The Gumbel Trick

Irene Chen

1 Definitions and useful lemmas

Definition 1 (Gumbel distribution)**.** A variable X drawn from $Gumbel(\mu)$ has the probability distribution

$$f_x(x) = \exp(-(z + \exp(-z))$$

where $z = x - \mu$. We set the scale parameter $\beta = 1$.

Lemma 1 (Change of variables)**.** *If X has probability density function f_X, u is monotonic,$v(x) := u^{-1}(x)$, then the probability density function of $Y = u(X)$ is*

$$f_y(y) = f_x(v(y)) \cdot |v'(y)|$$

Lemma 2. *If $X \sim Exp(\lambda)$ and $g(x) := -\ln x - c$, then $g(X) \sim Gumbel(-c + \ln \lambda)$.*

Proof. Using the change of variables equation (Lemma 1) and letting $h = g^{-1} = e^{-(y+c)}$

$$\begin{aligned}
f_y(y) &= f_x(h(y)) \cdot |v'(y)| \\
&= \lambda e^{-\lambda e^{-(y+c)}} e^{-(y+c)} \\
&= e^{-(y+c)+\ln\lambda+e^{-(y+c)+\ln\lambda)}} \\
&= e^{-(y+c-\ln\lambda)+e^{-(y+c-\ln\lambda)}}
\end{aligned}$$

This is in fact the probability distribution for $Gumbel(-c + \ln \lambda)$.

Lemma 3. *A variable $Y \sim Gumbel(-c + \ln \lambda)$ can also be represented as $\ln \lambda + \gamma$ where $\gamma \sim Gumbel(-c)$.*

Proof. Let $g(x) = x + \ln \lambda$, and $g^{-1}(x) = h(x) = y - \ln \lambda$. Using change of variables again, we see

$$\begin{aligned}
f_y(y) &= f_x(h(y)) \cdot |v'(y)| \\
&= e^{-(y-\ln\lambda-c+e^{-(y-\ln\lambda-c)})} \cdot 1
\end{aligned}$$

which is in fact the probability density function for $Gumbel(-c + \ln \lambda)$

2 Actual trick

The Gumbel trick allows us to estimate the log partition function $\ln Z$ but perturbing the potential functions with noise distributed according to the Gumbel distribution. For finite sample $\mathcal{X}$ of size N, we define an unnormalized mass function $\tilde{p} : \mathcal{X} \to [0, \infty)$ and let $Z := \sum_{x \in \mathcal{X}} \tilde{p}(x)$ be the normalizing partition function. We then define $\phi(x) := \ln \tilde{p}(x)$ as the log-unnormalized probabilities or the potential function.

We then seek to show that adding selective noise to our potential functions allows for easy recovery of the partition function or at least the log-partition function. We exploit the convenient property that finding the maximum a posteriori (MAP) is NP-hard but can be efficiently computed or estimated in practice. Note that the partition function is a harder probably, containing #P-hard problems.

We want to show the usefulness of the Gumbel trick then using the *Perturb-and-MAP* method, specifically

$$\max_{x \in \mathcal{X}} \{\phi(x) + \gamma(x)\} \sim \mathit{Gumbel}(-c + \ln Z)$$

In the case of exponential competing clocks, we consider N clocks $\{T_x\}_{c \in \mathcal{X}}$ with respective rates λ_x. Conveniently, we can use the following property about the smallest wait time until any of the exponential clocks goes off.

$$\min_{x \in \mathcal{X}} \{T_x\} \sim \mathit{Exp}(Z)$$

We then show the following

$$\begin{aligned}
\min_{x \in \mathcal{X}} \{T_x\} &\sim \mathit{Exp}(Z) && \text{(Given)} \\
g(\min_{x \in \mathcal{X}} \{T_x\}) &\sim g(\mathit{Exp}(Z)) && \text{(apply } g\text{)} \\
\max_{x \in \mathcal{X}} \{g(T_x)\} &\sim g(\mathit{Exp}(Z)) && (g \text{ is negative}) \\
\max_{x \in \mathcal{X}} \{g(T_x)\} &\sim \mathit{Gumbel}(-c + \ln Z) && \text{(Lemma 2)}
\end{aligned}$$

From here, we define for convenience $\beta(x) \sim \mathit{Gumbel}(-c + \ln \lambda_x)$ and $\gamma(x) \sim \mathit{Gumbel}(-c)$. Continuing onward, we see

$$\begin{aligned}
\max_{x \in \mathcal{X}} \{\beta(x)\} &\sim \mathit{Gumbel}(-c + \ln Z) && \text{(Lemma 2)} \\
\max_{x \in \mathcal{X}} \{\gamma(x) + \ln \lambda_x\} &\sim \mathit{Gumbel}(-c + \ln Z) && \text{(Lemma 3)} \\
\max_{x \in \mathcal{X}} \{\gamma(x) + \phi(x)\} &\sim \mathit{Gumbel}(-c + \ln Z) && \text{(definition of } \phi(x)\text{)}
\end{aligned}$$