EECE 571F: Deep Learning with Structures

Lecture 11: Stochastic Gradient Estimator for Discrete Latent Variable Models

Renjie Liao

University of British Columbia

Winter, Term 1, 2022

Course Scope

- Brief Intro to Deep Learning

- Geometric Deep Learning

  - Deep Learning Models for Sets and Sequences: Deep Sets & Transformers
  - Deep Learning Models for Graphs: Graph Convolution & Message Passing GNNs
  - Expressiveness & Generalizations of GNNs
  - Unsupervised/Self-supervised Graph Representation Learning

- Probabilistic Deep Learning

  - Deep Generative Models:
    Auto-regressive models, GANs, VAEs, Diffusion/Score based models
  - Discrete/Hybrid Latent Variable Models: RBMs, Latent Graph Models
  - Stochastic Gradient Estimation

Course Scope

- Brief Intro to Deep Learning

- Geometric Deep Learning

    - Deep Learning Models for Sets and Sequences: Deep Sets & Transformers
    - Deep Learning Models for Graphs: Graph Convolution & Message Passing GNNs
    - Expressiveness & Generalizations of GNNs
    - Unsupervised/Self-supervised Graph Representation Learning

- Probabilistic Deep Learning

    - Deep Generative Models:
      Auto-regressive models, GANs, VAEs, Diffusion/Score based models
    - Discrete/Hybrid Latent Variable Models: RBMs, Latent Graph Models
    - **Stochastic Gradient Estimation**

Sampling Discrete (Categorical) Random Variables

Given a probability mass function of a discrete RV, how to draw samples?

Sampling Discrete (Categorical) Random Variables

Given a probability mass function of a discrete RV, how to draw samples?

Inverse transform sampling

Sampling Discrete (Categorical) Random Variables

Given a probability mass function of a discrete RV, how to draw samples?

Inverse transform sampling

Binominal(10, 0.4)

Sampling Discrete (Categorical) Random Variables

Given a probability mass function of a discrete RV, how to draw samples?

Inverse transform sampling

Binominal(10, 0.4)

Probability Mass Function (PMF)

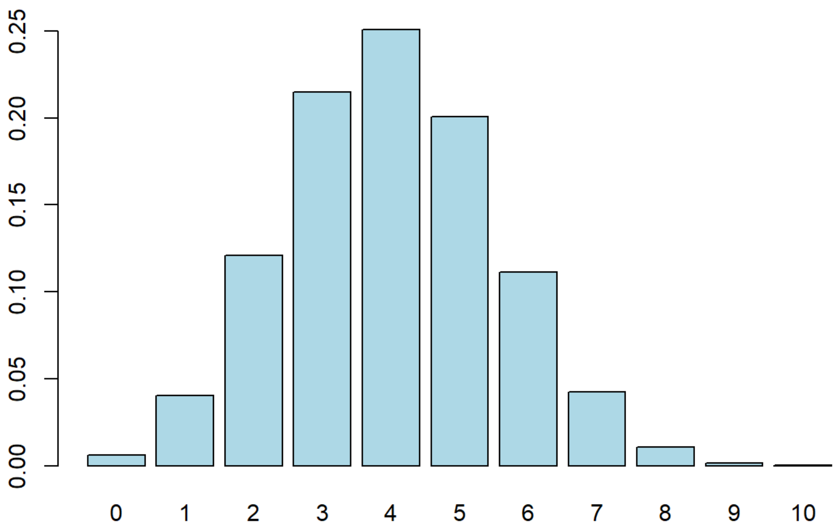


Image Credit: http://www.est.uc3m.es/icascos/eng/probability_notes/discrete-random-variables.html

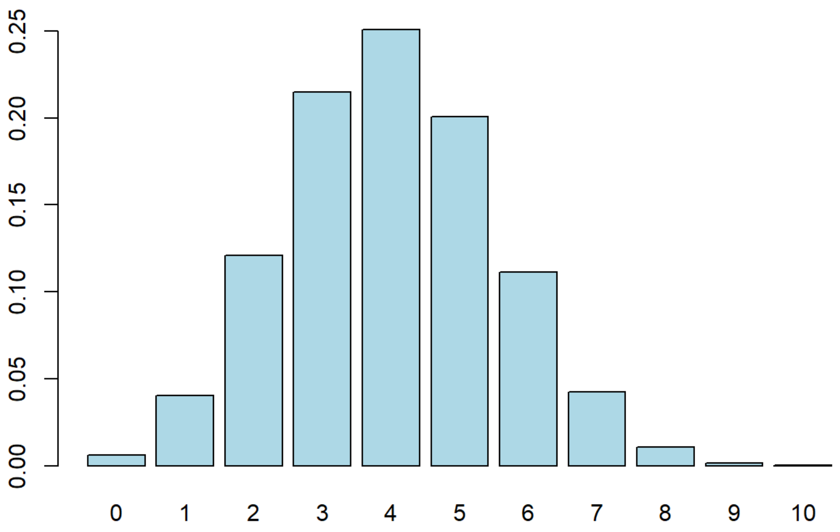
Sampling Discrete (Categorical) Random Variables

Given a probability mass function of a discrete RV, how to draw samples?

Inverse transform sampling

Binominal(10, 0.4)

Probability Mass Function (PMF)



Cumulative Distribution Function (CDF)

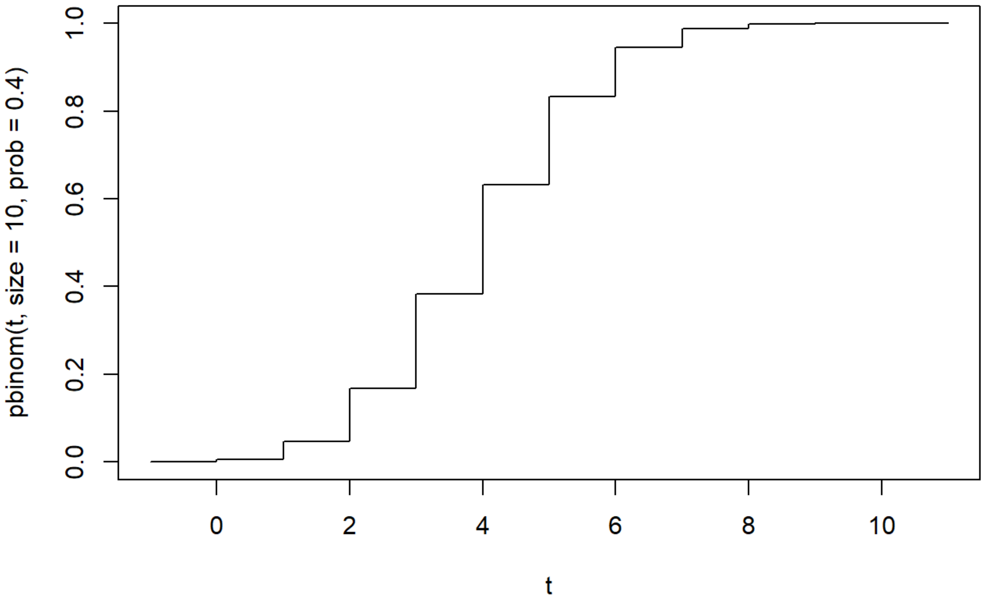


Image Credit: http://www.est.uc3m.es/icascos/eng/probability_notes/discrete-random-variables.html

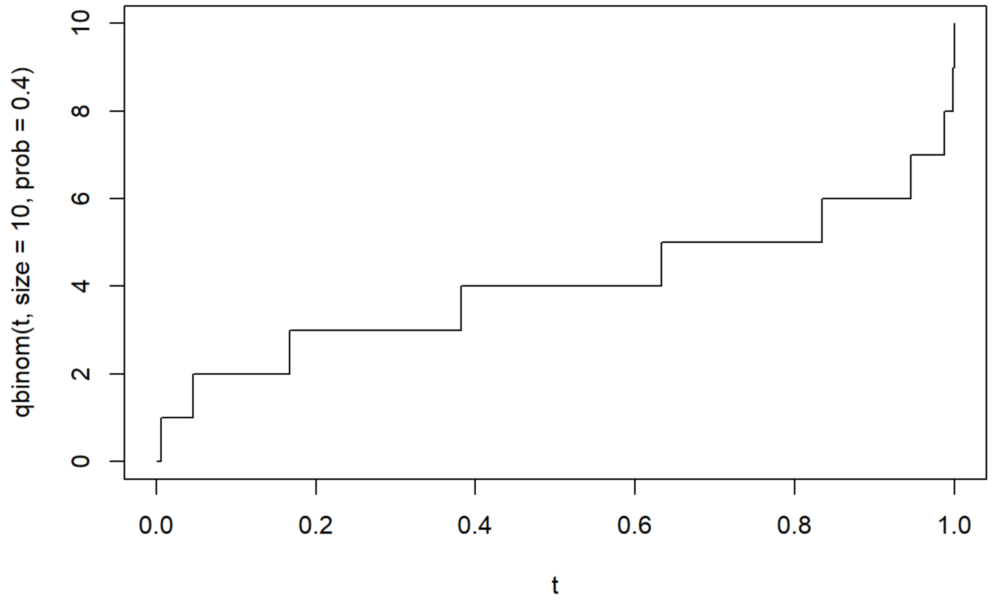
Sampling Discrete (Categorical) Random Variables

Given a probability mass function of a discrete RV, how to draw samples?

Inverse transform sampling

Binominal(10, 0.4)

Inverse CDF



Cumulative Distribution Function (CDF)

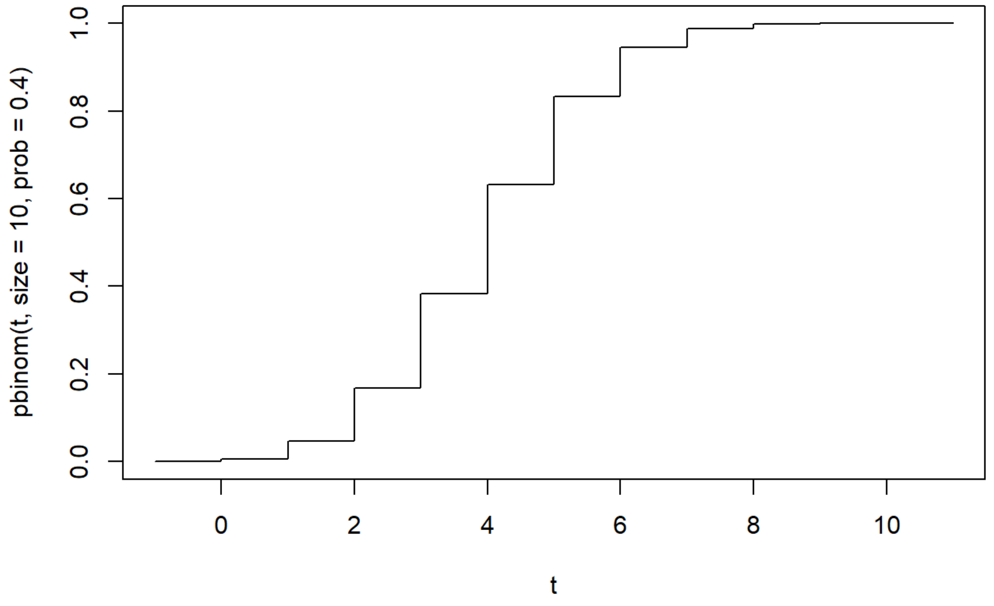


Image Credit: http://www.est.uc3m.es/icascos/eng/probability_notes/discrete-random-variables.html

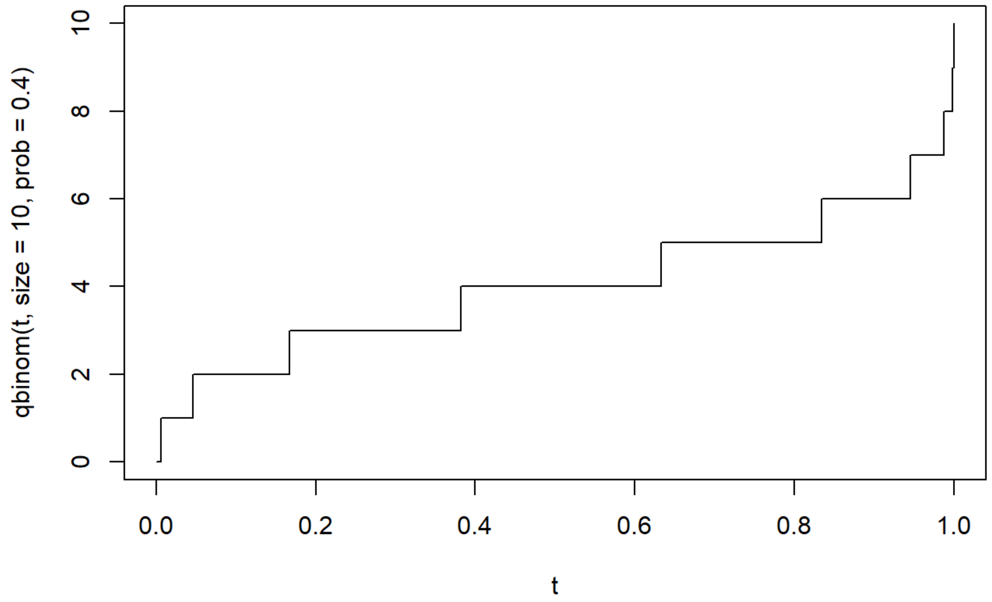
Sampling Discrete (Categorical) Random Variables

Given a probability mass function of a discrete RV, how to draw samples?

Inverse transform sampling

Binominal(10, 0.4)

Inverse CDF



$$u \sim \text{Uniform}(0, 1)$$
$$x = \text{InverseCDF}(u)$$

CDF and its inverse are discontinuous (piece-wise constant)!

Image Credit: http://www.est.uc3m.es/icascos/eng/probability_notes/discrete-random-variables.html

Sampling Discrete (Categorical) Random Variables

Given a probability mass function of a discrete RV, how to draw samples?

Gumbel-Max Trick

Sampling Discrete (Categorical) Random Variables

Given a probability mass function of a discrete RV, how to draw samples?

Gumbel-Max Trick

RV X follows a categorical distribution where $X \in \{1, 2, \cdots, K\}$ and

$$p(X = k) = \frac{\exp(x_k)}{\sum_{i=1}^{K} \exp(x_i)}$$

Sampling Discrete (Categorical) Random Variables

Given a probability mass function of a discrete RV, how to draw samples?

Gumbel-Max Trick

RV X follows a categorical distribution where $X \in \{1, 2, \cdots, K\}$ and

$$p(X = k) = \frac{\exp(x_k)}{\sum_{i=1}^{K} \exp(x_i)}$$

We have y follows the same distribution as X

$$y = \arg\max_{i \in \{1, \cdots, K\}} x_i + z_i$$

where

$$z_i \overset{iid}{\sim} \text{Gumbel}(0, 1)$$

Sampling Discrete (Categorical) Random Variables

Gumbel distribution $\text{Gumbel}(\mu, \beta)$

PDF:

$$p(x) = \frac{1}{\beta} \exp\left(- \left(\frac{x-\mu}{\beta} + \exp(-\frac{x-\mu}{\beta}) \right) \right)$$

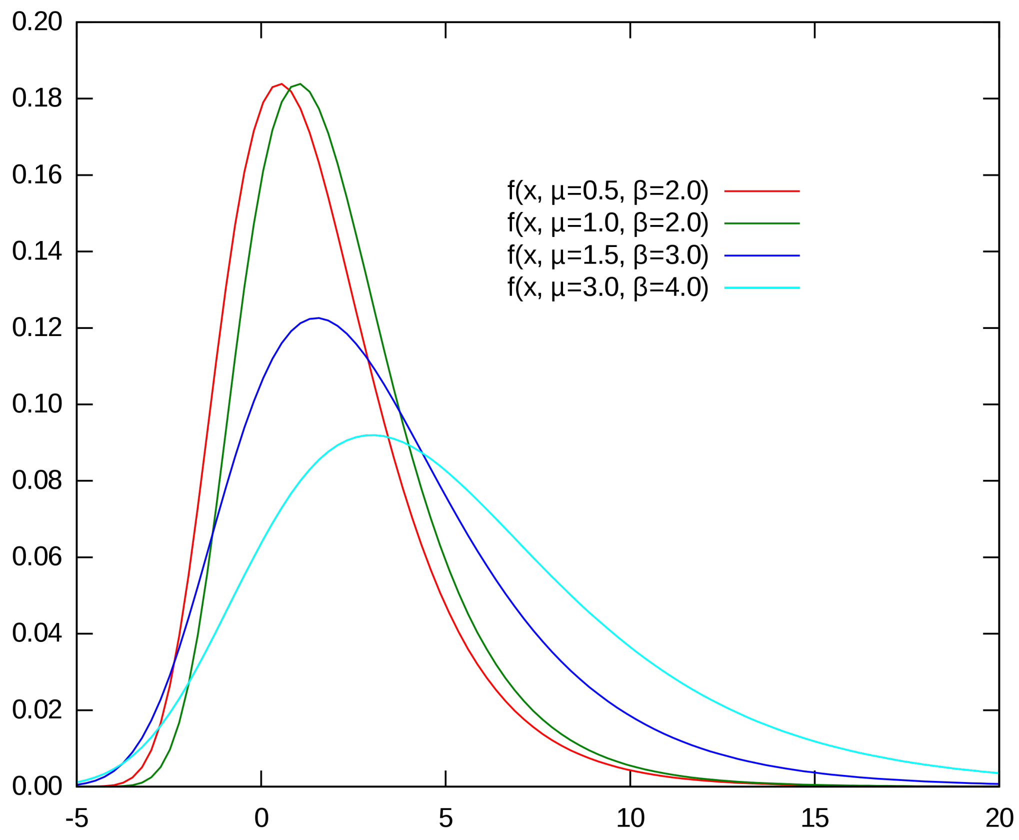


Image Credit: https://en.wikipedia.org/wiki/Gumbel_distribution

Sampling Discrete (Categorical) Random Variables

Gumbel distribution $\text{Gumbel}(\mu, \beta)$

PDF:

$$p(x) = \frac{1}{\beta} \exp\left(-\left(\frac{x-\mu}{\beta} + \exp(-\frac{x-\mu}{\beta})\right)\right)$$

CDF:

$$F(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right)$$

It is used to model the distribution of the maximum (or the minimum) of a number of samples of various distributions

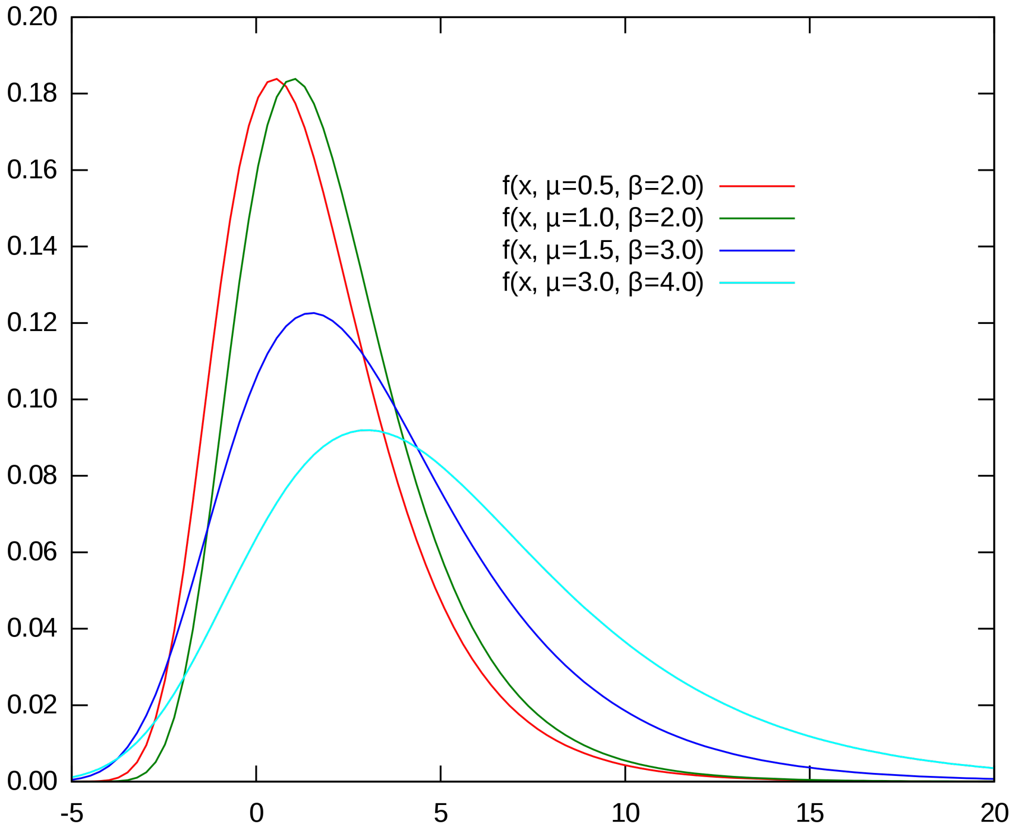


Image Credit: https://en.wikipedia.org/wiki/Gumbel_distribution

Sampling Discrete (Categorical) Random Variables

Gumbel distribution $\text{Gumbel}(0, 1)$

CDF:

$$F(x) = \exp\left(-\exp\left(-x\right)\right)$$

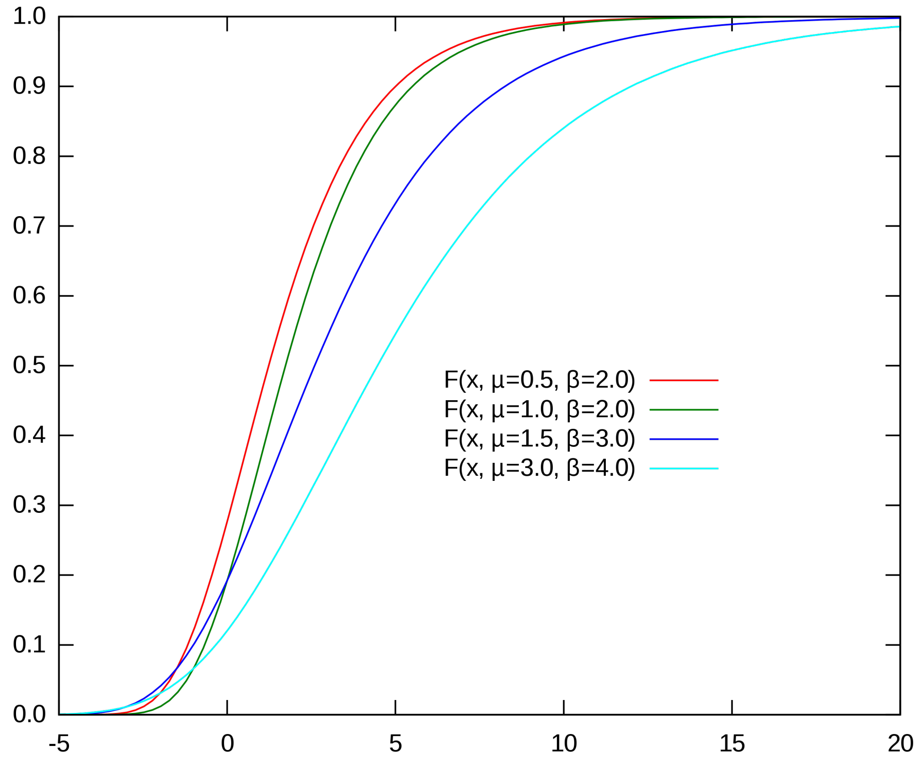


Image Credit: https://en.wikipedia.org/wiki/Gumbel_distribution

Sampling Discrete (Categorical) Random Variables

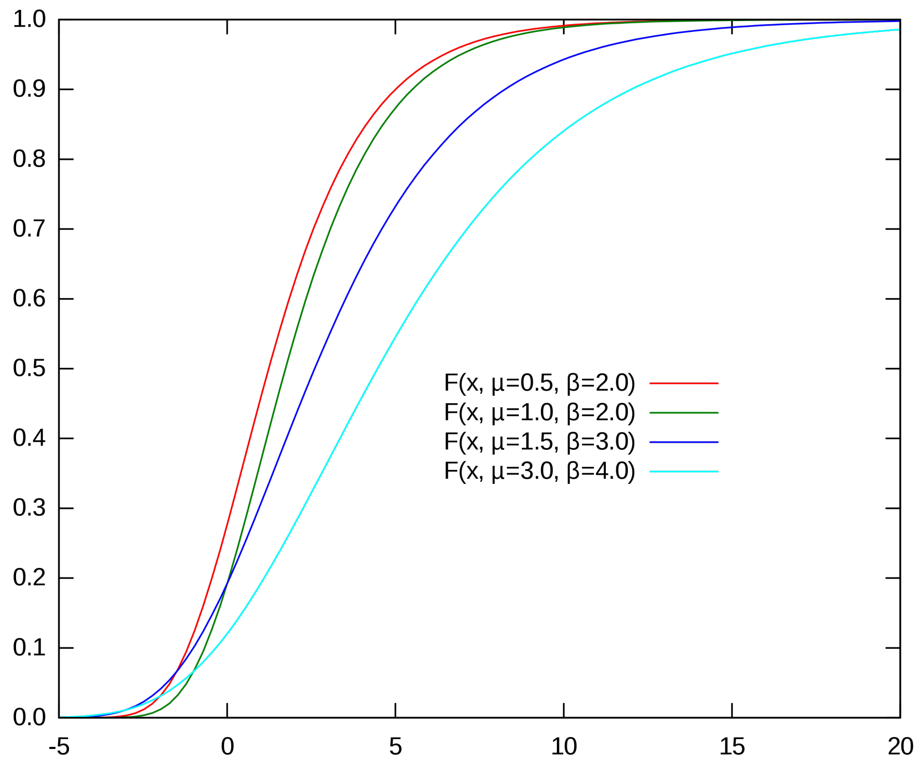
Gumbel distribution $\mathrm{Gumbel}(0, 1)$

CDF:

$$F(x) = \exp\left(-\exp\left(-x\right)\right)$$

Inverse CDF:

$$F^{-1}(u) = -\log\left(-\log\left(u\right)\right)$$



Inverse Transform Sampling is efficient

$$u \sim \mathrm{Uniform}(0, 1)$$
$$x = \mathrm{InverseCDF}(u)$$

Image Credit: https://en.wikipedia.org/wiki/Gumbel_distribution

Sampling Discrete (Categorical) Random Variables

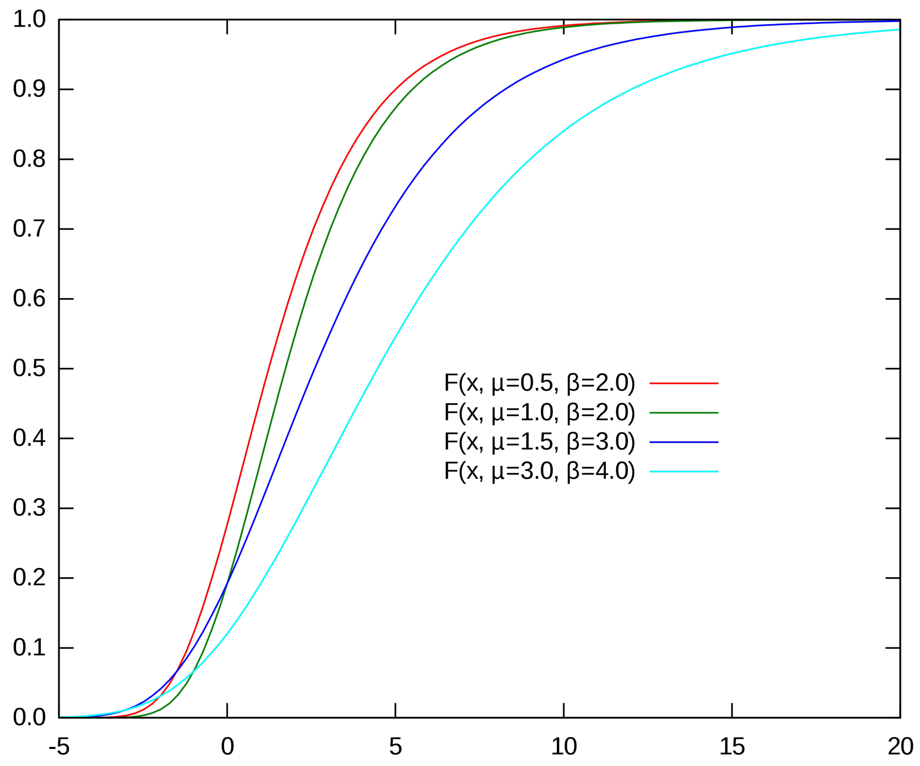
Gumbel distribution $\text{Gumbel}(0, 1)$

CDF:

$$F(x) = \exp\left(-\exp\left(-x\right)\right)$$

Inverse CDF:

$$F^{-1}(u) = -\log\left(-\log\left(u\right)\right)$$



Inverse Transform Sampling is efficient

$$u \sim \text{Uniform}(0, 1)$$
$$x = \text{InverseCDF}(u)$$

Image Credit: https://en.wikipedia.org/wiki/Gumbel_distribution

Sampling Discrete (Categorical) Random Variables

Gumbel Max Trick

Given any real numbers $\{x_i | i = 1, \cdots, K\}$, we sample $z_i \overset{iid}{\sim} \mathrm{Gumbel}(0, 1)$

Sampling Discrete (Categorical) Random Variables

Gumbel Max Trick

Given any real numbers $\{x_i | i = 1, \cdots, K\}$, we sample $z_i \overset{iid}{\sim} \text{Gumbel}(0, 1)$

We have $x_i + z_i \overset{iid}{\sim} \text{Gumbel}(x_i, 1)$

Sampling Discrete (Categorical) Random Variables

Gumbel Max Trick

Given any real numbers $\{x_i | i = 1, \cdots, K\}$, we sample $z_i \overset{iid}{\sim} \text{Gumbel}(0, 1)$

We have $x_i + z_i \overset{iid}{\sim} \text{Gumbel}(x_i, 1)$

The conditional probability that the index of maximum value of $\{x_i + z_i | i = 1, \cdots, K\}$ is i^* is

$$
\begin{aligned}
p\left(i^* = \underset{i \in \{1, \cdots, K\}}{\arg\max}\; x_i + z_i \,\middle|\, \{x_i + z_i | i = 1, \cdots, K\}\right) &= p\left(x_j + z_j \leq x_{i^*} + z_{i^*},\ \forall j \neq i^* | \{x_i + z_i | i = 1, \cdots, K\}\right) \\
&= \prod_{j \neq i^*} p\left(x_j + z_j \leq x_{i^*} + z_{i^*} | \{x_i + z_i | i = 1, \cdots, K\}\right) \\
&= \prod_{j \neq i^*} \text{CDF}_j\left(x_{i^*} + z_{i^*}\right) \\
&= \prod_{j \neq i^*} \exp\left(-\exp(-(x_{i^*} + z_{i^*} - x_j))\right)
\end{aligned}
$$

Sampling Discrete (Categorical) Random Variables

The marginal probability that the index of maximum value of $\{x_i + z_i | i = 1, \cdots, K\}$ is i^* is

$$
\begin{aligned}
&p\left(i^* = \operatorname*{arg\,max}_{i \in \{1,\cdots,K\}} x_i + z_i\right) \\
&= \int p\left(i^* = \operatorname*{arg\,max}_{i \in \{1,\cdots,K\}} x_i + z_i \middle| \{x_i + z_i | i = 1, \cdots, K\}\right) p(\{x_i + z_i\}) d\{x_i + z_i\} \\
&= \int \prod_{j \neq i^*} \exp\left(-\exp(-(x_{i^*} + z_{i^*} - x_j)\right) \prod_j \exp\left(-\left(z_j + \exp(-z_j)\right)\right) d\{z_i\} \\
&= \int \exp\left(-\left(z_{i^*} + \exp(-z_{i^*})\right)\right) \int \prod_{j \neq i^*} \exp\left(-\exp(-(x_{i^*} + z_{i^*} - x_j)\right) \\
&\qquad \exp\left(-\left(z_j + \exp(-z_j)\right)\right) d\{z_j | j \neq i^*\} dz_{i^*} \\
&= \int \exp\left(-\left(z_{i^*} + \exp(-z_{i^*})\right)\right) \prod_{j \neq i^*} \exp\left(-\exp(-(x_{i^*} + z_{i^*} - x_j)\right) dz_{i^*}
\end{aligned}
$$

Sampling Discrete (Categorical) Random Variables

The marginal probability that the index of maximum value of $\{x_i + z_i | i = 1, \cdots, K\}$ is i^* is

$$p\left(i^* = \underset{i \in \{1,\cdots,K\}}{\arg\max}\ x_i + z_i\right)$$

$$= \int p\left(i^* = \underset{i \in \{1,\cdots,K\}}{\arg\max}\ x_i + z_i \middle| \{x_i + z_i | i = 1, \cdots, K\}\right) p(\{x_i + z_i\}) d\{x_i + z_i\}$$

$$= \int \prod_{j \neq i^*} \exp\left(-\exp(-(x_{i^*} + z_{i^*} - x_j)\right) \prod_j \exp\left(-\left(z_j + \exp(-z_j)\right)\right) d\{z_i\}$$

$$= \int \exp\left(-\left(z_{i^*} + \exp(-z_{i^*})\right)\right) \int \prod_{j \neq i^*} \exp\left(-\exp(-(x_{i^*} + z_{i^*} - x_j)\right)$$

$$\exp\left(-\left(z_j + \exp(-z_j)\right)\right) d\{z_j | j \neq i^*\} dz_{i^*}$$

$$= \int \exp\left(-\left(z_{i^*} + \exp(-z_{i^*})\right)\right) \prod_{j \neq i^*} \exp\left(-\exp(-(x_{i^*} + z_{i^*} - x_j)\right) dz_{i^*}$$

$$\text{Gumbel}(0,1) \quad \int p(x) dx = \int \exp\left(-\left(x + \exp(-x)\right)\right) dx = 1$$

Sampling Discrete (Categorical) Random Variables

The marginal probability that the index of maximum value of $\{x_i + z_i | i = 1, \cdots, K\}$ is i^* is

$$
\begin{aligned}
p\left(i^* = \underset{i\in\{1,\cdots,K\}}{\arg\max}\; x_i + z_i\right) &= \int \exp\left(-\left(z_{i^*} + \exp(-z_{i^*})\right)\right) \prod_{j\neq i^*} \exp\left(-\exp(-(x_{i^*} + z_{i^*} - x_j)\right) dz_{i^*} \\
&= \int \exp\left(-z_{i^*} - \exp(-z_{i^*}) - \sum_{j\neq i^*} \exp(-(x_{i^*} + z_{i^*} - x_j)\right) dz_{i^*} \\
&= \int \exp\left(-z_{i^*} - \sum_{j} \exp(-(x_{i^*} + z_{i^*} - x_j)\right) dz_{i^*} \\
&= \int \exp\left(-z_{i^*} - \exp(-z_{i^*}) \sum_{j} \exp(-x_{i^*} + x_j)\right) dz_{i^*} \\
&= \left(\sum_j \exp(-x_{i^*} + x_j)\right)^{-1} \\
&= \frac{\exp(x_{i^*})}{\sum_j \exp(x_j)}
\end{aligned}
$$

Sampling Discrete (Categorical) Random Variables

The marginal probability that the index of maximum value of $\{x_i + z_i | i = 1, \cdots, K\}$ is i^* is

$$p\left(i^* = \operatorname*{arg\,max}_{i \in \{1,\cdots,K\}} x_i + z_i\right) = \int \exp\left(-\left(z_{i^*} + \exp(-z_{i^*})\right)\right) \prod_{j \neq i^*} \exp\left(-\exp(-(x_{i^*} + z_{i^*} - x_j)\right) dz_{i^*}$$

$$= \int \exp\left(-z_{i^*} - \exp(-z_{i^*}) - \sum_{j \neq i^*} \exp(-(x_{i^*} + z_{i^*} - x_j)\right) dz_{i^*}$$

$$= \int \exp\left(-z_{i^*} - \sum_{j} \exp(-(x_{i^*} + z_{i^*} - x_j)\right) dz_{i^*}$$

$$= \int \exp\left(-z_{i^*} - \exp(-z_{i^*}) \sum_{j} \exp(-x_{i^*} + x_j)\right) dz_{i^*}$$

$$= \left(\sum_{j} \exp(-x_{i^*} + x_j)\right)^{-1}$$

$$= \frac{\exp(x_{i^*})}{\sum_j \exp(x_j)}$$

$$\text{Gumbel}(\mu, \beta)$$

$$\int p(x) dx$$

$$= \int \frac{1}{\beta} \exp\left(-\left(\frac{x-\mu}{\beta} + \exp(-\frac{x-\mu}{\beta})\right)\right) dx$$

$$= 1$$

$$\mu = \log\left(\sum_j \exp(-x_{i^*} + x_j)\right)$$

$$\beta = 1$$

Sampling Discrete (Categorical) Random Variables

We have derived *Gumbel-Max Trick*

RV X follows a categorical distribution where $X \in \{1, 2, \cdots, K\}$ and

$$p(X = k) = \frac{\exp(x_k)}{\sum_{i=1}^{K} \exp(x_i)}$$

We have y follows the same distribution as X

$$y = \underset{i \in \{1, \cdots, K\}}{\arg\max} \; x_i + z_i$$

where $$z_i \overset{iid}{\sim} \text{Gumbel}(0, 1)$$

Sampling Discrete (Categorical) Random Variables

We have derived *Gumbel-Max Trick*

RV X follows a categorical distribution where $X \in \{1, 2, \cdots, K\}$ and

$$p(X = k) = \frac{\exp(x_k)}{\sum_{i=1}^{K} \exp(x_i)}$$

We have y follows the same distribution as X

$$y = \underset{i \in \{1, \cdots, K\}}{\arg\max} \; x_i + z_i$$

where $$z_i \overset{iid}{\sim} \text{Gumbel}(0, 1)$$

It can be used for efficient sampling from complicated probabilistic models like discrete MRFs [5]

Stochastic Gradient Estimation Again

Suppose we want to optimize $\mathcal{L}(\phi) = \mathbb{E}_{p_\phi(X)}[f(X)]$ and X is discrete

Stochastic Gradient Estimation Again

Suppose we want to optimize $\mathcal{L}(\phi) = \mathbb{E}_{p_\phi(X)}[f(X)]$ and X is discrete

Recall what we learned previously:

Reparameterization (path-derivative) gradient estimator $$\frac{\partial \mathcal{L}}{\partial \phi} = \frac{\partial}{\partial \phi}\mathbb{E}_{p(\epsilon)}[f(g(\phi, \epsilon))] = \mathbb{E}_{p(\epsilon)}\left[\frac{\partial f}{\partial g}\frac{\partial g}{\partial \phi}\right]$$

REINFORCE gradient estimator $$\frac{\partial \mathcal{L}}{\partial \phi} = \mathbb{E}_{p_\phi(X)}\left[\frac{\partial \log p_\phi(X)}{\partial \phi} f(X)\right]$$

Stochastic Gradient Estimation Again

Suppose we want to optimize $\mathcal{L}(\phi) = \mathbb{E}_{p_\phi(X)}[f(X)]$ and X is discrete

Recall what we learned previously:

Reparameterization (path-derivative) gradient estimator $\frac{\partial \mathcal{L}}{\partial \phi} = \frac{\partial}{\partial \phi}\mathbb{E}_{p(\epsilon)}[f(g(\phi, \epsilon))] = \mathbb{E}_{p(\epsilon)}\left[\frac{\partial f}{\partial g}\frac{\partial g}{\partial \phi}\right]$

We use the reparameterization $X = g(\phi, \epsilon)$

However, for discrete X , we can not differentiate $g(\phi, \epsilon)$!

Stochastic Gradient Estimation Again

Suppose we want to optimize $\mathcal{L}(\phi) = \mathbb{E}_{p_\phi(X)}[f(X)]$ and X is discrete

Recall what we learned previously:

Reparameterization (path-derivative) gradient estimator $\frac{\partial \mathcal{L}}{\partial \phi} = \frac{\partial}{\partial \phi}\mathbb{E}_{p(\epsilon)}[f(g(\phi, \epsilon))] = \mathbb{E}_{p(\epsilon)}\left[\frac{\partial f}{\partial g}\frac{\partial g}{\partial \phi}\right]$

We use the reparameterization $X = g(\phi, \epsilon)$

However, for discrete X, we can not differentiate $g(\phi, \epsilon)$!

Can we find some differentiable approximation?

Straight-Through Estimator

Consider the binary (Bernoulli) case

$$X = g(\phi, \epsilon) = \begin{cases} 1 & \text{if } \epsilon \geq \phi \\ 0 & \text{otherwise} \end{cases}$$

Straight-Through Estimator

Consider the binary (Bernoulli) case

$$X = g(\phi, \epsilon) = \begin{cases} 1 & \text{if } \epsilon \geq \phi \\ 0 & \text{otherwise} \end{cases}$$

Since step-function is non-differentiable, we can approximate it using identity [7, 1] and sigmoid [1]

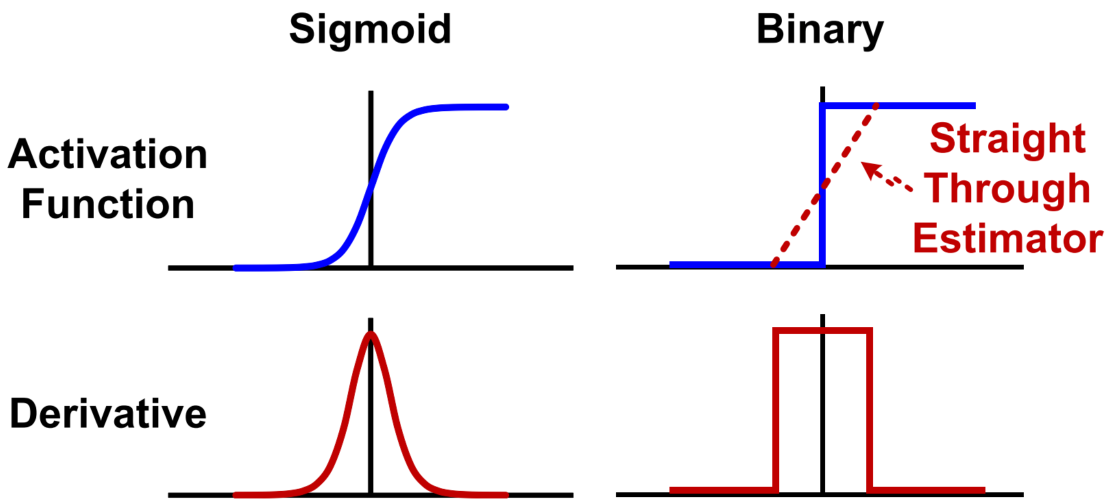


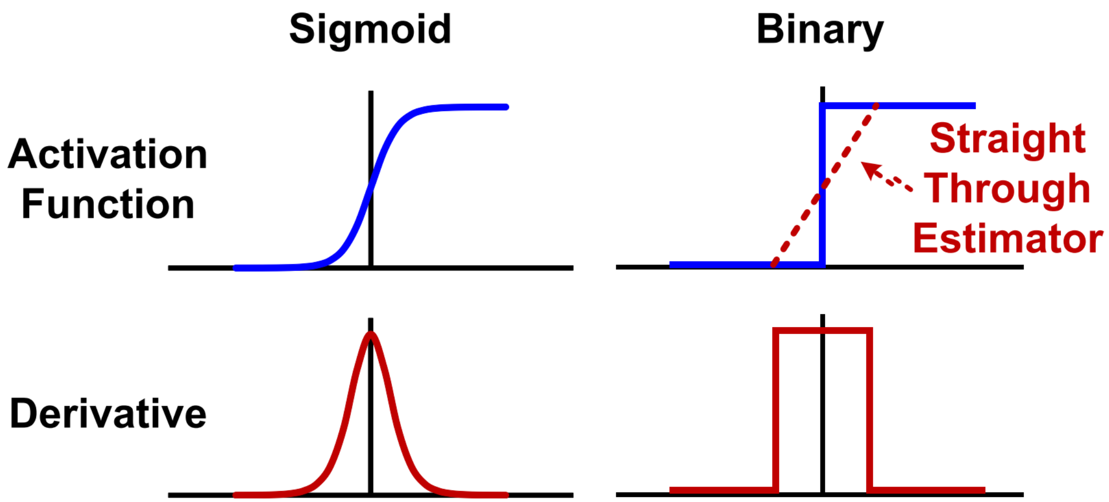
Image Credit: [8]

Straight-Through Estimator

Consider the binary (Bernoulli) case

$$X = g(\phi, \epsilon) = \begin{cases} 1 & \text{if } \epsilon \geq \phi \\ 0 & \text{otherwise} \end{cases}$$

Since step-function is non-differentiable, we can approximate it using identity [7, 1] and sigmoid [1]



We use discrete samples in forward pass and differentiable approximations in backward pass!

Image Credit: [8]

Gumbel-Softmax Estimator

We know *Gumbel-Max Trick* can sample from categorical distributions

RV X follows a categorical distribution where $X \in \{1, 2, \cdots, K\}$ and

$$p(X = k) = \frac{\exp(x_k)}{\sum_{i=1}^{K} \exp(x_i)}$$

We have y follows the same distribution as X

$$y = \underset{i \in \{1, \cdots, K\}}{\arg\max} \; x_i + z_i$$

where $$z_i \overset{iid}{\sim} \text{Gumbel}(0, 1)$$

Gumbel-Softmax Estimator

We know *Gumbel-Max Trick* can sample from categorical distributions

RV X follows a categorical distribution where $X \in \{1, 2, \cdots, K\}$ and

$$p(X = k) = \frac{\exp(x_k)}{\sum_{i=1}^{K} \exp(x_i)}$$

We have y follows the same distribution as X

$$y = \arg\max_{i \in \{1, \cdots, K\}} x_i + z_i$$

where $z_i \stackrel{iid}{\sim} \text{Gumbel}(0, 1)$

However, argmax is non-differentiable!

Softmax is a differentiable approximation of argmax!

Gumbel-Softmax Estimator

Recall in Gumbel-Max trick, we have $$y = \arg\max_{i \in \{1, \cdots, K\}} x_i + z_i$$

$$z_i \overset{iid}{\sim} \text{Gumbel}(0, 1)$$

Use softmax instead of argmax (adding temperature τ), we have

$$y = \frac{\exp\left((x_k + z_k)/\tau\right)}{\sum_{i=1}^{K} \exp\left((x_i + z_i)/\tau\right)}$$

Gumbel-Softmax Estimator

Recall in Gumbel-Max trick, we have $$y = \arg\max_{i \in \{1, \cdots, K\}} x_i + z_i$$

$$z_i \overset{iid}{\sim} \text{Gumbel}(0, 1)$$

Use softmax instead of argmax (adding temperature τ), we have

$$y = \frac{\exp\left((x_k + z_k)/\tau\right)}{\sum_{i=1}^{K} \exp\left((x_i + z_i)/\tau\right)}$$

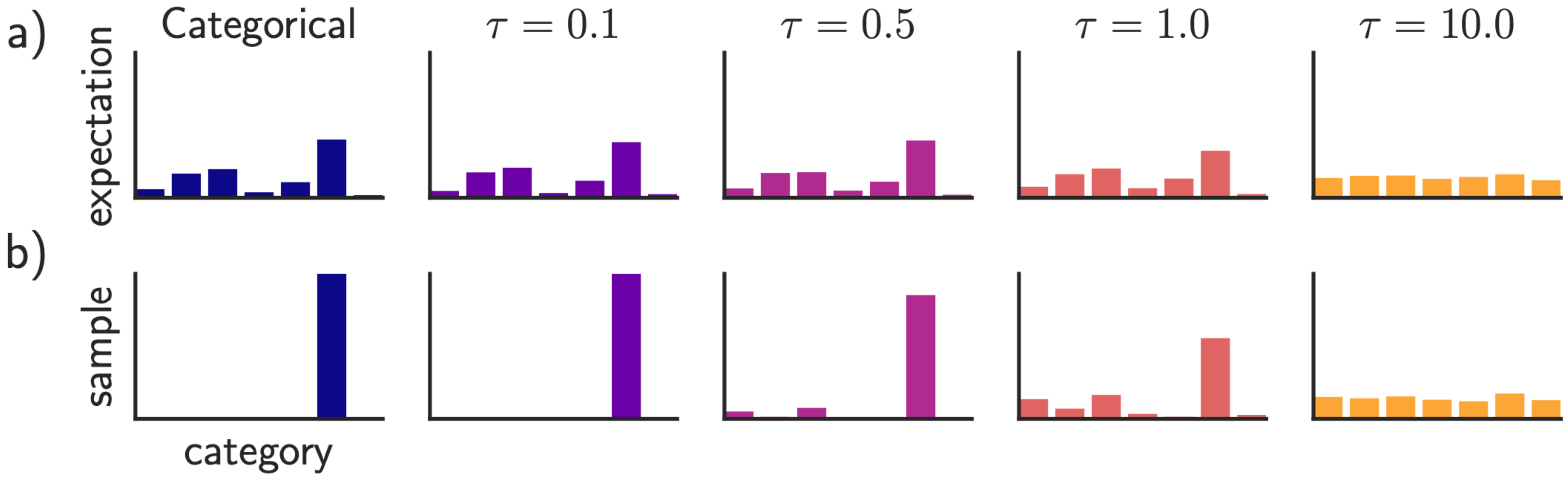
The Gumbel-Softmax distribution (concrete distribution) [2, 3] is then

$$p(y_1, \cdots, y_K) = \Gamma(K)\tau^{K-1} \left(\sum_{i=1}^{K} \frac{\pi_i}{y_i^{\tau}} \right)^{-K} \prod_{i=1}^{K} \frac{\pi_i}{y_i^{\tau+1}} \qquad \pi_i = \frac{\exp(x_k)}{\sum_{i=1}^{K} \exp(x_i)}$$

Gumbel-Softmax Estimator

The Gumbel-Softmax distribution (concrete distribution) [2, 3] is then

$$p(y_1, \cdots, y_K) = \Gamma(K)\tau^{K-1} \left(\sum_{i=1}^{K} \frac{\pi_i}{y_i^{\tau}} \right)^{-K} \prod_{i=1}^{K} \frac{\pi_i}{y_i^{\tau+1}}$$



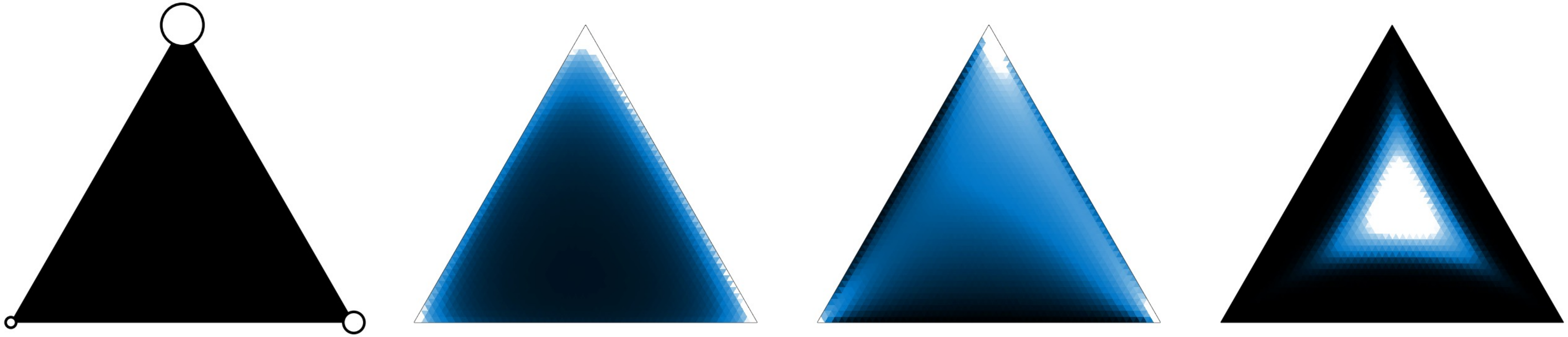
Histogram View

Image Credit: [2]

Gumbel-Softmax Estimator

The Gumbel-Softmax distribution (concrete distribution) [2, 3] is then

$$p(y_1, \cdots, y_K) = \Gamma(K)\tau^{K-1}\left(\sum_{i=1}^{K}\frac{\pi_i}{y_i^{\tau}}\right)^{-K}\prod_{i=1}^{K}\frac{\pi_i}{y_i^{\tau+1}}$$



(a) $\lambda = 0$ (b) $\lambda = 1/2$ (c) $\lambda = 1$ (d) $\lambda = 2$

Simplex View

Image Credit: [3]

Gumbel-Softmax Estimator

The Gumbel-Softmax distribution (concrete distribution) [2, 3] is then

$$p(y_1, \cdots, y_K) = \Gamma(K)\tau^{K-1}\left(\sum_{i=1}^{K}\frac{\pi_i}{y_i^{\tau}}\right)^{-K}\prod_{i=1}^{K}\frac{\pi_i}{y_i^{\tau+1}}$$

Sampling operator is

$$y = \frac{\exp\left((x_k + z_k)/\tau\right)}{\sum_{i=1}^{K}\exp\left((x_i + z_i)/\tau\right)}$$

Since samples are not discrete anymore, it is a biased estimator $\mathcal{L}(\phi) = \mathbb{E}_{p_\phi(X)}\left[f(X)\right]$

Image Credit: [3]

Gumbel-Softmax Estimator

The Gumbel-Softmax distribution (concrete distribution) [2, 3] is then

$$p(y_1, \cdots, y_K) = \Gamma(K)\tau^{K-1} \left(\sum_{i=1}^{K} \frac{\pi_i}{y_i^\tau} \right)^{-K} \prod_{i=1}^{K} \frac{\pi_i}{y_i^{\tau+1}}$$

Sampling operator is

$$y = \frac{\exp\left((x_k + z_k)/\tau\right)}{\sum_{i=1}^{K} \exp\left((x_i + z_i)/\tau\right)}$$

Since samples are not discrete anymore, it is a biased estimator $\mathcal{L}(\phi) = \mathbb{E}_{p_\phi(X)}\left[f(X)\right]$

But samples are differentiable, so we can use reparameterization trick!

Empirically, it has low variances with a reasonable temperature!

Gumbel-Top K for Sequences

For random processes like sequences of discrete random variables, we are interested in sampling most probable sequences, e.g., in language models

(Deterministic) beam search is typically used!

Gumbel Top K Trick:

$$y_1, \cdots, y_m = \text{argtop m}_{i \in \{1, \cdots, K\}} x_i + z_i$$

$$z_i \overset{iid}{\sim} \text{Gumbel}(0, 1)$$

We can generalize Gumbel-Softmax to Gumbel-TopK to construct stochastic beam search [4]!

Gumbel-Top K for Sequences

Stochastic beam search

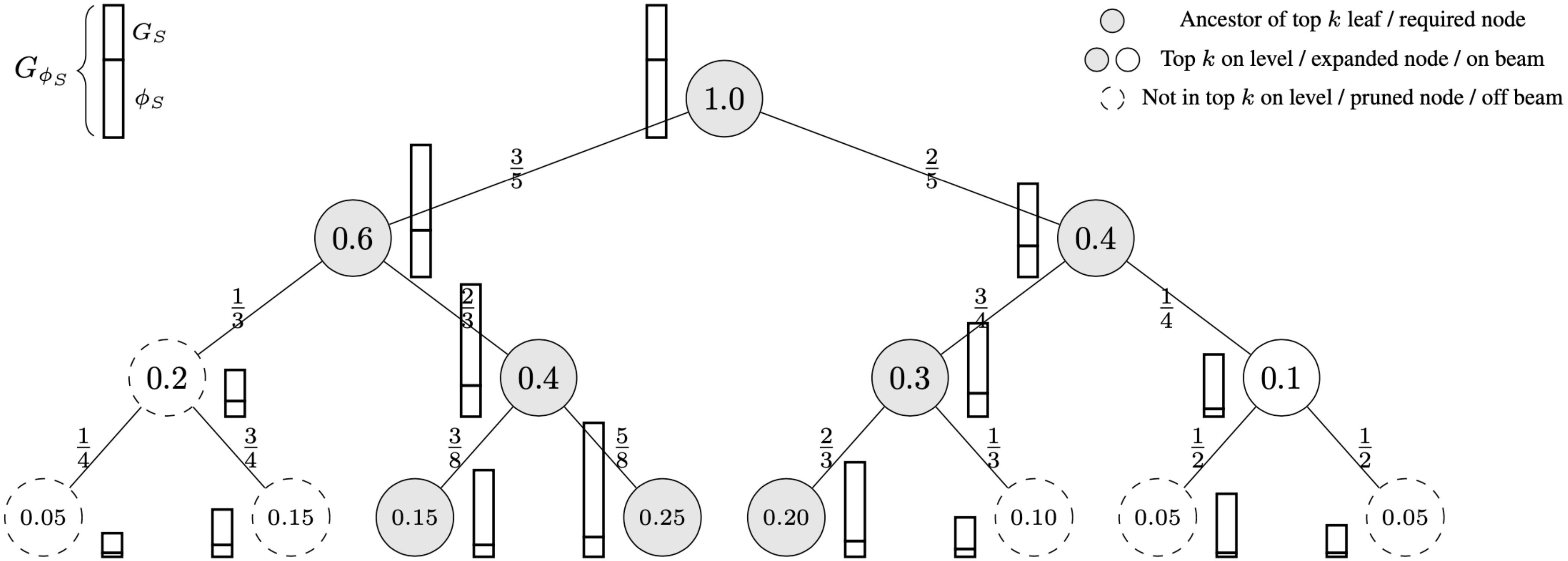


Image Credit: [4]

References

[1] Bengio, Y., Léonard, N. and Courville, A., 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432.

[2] Jang, E., Gu, S. and Poole, B., 2016. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144.

[3] Maddison, C.J., Mnih, A. and Teh, Y.W., 2016. The concrete distribution: A continuous relaxation of discrete random variables. arXiv preprint arXiv:1611.00712.

[4] Kool, W., Van Hoof, H. and Welling, M., 2019, May. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In International Conference on Machine Learning (pp. 3499-3508). PMLR.

[5] Papandreou, G. and Yuille, A.L., 2011, November. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In 2011 International Conference on Computer Vision (pp. 193-200). IEEE.

[6] Gumbel, E.J., 1954. Statistical theory of extreme values and some practical applications: a series of lectures (Vol. 33). US Government Printing Office.

[7] Hinton, G.E., 2012. Neural Networks for Machine Learning. Coursera, video lectures.

[8] Chen, G.K., Kumar, R., Sumbul, H.E., Knag, P.C. and Krishnamurthy, R.K., 2018. A 4096-neuron 1M-synapse 3.8-pJ/SOP spiking neural network with on-chip STDP learning and sparse weights in 10-nm FinFET CMOS. IEEE Journal of Solid-State Circuits, 54(4), pp.992-1002.

Questions?