APPENDIX: Lost Relatives of the Gumbel Trick

Here we provide proofs for the results stated in the main text, together with additional supporting lemmas required for these proofs.

A. Comparison of Gumbel and Exponential tricks

In Section 2.3.1 we analyzed the asymptotic efficiency of different estimators of Z by measuring their asymptotic variance. (As all our estimators in the full-rank perturbation setting are consistent, their bias is 0 in the limit of infinite data, and so this asymptotic variance equals the asymptotic MSE.) In the non-asymptotic regime, where an estimator $\hat{Z}$ is constructed from a finite set of M samples, we can analyze both the variance $\mathrm{var}(\hat{Z})$ and the bias $(\mathbb{E}[\hat{Z}] - Z)$ of the estimator. While in most cases these cannot be obtained analytically and there we can resort to an empirical evaluation, for the estimators stemming from the Gumbel and Exponential tricks analytical treatment turns out to be possible using standard methods.

A.1. Estimating Z

Gumbel trick The Gumbel trick yields an unbiased estimator for $\ln Z$, and we can turn it into a consistent estimator of Z by exponentiating it:

$$\hat{Z} := \exp\left(\frac{1}{M}\sum_{m=1}^{M} X_m\right) \qquad \text{where} \qquad X_1, \dots, X_M \overset{\text{iid}}{\sim} \mathrm{Gumbel}(-c + \ln Z).$$

Recalling that the moment generating function of a $\mathrm{Gumbel}(\mu)$ distribution is $G(t) = \Gamma(1-t)e^{\mu t}$, we can obtain by using independence of the samples:

$$\mathbb{E}[\hat{Z}] = \prod_{m=1}^{M} \mathbb{E}[e^{X_m/M}] = \left(\Gamma(1-1/M)e^{(\ln Z - c)/M}\right)^M = \Gamma(1-1/M)^M e^{-c} Z,$$
$$\mathbb{E}[\hat{Z}^2] = \prod_{m=1}^{M} \mathbb{E}[e^{2X_m/M}] = \left(\Gamma(1-2/M)e^{2(\ln Z - c)/M}\right)^M = \Gamma(1-2/M)^M e^{-2c} Z^2.$$

Therefore the squared bias, variance and MSE of the estimator $\hat{Z}$ are, respectively:

$$\begin{aligned}
\mathrm{bias}(\hat{Z})^2 &= (\mathbb{E}[\hat{Z}] - Z)^2 = Z^2\left(\Gamma(1-1/M)^M e^{-c} - 1\right), \\
\mathrm{var}(\hat{Z}) &= \mathbb{E}[\hat{Z}^2] - \mathbb{E}[\hat{Z}]^2 = Z^2\left(\Gamma(1-2/M)^M e^{-2c} - \Gamma(1-1/M)^{2M} e^{-2c}\right), \\
\mathrm{MSE}(\hat{Z}) &= \mathrm{bias}(\hat{Z})^2 + \mathrm{var}(\hat{Z}) = Z^2\left(\Gamma(1-2/M)^M e^{-2c} - 2\Gamma(1-1/M)^M e^{-c} + 1\right).
\end{aligned}$$

These formulas hold for $M > 2$ where the moment generating functions are defined. For $M = 1$ the estimator has infinite bias (and infinite variance), and for $M = 2$ it has infinite variance. Figure 1 (left) shows the functional dependence of $\mathrm{MSE}(\hat{Z})$ on the number of samples $M \geq 3$, in units of Z^2.

Exponential trick The Exponential trick yields an unbiased estimator of $1/Z$, and we can turn it into a consistent estimator of Z by inverting it:

$$\hat{Z} := \left(\frac{1}{M}\sum_{m=1}^{M} X_m\right)^{-1} \qquad \text{where} \qquad X_1, \dots, X_M \overset{\text{iid}}{\sim} \mathrm{Exp}(Z).$$

As $X_1, \dots, X_M$ are independent and exponentially distributed with identical rates Z, their sum follows the Gamma distribution with shape M and rate Z. Therefore the estimator $\hat{Z}$ can be written as $\hat{Z} = MY$, where $Y \sim \mathrm{InvGamma}(M, Z)$.

Recalling the mean and variance of the Inverse-Gamma distribution, we obtain:

$$\begin{aligned}
\mathrm{bias}(\hat{Z})^2 &= (\mathbb{E}[\hat{Z}] - Z)^2 = Z^2\left(\frac{M}{M-1} - 1\right) = Z^2\frac{1}{M-1},\\
\mathrm{var}(\hat{Z}) &= Z^2 M^2 \frac{1}{(M-1)^2(M-2)},\\
\mathrm{MSE}(\hat{Z}) &= \mathrm{bias}(\hat{Z})^2 + \mathrm{var}(\hat{Z}) = Z^2\frac{M-2+M^2}{(M-1)^2(M-2)} = Z^2\frac{M+2}{(M-1)(M-2)}.
\end{aligned}$$

Again these formulas hold for $M > 2$ where the relevant expectations are defined: for $M = 1$ the estimator has infinite bias, and for $M \in \{1, 2\}$ it has infinite variance. Figure 1 (left) shows the functional dependence of $\mathrm{MSE}(\hat{Z})$ on the number of samples $M \geq 3$, in units of Z^2. By inspecting the curves we observe that the Gumbel trick estimator requires roughly 45% more samples to yield the same MSE as the Exponential trick estimator.

A.2. Estimating $\ln Z$

A similar analysis can be performed for estimating $\ln Z$ rather than Z. In that case the Gumbel trick estimator of $\ln Z$ is unbiased and has variance (and thus MSE) equal to $\frac{1}{M}\frac{\pi^2}{6}$. On the other hand, the Exponential trick estimator is

$$\widehat{\ln Z} = -\ln\left(\frac{1}{M}\sum_{m=1}^{M} X_m\right) \qquad \text{where} \qquad X_1, \dots, X_M \overset{\text{iid}}{\sim} \mathrm{Exp}(Z).$$

Again $\sum_{m=1}^{M} X_m \sim \mathrm{Gamma}(M, Z)$ and by reference to properties of the Gamma distribution,

$$\begin{aligned}
\mathrm{bias}(\widehat{\ln Z})^2 &= (\mathbb{E}[\hat{Z}] - Z)^2 = (\ln M - (\psi(M) - \ln Z) - \ln Z)^2 = (\ln M - \psi(M))^2,\\
\mathrm{var}(\widehat{\ln Z}) &= \psi_1(M),\\
\mathrm{MSE}(\widehat{\ln Z}) &= \mathrm{bias}(\widehat{\ln Z})^2 + \mathrm{var}(\widehat{\ln Z}) = (\ln M - \psi(M))^2 + \psi_1(M),
\end{aligned}$$

where $\psi(\cdot)$ is the digamma function and $\psi_1(\cdot)$ is the trigamma function. Note that the estimator can be debiased by subtracting its bias $(\ln M - \psi(M))$. Figure 1 (right) compares the MSE of the Gumbel and Exponential trick estimators of $\ln Z$. We observe that the Gumbel trick estimator performs better only for $M = 1$, and even in that case the Exponential trick estimator is better when debiased.

B. Sum-unary perturbations

Recall that *sum-unary perturbations* refer to the setting where each variable's unary potentials are perturbed with Gumbel noise, and the perturbed potential of a configuration sums the perturbations from all variables (see Definition 3 in the main text). Using sum-unary perturbations we can derive a family $\mathcal{U}(\alpha)$ of upper bounds on the log partition function (Proposition 4) and construct sequential samplers for the Gibbs distribution (Algorithm 1). Here we provide proofs for the related results stated in Sections 3.1 and 3.2.

Notation We will write $\mathrm{pow}_\beta\, x$ for x^β, where $x, \beta \in \mathbb{R}$, when we find this increases clarity of our exposition.

Lemma 13 (Weibull and Fréchet tricks)**.** *For any finite set $\mathcal{Y}$ and any function h, we have*

$$\begin{aligned}
\mathop{\mathrm{pow}}_{-\alpha} \sum_{y\in\mathcal{Y}} \mathop{\mathrm{pow}}_{-1/\alpha} h(y) &= \mathbb{E}_W\left[\min_y\left\{h(y)\frac{W(y)}{\Gamma(1+\alpha)}\right\}\right] \quad \textit{where } \{W(y)\}_{y\in\mathcal{Y}} \overset{i.i.d.}{\sim} \mathrm{Weibull}(1, \alpha^{-1}) \quad \textit{for } \alpha \in (0, \infty),\\
\mathop{\mathrm{pow}}_{-\alpha} \sum_{y\in\mathcal{Y}} \mathop{\mathrm{pow}}_{-1/\alpha} h(y) &= \mathbb{E}_F\left[\max_y\left\{h(y)\frac{F(y)}{\Gamma(1+\alpha)}\right\}\right] \quad \textit{where } \{F(y)\}_{y\in\mathcal{Y}} \overset{i.i.d.}{\sim} \text{Fréchet}(1, -\alpha^{-1}) \quad \textit{for } \alpha \in (-1, 0).
\end{aligned}$$

Proof. This follows from setting up competing exponential clocks with rates $\lambda_y = h(y)^{-1/\alpha}$ and then applying the function $g(x) = x^\alpha$ as in Example 1 for the case of the Weibull trick. The case of the Fréchet trick is similar, except that g is strictly decreasing for $\alpha \in (-1, 0)$, hence the maximization in place of the minimization. $\square$

B.1. Upper bounds on the partition function

Proposition 4. For any $\alpha \in (-1,0) \cup (0,\infty)$, the upper bound $\ln Z \leq \mathcal{U}(\alpha)$ holds with

$$\mathcal{U}(\alpha) := n\frac{\ln\Gamma(1+\alpha)}{\alpha} + nc - \frac{1}{\alpha}\ln\mathbb{E}_\gamma\left[e^{-\alpha U}\right].$$

Proof. We show the result for $\alpha \in (0,\infty)$ using the Weibull trick; the case of $\alpha \in (-1,0)$ can be proved similarly using the Fréchet trick. The idea is to prove by induction on n that $Z^{-\alpha} \geq e^{-\alpha\mathcal{U}(\alpha)}$, so that the claimed result follows by applying the monotonically decreasing function $x \mapsto -\ln(x)/\alpha$.

The base case $n = 1$ is the Clamping Lemma 7 below with $j = n = 1$. Now assume the claim for $n-1 \geq 1$ and for $x_n \in \mathcal{X}_n$ define

$$\mathcal{U}_{n-1}(\alpha, x_1) := (n-1)\frac{\ln\Gamma(1+\alpha)}{\alpha} + (n-1)c - \frac{1}{\alpha}\ln\mathbb{E}_\gamma\left[\exp\left(-\alpha\max_{x_2,\dots,x_n}\left\{\phi(x) + \sum_{i=2}^{n}\gamma_i(x_i)\right\}\right)\right].$$

With this definition, the Clamping Lemma with $j = 1$ states that $\sum_{x_1} \operatorname{pow}_{-1/\alpha} e^{-\alpha\mathcal{U}_{n-1}(\alpha,x_1)} \leq \operatorname{pow}_{-1/\alpha} e^{-\alpha\mathcal{U}(\alpha)}$, so:

$$\begin{aligned}
Z^{-\alpha} &\geq \operatorname*{pow}_{-\alpha}\sum_{x_1\in\mathcal{X}_1}\operatorname*{pow}_{-1/\alpha} e^{-\alpha\mathcal{U}_{n-1}(\alpha,x_1)} && \text{[inductive hypothesis]}\\
&\geq \operatorname*{pow}_{-\alpha}\operatorname*{pow}_{-1/\alpha} e^{-\alpha\mathcal{U}(\alpha)} && \text{[Clamping Lemma]}\\
&= e^{-\alpha\mathcal{U}(\alpha)},
\end{aligned}$$

as required to complete the inductive step. □

Proposition 5. The limit of $\mathcal{U}(\alpha)$ as $\alpha \to 0$ exists and equals $\mathcal{U}(0) := \mathbb{E}[U]$, i.e. the Gumbel trick upper bound.

Proof. Recall that $\mathcal{U}(\alpha) = n\frac{\ln\Gamma(1+\alpha)}{\alpha} + nc - \frac{1}{\alpha}\ln\mathbb{E}\left[e^{-\alpha U}\right]$. The first term tends to $n\psi(1) = -cn$ as $\alpha \to 0$ by L'Hôpital's rule, where ψ is the digamma function. The second term is constant in α. In the last term, $\mathbb{E}\left[e^{-\alpha U}\right]$ is the moment generating function of U evaluated at $-\alpha$, and as such its derivative at $\alpha = 0$ exists and equals the negative of the mean of U. Hence by L'Hôpital's rule,

$$-\lim_{\alpha\to 0}\frac{1}{\alpha}\ln\mathbb{E}\left[e^{-\alpha U}\right] = -\lim_{\alpha\to 0}\frac{-\mathbb{E}[U]}{\mathbb{E}\left[e^{-\alpha U}\right]} = \mathbb{E}[U] = \mathcal{U}(0).$$

The claimed result then follows by the Algebra of Limits, as the contributions of the first two terms cancel. □

Proposition 6. The function $\mathcal{U}(\alpha)$ is differentiable at $\alpha = 0$ and the derivative equals

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathcal{U}(\alpha)\bigg|_{\alpha=0} = n\frac{\pi^2}{12} - \frac{\operatorname{var}(U)}{2}.$$

Proof. First we show that $\mathcal{U}(\alpha)$ is differentiable on $(-1,0)\cup(0,\infty)$, and that the limit of the derivative as $\alpha \to 0$ exists and equals $n\pi^2/12 - \operatorname{var}(U)/2$.

The first term of $\mathcal{U}(\alpha)$ is $n\frac{\ln\Gamma(1+\alpha)}{\alpha}$, which is differentiable for $\alpha \in (-1,0)\cup(0,\infty)$ by the Quotient Rule, and its derivative equals

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}n\frac{\ln\Gamma(1+\alpha)}{\alpha} = n\frac{\psi(1+\alpha)\alpha - \ln\Gamma(1+\alpha)}{\alpha^2},$$

where ψ is the digamma function (logarithmic derivative of the gamma function). Applying L'Hôpital's rule we note that

$$\lim_{\alpha\to 0}\frac{\mathrm{d}}{\mathrm{d}\alpha}n\frac{\ln\Gamma(1+\alpha)}{\alpha} = n\lim_{\alpha\to 0}\frac{\psi(1+\alpha) + \alpha\psi^{(1)}(1+\alpha) - \psi(1+\alpha)}{2\alpha} = n\frac{\psi^{(1)}(1)}{2} = n\frac{\zeta(2)}{2} = n\frac{\pi^2}{12},$$

where $\psi^{(1)}$ is the trigamma function (derivative of the digamma function), whose value at 1 is known to be $\zeta(2) = \pi^2/6$, the Riemann zeta function evaluated at 2.

The second term of $\mathcal{U}(\alpha)$ is constant in α. The last term can be written as $K(-\alpha)/(-\alpha)$, where K is the cumulant generating function (logarithm of the moment generating function) of the random variable U. The cumulant generating function is differentiable, and by the Quotient rule

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\frac{K(-\alpha)}{-\alpha} = -\frac{\alpha K'(-\alpha) - K(-\alpha)}{\alpha^2}.$$

Applying L'Hôpital's rule we note that

$$\lim_{\alpha\to 0}\frac{\mathrm{d}}{\mathrm{d}\alpha}\frac{K(-\alpha)}{-\alpha} = \lim_{\alpha\to 0}\frac{K'(-\alpha) + \alpha K''(-\alpha) - K'(-\alpha)}{2\alpha} = \frac{K''(0)}{2} = \frac{\mathrm{var}(U)}{2},$$

where we have used that the second derivative of the cumulant generating function is the variance.

As $\mathcal{U}(\alpha)$ is continuous at 0 by construction, the above implies that it has left and right derivatives at 0. As the values of these derivatives coincide, the function is differentiable at 0 and the derivative has the stated value. $\square$

Recall that for a variable index $j \in \{1, \dots, n\}$ we also defined *partial sum-unary perturbations*

$$U_j(x_1, \dots, x_{j-1}) := \max_{x_j,\dots,x_n}\left\{\phi(\mathbf{x}) + \sum_{i=j}^{n}\gamma_i(x_i)\right\},$$

which fix the variables $x_1, \dots, x_{j-1}$ and perturb the remaining ones.

Lemma 7 (Clamping Lemma)**.** For any $j \in \{1, \dots, n\}$ and any fixed partial variable assignment $(x_1, \dots, x_{j-1}) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_{j-1}$, the following inequality holds with any trick parameter $\alpha \in (-1, 0) \cup (0, \infty)$:

$$\begin{aligned}&\sum_{x_j\in\mathcal{X}_j}\mathbb{E}_\gamma\left[e^{-(n-j)\ln\Gamma(1+\alpha)-\alpha(n-j)c)}e^{-\alpha U_{j+1}(x_1,\dots,x_j)}\right]^{-1/\alpha}\\ &\leq \mathbb{E}_\gamma\left[e^{-(n-(j-1))\ln\Gamma(1+\alpha)-\alpha(n-(j-1))c)}e^{-\alpha U_j(x_1,\dots,x_{j-1})}\right]^{-1/\alpha}.\end{aligned}$$

Proof. For $\alpha > 0$, from the Weibull trick (Lemma 13), using independence of the perturbations and Jensen's inequality,

$$\begin{aligned}&\operatorname*{pow}_{-\alpha}\sum_{x_j\in\mathcal{X}_j}\operatorname*{pow}_{-1/\alpha}\mathbb{E}_W\left[\min_{x_{j+1},\dots,x_n}\tilde{p}(\mathbf{x})^{-\alpha}\prod_{i=j+1}^{n}\frac{W(x_i)}{\Gamma(1+\alpha)}\right]\\ &= \mathbb{E}_W\left[\min_{x_j\in\mathcal{X}_j}\left\{\mathbb{E}_W\left[\min_{x_{j+1},\dots,x_n}\tilde{p}(\mathbf{x})^{-\alpha}\prod_{i=j+1}^{n}\frac{W(x_i)}{\Gamma(1+\alpha)}\right]\frac{W(x_j)}{\Gamma(1+\alpha)}\right\}\right]\\ &\leq \mathbb{E}_W\left[\min_{x_j,\dots,x_n}\tilde{p}(\mathbf{x})^{-\alpha}\prod_{i=j}^{n}\frac{W(x_i)}{\Gamma(1+\alpha)}\right]\end{aligned}$$

Representing the Weibull random variables in terms of Gumbel random variables using the transformation $W = e^{-(\gamma+c)\alpha}$, where $\gamma \sim \mathrm{Gumbel}(-c)$, and manipulating the obtained expressions yields the claimed result. $\square$

B.2. Sequential samplers for the Gibbs distribution

The family of sequential samplers for the Gibbs distribution presented in the main text as Algorithm 1 has the same overall structure as the sequential sampler derived by Hazan et al. (2013) from the Gumbel trick upper bound $\mathcal{U}(0)$, and hence correctness can be argued similarly. Conditioned on accepting the sample, the probability that $\mathbf{x} = (x_1, \ldots, x_n)$ is returned is

$$\prod_{i=1}^{n} p_i(x_i) = \prod_{i=1}^{n} \frac{e^{-c}}{\Gamma(1+\alpha)^{1/\alpha}} \frac{\mathbb{E}_\gamma \left[e^{-\alpha U_{i+1}(x_1,\ldots,x_i)}\right]^{-1/\alpha}}{\mathbb{E}_\gamma \left[e^{-\alpha U_i(x_1,\ldots,x_{i-1})}\right]^{-1/\alpha}} = \frac{e^{-nc}}{\Gamma(1+\alpha)^{n/\alpha}} \frac{\left(e^{-\alpha\phi(x_1,\ldots,x_n)}\right)^{-1/\alpha}}{\mathbb{E}[e^{-\alpha U}]^{-1/\alpha}} \propto p(x),$$

as required to show that the produced samples follow the Gibbs distribution p. Note, however, that in practice one introduces an approximation by replacing expectations with sample averages.

B.3. Relationship between errors of sum-unary Gumbel perturbations

We write $\mathbf{x}^*$ for the (random) MAP configuration after sum-unary perturbation of the potential function, i.e.,

$$\mathbf{x}^* := \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}} \left\{ \phi(\mathbf{x}) + \sum_{i=1}^{n} \gamma_i(x_i) \right\}.$$

Let $q_{\text{sum}}(\mathbf{x}) := \mathbb{P}[\mathbf{x} = \mathbf{x}^*]$ be the probability mass function of $\mathbf{x}^*$.

The following results links together the errors acquired when using summed unary perturbations to upper bound the log partition function $\ln Z \leq \mathcal{U}(0)$ using the Gumbel trick upper bound by Hazan & Jaakkola (2012), to approximately sample from the Gibbs distribution by using q_{sum} instead, and to upper bound the entropy of the approximate distribution q_{sum} using the bound due to Maji et al. (2014).

Proposition 11. Writing p for the Gibbs distribution, we have

$$\underbrace{(\mathcal{U}(0) - \ln Z)}_{\text{error in } \ln Z \text{ bound}} + \underbrace{\mathrm{KL}(q_{\text{sum}} \parallel p)}_{\text{sampling error}} = \underbrace{\mathbb{E}_{\gamma_i}\left[\gamma_i(\mathbf{x}_i^*)\right] - H(q_{\text{sum}})}_{\text{error in entropy estimation}}.$$

Proof. By conditioning on the maximizing configuration $\mathbf{x}^*$, we can rewrite the Gumbel trick upper bound $\mathcal{U}(0)$ as follows:

$$\begin{aligned}
\mathcal{U}(0) &= \mathbb{E}_\gamma \left[\max_{\mathbf{x}\in\mathcal{X}} \left\{ \theta(\mathbf{x}) + \sum_{i=1}^{n} \gamma_i(x_i) \right\} \right] \\
&= \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{sum}}(\mathbf{x}) \left(\theta(\mathbf{x}) + \mathbb{E}_\gamma \left[\sum_{i=1}^{n} \gamma_i(x_i) \mid \mathbf{x} = \mathbf{x}^* \right] \right) \\
&= \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{sum}}(\mathbf{x})\theta(\mathbf{x}) + \sum_{i=1}^{n} \mathbb{E}_{\gamma_i}\left[\gamma_i(x_i^*)\right].
\end{aligned}$$

At the same time, the KL divergence between q_{sum} and the Gibbs distribution p generally expands as

$$\begin{aligned}
\mathrm{KL}(q_{\text{sum}} \parallel p) &= -H(q_{\text{sum}}) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{sum}}(\mathbf{x}) \ln \frac{\exp(\theta(\mathbf{x}))}{\sum_{\tilde{\mathbf{x}}\in\mathcal{X}} \exp(\theta(\tilde{\mathbf{x}}))} \\
&= -H(q_{\text{sum}}) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{sum}}(\mathbf{x})\theta(\mathbf{x}) + \ln Z.
\end{aligned}$$

Adding the two equations together and rearranging yields the claimed result. $\square$

C. Averaged unary perturbations

C.1. Lower bounds on the partition function

In the main text we stated the following two lower bounds on the log partition function $\ln Z$.

Proposition 9. Let $\alpha \in (-1,0) \cup (0,\infty)$. For any subset $S \subseteq \{1,\dots,n\}$ of the variables $x_1,\dots,x_n$ we have $\ln Z \geq$

$$c + \frac{\ln \Gamma(1+\alpha)}{\alpha} - \frac{1}{\alpha} \ln \mathbb{E}\left[e^{-\alpha \max_{\mathbf{x}}\{\phi(\mathbf{x})+\gamma_S(\mathbf{x}_S)\}}\right],$$

where $\mathbf{x}_S := \{x_i : i \in S\}$ and $\gamma_S(\mathbf{x}_S) \sim \mathrm{Gumbel}(-c)$ independently for each setting of $\mathbf{x}_S$.

Proof. Let $\bar{S} := \{1,\dots,n\} \setminus S$. First we handle the case $\alpha > 0$. We have trivially that

$$\mathrm{pow}_{-\alpha} Z = \mathrm{pow}_{-\alpha} \sum_{\mathbf{x}_S} \sum_{\mathbf{x}_{\bar{S}}} e^{\phi(\mathbf{x}_S,\mathbf{x}_{\bar{S}})} \leq \mathrm{pow}_{-\alpha} \sum_{\mathbf{x}_S} \max_{\mathbf{x}_{\bar{S}}} e^{\phi(\mathbf{x}_S,\mathbf{x}_{\bar{S}})}.$$

The Weibull trick tells us that $\mathrm{pow}_{-\alpha} \sum_y \mathrm{pow}_{-1/\alpha} h(y) = \mathbb{E}_W[\min_y \frac{h(y)}{\Gamma(1+\alpha)} W(y)]$ where $\{W(y)\}_y \overset{iid}{\sim} \mathrm{Weibull}(1,\alpha^{-1})$. Applying this to the summation over $\mathbf{x}_S$ on the right-hand side of the above inequality, we obtain

$$\mathrm{pow}_{-\alpha} Z \leq \mathbb{E}_W\left[\min_{\mathbf{x}_S} \frac{\mathrm{pow}_{-\alpha} \max_{\mathbf{x}_{\bar{S}}} e^{\phi(\mathbf{x}_S,\mathbf{x}_{\bar{S}})}}{\Gamma(1+\alpha)} W(\mathbf{x}_S)\right].$$

Expressing the Weibull random variable $W(\mathbf{x}_S)$ as $e^{-\alpha(\gamma_S(\mathbf{x}_S)+c)}$ with $\gamma_S(\mathbf{x}_S) \sim \mathrm{Gumbel}(-c)$, the right-hand side can be simplified as follows:

$$\begin{aligned}
\mathrm{pow}_{-\alpha} Z &\leq \frac{1}{\Gamma(1+\alpha)} \mathbb{E}_\gamma\left[\mathrm{pow}_{-\alpha} \max_{\mathbf{x}_S} \max_{\mathbf{x}_{\bar{S}}} e^{\phi(\mathbf{x}_S,\mathbf{x}_{\bar{S}})} e^{\gamma_S(\mathbf{x}_S)+c}\right] \\
&= \frac{e^{-\alpha c}}{\Gamma(1+\alpha)} \mathbb{E}_\gamma\left[\exp\left(-\alpha \max_{\mathbf{x}} \{\phi(\mathbf{x}) + \gamma_S(\mathbf{x}_S)\}\right)\right].
\end{aligned}$$

Taking the logarithm and dividing by $-\alpha < 0$ yields the claimed result for positive α. For $\alpha \in (-1,0)$ we proceed similarly, obtaining that

$$\begin{aligned}
\mathrm{pow}_{-\alpha} Z &\geq \mathrm{pow}_{-\alpha} \sum_{\mathbf{x}_S} \max_{\mathbf{x}_{\bar{S}}} e^{\phi(\mathbf{x}_S,\mathbf{x}_{\bar{S}})} \\
&= \mathbb{E}_F\left[\min_{\mathbf{x}_S} \frac{\mathrm{pow}_{-\alpha} \max_{\mathbf{x}_{\bar{S}}} e^{\phi(\mathbf{x}_S,\mathbf{x}_{\bar{S}})}}{\Gamma(1+\alpha)} F(\mathbf{x}_S)\right],
\end{aligned}$$

where $F(\mathbf{x}(S)) \sim \mathrm{Fréchet}(1,-\alpha^{-1})$. Representing these random variables as $e^{-\alpha(\gamma_S(\mathbf{x}_S)+c)}$ with $\gamma_S(\mathbf{x}_S) \sim \mathrm{Gumbel}(-c)$, simplifying as in the previous case and finally dividing the inequality by $-\alpha > 0$ yields the claimed result for $\alpha \in (-1,0)$. □

Corollary 10. For any $\alpha \in (-1,0) \cup (0,\infty)$, we have the lower bound $\ln Z \geq \mathcal{L}(\alpha)$, where

$$\mathcal{L}(\alpha) := c + \frac{\ln \Gamma(1+\alpha)}{\alpha} - \frac{1}{n\alpha} \ln \mathbb{E}\left[\exp\left(-n\alpha L\right)\right],$$

Proof. Applying Proposition 9 n times with all singleton sets $S = \{i\}$ and averaging the obtained lower bounds yields

$$\begin{aligned}
\ln Z &\geq c + \frac{\ln \Gamma(1+\alpha)}{\alpha} - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\alpha} \ln \mathbb{E}\left[\exp\left(-\alpha \max_{\mathbf{x}} \{\phi(\mathbf{x}) + \gamma_i(x_i)\}\right)\right] \\
&= c + \frac{\ln \Gamma(1+\alpha)}{\alpha} - \frac{1}{n\alpha} \ln \mathbb{E}\left[\exp\left(-\sum_{i=1}^{n} \alpha \max_{\mathbf{x}} \{\phi(\mathbf{x}) + \gamma_i(x_i)\}\right)\right] \\
&= c + \frac{\ln \Gamma(1+\alpha)}{\alpha} - \frac{1}{n\alpha} \ln \mathbb{E}\left[\exp\left(-n\alpha \frac{1}{n} \sum_{i=1}^{n} \max_{\mathbf{x}} \{\phi(\mathbf{x}) + \gamma_i(x_i)\}\right)\right],
\end{aligned}$$

where the first equality used the fact that the perturbations $\gamma_i(x_i)$ are mutually independent for different indices i to replace the product of expectations with the expectation of the product. The claimed result follows by applying Jensen's inequality to swap the summation and the convex $\max_{\mathbf{x}}$ function, noting that the inequality works out the right way for both positive and negative α. $\square$

Jensen's inequality can be used to relate the general lower bound $\mathcal{L}(\alpha)$ to the Gumbel trick lower bound $\mathcal{L}(0)$, showing that the former cannot be arbitrarily worse than the latter:

Proposition 14. *For all $\alpha \in (-1, 0)$, the lower bound $\mathcal{L}(\alpha)$ on $\ln Z$ satisfies*

$$\mathcal{L}(\alpha) \geq \mathcal{L}(0) + \frac{\ln \Gamma(1+\alpha)}{\alpha} + c$$

Proof. Apply Jensen's inequality with the convex function $x \mapsto e^{-n\alpha}$ to the last term in the definition of $\mathcal{L}(\alpha)$, noting that the inequality works out the stated way for $\alpha < 0$. $\square$

Note that $\frac{\ln \Gamma(1+\alpha)}{\alpha} + c \leq 0$ for $\alpha \in (-1, 0)$ so this result does *not* imply that the Fréchet lower bounds are tighter than the Gumbel lower bound $\mathcal{L}(0)$; it merely says that they cannot be arbitrarily worse than $\mathcal{L}(0)$.

C.2. Relationship between errors of averaged-unary Gumbel perturbations

In this section we write $\mathbf{x}^*$ for the (random) MAP configuration after average-unary perturbation of the potential function, i.e.,

$$\mathbf{x}^* := \operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} \left\{ \phi(\mathbf{x}) + \frac{1}{n} \sum_{i=1}^{n} \gamma_i(x_i) \right\}.$$

where $\{\gamma_i(x_i) \mid x_i \in \mathcal{X}_i, 1 \leq i \leq n\} \overset{\text{i.i.d.}}{\sim} \mathrm{Gumbel}(-c)$. Let $q_{\text{avg}}(\mathbf{x}) := \mathbb{P}[\mathbf{x} = \mathbf{x}^*]$ be the probability mass function of $\mathbf{x}^*$. The Gumbel trick lower bound on the log partition function $\ln Z$ due to Hazan et al. (2013) is:

$$\ln Z \geq \mathcal{L}(0) = \mathcal{L}_\phi(0) := \mathbb{E}_\gamma \left[\min_{\mathbf{x} \in \mathcal{X}} \left\{ \phi(\mathbf{x}) + \frac{1}{n} \sum_{i=1}^{n} \gamma_i(x_i) \right\} \right]. \tag{3}$$

We show that the gap of this Gumbel trick lower bound on $\ln Z$ upper bounds the KL divergence between the approximate distribution q_{avg} and the Gibbs distribution p. To this end, we first need an entropy bound for q_{avg} analogous to Theorem 1 of (Maji et al., 2014).

Theorem 15. *The entropy of q_{avg} can be lower bounded using expected values of max-perturbations as follows:*

$$H(q_{avg}) \geq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\gamma_i} \left[\gamma_i(x_i^*) \right]$$

Remark. Theorem 1 of (Maji et al., 2014) and this Theorem 15 differ in three aspects: (1) the former is an upper bound and the latter is a lower bound, (2) the former sums the expectations while the latter averages them, and (3) the distributions q_{sum} and q_{avg} of $\mathbf{x}^*$ in the two theorems are different.

Proof. By the duality relation between negative entropy and the log partition function (Wainwright & Jordan, 2008), the entropy $H(q_{\text{avg}})$ of the unary-avg perturb-max distribution q_{avg} can be expressed as

$$H(q_{\text{avg}}) = \inf_{\varphi} \left\{ \ln Z_\varphi - \sum_{\mathbf{x} \in \mathcal{X}} q_{\text{avg}}(\mathbf{x}) \varphi(\mathbf{x}) \right\},$$

where the variable φ ranges over all potential functions on $\mathcal{X}$, and $Z_\varphi = \sum_{\mathbf{x} \in \mathcal{X}} \exp \varphi(\mathbf{x})$. Applying the Gumbel trick lower bound on the log partition function gives

$$H(q_{\text{avg}}) \geq \inf_{\varphi} \left\{ \mathcal{L}_\varphi(0) - \sum_{\mathbf{x} \in \mathcal{X}} q_{\text{avg}}(\mathbf{x}) \varphi(\mathbf{x}) \right\},$$

Proposition 16 in Appendix D shows that $\mathcal{L}_\varphi(0)$ is a convex function of φ. The expression $-\sum_{\mathbf{x}\in\mathcal{X}} q(\mathbf{x})\varphi(\mathbf{x})$ is a linear function of φ, so also convex, and thus as a sum of two convex functions, the quantity $\mathcal{L}_\varphi(0) - \sum_{\mathbf{x}\in\mathcal{X}} q(\mathbf{x})\varphi(\mathbf{x})$ within the infimum is a convex function of φ. Moreover, Proposition 17 in Appendix D tells us that the partial derivatives can be computed as

$$\frac{\partial}{\partial\varphi(\mathbf{x})}\left(\mathcal{L}_\varphi(0) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x})\varphi(\mathbf{x})\right) = q_\varphi(\mathbf{x}) - q_{\text{avg}}(\mathbf{x})$$

where $q_\varphi(\mathbf{x})$ is the unary-avg perturb-max distribution associated with the potential function φ. Proposition 18 in Appendix D confirms that these partial derivatives are continuous, so we observe that as a function of φ, the expression $\mathcal{L}_\varphi(0) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x})\varphi(\mathbf{x})$ is a convex function with continuous partial derivatives, so it is a differentiable convex function. This is sufficient to establish that the point $\varphi = \phi$ is a global minimum of this function (Wright & Nocedal, 1999). Hence

$$\begin{aligned}
H(q_{\text{avg}}) &\geq \inf_\varphi \left\{\mathcal{L}_\varphi(0) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x})\varphi(\mathbf{x})\right\} \\
&= \mathcal{L}_\phi(0) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x})\phi(\mathbf{x}) \\
&= \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x})\mathbb{E}_\gamma\left[\phi(\mathbf{x}) + \frac{1}{n}\sum_{i=1}^n \gamma_i(x_i) \mid \mathbf{x} = \mathbf{x}^*\right] - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x})\phi(\mathbf{x}) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\gamma_i}\left[\gamma_i(x_i^*)\right]
\end{aligned}$$

where we conditioned on the maximizing configuration $\mathbf{x}^*$ when expanding $\mathcal{L}_\phi(0)$. □

Remark. This proof proceeded in the same way as the proof of Maji et al. (2014) for the upper bound, except that establishing the minimizing configuration of the infimum is a non-trivial step that is actually required in this case. The second revision of (Hazan et al., 2016) computes the derivative of $\mathcal{U}_\varphi(0) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{sum}}(\mathbf{x})\varphi(\mathbf{x})$, which is similar to our $\mathcal{L}_\varphi(0) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x})\varphi(\mathbf{x})$, by differentiating under the expectation.

Equipped with Theorem 15, we can now show a link between the approximation "errors" of the averaged-unary perturbation MAP configuration distribution q_{avg} (to the Gibbs distribution p) and estimate $\mathcal{L}(0)$ (to $\ln Z$).

Proposition 12. Let p be the Gibbs distribution on $\mathcal{X}$. Then

$$\underbrace{\ln Z - \mathcal{L}(0)}_{\text{error in } \ln Z \text{ bound}} \geq \underbrace{\mathrm{KL}(q_{\text{avg}} \parallel p)}_{\text{sampling error}} \geq 0$$

Remark. While we knew from Hazan et al. (2013) that $\ln Z - \mathcal{L}(0) \geq 0$ (i.e. that $\mathcal{L}(0)$ is a lower bound on $\ln Z$), this is a stronger result showing that the size of the gap is an upper bound on the KL divergence between the average-unary perturbation MAP distribution q_{avg} and the Gibbs distribution p.

Proof. The Kullback-Leibler divergence in question expands as

$$\mathrm{KL}(q_{\text{avg}} \parallel p) = -H(q_{\text{avg}}) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x}) \ln \frac{\exp\phi(\mathbf{x})}{\sum_{\tilde{\mathbf{x}}\in\mathcal{X}}\exp\phi(\tilde{\mathbf{x}})} = -H(q_{\text{avg}}) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x})\phi(\mathbf{x}) + \ln Z.$$

From the proof of Theorem 15 we know that $H(q_{\text{avg}}) \geq \mathcal{L}(0) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x})\phi(\mathbf{x})$, so

$$\mathrm{KL}(q_{\text{avg}} \parallel p) \leq -\mathcal{L}(0) + \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x})\phi(\mathbf{x}) - \sum_{\mathbf{x}\in\mathcal{X}} q_{\text{avg}}(\mathbf{x})\phi(\mathbf{x}) + \ln Z = \ln Z - \mathcal{L}(0).$$ □

D. Technical results

In this section we write $\mathcal{L}(\phi)$ instead of $\mathcal{L}_\phi(0)$ for the Gumbel trick lower bound on $\ln Z$ associated with the potential function ϕ, see equation (3).

Proposition 16. *The Gumbel trick lower bound $\mathcal{L}(\phi)$, viewed as a function of the potentials ϕ, is convex.*

Proof. Convexity can be proved directly from definition. Let ϕ_1 and ϕ_2 be two arbitrary potential functions on a discrete product space $\mathcal{X}$, and let $\lambda \in [0, 1]$. Then

$$\begin{aligned}
&\mathcal{L}(\lambda\phi_1 + (1-\lambda)\phi_2) \\
&= \mathbb{E}_\gamma\left[\max_{\mathbf{x}\in\mathcal{X}}\left\{\lambda\phi_1(\mathbf{x}) + (1-\lambda)\phi_2(\mathbf{x}) + \frac{1}{n}\sum_{i=1}^n \gamma_i(x_i)\right\}\right] \\
&= \mathbb{E}_\gamma\left[\max_{\mathbf{x}\in\mathcal{X}}\left\{\lambda\left(\phi_1(\mathbf{x}) + \frac{1}{n}\sum_{i=1}^n \gamma_i(x_i)\right) + (1-\lambda)\left(\phi_2(\mathbf{x}) + \frac{1}{n}\sum_{i=1}^n \gamma_i(x_i)\right)\right\}\right] \\
&\le \mathbb{E}_\gamma\left[\lambda\max_{\mathbf{x}\in\mathcal{X}}\left\{\phi_1(\mathbf{x}) + \frac{1}{n}\sum_{i=1}^n \gamma_i(x_i)\right\} + (1-\lambda)\max_{\mathbf{x}\in\mathcal{X}}\left\{\phi_2(\mathbf{x}) + \frac{1}{n}\sum_{i=1}^n \gamma_i(x_i)\right\}\right] \\
&= \lambda\mathcal{L}(\phi_1) + (1-\lambda)\mathcal{L}(\phi_2),
\end{aligned}$$

where we have used convexity of the $\max$ function to obtain the inequality, and linearity of expectation to arrive at the final equality. $\square$

Remark. This convexity proof goes through for other (low-dimensional) perturbations as well, e.g. it also works for $\mathcal{U}_\phi(0)$.

Proposition 17. *The Gumbel trick lower bound $\mathcal{L}(\phi)$, viewed as a function of the potentials ϕ, has partial derivatives*

$$\frac{\partial}{\partial\phi(\tilde{\mathbf{x}})}\mathcal{L}(\phi) = q_\phi(\tilde{\mathbf{x}})$$

where q_ϕ is the probability mass function of the average-unary perturbation MAP configuration's distribution associated with the potential function ϕ.

Proof. Let $\tilde{\mathbf{x}} \in \mathcal{X}$, so that $\phi(\tilde{\mathbf{x}})$ is a general component of ϕ, and let $e_{\tilde{\mathbf{x}}}$ be the indicator vector of $\tilde{\mathbf{x}}$. For any $\delta \in \mathbb{R}$, the change in the lower bound $\mathcal{L}$ due to replacing $\phi(\tilde{\mathbf{x}})$ with $\phi(\tilde{\mathbf{x}}) + \delta$ is

$$\begin{aligned}
\mathcal{L}(\phi + \delta e_{\tilde{\mathbf{x}}}) - \mathcal{L}(\phi) &= \mathbb{E}_\gamma\left[\max_{\mathbf{x}\in\mathcal{X}}\left\{\phi(\mathbf{x}) + \delta\mathbb{1}\{\mathbf{x} = \tilde{\mathbf{x}}\} + \frac{1}{n}\sum_{i=1}^n \gamma_i(x_i)\right\}\right] - \mathbb{E}_\gamma\left[\max_{\mathbf{x}\in\mathcal{X}}\left\{\phi(\mathbf{x}) + \frac{1}{n}\sum_{i=1}^n \gamma_i(x_i)\right\}\right] \\
&= \mathbb{E}_\gamma\left[\max_{\mathbf{x}\in\mathcal{X}}\left\{\phi(\mathbf{x}) + \delta\mathbb{1}\{\mathbf{x} = \tilde{\mathbf{x}}\} + \frac{1}{n}\sum_{i=1}^n \gamma_i(x_i)\right\} - \max_{\mathbf{x}\in\mathcal{X}}\left\{\phi(\mathbf{x}) + \frac{1}{n}\sum_{i=1}^n \gamma_i(x_i)\right\}\right] \\
&= \mathbb{E}_\gamma\left[\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma)\right]
\end{aligned}$$

by linearity of expectation, where we have denoted by $\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma)$ the change in maximum due to replacing the potential $\phi(\tilde{\mathbf{x}})$ with $\phi(\tilde{\mathbf{x}}) + \delta$. Let's condition on the argmax before modifying ϕ:

$$\mathcal{L}(\phi + \delta e_{\tilde{\mathbf{x}}}) - \mathcal{L}(\phi) = \mathbb{E}_\gamma\left[\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma)\right] = \sum_{\mathbf{x}\in\mathcal{X}} q_\phi(\mathbf{x})\mathbb{E}_\gamma\left[\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma) \mid \mathbf{x} \text{ is the original argmax}\right]$$

Now let's condition on the size of the gap G between the maximum and the runner-up:

$$\begin{aligned}
\mathbb{E}_\gamma\left[\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma) \mid \mathbf{x} \text{ is the original argmax}\right] &= \mathbb{P}(G \le |\delta|)\mathbb{E}_\gamma\left[\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma) \mid \mathbf{x} \text{ is the original argmax}, G \le |\delta|\right] \\
&\quad + \mathbb{P}(G > |\delta|)\mathbb{E}_\gamma\left[\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma) \mid \mathbf{x} \text{ is the original argmax}, G > |\delta|\right]
\end{aligned}$$

Let's examine all four terms on the right-hand side one by one:

1. $\mathbb{P}(G \leq |\delta|) \to \mathbb{P}(G = 0) = 0$ as $\delta \to 0$ by monotonicity of measure.
2. $\mathbb{E}_\gamma \left[\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma) \mid \mathbf{x} \text{ is the original argmax}, G \leq |\delta|\right] \leq \delta$ since $|\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma)| \leq |\delta|$ always holds.
3. $\mathbb{P}(G > |\delta|) \to \mathbb{P}(G \geq 0) = 1$ as $\delta \to 0$ by monotonicity of measure.
4. $\mathbb{E}_\gamma \left[\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma) \mid \mathbf{x} \text{ is the original argmax}, G > |\delta|\right] = \delta \mathbb{1}\{\mathbf{x} = \tilde{\mathbf{x}}\}$ since in this case both maximizations in the definition of $\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma)$ are maximized at $\mathbf{x}$.

Therefore, as $\delta \to 0$,

$$\mathbb{E}_\gamma \left[\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma) \mid \mathbf{x} \text{ is the original argmax}\right] = o(1)o(\delta) + (1 + o(1))\delta \mathbb{1}\{\mathbf{x} = \tilde{\mathbf{x}}\}$$

Putting things together, we have

$$\begin{aligned}
\lim_{\delta \to 0} \frac{\mathcal{L}(\phi + \delta e_{\tilde{\mathbf{x}}}) - \mathcal{L}(\phi)}{\delta} &= \sum_{\mathbf{x} \in \mathcal{X}} q_\phi(\mathbf{x}) \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{E}_\gamma \left[\Delta(\phi, \delta, \tilde{\mathbf{x}}, \gamma) \mid \mathbf{x} \text{ is the original argmax}\right] \\
&= \sum_{\mathbf{x} \in \mathcal{X}} q_\phi(\mathbf{x}) \mathbb{1}\{\mathbf{x} = \tilde{\mathbf{x}}\} \\
&= q_\phi(\tilde{\mathbf{x}}),
\end{aligned}$$

which proves the stated claim directly from definition of a partial derivative. □

Proposition 18. *The probability mass function q_ϕ of the average-unary perturbation MAP configuration's distribution associated with a potential function ϕ is continuous in ϕ.*

Proof. For any $\mathbf{x}^* \in \mathcal{X}$ we have from definition

$$\begin{aligned}
q_\phi(\mathbf{x}^*) &= \mathbb{P}\left[\mathbf{x}^* = \operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} \left\{\phi(\mathbf{x}) + \frac{1}{n} \sum_{i=1}^n \gamma_i(x_i)\right\}\right] \\
&= \mathbb{P}\left[\phi(\mathbf{x}^*) + \frac{1}{n} \sum_{i=1}^n \gamma_i(x_i^*) > \max_{\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{x}^*\}} \left\{\phi(\mathbf{x}) + \frac{1}{n} \sum_{i=1}^n \gamma_i(x_i)\right\}\right] \\
&= \mathbb{E}\left[\mathbb{1}\left\{\phi(\mathbf{x}^*) + \frac{1}{n} \sum_{i=1}^n \gamma_i(x_i^*) > \max_{\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{x}^*\}} \left\{\phi(\mathbf{x}) + \frac{1}{n} \sum_{i=1}^n \gamma_i(x_i)\right\}\right\}\right]
\end{aligned}$$

which is continuous in ϕ by continuity of $\max$, of $\mathbb{1}\{\cdot > \cdot\}$ (as a function of ϕ) and by the Bounded Convergence Theorem. □

Remark. The results above show that the Gumbel trick lower bound $\mathcal{L}(\phi)$, viewed as a function of the potentials ϕ, is convex and has continuous partial derivatives.