Learning Randomly Perturbed Structured Predictors for Direct Loss Minimization Supplementary Material

1 Connecting variance of Gumbel random variables and temperature of Gibbs models

We focus on the Gumbel distribution with zero mean, which is described by its a double exponential cumulative distribution function

$$G(t) = P(\gamma(y) \leq t) = e^{-e^{-(t+c)}} \tag{1}$$

where $c \approx 0.5772$ is the Euler-Mascheroni constant.

We will show that then one obtains the following identity:

$$e^{\frac{\mu_w(x,y)}{\sigma(x)}} = P_{\gamma \sim g}[y^* = y] \tag{2}$$

, when

$$y^* = \arg\max_{\hat{y}} \{\mu_w(x, \hat{y}) + \sigma(x)\gamma(\hat{y})\}. \tag{3}$$

Let us define: $\hat{Z}(\mu, \sigma) = \sum_{y \in Y} e^{\frac{\mu(x,y)}{\sigma(x)}}$ and $Z(\mu) = \sum_{y \in Y} e^{\mu(x,y)}$

Theorem 1. *Let $\gamma = \{\gamma(y) : y \in Y\}$ be a collection of i.i.d. Gumbel random variables with cumulative distribution function (1). Then, the random variable $\max_{y \in Y}\{\frac{\mu(x,y)}{\sigma(x)} + \gamma(y)\}$ is distributed according to the Gumbel distribution whose mean is the log-partition function $log\hat{Z}(\mu, \sigma)$.*

Proof. We denote by $F(t) = P(\gamma(y) \leq t)$ the cumulative distribution function of $\gamma(y)$.

The independence of $\gamma(y)$ across $y \in Y$ implies that:

$$\begin{aligned} P_\gamma(\max_{y \in Y}\{\frac{\mu(x,y)}{\sigma(x)} + \gamma(y)\} \leq t) &= P_\gamma(\forall_{y \in Y}\{\frac{\mu(x,y)}{\sigma(x)} + \gamma(y)\} \leq t) \\ &= P_\gamma(\forall_{y \in Y}\{\gamma(y)\} \leq t - \frac{\mu(x,y)}{\sigma(x)}) \\ &= \prod_{y \in Y} F(t - \frac{\mu(x,y)}{\sigma(x)}) \end{aligned}$$

The Gumbel, Frechet, and Weibull distributions, used in extremal statistics, are max-stable distributions: the product $\prod_{y\in Y} F(t - \frac{\mu(x,y)}{\sigma(x)})$ can be described in terms of F$(\cdot)$ itself. Under the said setting, the double exponential form of the Gumbel distribution yields the result:

$$\begin{aligned}\prod_{y\in Y} F(t - \frac{\mu(x,y)}{\sigma(x)}) &= e^{-\sum_{y\in Y} e^{-(t-\frac{\mu(x,y)}{\sigma(x)}+c)}} \\ &= e^{-e^{-(t+c-\log \hat{Z}(\mu,\sigma))}} \\ &= F(t - \log \hat{Z}(\mu,\sigma))\end{aligned}$$

□

Corollary 1. *Let $\gamma = \{\gamma(y) : y \in Y\}$ be a collection of i.i.d. Gumbel random variables with cumulative distribution function (1). Then, for all $\hat{y}$:*

$$\frac{e^{\frac{\mu(x,y)}{\sigma(x)}}}{\hat{Z}(\mu,\sigma)} = P_\gamma(\hat{y} = \arg\max_{y\in Y}\{\mu(x,y) + \sigma(x)\gamma(y)\})$$

Proof. For Gumbel random variables with cumulative distribution function (1) it holds:

$$G'(t) = e^{-t}G(t) = e^{-t}e^{-e^{-t}} = e^{-t-e^{-t}} = e^{-(t+e^{-t})} = g(t) \tag{4}$$

g(t) is the probability density function of the standard Gumbel distribution.

We note that:

$$\begin{aligned}P_\gamma(\max_{y\in Y}\{\mu(x,y) + \sigma(x)\gamma(y)\}) &= \frac{\sigma(x)}{\sigma(x)} P_\gamma(\max_{y\in Y}\{\mu(x,y) + \sigma(x)\gamma(y)\}) \\ &= \sigma(x) P_\gamma(\max_{y\in Y}\{\frac{\mu(x,y)}{\sigma(x)} + \gamma(y)\})\end{aligned}$$

From Theorem 1, we have $\mathbb{E}_\gamma[\max_{y\in Y}\{\frac{\mu(x,y)}{\sigma(x)} + \gamma(y)\}] = \log \hat{Z}(\mu,\sigma)$

Putting it together we have that: $\mathbb{E}_\gamma(\max_{y\in Y}\{\mu(x,y)+\sigma(x)\gamma(y)\}) = \sigma(x)\log \hat{Z}(\mu,\sigma)$. We can derive w.r.t. some $\mu'(x,y)$.

We note that by differentiating the right hand side we get:

$$\frac{\partial(\sigma(x)\log \hat{Z}(\mu,\sigma))}{\partial\mu'(x,y)} = \frac{e^{\frac{\mu'(x,y)}{\sigma(x)}}}{\hat{Z}(\mu,\sigma)}$$

Differentiate the left hand side: First, we can differentiate under the integral sign:

$$\frac{\partial}{\partial\mu'(x,y)} \int_{\mathbb{R}^{|Y|}} \max_{y\in Y}\{\mu(x,y)+\sigma(x)\gamma(y)\} d\gamma = \int_{\mathbb{R}^{|Y|}} \frac{\partial}{\partial\mu'(x,y)} \max_{y\in Y}\{\mu(x,y)+\sigma(x)\gamma(y)\} d\gamma$$

We can write a subgradient of the max-function using an indicator function (an application of Danskin's Theorem):

$$\frac{\partial}{\partial \mu'(x,y)} max_{y\in Y}\{\mu(x,y)+\sigma(x)\gamma(y)\} = \mathbb{1}(\hat{y} = \arg\max_{y\in Y}\{\mu(x,y)+\sigma(x)\gamma(y)\})$$

The corollary then follows by applying the expectation to both sides of the last equation. □

An alternative proof of the preceding corollary can also be made. We begin by noting that: $P_\gamma(\hat{y} = \arg\max_{y\in Y}\{\mu(x,y)+\sigma(x)\gamma(y)\}) = P_\gamma(\hat{y} = \arg\max_{y\in Y}\{\frac{\mu(x,y)}{\sigma(x)}+\gamma(y)\})$. Then,

$$\begin{aligned}
P_\gamma(\hat{y} = \arg\max_{y\in Y}\{\mu(x,y)+\sigma(x)\gamma(y)\}) &= \int_t G'(t-\frac{\mu(x,y)}{\sigma(x)})\prod_{\hat{y}\neq y} G(t-\frac{\mu(x,\hat{y})}{\sigma(x)})dt \\
&\star = \int_t e^{-(t-\frac{\mu(x,y)}{\sigma(x)})} G(t-\frac{\mu(x,y)}{\sigma(x)})\prod_{\hat{y}\neq y} G(t-\frac{\mu(x,\hat{y})}{\sigma(x)})dt \\
&= e^{\frac{\mu(x,y)}{\sigma(x)}}\int_t e^{-t}\prod_{\hat{y}} G(t-\frac{\mu(x,\hat{y})}{\sigma(x)})dt \\
&= e^{\frac{\mu(x,y)}{\sigma(x)}} * constant
\end{aligned}$$

Therefore the probability that $\hat{y}$ maximizes $\mu(x,y)+\sigma(x)\gamma(y)$ is proportional to $e^{\frac{\mu(x,y)}{\sigma(x)}}$.

$\star$ is due to the probability density function of the Gumbel distribution as shown in (4).

2 Proof of Corollary 2

Recall that we defined the prediction $y^*_{w,\gamma} =$

$$\arg\max_{\hat{y}}\Big\{\sum_{\alpha\in\mathcal{A}}\mu_{u,\alpha}(x,\hat{y}_\alpha)+\sum_{i=1}^{n}\sigma_v(x)\gamma_i(\hat{y}_i)\Big\} \tag{5}$$

The loss-perturbed prediction $y^*_{w,\gamma}(\epsilon) =$

$$\arg\max_{\hat{y}}\Big\{\sum_{\alpha\in\mathcal{A}}\mu_{u,\alpha}(x,\hat{y}_\alpha)+\sum_{i=1}^{n}\sigma_v(x)\gamma_i(\hat{y}_i)+\epsilon\ell(y,\hat{y})\Big\} \tag{6}$$

$w=(u,v)$ are the learned parameters.

Our aim is to prove the following gradient steps: $\frac{\partial}{\partial u}\mathbb{E}_\gamma[\ell(y,y^*_{w,\gamma})] =$

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\mathbb{E}_\gamma\Big[\sum_{\alpha\in\mathcal{A}}(\nabla\mu_{u,\alpha}(x,y^*_\alpha(\epsilon))-\nabla\mu_{u,\alpha}(x,y^*_\alpha))\Big] \tag{7}$$

and $\frac{\partial}{\partial v}\mathbb{E}_\gamma[\ell(y, y^*_{w,\gamma})] =$

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\mathbb{E}_\gamma\Big[\sum_{i=1}^{n}\nabla\sigma_v(x)\Big(\gamma_i(y_i^*(\epsilon)) - \gamma_i(y_i^*)\Big)\Big]. \tag{8}$$

When we use the shorthand notation $y^*_\alpha = y^*_{w,\gamma,\alpha}$ and $y^*_i = y^*_{i,w,\gamma}$ and similarly $y^*_\alpha(\epsilon) = y^*_{w,\gamma,\alpha}(\epsilon)$ and $y^*_i(\epsilon) = y^*_{i,w,\gamma}(\epsilon)$ and recall that w refers to u and v.

The main challenge is to show that $G(u, v, \epsilon)$, as defined in Equation (11), is differentiable, i.e., there exists a vector $\sum_\alpha \nabla_u\mu_{u,\alpha}(x, y^*_{u,v,\gamma,\alpha}(\epsilon))$ such that for any direction z, its corresponding directional derivative $\lim_{h\to 0}\frac{G(u+hz,v,\epsilon)-G(u,v,\epsilon)}{h}$ equals $\mathbb{E}_{\gamma\sim\mathcal{G}}[\sum_\alpha \nabla_u\mu_{u,\alpha}(x, y^*_{u,v,\gamma,\alpha}(\epsilon))^\top z]$.

Similarly, we will show that there exists a vector $\sum_{i=1}^n \nabla_v\sigma_v(x)\gamma_i(y^*_{i,v,u,\gamma}(\epsilon))$ such that for any direction z, its corresponding directional derivative $\lim_{h\to 0}\frac{G(u,v+hz,\epsilon)-G(u,v,\epsilon)}{h}$ equals $\mathbb{E}_{\gamma\sim\mathcal{G}}[\sum_{i=1}^n \nabla_v\sigma_v(x)\gamma_i(y^*_{i,u,v,\gamma}(\epsilon))^\top z]$.

This challenge is addressed in Theorem 2, which also utilizes Lemma 1. This lemma relies of the discrete nature of the label space, ensuring that the optimal label does not change in the vicinity of $y^*_{u,v,\gamma}(\epsilon)$. The proof concludes by the Hessian of $G(u, v, \epsilon)$ symmetric entries in Corollary 2.

Lemma 1. *Assume $\{\mu_{u,\alpha}(x, y_\alpha)\}$ is a set of continuous functions of u for $\alpha \in \mathcal{A}$ and assume $\sigma_v(x)$ is a smooth function of v. Let $\gamma_i(y_i)$ be i.i.d. random variables with a smooth probability density function $\mathcal{G}$. Assume that the loss-perturbed maximal arguments $y^*_{u+\frac{1}{n}z,v,\gamma}(\epsilon)$ and $y^*_{u,v+\frac{1}{n}z,\gamma}(\epsilon)$, as defined in Equation (6), are unique for any z and n. Then, there exists n_0 such that for $n \geq n_0$ there holds*

$$y^*_{u+\frac{1}{n}z,v,\gamma}(\epsilon) = y^*_{u,v,\gamma}(\epsilon) \tag{9}$$

and there exists n_1 such that for $n \geq n_1$ there holds

$$y^*_{u,v+\frac{1}{n}z,\gamma}(\epsilon) = y^*_{u,v,\gamma}(\epsilon). \tag{10}$$

Proof. We will first prove Equation (9). Let $f_n(\hat{y}) = \sum_\alpha \mu_{u+\frac{1}{n}z,\alpha}(x, \hat{y}_\alpha) + \sum_{i=1}^n \sigma_v(x)\gamma_i(\hat{y}_i) + \epsilon\ell(y, \hat{y})$ so that $y^*_{u+\frac{1}{n}z,v,\gamma}(\epsilon) = \arg\max_{\hat{y}} f_n(\hat{y})$. Also, let $f_\infty(\hat{y}) = \sum_\alpha \mu_{u,\alpha}(x, \hat{y}_\alpha)+\sum_{i=1}^n \sigma_v(x)\gamma_i(\hat{y}_i)+\epsilon\ell(y, \hat{y})$ so that $y^*_{u,v,\gamma}(\epsilon) = \arg\max_{\hat{y}} f_\infty(\hat{y})$. Since $f_n(\hat{y})$ is a continuous function then $\max_{\hat{y}} f_n(\hat{y})$ is also a continuous function and $\lim_{n\to\infty}\max_{\hat{y}} f_n(\hat{y}) = \max_{\hat{y}} f_\infty(\hat{y})$. Since $\max_{\hat{y}} f_n(\hat{y}) = f_n(y^*_{u+\frac{1}{n}z,v,\gamma}(\epsilon))$ is arbitrarily close to $\max_{\hat{y}} f_\infty(\hat{y}) = f_\infty(y^*_{u,v,\gamma}(\epsilon))$, and $y^*_{u,v,\gamma}(\epsilon), y^*_{u+\frac{1}{n}z,v,\gamma}(\epsilon)$ are unique then for any $n \geq n_0$ these two arguments must be the same, otherwise there is a $\delta > 0$ for which $|f_\infty(y^*_{u,v,\gamma}(\epsilon)) - f_n(y^*_{u+\frac{1}{n}z,v,\gamma}(\epsilon))| \geq \delta$.

To prove Equation (10), one can define

$f'_n(\hat{y}) = \sum_\alpha \mu_{u,\alpha}(x, \hat{y}_\alpha) + \sum_{i=1}^n \sigma_{v+\frac{1}{n}z}(x)\gamma_i(\hat{y}_i) + \epsilon\ell(y, \hat{y})$ and follow the same steps to show that for any $n \geq n_1$ it holds that $f'_\infty(y^*_{u,v,\gamma}(\epsilon))$ and $f'_n(y^*_{u,v+\frac{1}{n}z,\gamma}(\epsilon))$ are arbitrarily close and since $y^*_{u,v+\frac{1}{n}z,\gamma}(\epsilon)$ and $y^*_{u,v,\gamma}(\epsilon)$ are unique then $y^*_{u,v+\frac{1}{n}z,\gamma}(\epsilon) = y^*_{u,v,\gamma}(\epsilon)$. □

Theorem 2. *Assume that* $E_{\gamma\sim\mathcal{G}}\|\nabla_u \mu_{u,\alpha}(x,y_\alpha)\| \leq \infty$, *and that* $E_{\gamma\sim\mathcal{G}}\|\nabla_v \sigma_v(x)\| \leq \infty$. *Define the prediction generating function* $G(u,v,\epsilon) =$

$$\mathbb{E}_{\gamma\sim\mathcal{G}}\Big[\max_{\hat{y}\in Y}\Big\{\sum_{\alpha\in\mathcal{A}}\mu_{u,\alpha}(x,\hat{y}_\alpha)+\sum_{i=1}^{n}\sigma_v(x)\gamma_i(\hat{y}_i)+\epsilon\ell(y,\hat{y})\Big\}\Big]. \tag{11}$$

If the conditions of Lemma 1 hold then $G(u,v,\epsilon)$ *as defined in Equation (11) is differentiable and*

$$\frac{\partial G(u,v,\epsilon)}{\partial u} = \mathbb{E}_{\gamma}\Big[\sum_{\alpha}\nabla_u\mu_{u,\alpha}(x,y^*_\alpha(\epsilon))\Big] \tag{12}$$

$$\frac{\partial G(u,v,\epsilon)}{\partial v} = \mathbb{E}_{\gamma}\Big[\sum_{i=1}^{n}\nabla\sigma_v(x)\gamma_i(y^*_i(\epsilon))\Big] \tag{13}$$

Proof. We will first prove Equation (12). Let $f_n(\hat{y}) = \sum_\alpha \mu_{u+\frac{1}{n}z,\alpha}(x,\hat{y}_\alpha) + \sum_{i=1}^n \sigma_v(x)\gamma_i(\hat{y}_i) + \epsilon\ell(y,\hat{y})$ as in Lemma 1.

The proof builds a sequence of functions $\{g_n(z)\}_{n=1}^{\infty}$ that satisfies

$$\lim_{h\to 0}\frac{G(u+hz,v,\epsilon)-G(u,v,\epsilon)}{h} = \lim_{n\to\infty}\mathbb{E}_{\gamma\sim\mathcal{G}}[g_n(z)] \tag{14}$$

$$\mathbb{E}_{\gamma\sim\mathcal{G}}[\lim_{n\to\infty}g_n(z)] = \mathbb{E}_{\gamma\sim\mathcal{G}}[\sum_{\alpha}\nabla_u\mu_{u,\alpha}(x,y^*_{u,v,\gamma,\alpha}(\epsilon))^\top z]. \tag{15}$$

The functions $g_n(z)$ correspond to the loss perturbed prediction $y^*_{u,v,\gamma}(\epsilon)$ through the quantity $\sum_\alpha \mu_{u+\frac{1}{n}z,\alpha}(x,\hat{y}_\alpha) + \sum_{i=1}^n \sigma_v(x)\gamma_i(\hat{y}_i) + \epsilon\ell(y,\hat{y})$. The key idea we are exploiting is that there exists n_0 such that for any $n \geq n_0$ the maximal argument $y^*_{u+\frac{1}{n}z,v,\gamma}(\epsilon)$ does not change.

Thus, let

$$g_n(z) \triangleq \frac{\max_{\hat{y}\in Y} f_n(\hat{y}) - \max_{\hat{y}\in Y} f_\infty(\hat{y})}{1/n} \tag{16}$$

We apply the dominated convergence theorem on $g_n(z)$, so that $\lim_{n\to\infty}\mathbb{E}_{\gamma\sim\mathcal{G}}[g_n(z)] = \mathbb{E}_{\gamma\sim\mathcal{G}}[\lim_{n\to\infty} g_n(z)]$ in order to prove Equations (14,15). We note that we may apply the dominated convergence theorem, since the conditions $E_{\gamma\sim\mathcal{G}}\|\nabla_u\mu_{u,\alpha}(x,y_\alpha)\| \leq \infty$, and $E_{\gamma\sim\mathcal{G}}\|\nabla_v\sigma_v(x)\| \leq \infty$ imply that the expected value of g_n is finite (We recall that f_n is a measurable function, and note that since $\hat{y}\in Y$ is an element from a discrete set Y, then g_n is also a measurable function.).

From Lemma 1, the terms $\ell(y,y^*)$ and $\sum_{i=1}^n \sigma_v(x)\gamma_i(y^*_i)$ are identical in both $\max_{\hat{y}\in Y} f_n(\hat{y})$ and $\max_{\hat{y}\in Y} f_\infty(\hat{y})$. Therefore, they cancel out when computing the difference $\max_{\hat{y}\in Y} f_n(\hat{y}) - \max_{\hat{y}\in Y} f_\infty(\hat{y})$. Then, for $n \geq n_0$:

$$\max_{\hat{y}\in Y} f_n(\hat{y}) - \max_{\hat{y}\in Y} f_\infty(\hat{y}) = \sum_\alpha \mu_{u+\frac{1}{n}z,\alpha}(x,y^*_\alpha(\epsilon)) - \sum_\alpha \mu_{u,\alpha}(x,y^*_\alpha(\epsilon))$$

and Equation (16) becomes:

$$g_n(u) = \frac{\sum_\alpha \mu_{u+\frac{1}{n}z,\alpha}(x,y^*_\alpha(\epsilon)) - \sum_\alpha \mu_{u,\alpha}(x,y^*_\alpha(\epsilon))}{1/n}. \tag{17}$$

Since $\{\mu_{u,\alpha}(x, y_\alpha)\}$ is a set of continuous functions of u , then $\lim_{n\to\infty} g_n(z)$ is composed of the derivatives of $\mu_{u,\alpha}(x, y^*_\alpha(\epsilon))$ in direction z, namely, $\lim_{n\to\infty} g_n(z) = \sum_\alpha \nabla_u \mu_{u,\alpha}(x, y^*_\alpha(\epsilon))^\top z$.

We now turn to prove Equation (13).

Let $f'_n(\hat{y}) = \sum_\alpha \mu_{u,\alpha}(x, \hat{y}_\alpha) + \sum_{i=1}^n \sigma_{v+\frac{1}{n}z}(x)\gamma_i(\hat{y}_i) + \epsilon\ell(y, \hat{y})$ as in Lemma 1.

The proof builds a sequence of functions $\{g'_n(z)\}_{n=1}^\infty$ that satisfies

$$\lim_{h\to 0} \frac{G(u, v + hz, \epsilon) - G(u, v, \epsilon)}{h} = \lim_{n\to\infty} \mathbb{E}_{\gamma\sim\mathcal{G}}[g'_n(z)] \tag{18}$$

$$\mathbb{E}_{\gamma\sim\mathcal{G}}[\lim_{n\to\infty} g'_n(z)] = \mathbb{E}_{\gamma\sim\mathcal{G}}[\sum_{i=1}^n \nabla_v \sigma_v(x)\gamma_i(y^*_{i,u,v,\gamma}(\epsilon))^\top z]. \tag{19}$$

The functions $g'_n(z)$ correspond to the loss perturbed prediction $y^*_{u,v,\gamma}(\epsilon)$ through the quantity $\sum_\alpha \mu_{u,\alpha}(x, \hat{y}_\alpha) + \sum_{i=1}^n \sigma_{v+\frac{1}{n}z}(x)\gamma_i(\hat{y}_i) + \epsilon\ell(y, \hat{y})$.

As before, we are exploiting the fact that there exists n_1 such that for any $n \geq n_1$ the maximal argument $y^*_{u,v+\frac{1}{n}z,\gamma}(\epsilon)$ does not change.

Thus, let

$$g'_n(z) \triangleq \frac{\max_{\hat{y}\in Y} f'_n(\hat{y}) - \max_{\hat{y}\in Y} f'_\infty(\hat{y})}{1/n} \tag{20}$$

We apply the dominated convergence theorem on $g'_n(z)$, so that $\lim_{n\to\infty} \mathbb{E}_{\gamma\sim\mathcal{G}}[g'_n(z)] = \mathbb{E}_{\gamma\sim\mathcal{G}}[\lim_{n\to\infty} g'_n(z)]$ in order to prove Equations (18,19), with the same justification as before for the expected value of $g'_n(z)$ being finite.

From Lemma 1, the terms $\ell(y, y^*)$ and $\sum_\alpha \mu_{u,\alpha}(x, y^*_\alpha)$ are identical in both $\max_{\hat{y}\in Y} f'_n(\hat{y})$ and $\max_{\hat{y}\in Y} f'_\infty(\hat{y})$.

Therefore, they cancel out when computing the difference $\max_{\hat{y}\in Y} f'_n(\hat{y}) - \max_{\hat{y}\in Y} f'_\infty(\hat{y})$. Then, for $n \geq n_1$ Equation (20) becomes:

$$g'_n(u) = \frac{\sum_{i=1}^n \sigma_{v+\frac{1}{n}z}(x)\gamma_i(y^*_i(\epsilon)) - \sum_{i=1}^n \sigma_v(x)\gamma_i(y^*_i(\epsilon))}{1/n}. \tag{21}$$

Since $\sigma_v(x)$ is a smooth function of v, then $\lim_{n\to\infty} g'_n(z)$ is composed of the derivatives of $\sigma_v(x)\gamma_i(y^*_i(\epsilon))$ in direction z, namely, $\lim_{n\to\infty} g'_n(z) = \sum_{i=1}^n \nabla_v \sigma_v(x)\gamma_i(y^*_i(\epsilon))^\top z$.

□

Corollary 2. *Under the conditions of Theorem 2, $G(u, v, \epsilon)$ as defined in Equation (11), is a smooth function and $\frac{\partial}{\partial u}\mathbb{E}_\gamma[\ell(y, y^*_{w,\gamma})] =$*

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}\mathbb{E}_\gamma\Big[\sum_{\alpha\in\mathcal{A}} (\nabla\mu_{u,\alpha}(x, y^*_\alpha(\epsilon)) - \nabla\mu_{u,\alpha}(x, y^*_\alpha))\Big] \tag{22}$$

and $\frac{\partial}{\partial v}\mathbb{E}_\gamma[\ell(y, y^*_{w,\gamma})] =$

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}\mathbb{E}_\gamma\Big[\sum_{i=1}^n \nabla\sigma_v(x)\Big(\gamma_i(y^*_i(\epsilon)) - \gamma_i(y^*_i)\Big)\Big]. \tag{23}$$

Proof. We will first prove Equation (22). Since Theorem 2 holds for every direction z:

$$\frac{\partial G(u,v,\epsilon)}{\partial u} = \mathbb{E}_\gamma\Big[\sum_\alpha \nabla_u \mu_{u,\alpha}(x, y^*_\alpha(\epsilon))\Big].$$

Adding a derivative with respect to ϵ we get:

$$\frac{\partial}{\partial \epsilon}\frac{\partial G(u,v,0)}{\partial u} = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\mathbb{E}_\gamma\Big[\sum_\alpha \nabla_u \mu_{u,\alpha}(x, y^*_\alpha(\epsilon)) - \sum_\alpha \nabla_u \mu_{u,\alpha}(x, y^*_\alpha)\Big]$$

The proof follows by showing that the gradient computation is apparent in the Hessian, namely Equation (22) is attained by the identity $\frac{\partial}{\partial u}\frac{\partial G(u,v,0)}{\partial \epsilon} = \frac{\partial}{\partial \epsilon}\frac{\partial G(u,v,0)}{\partial u}$.

Now we turn to show that $\frac{\partial}{\partial u}\frac{\partial G(u,v,0)}{\partial \epsilon} = \nabla_u \mathbb{E}_\gamma[\ell(y, y^*_{w,\gamma})]$. Since ϵ is a real valued number rather than a vector, we do not need to consider the directional derivative, which greatly simplifies the mathematical derivations. We define $f_n(\gamma, \hat{y}) \triangleq \sum_\alpha \mu_{u,\alpha}(x, \hat{y}_\alpha) + \sigma_v(x)\sum_{i=1}^n \gamma_i(\hat{y}_i) + \frac{1}{n}\ell(y, \hat{y})$ and follow the same derivation as above to show that $\frac{\partial G(u,v,0)}{\partial \epsilon} = \mathbb{E}_\gamma[\ell(y, y^*_{w,\gamma})]$. Therefore $\frac{\partial}{\partial u}\frac{\partial G(u,v,0)}{\partial \epsilon} = \nabla_u \mathbb{E}_\gamma[\ell(y, y^*_{w,\gamma})]$.

We now turn to prove Equation (23).

Since Theorem 2 holds for every direction z:

$$\frac{\partial G(u,v,\epsilon)}{\partial v} = \mathbb{E}_\gamma\Big[\sum_{i=1}^n \nabla_v \sigma_v(x)\gamma_i(y^*_i(\epsilon))\Big]. \tag{24}$$

Adding a derivative with respect to ϵ we get:

$$\frac{\partial}{\partial \epsilon}\frac{\partial G(u,v,0)}{\partial v} = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\mathbb{E}_\gamma\Big[\sum_{i=1}^n \nabla_v \sigma_v(x)\gamma_i(y^*_i(\epsilon)) - \sum_{i=1}^n \nabla_v \sigma_v(x)\gamma_i(y^*_i)\Big] \tag{25}$$

Following the above steps, it holds that $\frac{\partial}{\partial v}\frac{\partial G(u,v,0)}{\partial \epsilon} = \nabla_v \mathbb{E}_\gamma[\ell(y, y^*_{w,\gamma})]$. Equation (23) is attained by the Hessian symmetric entries $\frac{\partial}{\partial v}\frac{\partial G(u,v,0)}{\partial \epsilon} = \frac{\partial}{\partial \epsilon}\frac{\partial G(u,v,0)}{\partial v}$, when considering Equation (25).

□

3 Training and architecture details

Both experiments are run on NVIDIA Tesla K80 standard machine.

3.1 Bipartite matching

Training We set maximum of 2000 training epochs, and deploy early stopping with patience of 50 epochs.

Our signal embedding network μ is trained with ADAM optimizer with learning rate (lr) = 0.1 and default parameters. The noise variance network σ is trained with Stochastic Gradient Descent optimizer, with lr=1e-6.

Hyper-parameters We set epsilon to -12. To escape zero gradients when loss is positive, we attempt increasing epsilon by 10%.

We learn from five noise perturbations for each permutation representation.

Signal embedding architecture The network μ has a first fully connected layer that links the sets of samples to an intermediate representation (with 32 neurons), and a second (fully connected) layer that turns those representations into batches of latent permutation matrices of dimension d by d each.

Noise variance architecture The network σ has a single layer connecting input sample sequences to a single output which is then activated by a softplus activation. We have chosen such an activation to enforce a positive sigma value.

3.2 k-nn for image classification

Datasets We consider three benchmark datasetes: MNIST dataset of handwritten digits, Fashion-MNIST dataset of fashion apparel, and CIFAR-10 dataset of natural images (no data augmentation) with the canonical splits for training and testing.

Training We train for 220 epochs.

Our signal embedding network μ is trained with ADAM optimizer, with learning rate set to 0.001 in all experiments.

The noise variance network σ is trained with Stochastic Gradient Descent optimizer. We perform a grid search over a small number of learning rates of the noise variance network σ. For MNIST and Fashion-MNIST datasets $lr \in \{1e-05, 1e-06\}$, and for CIFAR-10 dataset $lr \in \{1e-06, 1e-07\}$.

Hyper-parameters We set the number of candidate image to 800 for MNIST and Fashoin-MNIST and to 600 for CIFAR-10. Generally, our method is benefited from an increased number of candidate images due to the sparsity of the gradients resulting from the max predictors nature. The number of query images in a batch is 100 in all experiments.

We grid search over a small number of ϵ values. For MNIST dataset $\epsilon \in \{-0.05, -0.1, -0.2\}$, for Fashion-MNIST dataset $\epsilon \in \{-0.1, -0.2\}$, for CIFAR-10 dataset $\epsilon = -0.2$. To escape zero gradients when loss is positive, we attempt increasing epsilon by 10% up to a threshold of -0.9999.

In our 'Direct Stochastic Learning' settings, we attempt a single perturbation as well as five perturbations, though in almost all cases, a single perturbation is better for k $>$ 1, while five perturbations are better for k=1.

Signal embedding architecture For MNIST dataset the following embedding μ network is deployed:

Conv[Kernel: 5x5, Stride: 1, Output: 24x24x20, Activation: Relu]
Pool[Stride: 2, Output: 12x12x20]
Conv[Kernel: 5x5, Stride: 1, Output: 8x8x50, Activation: Relu]
Pool[Stride: 2, Output: 4x4x50]
FC[Units: 500, Activation: Relu]

For the Fashion-MNIST and CIFAR datasets embedding networks μ, we use the ResNet18 architecture as described in `https://github.com/kuangliu/pytorch-cifar`.

Noise variance architecture For MNIST and Fashion-MNIST datasets, the noise learning network σ is as follows:

Conv[Kernel: 5x5, Stride: 1, Output: 24x24x20, Activation: Relu]
Pool[Stride: 2, Output: 12x12x20]
Conv[Kernel: 5x5, Stride: 1, Output: 8x8x50, Activation: Relu]
Pool[Stride: 2, Output: 4x4x50]
FC[Units: 500, Activation: Relu]
FC[Units: 1, Activation: Softplus]

For CIFAR-10 dataset, the noise learning network σ is as follows:

Conv[Kernel: 5x5, Stride: 1, Output: 28x28x20, Activation: Relu]
Pool[Stride: 2, Output: 14x14x20]
FC[Units: 1, Activation: Softplus]

We have chosen the Softplus activation to enforce a positive sigma value.