17 Maximum Entropy Distributions: Maximally Ignorant Priors

As we noted in our development of model capacity, the least informative prior is the solution to a potentially high-dimensional convex optimization problem, and one that depends on $\mathcal{Y}$, which can be challenging in practice. At the same time, we noted that the detailed structure of this prior is not critical to achieving good performance; what is most important is that it be a strictly positive distribution.

In this section, we take advantage of this flexibility to choose a prior that is both simpler to design and independent of $\mathcal{Y}$, and thus is an attractive alternative to the least informative prior. In particular, we can choose the prior (weights) for which the expected inference loss in the absence of data is as large as possible. We refer to this as the "maximally ignorant prior." From our development of inference as a decision, this is the prior p with the largest possible entropy $H(p)$. In essence, this is the most random distribution, and reflects the maximum uncertainty about the correct model in the class.

We now show how to obtain such *maximum entropy distributions* subject to any number of linear constraints of interest, which can be used to capture additional knowledge that may be available in the problem of interest. Conveniently, these distributions are readily obtained from information geometric analysis.

17.1 Finite Alphabet Case

We start with the observation that for a finite alphabet $\mathcal{Y}$,

$$D(p\|\mathtt{U}) = \sum_y p(y)\log p(y) + \log|\mathcal{Y}| = \log|\mathcal{Y}| - H(p), \tag{1}$$

where, as per our convention, $\mathtt{U}$ denotes the uniform distribution (over $\mathcal{Y}$). Hence, maximizing entropy is equivalent to minimizing information divergence from the uniform distribution.

Next, we capture the desired constraints via a linear family. In particular, we express the constraints in the form

$$\mathbb{E}_p\left[t_k(\mathsf{y})\right] = \bar{t}_k, \quad k = 1, 2, \ldots, K \tag{2}$$

for some K, $\mathbf{t}(\cdot) = [t_1(\cdot), \ldots, t_K(\cdot)]^{\mathrm{T}}$, and $\bar{\mathbf{t}} = [\bar{t}_1, \ldots, \bar{t}_K]^{\mathrm{T}}$. With this notation, the linear family is, as in our development of information geometry,

$$\mathcal{L}_{\mathbf{t}}(p^*) = \left\{p \in \mathcal{P}^{\mathcal{Y}} \colon\ \mathbb{E}_p\left[t_k(\mathsf{y})\right] = \bar{t}_k, \quad k = 1, 2, \ldots, K\right\}. \tag{3}$$

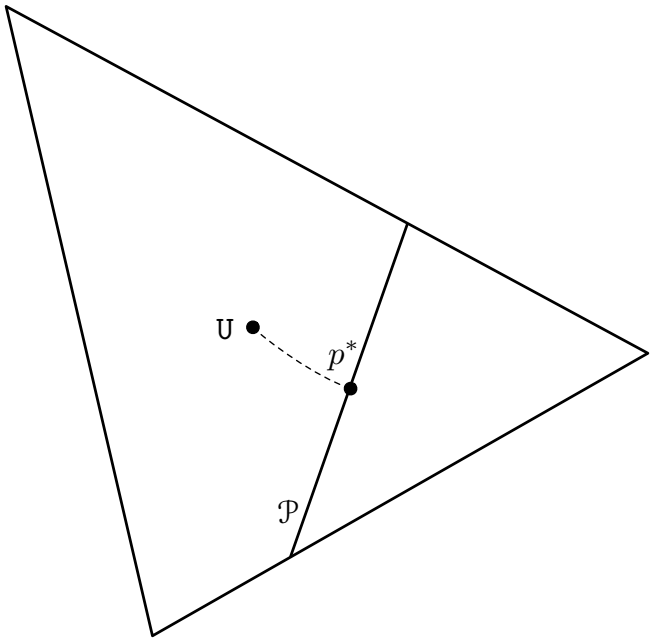

Figure 1: The maximum entropy distribution corresponds to the I-projection the uniform distribution $\mathtt{U}$ (located at the center of the simplex) onto the linear family $\mathcal{P}$ describing the linear constraints. In this example, $\mathcal{Y} = 1, 2, 3$ and there is a single linear constraint, so $\mathcal{P}$ is a line.

where p^* denotes the distribution of maximum entropy in this family.

Via (1) and (3), this maximum entropy distribution can thus be expressed in the form

$$p^* = \arg\max_{p \in \mathcal{L}_{\mathbf{t}}} H(p) = \arg\min_{p \in \mathcal{L}_{\mathbf{t}}} D(p \| \mathtt{U}), \tag{4}$$

i.e., the maximum entropy distribution p^* is the distribution in the family $\mathcal{L}_{\mathbf{t}}(p^*)$ that is the closest to the uniform distribution in the sense of information divergence. This relationship is depicted in Fig. 1.

From the characterization (4), we obtain via our I-projection results that the maximum entropy distribution p^* can be obtained as the member of the exponential family $\mathcal{E}_{\mathbf{t}}(\mathtt{U})$, which includes the uniform distribution and is orthogonal to the linear family $\mathcal{L}_{\mathbf{t}}(p^*)$. Since $\mathtt{U}$ can be expressed in the form $q(y) \propto e^{\beta(y)}$ with $\beta \equiv 0$, this means that p^* takes the form

$$p^*(y) = \exp\left\{ \sum_{i=1}^{K} x_i t_i(y) - \alpha(\mathbf{x}) \right\}, \tag{5}$$

where $\mathbf{x} = [x_1, \ldots, x_K]^{\mathrm{T}}$ is chosen to satisfy the constraints (2).

We illustrate the mechanics of this methodology through a variety of examples.

Example 1. Consider the case $K = 0$, so $\mathcal{L}_{\mathbf{t}}(p^*)$ is the whole probability simplex. Then the corresponding exponential family $\mathcal{E}_{\mathbf{t}}(\mathtt{U})$ is the set of distributions such that $p(y)$ is constant. This is another way to confirm a result easily proved by brute-force optimization—that the maximum entropy distribution for the finite alphabet case is uniform.

Example 2. Consider a discrete case with alphabet $\mathcal{Y} = \{1, \ldots, M\}$ and $M > 2$, where the constraint is that $\mathbb{P}(\mathsf{y} = 1) = 1/2$. This constraint can be expressed in the form (2) with $K = 1$, $t_1(y) = \mathbb{1}_{y=1}$, and $\bar{t}_1 = 1/2$. Hence,

$$p^*(y) = e^{x\,\mathbb{1}_{y=1} - \alpha(x)} \propto \begin{cases} e^x & y = 1, \\ 1 & y \neq 1. \end{cases} \tag{6}$$

Note that (6) establishes all values of y except 1 are equiprobable. Thus we choose x so that $p^*(1) = 1/2$ and distribute the remaining probability uniformly over the other symbols in the alphabet $\mathcal{Y}$.

Even richer examples are possible when we consider multivariate distributions. Two examples are as follows.

Example 3. Consider a pair of jointly distributed random variables $\mathsf{y}_1, \mathsf{y}_2$, each defined over the alphabet $\mathcal{Y} = \{1, \ldots, M\}$, where the constraints are that the marginal distributions $p_{\mathsf{y}_1}(\cdot)$ and $p_{\mathsf{y}_2}(\cdot)$ are prescribed and strictly positive.

In this case, the constraints can be expressed in the form

$$t_i^{(1)}(y_1, y_2) = \mathbb{1}_{y_1 = i}, \quad \bar{t}_i^{(1)} = p_{\mathsf{y}_1}(i), \quad i = 1, \ldots, M, \tag{7}$$

$$t_i^{(2)}(y_1, y_2) = \mathbb{1}_{y_2 = i}, \quad \bar{t}_i^{(2)} = p_{\mathsf{y}_2}(i), \quad i = 1, \ldots, M. \tag{8}$$

As a result, the joint distribution of maximum entropy has the form

$$p^*(y_1, y_2) \propto \exp\left\{ \sum_{i=1}^{M} x_i^{(1)} t_i^{(1)}(y_1, y_2) + \sum_{i=1}^{M} x_i^{(2)} t_i^{(2)}(y_1, y_2) \right\} \tag{9}$$

$$= \exp\left\{ \sum_{i=1}^{M} x_i^{(1)} \mathbb{1}_{y_1 = i} + \sum_{i=1}^{M} x_i^{(2)} \mathbb{1}_{y_2 = i} \right\} \tag{10}$$

$$= \exp\left\{x_{y_1}^{(1)}\right\} \exp\left\{x_{y_2}^{(2)}\right\} \tag{11}$$

Note that since (11) takes the form of a product of a function of y_1 and a function of y_2, it follows immediately that the maximum entropy distribution is the one in which y_1 and y_2 are independent and have the given marginals.

To develop this conclusion systematically from (11), it suffices to solve for parameters that ensure the constraints are satisfied. Specifically

$$p^*(y_1) = \sum_{y_2} p^*(y_1, y_2) \propto \exp\left\{x_{y_1}^{(1)}\right\} \sum_{y_2} \exp\left\{x_{y_2}^{(2)}\right\} \propto \exp\left\{x_{y_1}^{(1)}\right\} \tag{12}$$

$$p^*(y_2) = \sum_{y_1} p^*(y_1, y_2) \propto \exp\left\{x_{y_2}^{(2)}\right\} \sum_{y_1} \exp\left\{x_{y_1}^{(1)}\right\} \propto \exp\left\{x_{y_2}^{(2)}\right\}, \tag{13}$$

so choosing

$$x_{y_1}^{(1)} = \ln p_{\mathsf{y}_1}(y_1)$$
$$x_{y_2}^{(2)} = \ln p_{\mathsf{y}_2}(y_2)$$

yields

$$p^*(y_1, y_2) = p_{\mathsf{y}_1}(y_1)\, p_{\mathsf{y}_2}(y_2).$$

Example 4. Consider a triple of jointly distributed random variables $\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3$, each defined over the alphabet $\mathcal{Y} = \{1, \ldots, M\}$, where the constraints are that the pairwise marginal distributions $p_{\mathsf{y}_1,\mathsf{y}_2}(\cdot,\cdot)$ and $p_{\mathsf{y}_2,\mathsf{y}_3}(\cdot,\cdot)$ are prescribed and strictly positive.

In this case, the constraints can be expressed in the form

$$t_{i,j}^{(\mathrm{A})}(y_1, y_2, y_3) = \mathbb{1}_{y_1=i, y_2=j}, \quad \bar{t}_{i,j}^{(\mathrm{A})} = p_{\mathsf{y}_1,\mathsf{y}_2}(i,j), \quad i,j = 1, \ldots, M, \tag{14}$$
$$t_{k,l}^{(\mathrm{B})}(y_1, y_2, y_3) = \mathbb{1}_{y_2=k, y_3=l}, \quad \bar{t}_{k,l}^{(\mathrm{B})} = p_{\mathsf{y}_2,\mathsf{y}_3}(k,l), \quad k,l = 1, \ldots, M. \tag{15}$$

As a result, the joint distribution of maximum entropy has the form

$$p^*(y_1, y_2, y_3) \propto \exp\left\{ \sum_{i,j=1}^{M} x_{i,j}^{(\mathrm{A})} t_{i,j}^{(\mathrm{A})}(y_1, y_2, y_3) + \sum_{k,l=1}^{M} x_{k,l}^{(\mathrm{B})} t_{k,l}^{(\mathrm{B})}(y_1, y_2, y_3) \right\} \tag{16}$$
$$= \exp\left\{ \sum_{i,j=1}^{M} x_{i,j}^{(\mathrm{A})} \mathbb{1}_{y_1=i, y_2=j} + \sum_{k,l=1}^{M} x_{k,l}^{(\mathrm{B})} \mathbb{1}_{y_2=k, y_2=l} \right\} \tag{17}$$
$$= \underbrace{\exp\left\{x_{y_1,y_2}^{(\mathrm{A})}\right\}}_{\triangleq \phi_{\mathrm{A}}(y_1,y_2)} \underbrace{\exp\left\{x_{y_2,y_3}^{(\mathrm{B})}\right\}}_{\triangleq \phi_{\mathrm{B}}(y_2,y_3)}. \tag{18}$$

The factored form (18) implies that the maximum entropy distribution is the one such that $\mathsf{y}_1 \leftrightarrow \mathsf{y}_2 \leftrightarrow \mathsf{y}_3$ form a Markov chain. To see this, it suffices to note that

$$p^*(y_3|y_2, y_1) = \frac{p^*(y_1, y_2, y_3)}{\sum_{y_3'} p^*(y_1, y_2, y_3')} \propto \frac{\phi_{\mathrm{A}}(y_1, y_2)\, \phi_{\mathrm{B}}(y_2, y_3)}{\phi_{\mathrm{A}}(y_1, y_2) \sum_{y_3'} \phi_{\mathrm{B}}(y_2, y_3')} = \frac{\phi_{\mathrm{B}}(y_2, y_3)}{\sum_{y_3'} \phi_{\mathrm{B}}(y_2, y_3')}, \tag{19}$$

which we note does not depend on y_1, whence

$$p^*(y_3|y_2, y_1) = p^*(y_3|y_2).$$

In turn, since the Markov relationship implies that y_1 and y_3 are independent conditioned on y_2, we can express p^* in terms of the constraints as follows (among other possibilities):

$$p^*(y_1, y_2, y_3) = \underbrace{p_{\mathsf{y}_1,\mathsf{y}_2}(y_1, y_2)}_{=\phi_{\mathrm{A}}(y_1,y_2)} \cdot \underbrace{\frac{p_{\mathsf{y}_2,\mathsf{y}_3}(y_2, y_3)}{\sum_{y_3'} p_{\mathsf{y}_2,\mathsf{y}_3}(y_2, y_3')}}_{=\phi_{\mathrm{B}}(y_2,y_3)} \quad \text{or} \quad \underbrace{\frac{p_{\mathsf{y}_1,\mathsf{y}_2}(y_1, y_2)}{\sum_{y_1'} p_{\mathsf{y}_1,\mathsf{y}_2}(y_1', y_2)}}_{=\phi_{\mathrm{A}}(y_1,y_2)} \cdot \underbrace{p_{\mathsf{y}_2,\mathsf{y}_3}(y_2, y_3)}_{=\phi_{\mathrm{B}}(y_2,y_3)}.$$

17.2 Infinite Alphabet Case

The derivation in Section 17.1 cannot be used for the case of continuous random variables both because we haven't developed information geometry for this case, and because the uniform distribution need not be well-defined (such as when $\mathcal{Y} = \mathbb{R}$). Nevertheless, analogous results on maximum entropy distributions *do* hold for the continuous case, where we replace the entropy $H(p)$ with the differential entropy $h(p)$. Specifically, (5) is the form of the maximum differential entropy distribution, as the following claim establishes.

Claim 1. *For distributions over an alphabet $\mathcal{Y} \subset \mathbb{R}$, the distribution p^* given by* (5)*, when it exists,*[1] *is the unique distribution in the linear family* (3) *with maximum differential entropy.*

Proof. It suffices to note that for any $p \in \mathcal{L}_{\mathbf{t}}(p^*)$ we have

$$\begin{aligned}
h(p) &= -\int p(y) \log p(y) \, \mathrm{d}y \\
&= \underbrace{-\int p(y) \log \frac{p(y)}{p^*(y)} \, \mathrm{d}y}_{=-D(p\|p^*)} - \int p(y) \log p^*(y) \, \mathrm{d}y \\
&\leq -\int p(y) \log p^*(y) \, \mathrm{d}y && (20) \\
&= -\int p(y) \left[\sum_{i=1}^{K} x_i t_i(y) - \alpha(\mathbf{x}) \right] \mathrm{d}y && (21) \\
&= -\sum_{i=1}^{K} x_i \, \mathbb{E}_p \left[t_i(\mathsf{y}) \right] - \alpha(\mathbf{x}) \\
&= -\sum_{i=1}^{K} x_i \, \mathbb{E}_{p^*} \left[t_i(\mathsf{y}) \right] - \alpha(\mathbf{x}) && (22) \\
&= -\int p^*(y) \left[\sum_{i=1}^{K} x_i t_i(y) - \alpha(\mathbf{x}) \right] \mathrm{d}y \\
&= -\int p^*(y) \log p^*(y) \, \mathrm{d}y && (23) \\
&= h(p^*),
\end{aligned}$$

where to obtain (20) we have used that information divergence is nonnegative, where to obtain (21) we have used (5), where to obtain (22) we have used that since $p, p^* \in \mathcal{L}_{\mathbf{t}}(p^*)$ we have, via (3), $\mathbb{E}_p \left[t_i(\mathsf{y}) \right] = \mathbb{E}_{p^*} \left[t_i(y) \right] = \bar{t}_i$ for $i = 1, \ldots, K$, and where to

[1] Specifically, it may or may not be possible to normalize (5).

obtain (23) we have again used (5). The uniqueness of p^* follows from the fact that the inequality in (20) holds with equality if and only if $D(p\|p^*) = 0$, which happens if and only if $p = p^*$. □

Note that it is straightforward to adapt the preceding result to the case of discrete distributions over infinite alphabets. In particular, a directly analogous argument can be used to establish the following version of Claim 1, which generalizes our result for finite alphabets.

Claim 2. *For discrete distributions, the distribution p^* given by* (5)*, when it exists, is the unique distribution in the linear family* (3) *with maximum entropy.*

The following examples demonstrate some applications.

Example 5. Let us revisit Example 1 in the discrete case for the infinite alphabet $\mathcal{Y} = \{1, 2, \dots\}$. Since there are no constraints, (5) tells us that if p^* exists, it must be constant for all $y \in \mathcal{Y}$. However, such a p^* cannot be normalized, and thus p^* does not exist.[2] That no maximum entropy distribution exists in this case is reflected in the fact that the sequence of distributions

$$p^{(L)}(y) = \begin{cases} 1/L & 1 \le y \le L \\ 0 & \text{otherwise} \end{cases}, \qquad L = 1, 2, \dots$$

has entropy that increases without bound in L:

$$H(p^{(L)}) = \log(L) \to \infty, \quad \text{as } L \to \infty.$$

Example 6. Consider a continuous alphabet case for which

$$\mathcal{L}_{\mathbf{t}}(p^*) = \left\{p\colon\ \mathbb{E}\left[\mathsf{y}\right] = \mu, \quad \mathbb{E}\left[\mathsf{y}^2\right] = \sigma^2 + \mu^2\right\}.$$

Here $K = 2$ with $t_1(y) = y$ and $t_2(y) = y^2$. Here it follows that p^* must be of the exponential-quadratic form

$$p^*(y) = e^{x_1 y + x_2 y^2 - \alpha(x_1, x_2)},$$

i.e., y must be Gaussian. We solve for x_1 and x_2 from the mean and variance constraints, and $\alpha(x_1, x_2)$ is the normalization; specifically,

$$p^*(y) = \mathtt{N}(y; \mu, \sigma^2).$$

It is also straightforward to determine the resulting differential entropy via

$$\begin{aligned} h(p^*) &= -\int_{-\infty}^{+\infty} \mathtt{N}(y; \mu, \sigma^2) \left[-\log(e)\frac{1}{2\sigma^2}(y-\mu)^2 - \frac{1}{2}\log(2\pi\sigma^2)\right] \mathrm{d}y \\ &= \log(e)\frac{1}{2\sigma^2}\sigma^2 + \frac{1}{2}\log(2\pi\sigma^2) \\ &= \frac{1}{2}\log(2\pi e\sigma^2). \end{aligned} \tag{24}$$

[2]This is an instance of what is sometimes referred to as an *improper* distribution.

Example 7. Now consider a continuous-alphabet case with support constraints, i.e.,

$$\mathcal{L}_{\mathbf{t}}(p^*) = \{p\colon p(y) = 0 \text{ for } y < 0, \text{ and } \mathbb{E}[\mathsf{y}] = \mu\}.$$

Here $K = 2$ with

$$\begin{aligned} t_1(y) &= y, \quad \bar{t}_1 = \mu \\ t_2(y) &= \mathbb{1}_{y<0}, \quad \bar{t}_2 = 0. \end{aligned}$$

It follows that p^* must be of the form

$$p^*(y) = \begin{cases} e^{xy-\alpha(x)} & y \geq 0, \\ 0 & y < 0. \end{cases}$$

i.e., y must be an exponential random variable, where x is determined by the mean constraint; specifically,

$$p^*(y) = \frac{1}{\mu} e^{-y/\mu}, \quad y \geq 0.$$

The resulting differential entropy is then

$$\begin{aligned} h(p^*) &= -\int_{-\infty}^{+\infty} \frac{1}{\mu} e^{-y/\mu} \left[-\log(e)\frac{y}{\mu} - \log \mu \right] \\ &= \log(e)\frac{\mu}{\mu} + \log \mu \\ &= \log(\mu e). \end{aligned}$$

Note that if we did not have the support constraint, the optimization would have no solution, corresponding to an improper distribution.

Finally, we develop a multivariate example.

Example 8. Consider a collection of continuous random variables $\mathbf{y} = [\mathsf{y}_1, \ldots, \mathsf{y}_N]^{\mathrm{T}}$ with $\mathcal{Y} = \mathbb{R}^N$. If the constraints are $\mathbb{E}\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right] = \mathbf{K} > 0$,[3] i.e., $\mathbb{E}[\mathsf{y}_i \mathsf{y}_j] = [\mathbf{K}]_{ij}$ for $1 \leq i, j, \leq N$, then

$$t_{i,j}(y_1, \ldots, y_N) = y_i y_j, \quad i, j \in \{1, \ldots, N\}$$

so (5) implies the maximum differential entropy distribution for $\mathbf{y}$ must be of the form

$$p^*(\mathbf{y}) = \exp\left\{ \sum_{i,j=1}^{N} x_{i,j}\, y_i y_j - \alpha(\mathbf{X}) \right\} = \exp\left\{ \mathbf{y}^{\mathrm{T}} \mathbf{X} \mathbf{y} - \alpha(\mathbf{X}) \right\}, \tag{25}$$

[3]Recall that $\mathbf{K} > 0$ means $\mathbf{K}$ is positive definite and thus invertible.

where $[\mathbf{X}]_{i,j} = x_{i,j}$. However, the exponent in (25) is a quadratic form in $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$, so p^* must be a multivariate Gaussian. Moreover, comparing (25) to the general form of a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Lambda}$, viz.,

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]}{|2\pi\boldsymbol{\Lambda}|^{1/2}},$$

we see $\boldsymbol{\mu} = \mathbf{0}$, and thus $\boldsymbol{\Lambda} = \mathbf{K}$, which we obtain by the choice $\mathbf{X} = \mathbf{K}^{-1}/2$.

Finally, for the resulting $p^*(\mathbf{y}) = \mathtt{N}(\mathbf{y}; \mathbf{0}, \mathbf{K})$ we have

$$\begin{aligned}
h(p^*) &= \int_{-\infty}^{+\infty} p^*(\mathbf{y}) \left[-\frac{1}{2}\log(e)\,\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}\log|2\pi\mathbf{K}|\right] \mathrm{d}\mathbf{y} \\
&= -\frac{1}{2}\log(e)\,\mathbb{E}_{p^*}\left[\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}\right] - \frac{1}{2}\log|2\pi\mathbf{K}| \\
&= -\frac{1}{2}N\log(e) - \frac{1}{2}\log|2\pi\mathbf{K}| \qquad (26) \\
&= \frac{1}{2}\log\left(|2\pi e\mathbf{K}|\right), \qquad (27)
\end{aligned}$$

where to obtain (26) we have used the matrix trace identity $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$ whenever the dimensions of $\mathbf{A}$ and $\mathbf{B}$ are such that both traces are defined. In particular, we use this trace identity to see

$$\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} = \mathrm{tr}\left(\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}\right) = \mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{y}\mathbf{y}^{\mathrm{T}}\right),$$

from which we obtain

$$\mathbb{E}_{p^*}\left[\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}\right] = \mathbb{E}_{p^*}\left[\mathrm{tr}(\mathbf{K}^{-1}\mathbf{y}\mathbf{y}^{\mathrm{T}})\right] = \mathrm{tr}\left(\mathbf{K}^{-1}\,\mathbb{E}_{p^*}\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right]\right) = \mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{K}\right) = \mathrm{tr}(\mathbf{I}) = N.$$