A Derivation of Eq. (10)

To show that Eq. (10) does indeed follow from Eq. (8), we need to compute the mean and covariance of $\delta\mu_i$, and the derivatives of $S_q(\boldsymbol{\mu})$ with respect to μ_i. We start with the former. The mean of $\delta\mu_i$, which is given by (see Eq. (7) and (9))

$$\langle \delta\mu_i \rangle = \frac{1}{K} \sum_k \langle g_i(\mathbf{x}^{(k)}) \rangle_p - \langle g_i(\mathbf{x}) \rangle_p = 0 \,. \tag{A.1}$$

The covariance can be computed by noting that $\delta\mu_i$ is the mean of K uncorrelated, zero mean random variables (see Eq. (9)), which implies that

$$\langle \delta g_i \delta g_j \rangle_p = \frac{1}{K} \left[\langle g_i(\mathbf{x}) g_j(\mathbf{x}) \rangle_p - \langle g_i(\mathbf{x}) \rangle_p \langle g_j(\mathbf{x}) \rangle_p \right] = \frac{C^p_{ij}}{K} \tag{A.2}$$

where the last equality follows from the definition given in Eq. (11a).

We next compute derivatives of the entropy with respect to the μ_i. Using Eq. (6) for the entropy, we have

$$\frac{\partial S_q(\boldsymbol{\mu})}{\partial \mu_i} = \frac{\partial \log Z(\boldsymbol{\mu})}{\partial \mu_j} - \lambda_i - \sum_j \mu_j \frac{\partial \lambda_j}{\partial \mu_i} \,. \tag{A.3}$$

From the definition of $\log Z(\boldsymbol{\mu})$, Eq. (5), it is straightforward to show that

$$\frac{\partial \log Z(\boldsymbol{\mu})}{\partial \mu_i} = \sum_j \mu_j \frac{\partial \lambda_j}{\partial \mu_i} \tag{A.4}$$

Inserting Eq. (A.4) into (A.3), the first and third terms cancel, and we are left with

$$\frac{\partial S_q(\boldsymbol{\mu})}{\partial \mu_i} = -\lambda_i \,. \tag{A.5}$$

The second derivative of the entropy is thus trivial,

$$\frac{\partial^2 S_q(\boldsymbol{\mu})}{\partial \mu_i \partial \mu_j} = -\frac{\partial \lambda_i}{\partial \mu_j} \,. \tag{A.6}$$

This quantity is hard to compute, so instead we compute its inverse, $\partial\mu_j/\partial\lambda_i$. Using the definition of μ_j,

$$\mu_j = \sum_{\mathbf{x}} g_j(\mathbf{x}) \frac{\exp\left[\sum_i \lambda_i g_i(\mathbf{x})\right]}{Z(\boldsymbol{\mu})} \,, \tag{A.7}$$

differentiating both sides with respect to λ_i, and applying Eq. (A.4), we find that

$$\frac{\partial \mu_j}{\partial \lambda_i} = \langle g_i(\mathbf{x}) g_j(\mathbf{x}) \rangle_{q(\mathbf{x}|\boldsymbol{\mu})} - \langle g_i(\mathbf{x}) \rangle_{q(\mathbf{x}|\boldsymbol{\mu})} \langle g_j(\mathbf{x}) \rangle_{q(\mathbf{x}|\boldsymbol{\mu})} = C^q_{ij} \,. \tag{A.8}$$

The right hand side is the covariance matrix within the model class.

Combining Eq. (A.6) with (A.8) and noting that

$$\frac{\partial \lambda_i}{\partial \lambda_{i'}} = \sum_j \frac{\partial \lambda_i}{\partial \mu_j} \frac{\partial \mu_j}{\partial \lambda_{i'}} = \delta_{ii'} \quad \Rightarrow \quad \frac{\partial \lambda_i}{\partial \mu_j} = C^{q^{-1}}_{ij} \,, \tag{A.9}$$

we have

$$\frac{\partial^2 S_q(\boldsymbol{\mu})}{\partial \mu_i \partial \mu_j} = -C^{q^{-1}}_{ij} \,. \tag{A.10}$$

Inserting Eqs. (A.1), (A.1), (A.5) and (A.10) into (8), we arrive at Eq. (10).

B Alternative derivation of the within-model class bias

We present a brief alternative derivation of the within-class bias from classical results about the asymptotic distribution of maximum likelihood estimators. Suppose that $X_K = \{\mathbf{x}^k\}_{k=1,\dots K}$ is a sample of size K from the model $q(\mathbf{x}|\boldsymbol{\lambda})$ with true parameter $\boldsymbol{\lambda}$, and that $L(\boldsymbol{\lambda}') = \sum_k \log q(\mathbf{x}^k|\boldsymbol{\lambda}')$ is the likelihood of some parameters $\boldsymbol{\lambda}'$ given the data. Then, it can be shown that the asymptotic distribution of (twice) the difference between the true log-likelihood $L(\boldsymbol{\lambda})$ and the log-likelihood of a maximum likelihood-estimate $\hat{\boldsymbol{\lambda}} = \text{argmax}_{\lambda'} L(\boldsymbol{\lambda}')$ has a Chi-square distribution with m degrees of freedom (where m is the number of parameters, the dimensionality of the vector $\boldsymbol{\lambda}$) [20],

$$2\left(L(\hat{\boldsymbol{\lambda}}) - L(\boldsymbol{\lambda})\right) \sim \chi^2_m. \tag{B.1}$$

As the mean of a random variable with distribution χ^2_m is simply m, this implies that the bias in the estimate of the log-likelihood is $\langle (L(\hat{\boldsymbol{\lambda}}) - L(\boldsymbol{\lambda})\rangle_{q(\mathbf{x}|\lambda} = \frac{1}{2}m$. Using the duality between maximum-entropy estimation and maximum likelihood estimation in exponential family models, we can now derive the entropy bias from the likelihood bias: maximizing the entropy subject to the empirically measured moments $\hat{\boldsymbol{\mu}}$ is equivalent to maximizing the likelihood of model (4).

This means that maximum entropy model $q(\mathbf{x}|\boldsymbol{\mu})$, which matches the empirical means $\hat{\boldsymbol{\mu}}$ in the data-set, is the same model whose parameters $\hat{\boldsymbol{\lambda}}$ maximize the likelihood $L(\boldsymbol{\lambda}')$, and here therefore we slightly abuse notation to use $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\mu}}$ interchangeably,

$$\begin{aligned}
\frac{1}{2}m &= \left\langle L(\hat{\boldsymbol{\lambda}}) - L(\boldsymbol{\lambda})\right\rangle_q \\
&= \left\langle \sum_k \log q(\mathbf{x}_k|\hat{\boldsymbol{\lambda}})\right\rangle_q - K\sum_x q(\mathbf{x}|\boldsymbol{\lambda})\log q(\mathbf{x}|\boldsymbol{\lambda}) \\
&= KS_q(\boldsymbol{\lambda}) + \left\langle \sum_k \hat{\boldsymbol{\lambda}}^\top g(\mathbf{x}_k) - \log(Z(\hat{\boldsymbol{\lambda}})\right\rangle_q \\
&= KS_q(\boldsymbol{\lambda}) - K\left\langle \log(Z(\hat{\boldsymbol{\lambda}}) - \hat{\boldsymbol{\lambda}}^\top\hat{\boldsymbol{\mu}}\right\rangle_q \\
&= K\langle S_q(\boldsymbol{\lambda}) - S_q(\hat{\boldsymbol{\lambda}})\rangle_q
\end{aligned} \tag{B.2}$$

Rearranging terms, we recover our result that $\text{Bias}[S] = -m/2K$.

C Calculating $b'(0)$

Here we compute $b'(0)$ (as in the main text, primes denote derivatives with respect to β). Recalling that $b(\beta) = \langle B(\mathbf{x})\rangle_{p(\mathbf{x}|\boldsymbol{\mu},\beta)}$, using the definition of $p(\mathbf{x}|\boldsymbol{\mu},\beta)$ given in Eq. (18), and making use of the relationship $\log Z'(\boldsymbol{\mu},\beta) = b + \sum_i \mu_i\lambda'_i(\boldsymbol{\mu},\beta)$, we have

$$b'(\beta) = \text{Var}[B]_{p(\mathbf{x}|\boldsymbol{\mu},\beta)} + \sum_{i-1}^{m}\langle B(\mathbf{x})\delta g_i(\mathbf{x})\rangle_{p(\mathbf{x}|\boldsymbol{\mu},\beta)}\lambda'_i(\boldsymbol{\mu},\beta) \tag{C.1}$$

where $\lambda'_i(\boldsymbol{\mu},\beta)$ denotes a derivative with respect to β.

To compute $\lambda'_i(\boldsymbol{\mu},\beta)$, we use the fact that $\langle g_i(\mathbf{x})\rangle_{p(\mathbf{x}|\boldsymbol{\mu},\beta)})$ is independent of β, which implies that

$$0 = \frac{d\langle g_i(\mathbf{x})\rangle_{p(\mathbf{x}|\boldsymbol{\mu},\beta)}}{d\beta} = \langle \delta g_i(\mathbf{x})B(\mathbf{x})\rangle_{p(\mathbf{x}|\boldsymbol{\mu},\beta)} + \sum_j \langle \delta g_i(\mathbf{x})\delta g_j(\mathbf{x})\rangle_{p(\mathbf{x}|\boldsymbol{\mu},\beta)}\lambda'_j(\beta)\,. \tag{C.2}$$

While we can't invert the matrix $\langle \delta g_i(\mathbf{x})\delta g_j(\mathbf{x})\rangle_{p(\mathbf{x}|\boldsymbol{\mu},\beta)}$ for arbitrary β, we can invert it when $\beta = 0$, since $\langle \delta g_i(\mathbf{x})\delta g_j(\mathbf{x})\rangle_{\beta=0} = C^q_{ij}$. Setting β to 0 in Eq. (C.2), we have

$$\lambda'_i(\boldsymbol{\mu},0) = -\sum_j C^{q^{-1}}_{ij}\langle \delta g_j(\mathbf{x})B(\mathbf{x})\rangle_{q(\mathbf{x}|\boldsymbol{\mu})} \tag{C.3}$$

where we used the fact that $p(\mathbf{x}|\boldsymbol{\mu},0) = q(\mathbf{x}|\boldsymbol{\mu})$. Inserting this expression into Eq. (C.1), setting β to zero, and replacing $p(\mathbf{x}|\boldsymbol{\mu},0)$ with $q(\mathbf{x}|\boldsymbol{\mu})$, we recover Eq. (23).