# Gradient Estimation with Stochastic Softmax Tricks

# Problem statement

1) В некоторых ситуациях выходы модели(например, автокодировщика) удобнее или нужно хранить в виде ohe-векторов, т.е. выход представлен в виде вектора (0,0,…., 1, 0, 0), например argmin по выходу из слоя в виде ohe, где 1 на месте аргумента, чье значение минимально

2) возникает проблема обучения модели(градиенты всегда будут нулевые)

# Возможное решение проблемы

1) Представим что мы генерируем выход ohe векторов как распределение, которое было до применения ohe, например это мог быть слой активаций, целевая функция – матожидание исходной целевой функции по распределению ohe.

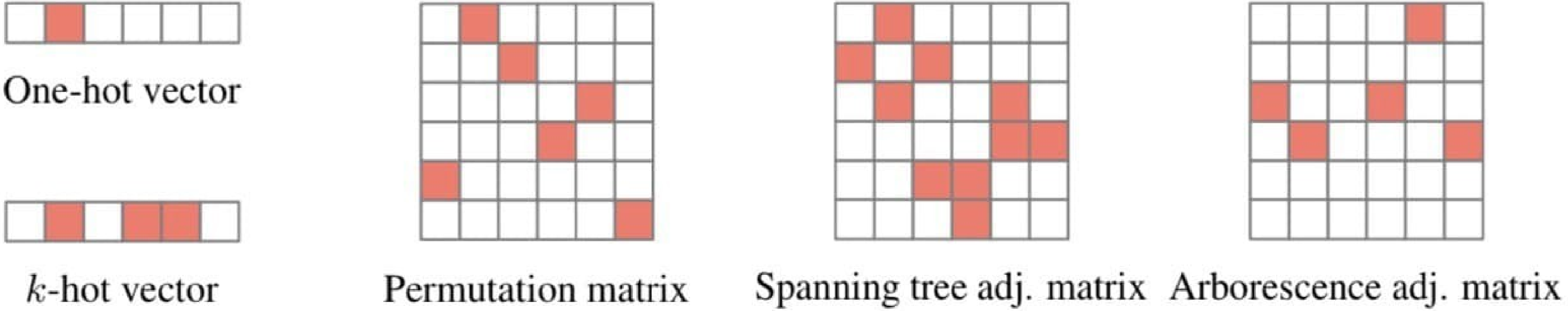
2) теперь нужно уметь считать матожидание, что затруднительно

Given a probability mass function $p_\theta : \mathcal{X} \to (0, 1]$ that is differentiable in $\theta \in \mathbb{R}^m$, a loss function $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$, and $X \sim p_\theta$, our ultimate goal is gradient-based optimization of $\mathbb{E}[\mathcal{L}(X)]$. Thus, we are concerned in this paper with the problem of estimating the derivatives of the expected loss,

$$\frac{d}{d\theta}\mathbb{E}[\mathcal{L}(X)] = \frac{d}{d\theta}\left(\sum\nolimits_{x\in\mathcal{X}} \mathcal{L}(x)p_\theta(x)\right). \tag{1}$$

# Gumbel max and softmax tricks

**Пусть наш выход из модели (из сети) есть некое дискретное конечное множество, которое мы представили в виде ohe или матриц, если это более сложные объекты:**



# Gumbel max and softmax tricks

**Gumbel max trick:**

The GMT is the following identity: for $X \sim p_\theta$ and $G_i + \theta_i \sim \text{Gumbel}(\theta_i)$ indep.,

$$X \stackrel{d}{=} \arg\max_{x \in \mathcal{X}} (G + \theta)^T x.$$

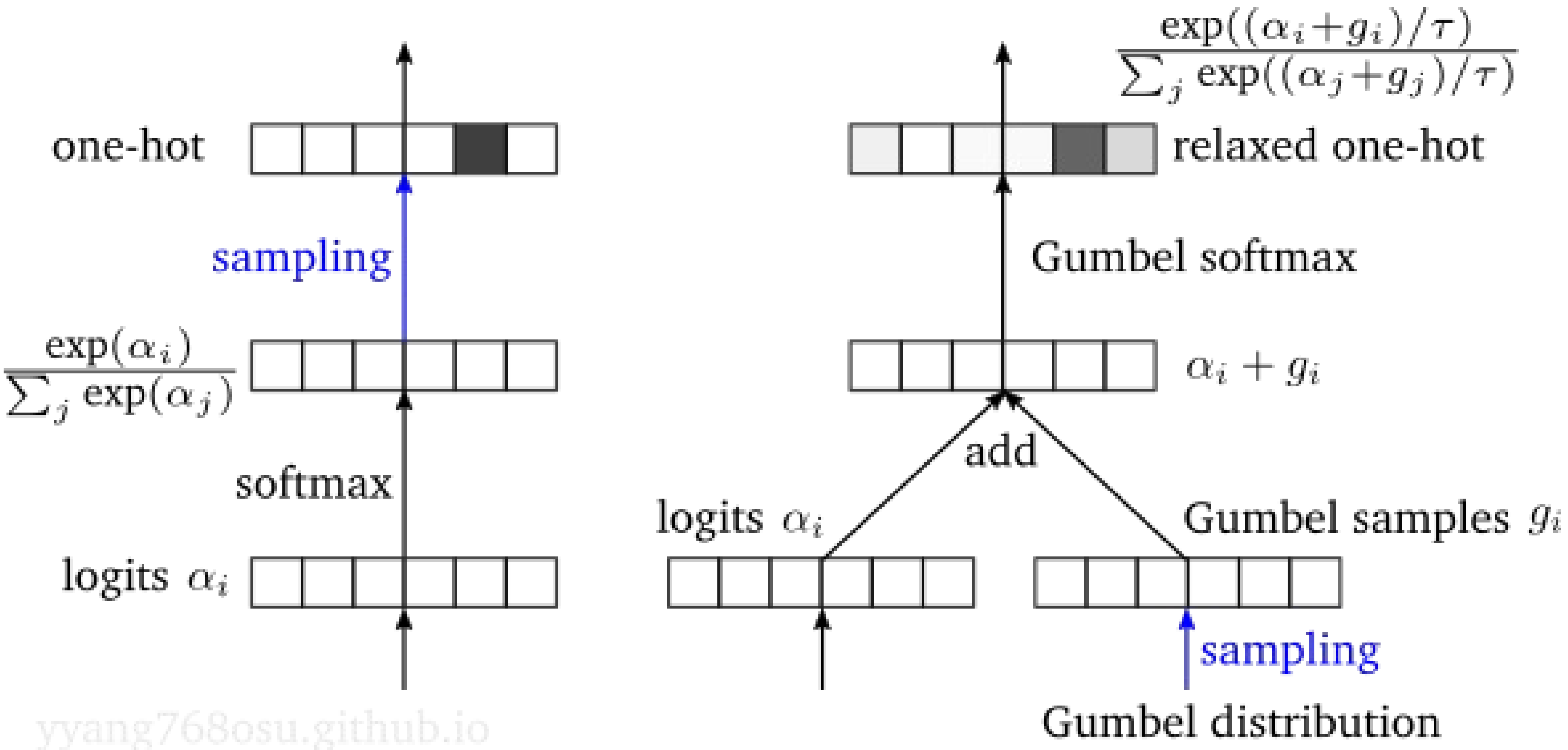
Ideally, one would have a reparameterization estimator, $\mathbb{E}[d\mathcal{L}(X)/d\theta] = d\,\mathbb{E}[\mathcal{L}(X)]/d\theta$

**Gumbel softmax trick:**

$\text{softmax}_t(u)_i = \exp(u_i/t)/\sum_{j=1}^{n} \exp(u_j/t)$ for $u \in \mathbb{R}^n, t > 0$, to continuously approximate $X$,

$$X_t = \text{softmax}_t(G + \theta). \tag{3}$$

# Gumbel softmax trick



# Stochastic argmax and softmax trick

## Stochastic argmax trick

**Definition 1.** *Given a non-empty, convex independent, finite set $\mathcal{X} \subseteq \mathbb{R}^n$ and a random utility $U$ whose distribution is parameterized by $\theta \in \mathbb{R}^m$, a* stochastic argmax trick *for $X$ is the linear program,*

$$X = \arg\max_{x \in \mathcal{X}} U^T x. \tag{4}$$

## Stochastic softmax trick

Given an SMT, an SST incorporates a strongly convex regularizer to the linear objective, and expands the state space to the convex hull of the embeddings $\mathcal{X} = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}^n$,

$$P := \mathrm{conv}(\mathcal{X}) := \left\{ \sum_{i=1}^{m} \lambda_i x_i \,\middle|\, \lambda_i \geq 0, \sum_{i=1}^{m} \lambda_i = 1 \right\}. \tag{5}$$

Expanding the state space to a convex polytope makes it path-connected, and the strongly convex regularizer ensures that the solutions are continuous over the polytope.

**Definition 2.** *Given a stochastic argmax trick $(\mathcal{X}, U)$ where $P := \mathrm{conv}(\mathcal{X})$ and a proper, closed, strongly convex function $f : \mathbb{R}^n \to \{\mathbb{R}, \infty\}$ whose domain contains the relative interior of $P$, a* stochastic softmax trick *for $X$ at temperature $t > 0$ is the convex program,*

$$X_t = \arg\max_{x \in P} U^T x - t f(x) \tag{6}$$

# Main trick with mean function

**С помощью трюка мы можем оценить мат ожидание без трудоемких вычислений**

**Proposition 1.** *If $X$ in Def. 1 is a.s. unique, then for $X_t$ in Def. 2, $\lim_{t \to 0^+} X_t = X$ a.s. If additionally $\mathcal{L} : P \to \mathbb{R}$ is bounded and continuous, then $\lim_{t \to 0^+} \mathbb{E}[\mathcal{L}(X_t)] = \mathbb{E}[\mathcal{L}(X)]$.*

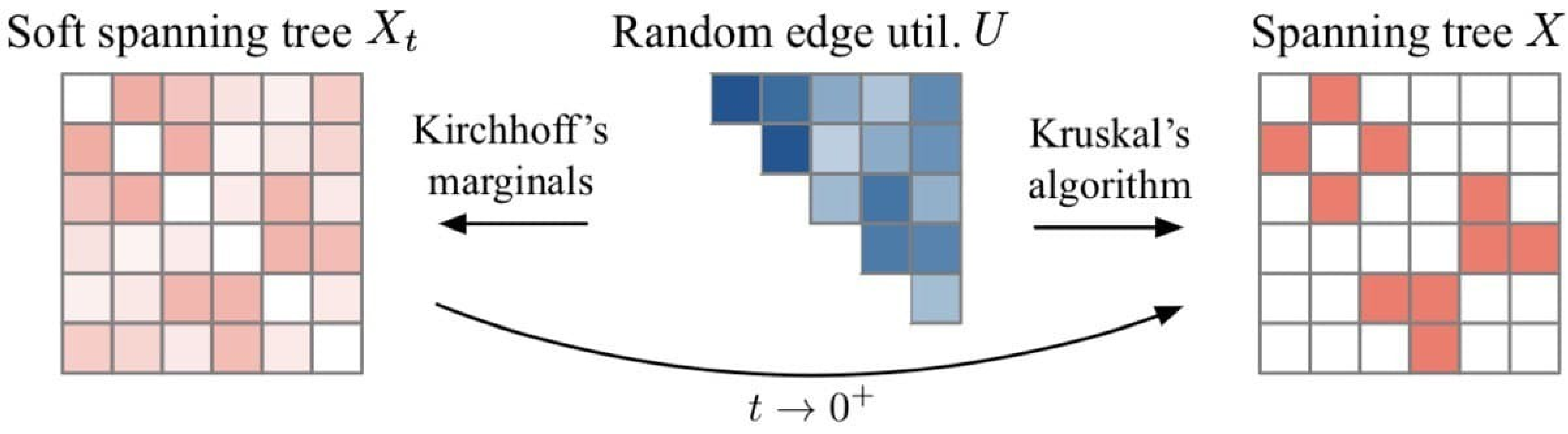
It is common to consider temperature parameters that interpolate between marginal inference and a deterministic, most probable state. While superficially similar, our relaxation framework is different; as $t \to 0^+$, an SST approaches *a sample from the SMT model* as opposed to a deterministic state.

$X_t$ also admits a reparameterization trick. The SST reparameterization gradient estimator given by,

$$\frac{d\mathcal{L}(X_t)}{d\theta} = \frac{\partial \mathcal{L}(X_t)}{\partial X_t} \frac{\partial X_t}{\partial U} \frac{dU}{d\theta}. \tag{7}$$

If $\mathcal{L}$ is differentiable on $P$, then this is an unbiased estimator[6] of the gradient $d\mathbb{E}[\mathcal{L}(X_t)]/d\theta$, because $X_t$ is continuous and a.e. differentiable:

# Пример с spanning tree



# Вопросы:

1) проблема оценки градиента (слайд 2)

2) выписать формулы GMT GST (слайд 5)

3) выписать формулы SMT SST (слайд 7)