infty$ and applying (22),

$$\begin{aligned}
\lambda \int_0^\infty \exp(-\lambda x)(-\lambda x) dx &= \int_{-\infty}^0 \exp(u) u du \\
&= \exp(u)u\big|_{-\infty}^0 - \int_{-\infty}^0 \exp(u) du && \text{(integration by parts (20))} \\
&= (0 - 0) - \exp(u)\big|_{-\infty}^0 && (\lim_{u \to -\infty} \exp(u)u = 0) \\
&= -1
\end{aligned}$$

□

Lemma E.6. Define $\Delta := \mu' - \mu$. Then

$$H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) = \frac{1}{2}\Delta^\top \Sigma^{-1}\Delta + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma'\right) + \frac{1}{2}\log((2\pi)^d \det(\Sigma))$$

Proof.

$$\begin{aligned} H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) &:= \mathop{\mathbf{E}}_{x\sim\mathcal{N}(\mu',\Sigma')}\left[-\log \mathcal{N}(\mu, \Sigma)(x)\right] \\ &= \frac{1}{2}\mathop{\mathbf{E}}_{x\sim\mathcal{N}(\mu',\Sigma')}\left[(x-\mu)^\top \Sigma^{-1}(x-\mu)\right] + \frac{1}{2}\log((2\pi)^d \det(\Sigma)) \end{aligned}$$

By the cyclic property and the linearity of trace,

$$\begin{aligned} \mathop{\mathbf{E}}_{x\sim\mathcal{N}(\mu',\Sigma')}\left[(x-\mu)^\top \Sigma^{-1}(X-\mu)\right] &= \mathop{\mathbf{E}}_{x\sim\mathcal{N}(\mu',\Sigma')}\left[\mathrm{tr}\left((x-\mu)^\top \Sigma^{-1}(x-\mu)\right)\right] \\ &= \mathop{\mathbf{E}}_{x\sim\mathcal{N}(\mu',\Sigma')}\left[\mathrm{tr}\left(\Sigma^{-1}(x-\mu)(x-\mu)^\top\right)\right] \\ &= \mathrm{tr}\left(\Sigma^{-1}\mathop{\mathbf{E}}_{x\sim\mathcal{N}(\mu',\Sigma')}\left[(x-\mu)(x-\mu)^\top\right]\right) \end{aligned}$$

Rewriting the expectation,

$$\begin{aligned} \mathop{\mathbf{E}}_{x\sim\mathcal{N}(\mu',\Sigma')}\left[(x-\mu)(x-\mu)^\top\right] &= \mathop{\mathbf{E}}_{x\sim\mathcal{N}(\mu',\Sigma')}\left[(x-\mu'+\Delta)(x-\mu'+\Delta)^\top\right] \\ &= \mathop{\mathbf{E}}_{x\sim\mathcal{N}(\mu',\Sigma')}\left[(x-\mu')(x-\mu')^\top + (x-\mu')\Delta^\top + \Delta(x-\mu')^\top + \Delta\Delta^\top\right] \\ &= \Sigma' + \Delta\Delta^\top \end{aligned}$$

Therefore we have

$$\begin{aligned} H(\mathcal{N}(\mu', \Sigma'), \mathcal{N}(\mu, \Sigma)) &= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma' + \Sigma^{-1}\Delta\Delta^\top\right) + \frac{1}{2}\log((2\pi)^d \det(\Sigma)) \\ &= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma'\right) + \frac{1}{2}\Delta^\top\Sigma^{-1}\Delta + \frac{1}{2}\log((2\pi)^d \det(\Sigma)) \end{aligned}$$

□

Corollary E.7 (Of Lemma E.6)**.**

$$H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2}\log\left((2\pi e)^d \det(\Sigma)\right)$$

Corollary E.8 (Of Lemma E.6 and Corollary E.7)**.** Define $\Delta := \mu' - \mu$. Then

$$D_{\mathrm{KL}}(\mathcal{N}(\mu', \Sigma')||\mathcal{N}(\mu, \Sigma)) = \frac{1}{2}\Delta^\top\Sigma^{-1}\Delta + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma' - I_{d\times d}\right) + \frac{1}{2}\log\frac{\det(\Sigma)}{\det(\Sigma')}$$

Lemma E.9. Let $A \in \mathbb{R}^{d\times d}$. The **main-diagonal block matrix** of A at index $k \in \{1 \dots d\}$ with size n is a matrix $B(k, n) \in \mathbb{R}^{n\times n}$ with entries $B_{i,j}(k, n) = A_{k+i-1,k+j-1}$ for $i, j \in \{1 \dots n\}$. If $A \succ 0$, then $B(k, n) \succ 0$ for all valid k, n.

Proof. Suppose $u^\top B(k, n)u \leq 0$ for some nonzero $u \in \mathbb{R}^n$. Define $v \in \mathbb{R}^d$ where $v_{k+i-1} = u_i$ for $i = 1 \dots n$ and other entries are zero. Then v is nonzero and $v^\top Av = u^\top B(k, n)u \leq 0$, contradicting the premise that $A \succ 0$. □

Lemma E.10. Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ be jointly normal with parameters (μ, Σ). Assume that $\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ is invertible. Then for any $z = (x, y) \in \mathbb{R}^{d+d'}$,

$$\frac{1}{(\sqrt{2\pi})^{d+d'}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(z-\mu)^\top \Sigma^{-1}(z-\mu)\right) = \frac{1}{(\sqrt{2\pi})^{d}\sqrt{\det(\Sigma_X)}} \exp\left(-\frac{1}{2}(x-\mu_X)^\top \Sigma_X^{-1}(x-\mu_X)\right) \times \frac{1}{(\sqrt{2\pi})^{d'}\sqrt{\det(\Omega)}} \exp\left(-\frac{1}{2}(y-\phi(x))^\top \Omega^{-1}(y-\phi(x))\right) \quad (60)$$

where $\Omega \in \mathbb{R}^{d'\times d'}$ and $\phi(x) \in \mathbb{R}^{d'}$ are defined as

$$\Omega := \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY} \quad (61)$$

$$\phi(x) := \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x-\mu_X) \qquad \forall x \in \mathbb{R}^d \quad (62)$$

Proof. By block matrix inversion and abbreviating $O = \Sigma_X^{-1}\Sigma_{XY}$,

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma_X^{-1} + O\Omega^{-1}O^\top & -O\Omega^{-1} \\ -\Omega^{-1}O^\top & \Omega^{-1} \end{bmatrix}$$

Abbreviating $u = x - \mu_X$ and $v = y - \mu_Y$,

$$\begin{aligned}(z-\mu)^\top\Sigma^{-1}(z-\mu) &= u^\top\left(\Sigma_X^{-1} + O\Omega^{-1}O^\top\right)u - u^\top O\Omega^{-1}v - v^\top\Omega^{-1}O^\top u + v^\top\Omega^{-1}v \\ &= u^\top\Sigma_X^{-1}u + u^\top O\Omega^{-1}O^\top u - 2u^\top O\Omega^{-1}v + v^\top\Omega^{-1}v \\ &= u^\top\Sigma_X^{-1}u + (v - O^\top u)^\top\Omega^{-1}(v - O^\top u) \\ &= (x-\mu_X)^\top\Sigma_X^{-1}(x-\mu_X) + (y-\phi(x))^\top\Omega^{-1}(y-\phi(x))\end{aligned}$$

where we use the fact that Ω is symmetric. By the determinant identity of a block matrix, we have $\det(\Sigma) = \det(\Sigma_X\Omega) = \det(\Sigma_X)\det(\Omega)$. Applying these identities to the LHS of (60) yields the RHS. □

Lemma E.11. Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ be jointly normal with parameters (μ, Σ). Assume that $\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ is invertible. Then

$$H(Y|X) = \frac{1}{2}\log\left((2\pi e)^{d'}\det(\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY})\right) \quad (63)$$

$$I(X,Y) = \frac{1}{2}\log\left(\frac{\det(\Sigma_X)\det(\Sigma_Y)}{\det(\Sigma)}\right) \quad (64)$$

Proof. By Lemma E.10, $Y|X = x$ is distributed as $\mathcal{N}(\phi(x), \Omega)$ for any $x \in \mathbb{R}^d$ where $\phi(x) := \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x-\mu_X)$ and $\Omega := \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$. Thus

$$H(Y|X=x) = \mathbf{E}\left[-\log\Pr(Y|X=x)\right] = \frac{1}{2}\mathbf{E}\left[(Y-\phi(x))^\top\Omega^{-1}(Y-\phi(x))\right] + \frac{1}{2}\log((2\pi)^{d'}\det(\Omega))$$

Using trace similarly as in the proof of Lemma E.6, we can verify

$$\mathbf{E}\left[(Y-\phi(x))^\top\Omega^{-1}(Y-\phi(x))\right] = \Omega^{-1}(\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}(x-\mu_X)(x-\mu_X)^\top\Sigma_X^{-1}\Sigma_{XY})$$

Taking the expectation over x yields $I_{d'\times d'}$. This shows (63). To show (64), we have

$$\begin{aligned}I(X,Y) &= H(Y) - H(Y|X) \\ &= \frac{1}{2}\log\left((2\pi e)^{d'}\det(\Sigma_Y)\right) - \frac{1}{2}\log\left((2\pi e)^{d'}\det(\Omega)\right) \\ &= \frac{1}{2}\log\left(\frac{\det(\Sigma_Y)}{\det(\Omega)}\right) \\ &= \frac{1}{2}\log\left(\frac{\det(\Sigma_X)\det(\Sigma_Y)}{\det(\Sigma)}\right)\end{aligned}$$

where for the last equality we use the fact that $\det(\Sigma) = \det(\Sigma_X\Omega) = \det(\Sigma_X)\det(\Omega)$. □

Lemma E.12. The following statements about $X \in \mathbb{R}^d$ are equivalent.

1. $X \sim \mathcal{N}(\mu, \Sigma)$.
2. $M_X(t) = \exp(t^\top \mu + \frac{1}{2} t^\top \Sigma t)$ for all $t \in \mathbb{R}^d$.
3. $X = \Sigma^{1/2} Z + \mu$ where $Z \sim \mathcal{N}(0_d, I_{d \times d})$.
4. $Y = a^\top X$ has the density $\mathcal{N}(a^\top \mu, a^\top \Sigma a)$ for all nonzero $a \in \mathbb{R}^d$.

Proof. Lemma C.1 gives 1 ≡ 2. To show 2 ≡ 3 we note that by (42)

$$M_{\Sigma^{1/2} Z + \mu}(t) = \exp\left(t^\top \mu\right) M_Z(\Sigma^{1/2} t) = \exp\left(t^\top \mu + \frac{1}{2} t^\top \Sigma t\right) = M_X(t)$$

We have 1 ⇒ 4 since the density of Y is $\mathcal{N}(a^\top \mu, a^\top \Sigma a)$ by Lemma C.2. To show 4 ⇒ 2, pick any nonzero $a \in \mathbb{R}^d$. For all $t \in \mathbb{R}$

$$M_X(ta) = M_{a^\top X}(t) = \exp\left(t a^\top \mu + \frac{1}{2} t^2 a^\top \Sigma a\right)$$

where the first equality uses (42) and the second equality uses Lemma C.1. Setting $t = 1$ gives $M_X(a) = \exp\left(a^\top \mu + \frac{1}{2} a^\top \Sigma a\right)$. Additionally, $M_X(0_d) = 1 = \exp\left(0_d^\top \mu + \frac{1}{2} 0_d^\top \Sigma 0_d\right)$. Thus $M_X(t) = \exp(t^\top \mu + \frac{1}{2} t^\top \Sigma t)$ for all $t \in \mathbb{R}^d$. □

Lemma E.13 (Popoviciu's inequality)**.** For any bounded scalar random variable $X \in [a, b]$,

$$\operatorname{Var}(X) \leq \frac{(b-a)^2}{4}$$

with equality iff $\Pr(X = a) = \Pr(X = b) = \frac{1}{2}$.

Proof. For any constant $c \in \mathbb{R}$, $\mathbf{E}[(X-c)^2] = \mathbf{E}[(X - \mathbf{E}[X] + \mathbf{E}[X] - c)^2] \geq \operatorname{Var}(X)$. Choosing $c = \frac{b-a}{2}$ and using the fact that $\left|X - \frac{b-a}{2}\right| \leq \frac{b-a}{2}$, we have $\operatorname{Var}(X) \leq \mathbf{E}[(X - \frac{b-a}{2})^2] \leq \frac{(b-a)^2}{4}$. □

Lemma E.14 (Markov's inequality)**.** For any nonnegative scalar random variable $X \geq 0$, for any $\epsilon > 0$:

$$\Pr(X \geq \epsilon) \leq \frac{\mathbf{E}[X]}{\epsilon}$$

Proof.

$$\begin{aligned}
\mathbf{E}[X] &= \int_0^\infty \Pr(X = x) x \, dx && \text{(proof similar if } X \text{ is discrete)} \\
&\geq \int_\epsilon^\infty \Pr(X = x) x \, dx \\
&\geq \int_\epsilon^\infty \Pr(X = x) \epsilon \, dx \\
&\geq \epsilon \Pr(X \geq \epsilon)
\end{aligned}$$

□

Lemma E.15 (Chernoff's inequality)**.** For any scalar random variable $X \in \mathbb{R}$ and $\epsilon \geq \mathbf{E}[X]$,

$$\Pr(X \geq \epsilon) \leq e^{-\psi_X^*(\epsilon)}$$

where $\psi_X^*(\epsilon) = \sup_{t \in \mathbb{R}} t\epsilon - \psi_X(t)$ is the Legendre transform of the CGF $\psi_X(t) = \log \mathbf{E}[e^{tX}]$.

Proof.

$$\begin{aligned}
\Pr(X \geq \epsilon) &\leq \Pr(tX \geq t\epsilon) && \forall t \geq 0 \\
&= \Pr(e^{tX} \geq e^{t\epsilon}) \\
&= \frac{\mathbf{E}[e^{tX}]}{e^{t\epsilon}} && \text{(Markov's inequality, since } e^{tX} \geq 0 \text{ and } e^{t\epsilon} > 0\text{)} \\
&= e^{-(t\epsilon - \psi_X(t))}
\end{aligned}$$

In particular,

$$\begin{aligned}
\Pr(X \geq \epsilon) &\leq \inf_{t \geq 0} \; e^{-(t\epsilon - \psi_X(t))} \\
&= e^{-\left(\sup_{t \geq 0} \; t\epsilon - \psi_X(t)\right)} \\
&= e^{-\left(\sup_{t \in \mathbb{R}} \; t\epsilon - \psi_X(t)\right)} \\
&= e^{-\psi_X^*(\epsilon)}
\end{aligned} \tag{65}$$

The step (65) uses the following lemma.

Lemma. Let $J(t) := t\epsilon - \psi_X(t)$ and $J^* = \sup_{t \in \mathbb{R}} \; J(t)$. Then $J^* \geq J(0)$.

Proof.

$$\begin{aligned}
J(t) &= t\epsilon - \log \mathbf{E}[e^{tX}] \\
&\leq t\epsilon - t\mathbf{E}[X] && \text{(Jensen's inequality: } \log \mathbf{E}[X] \geq \mathbf{E}[\log X]\text{)} \\
&= t \underbrace{(\epsilon - \mathbf{E}[X])}_{\geq 0}
\end{aligned}$$

Thus $J(t) \leq 0$ for all $t < 0$. The lemma follows from the fact that $J(0) = 0$.

□

Theorem E.16 (Factorization Theorem)**.** Assume a joint distribution

$$p_{\Theta XT}(\theta, x, t) = p_\Theta(\theta) \times p_{X|\Theta}(x|\theta) \times [[\tau(x) = t]]$$

where $X \in \mathcal{X}$ is a sample from a distribution parametrized by $\Theta \in \mathcal{H}$, and $T = \tau(X) \in \mathcal{T}$ is the sample statistic for some function $\tau : \mathcal{X} \to \mathcal{T}$. The following statements about τ are equivalent: if any holds, we say τ is a **sufficient statistic** for Θ.

- X is conditionally independent of Θ given $T = t$:

$$p_{X|T}(x|t) = p_{X|T\Theta}(x|t, \theta) \tag{66}$$

- There exist $f_T : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{T} \times \mathcal{H} \to \mathbb{R}$ such that

$$p_{X|\Theta}(x|\theta) = f_T(x) \times g(\tau(x), \theta) \tag{67}$$

Proof. (67)⇒(66): For any t, θ,

$$\begin{aligned}
p_{T|\Theta}(t|\theta) &= \sum_{x \in \mathcal{X}:\; \tau(x) = t} p_{X|\Theta}(x|\theta) && \text{(proof similar if } X \text{ is continuous)} \\
&= \sum_{x \in \mathcal{X}:\; \tau(x) = t} f_T(x) \times g(\tau(x), \theta) && (67) \\
&= \left(\sum_{x \in \mathcal{X}:\; \tau(x) = t} f_T(x) \right) \times g(t, \theta)
\end{aligned}$$

thus for any x satisfying $\tau(x) = t$,

$$p_{X|T\Theta}(x|t,\theta) = \frac{p_{XT|\Theta}(x,t|\theta)}{p_{T|\Theta}(t|\theta)} = \frac{f_T(x) \times g(t,\theta)}{\left(\sum_{x\in\mathcal{X}:\,\tau(x)=t} f_T(x)\right) \times g(t,\theta)} = \frac{f_T(x)}{\sum_{x\in\mathcal{X}:\,\tau(x)=t} f_T(x)}$$

and $p_{X|T\Theta}(x|t,\theta) = 0$ for x such that $\tau(x) \neq t$. This implies $p_{X|T}(x|t) = p_{X|T\Theta}(x|t,\theta)$ for all θ.

(66)⇒(67): Define $f_T(x) = p_{X|T}(x|\tau(x))$ and $g(t,\theta) = p_{T|\Theta}(t|\theta)$. Then

$$\begin{aligned} p_{X|\Theta}(x|\theta) &= p_{XT|\Theta}(x,\tau(x)|\theta) \\ &= p_{X|T\Theta}(x|\tau(x),\theta) \times p_{T|\theta}(\tau(x)|\theta) \\ &= p_{X|T}(x|\tau(x)) \times p_{T|\Theta}(\tau(x)|\theta) \qquad (66) \\ &= f_T(x) \times g(\tau(x),\theta) \end{aligned}$$

□

Lemma E.17. Let $X \in \mathcal{X}$ be a random variable and $\tau : \mathcal{X} \to \mathbb{R}^m$ be a function such that

$$B_{p,\tau}(\theta) := \log \mathbf{E}\left[e^{\theta^\top \tau(X)}\right]$$

exists for all $\theta \in \mathbb{R}^m$. Then $B_{p,\tau} : \mathbb{R}^m \to \mathbb{R}$ is convex.

Proof. We use Hölder's inequality which states that $\mathbf{E}[|XY|] \leq \mathbf{E}[|X|^p]^{\frac{1}{p}}\mathbf{E}[|Y|^q]^{\frac{1}{q}}$ for any $p, q \geq 1$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$. For any $\alpha \in [0,1]$ and $\theta, \omega \in \mathbb{R}^m$:

$$\begin{aligned} \exp\left(B_{p,\tau}(\alpha\theta + (1-\alpha)\omega)\right) &= \mathbf{E}\left[e^{\alpha\theta^\top\tau(X)+(1-\alpha)\omega^\top\tau(X)}\right] \\ &= \mathbf{E}\left[\left|e^{\alpha\theta^\top\tau(X)}\right|\left|e^{(1-\alpha)\omega^\top\tau(X)}\right|\right] \\ &\leq \mathbf{E}\left[\left|e^{\alpha\theta^\top\tau(X)}\right|^{\frac{1}{\alpha}}\right]^{\alpha} \mathbf{E}\left[\left|e^{(1-\alpha)\omega^\top\tau(X)}\right|^{\frac{1}{1-\alpha}}\right]^{(1-\alpha)} \qquad \left(p = \frac{1}{\alpha}, q = \frac{1}{1-\alpha}\right) \\ &= \mathbf{E}\left[e^{\theta^\top\tau(X)}\right]^{\alpha} \mathbf{E}\left[e^{\omega^\top\tau(X)}\right]^{(1-\alpha)} \\ &= \exp\left(B_{p,\tau}(\theta)\right)^{\alpha} \exp\left(B_{p,\tau}(\omega)\right)^{(1-\alpha)} \end{aligned}$$

Taking the log on both sides yields $B_{p,\tau}(\alpha\theta + (1-\alpha)\omega) \leq \alpha B_{p,\tau}(\theta) + (1-\alpha)B_{p,\tau}(\omega)$. □

Lemma E.18. Let p be a distribution over $\mathbb{R}^d$ and define $q_{p,\tau,\theta}(x) := \frac{e^{\theta^\top\tau(x)}p(x)}{\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]}$ for function $\tau : \mathbb{R}^d \to \mathbb{R}^m$ and $\theta \in \mathbb{R}^m$ where $\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]$ exists. Let $B_{p,\tau}(\theta) := \log \mathbf{E}_{x\sim p}[e^{\theta^\top\tau(x)}]$. Then

$$\begin{aligned} \nabla B_{p,\tau}(\theta) &= \mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}[\tau(x)] \\ \nabla^2 B_{p,\tau}(\theta) &= \mathop{\mathrm{Cov}}_{x\sim q_{p,\tau,\theta}}(\tau(x)) \end{aligned}$$

Proof.

$$\begin{aligned} \nabla B_{p,\tau}(\theta) &= \frac{\mathbf{E}_{x\sim p}[e^{\theta^\top\tau(x)}\tau(x)]}{\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]} \\ \nabla^2 B_{p,\tau}(\theta) &= \frac{\mathbf{E}_{x\sim p}[e^{\theta^\top\tau(x)}\tau(x)\tau(x)^\top]}{\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]} - \left(\frac{\mathbf{E}_{x\sim p}[e^{\theta^\top\tau(x)}\tau(x)]}{\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]}\right)\left(\frac{\mathbf{E}_{x\sim p}[e^{\theta^\top\tau(x)}\tau(x)]}{\mathbf{E}_{x'\sim p}[e^{\theta^\top\tau(x')}]}\right)^\top \end{aligned}$$

Thus by the definition of $q_{p,\tau,\theta}$

$$\begin{aligned} \nabla B_{p,\tau}(\theta) &= \mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}[\tau(x)] \\ \nabla^2 B_{p,\tau}(\theta) &= \mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}[\tau(x)\tau(x)^\top] - \left(\mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}[\tau(x)]\right)\left(\mathop{\mathbf{E}}_{x\sim q_{p,\tau,\theta}}[\tau(x)]\right)^\top \end{aligned}$$

□

Lemma E.19. $\mathcal{N}(\mu, \Sigma)$ is in the exponential family, with one parameterization given by

$$\begin{aligned} h(x) &= \frac{1}{(\sqrt{2\pi})^d} && \text{(base measure)} \\ \theta &= \begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\text{vec}\left(\Sigma^{-1}\right) \end{bmatrix} \in \mathbb{R}^{d(d+1)} && \text{(natural parameter)} \\ \tau(x) &= \begin{bmatrix} x \\ \text{vec}\left(xx^\top\right) \end{bmatrix} \in \mathbb{R}^{d(d+1)} && \text{(sufficient statistic)} \\ A_{h,\tau}(\theta) &= \frac{1}{2}\left(\mu^\top\Sigma^{-1}\mu + \log\left(\det(\Sigma)\right)\right) && \text{(log-partition function)} \end{aligned}$$

where $\text{vec}(M) \in \mathbb{R}^{n^2}$ is the vector form of matrix $M \in \mathbb{R}^{n\times n}$ with $[\text{vec}(M)]_{(i-1)n+j} = M_{i,j}$.

Proof.

$$\begin{aligned} \mathcal{N}(\mu,\Sigma)(x) &= \frac{1}{(\sqrt{2\pi})^d\sqrt{\det(\Sigma)}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right) \\ &= \frac{1}{(\sqrt{2\pi})^d}\exp\left(\mu^\top\Sigma^{-1}x - \frac{1}{2}x^\top\Sigma^{-1}x - \frac{1}{2}\mu^\top\Sigma^{-1}\mu - \frac{1}{2}\log\left(\det(\Sigma)\right)\right) \\ &= \frac{1}{(\sqrt{2\pi})^d}\exp\left(\begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\text{vec}\left(\Sigma^{-1}\right) \end{bmatrix}^\top \begin{bmatrix} x \\ \text{vec}\left(xx^\top\right) \end{bmatrix} - \frac{1}{2}\left(\mu^\top\Sigma^{-1}\mu + \log\left(\det(\Sigma)\right)\right)\right) \end{aligned}$$

where we use the fact that $u^\top Mv = \text{vec}(M)^\top\text{vec}(uv^\top)$. □

Lemma E.20. Let $q_{h,\tau,\theta}(x) = h(x)\exp\left(\theta^\top\tau(x) - A_{h,\tau}(\theta)\right)$ with $A_{h,\tau}(\theta) = \log\left(\int_{x\in\mathbb{R}^d} h(x)\exp\left(\theta^\top\tau(x)\right)dx\right)$ denote an exponential family. The log-MGF of the sufficient statistic $\tau(x)$ is given by

$$\psi_{\tau(X)}(t) = A_{h,\tau}(\theta+t) - A_{h,\tau}(\theta)$$

Proof.

$$\begin{aligned} M_{\tau(X)}(t) &= \mathop{\mathbf{E}}_{x\sim q_{h,\tau,\theta}}\left[\exp(t^\top\tau(x))\right] \\ &= \int_{x\in\mathbb{R}^d} h(x)\exp\left(\theta^\top\tau(x) - A_{h,\tau}(\theta)\right)\exp(t^\top\tau(x))dx \\ &= \exp\left(-A_{h,\tau}(\theta)\right)\int_{x\in\mathbb{R}^d} h(x)\exp\left((\theta+t)^\top\tau(x)\right)dx \\ &= \exp\left(A_{h,\tau}(\theta+t) - A_{h,\tau}(\theta)\right) \end{aligned}$$

□

Lemma E.21. Let $q_{h,\tau,\theta}(x) = h(x)\exp\left(\theta^\top\tau(x) - A_{h,\tau}(\theta)\right)$ with $A_{h,\tau}(\theta) = \log\left(\int_{x\in\mathbb{R}^d} h(x)\exp\left(\theta^\top\tau(x)\right)dx\right)$ denote an exponential family. Define a distribution over $\theta\in\mathbb{R}^m$ by

$$\pi_{h,\tau}(\theta;\alpha,\beta) := \frac{1}{Z_{h,\tau}(\alpha,\beta)}\exp\left(\theta^\top\alpha - \beta A_{h,\tau}(\theta)\right)$$

for $\alpha\in\mathbb{R}^m$ and $\beta\in\mathbb{R}$ such that $Z_{h,\tau}(\alpha,\beta) := \int_{\theta\in\mathbb{R}^m}\exp\left(\theta^\top\alpha - \beta A_{h,\tau}(\theta)\right)d\theta$ exists. Then the conditional distribution over θ given x is

$$\kappa_{h,\tau}(\theta|x;\alpha,\beta) = \pi_{h,\tau}(\theta;\tau(x)+\alpha,1+\beta)$$

Proof. By Bayes' rule,

$$\begin{aligned}\kappa_{h,\tau}(\theta|x;\alpha,\beta) &\propto \pi_{h,\tau}(\theta;\alpha,\beta)\times q_{h,\tau,\theta}(x)\\ &= \frac{1}{Z_{h,\tau}(\alpha,\beta)}\exp\left(\theta^\top\alpha-\beta A_{h,\tau}(\theta)\right)\times h(x)\exp\left(\theta^\top\tau(x)-A_{h,\tau}(\theta)\right)\\ &\propto \exp\left(\theta^\top(\tau(x)+\alpha)-(1+\beta)A_{h,\tau}(\theta)\right)\end{aligned}$$

This implies $\kappa_{h,\tau}(\theta|x;\alpha,\beta)=\pi_{h,\tau}(\theta;\tau(x)+\alpha,1+\beta)$. □

Lemma E.22. Let X_t denote the t-tilted $X\sim\mathcal{N}(\mu,\Sigma)$ using $\tau(x)=x$. Then

$$X_t\sim\mathcal{N}(\mu+\Sigma t,\Sigma)$$

Proof. We can directly verify this claim using the fact that the CGF of X is $\mu^\top t+\frac{1}{2}t^\top\Sigma t$:

$$\begin{aligned}\Pr(X_t=x) &= \frac{e^{t^\top x}}{\mathbf{E}_{x'\sim\mathcal{N}(\mu,\Sigma)}[e^{t^\top x'}]}\mathcal{N}(\mu,\Sigma)(x)\\ &= \frac{1}{(\sqrt{2\pi})^d\sqrt{\det(\Sigma)}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)+t^\top x-\mu^\top t-\frac{1}{2}t^\top\Sigma t\right)\\ &= \frac{1}{(\sqrt{2\pi})^d\sqrt{\det(\Sigma)}}\exp\left(-\frac{1}{2}(x-\mu-\Sigma t)^\top\Sigma^{-1}(x-\mu-\Sigma t)\right)\end{aligned}$$

□

Lemma E.23 (Hoeffding's lemma)**.** Let $X\in[a,b]$ be a bounded scalar random variable. Then

$$\psi_{X-\mathbf{E}[X]}(t)\le\frac{(b-a)^2t^2}{8}$$

Proof. For any $t\in\mathbb{R}$, by Taylor's approximation of ψ_X around 0, for some η between 0 and t:

$$\psi_X(t)=\underbrace{\psi_X(0)}_{0}+\underbrace{\psi_X'(0)}_{\mathbf{E}[X]}t+\frac{1}{2}\underbrace{\psi_X''(\eta)}_{\mathrm{Var}(X_\eta)}t^2 \qquad\Leftrightarrow\qquad \psi_{X-\mathbf{E}[X]}(t)=\frac{\mathrm{Var}\left(X_\eta\right)t^2}{2}$$

where $X_\eta\in[a,b]$ is the η-tilted X (44). By Popoviciu's inequality (Lemma E.13), $\mathrm{Var}\left(X_\eta\right)\le\frac{(b-a)^2}{4}$. □

Lemma E.24. Let $\psi_X^*(t):=\sup_{\lambda\in\mathbb{R}^d}\ \lambda^\top t-\psi_X(\lambda)$ denote the Legendre transform of ψ_X. If $X\sim\mathcal{N}(\mu,\Sigma)$,

$$\psi_X^*(t)=\frac{1}{2}(t-\mu)^\top\Sigma^{-1}(t-\mu)$$

Proof. $J(\lambda)=\lambda^\top t-\psi_X(\lambda)$ is concave in $\lambda\in\mathbb{R}^d$ since ψ_X is convex. The stationary condition is

$$\nabla J(\lambda)=t-\nabla\psi_X(\lambda)=t-\mu-\Sigma\lambda=0_d$$

Thus $\lambda^*=\Sigma^{-1}(t-\mu)$ is the maximizer of J. Then

$$\begin{aligned}\psi_X^*(t) &= (\lambda^*)^\top t-\psi_X(\lambda^*)\\ &= (\lambda^*)^\top t-(\lambda^*)^\top\mu-\frac{1}{2}(\lambda^*)^\top\Sigma\lambda^*\\ &= (t-\mu)^\top\Sigma^{-1}t-(t-\mu)^\top\Sigma^{-1}\mu-\frac{1}{2}(t-\mu)^\top\Sigma^{-1}(t-\mu)\\ &= \frac{1}{2}(t-\mu)^\top\Sigma^{-1}(t-\mu)\end{aligned}$$

□

Lemma E.25. If $X \sim \mathcal{G}(\sigma^2)$, then $\mathrm{Var}\,(X) \le \sigma^2$.

Proof. By the Taylor series of $e^z = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \cdots$,

$$f(t) := \mathbf{E}[e^{tX}] = \mathbf{E}\left[1 + tX + \frac{t^2X^2}{2} + \frac{t^3X^3}{6} + \cdots\right] = 1 + \frac{t^2\mathbf{E}\left[X^2\right]}{2} + t^3P_1(t)$$
$$g(t) := \mathbf{E}[e^{\frac{\sigma^2t^2}{2}}] = 1 + \frac{\sigma^2t^2}{2} + \frac{\sigma^4t^4}{4} + \cdots = 1 + \frac{\sigma^2t^2}{2} + t^3P_2(t)$$

where P_1, P_2 are some polynomials. By premise, for all $t \in \mathbb{R}$

$$\begin{aligned} f(t) \le g(t) \quad &\Leftrightarrow \quad \frac{t^2\mathbf{E}\left[X^2\right]}{2} + t^3P_1(t) \le \frac{\sigma^2t^2}{2} + t^3P_2(t) \\ &\Leftrightarrow \quad \mathbf{E}[X^2] - \sigma^2 \le tG(t) \end{aligned}$$

where G is again some polynomial. Thus

$$\mathbf{E}[X^2] - \sigma^2 \le \lim_{t\to 0} tG(t) = 0 \quad \Leftrightarrow \quad \mathbf{E}[X^2] \le \sigma^2$$

□

Lemma E.26. If $X, Z \in \mathbb{R}$ are random variables with the CGFs $\psi_X, \phi_Z : \mathbb{R} \to \mathbb{R}$,

$$\psi_X(t) \le \phi_Z(t) \quad \forall t \in \mathbb{R} \quad \Rightarrow \quad \exp(-\psi_X^*(t)) \le \exp(-\phi_Z^*(t)) \quad \forall t \in \mathbb{R}$$

where $\psi_X^*(t) = \sup_{\lambda\in\mathbb{R}} \lambda t - \psi_X(t)$ is the Legendre transform of ψ_X (similarly for $\psi_Z^*(t)$).

Proof.

$$\begin{aligned} \psi_X(t) \le \phi_Z(t) \quad &\Leftrightarrow \quad -\psi_X(t) \ge -\phi_Z(t) \\ &\Leftrightarrow \quad \lambda t - \psi_X(t) \ge \lambda t - \phi_Z(t) \qquad \forall \lambda \in \mathbb{R} \\ &\Rightarrow \quad \sup_{\lambda\in\mathbb{R}} \lambda t - \psi_X(t) \ge \sup_{\lambda\in\mathbb{R}} \lambda t - \phi_Z(t) \\ &\Leftrightarrow \quad \psi_X^*(t) \ge \phi_Z^*(t) \\ &\Leftrightarrow \quad -\psi_X^*(t) \le -\phi_Z^*(t) \\ &\Leftrightarrow \quad \exp(-\psi_X^*(t)) \le \exp(-\phi_Z^*(t)) \end{aligned}$$

□

Lemma E.27. If $X_1 \dots X_N$ are independently sub-Gaussian with $X_i \sim \mathcal{G}(\sigma_i^2)$, then for all $\epsilon \ge 0$:

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^N X_i\right| \ge \epsilon\right) \le 2\exp\left(-\frac{N^2\epsilon^2}{2\left(\sum_{i=1}^N \sigma_i^2\right)}\right)$$

Proof. Let $\widetilde{X} := \sum_{i=1}^N X_i$. Note that $\widetilde{X} \sim \mathcal{G}(\sum_{i=1}^N \sigma_i^2)$ (4) and $-\widetilde{X} \sim \mathcal{G}(\sum_{i=1}^N \sigma_i^2)$ (2). Thus

$$\begin{aligned} \Pr\left(\left|\frac{1}{N}\widetilde{X}\right| \ge \epsilon\right) &= \Pr\left(\frac{1}{N}\widetilde{X} \le -\epsilon \vee \frac{1}{N}\widetilde{X} \ge \epsilon\right) \\ &\le \Pr\left(\frac{1}{N}\widetilde{X} \le -\epsilon\right) + \Pr\left(\frac{1}{N}\widetilde{X} \ge \epsilon\right) \qquad \text{(union bound)} \\ &= \Pr\left(-\widetilde{X} \ge N\epsilon\right) + \Pr\left(\widetilde{X} \ge N\epsilon\right) \\ &\le 2\exp\left(-\frac{N^2\epsilon^2}{2\left(\sum_{i=1}^N \sigma_i^2\right)}\right) \qquad (3) \end{aligned}$$

□

F Individually Normal But Not Jointly Normal

This is an example from Wikipedia. Let $X \sim \mathcal{N}(0,1)$ and, independently, $\epsilon \sim R$ where R denotes the Rademacher distribution. Let $Y = \epsilon X$. By the symmetry of the distribution of X, we have $Y \sim \mathcal{N}(0,1)$. More formally,

$$\begin{aligned}\Pr(Y \le x) &= \Pr(\epsilon = 1)\Pr(X \le x) + \Pr(\epsilon = -1)\Pr(X \ge -x) \\ &= \Pr(\epsilon = 1)\Pr(X \le x) + \Pr(\epsilon = -1)\Pr(-X \le x) \\ &= \frac{1}{2}\Pr(X \le x) + \frac{1}{2}\Pr(X \le x) \\ &= \Pr(X \le x)\end{aligned}$$

Let $Z = X + Y$. Then $Z = 0$ with probabilty $\frac{1}{2}$ and $Z = 2X$ with probability $\frac{1}{2}$, so

$$\Pr(Z = z) = \frac{1}{2}\left([[z = 0]] + \mathcal{N}(0,1)\left(\frac{z}{2}\right)\right) \tag{68}$$

which is not a normal distribution. Then by definition 4, $(X, Y) \in \mathbb{R}^2$ is not normally distributed. Thus X and Y are not jointly normal, even though they are individually normal.

Mutual information. X and Y are uncorrelated. More formally,

$$\operatorname{Cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y] = \mathbf{E}\left[\epsilon X^2\right] = \mathbf{E}[\epsilon]\,\mathbf{E}\left[X^2\right] = 0$$

Thus $\operatorname{cor}(X,Y) = 0$. But X and Y are not independent. Specifically, $\Pr(Y = x | X = x) = \frac{1}{2}$ is not equal to $\Pr(Y = x) = \mathcal{N}(0,1)(x)$ for any $x \in \mathbb{R}$. This illustrates the limitation of linear correlation. On the other hand, the mutual information between X and Y is positive:

$$I(X,Y) = H(X) - H(X|Y) = H(X) - \log(2) = \log\sqrt{\frac{\pi e}{2}} \approx 0.73$$