REINTRODUCING STRAIGHT-THROUGH ESTIMATORS AS PRINCIPLED METHODS FOR STOCHASTIC BINARY NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training neural networks with binary weights and activations is a challenging problem due to the lack of gradients and difficulty of optimization over discrete weights. Many successful experimental results have been achieved with empirical straight-through (ST) approaches, proposing a variety of ad-hoc rules for propagating gradients through non-differentiable activations and updating discrete weights. At the same time, ST methods can be truly derived as estimators in the stochastic binary network (SBN) model with Bernoulli weights. We advance these derivations to a more complete and systematic study. We analyze properties, estimation accuracy, obtains different forms of correct ST estimators for activations and weights, explains existing empirical approaches and their shortcomings, explain how latent weights arise from the mirror descent method when optimizing over probabilities. This allows to reintroduce, once empirical, ST methods as sound approximations, apply them with clarity and develop further improvements.

1 INTRODUCTION

Neural networks with binary weights and activations have much lower computation costs and memory consumption than their real-valued counterparts (Horowitz, 2014; Esser et al., 2016; Rastegari et al., 2016). They are therefore very attractive for applications in mobile devices, robotics and other resource-limited settings, in particular for solving vision and speech recognition problems (Bulat & Tzimiropoulos, 2017; Xiang et al., 2017).

The state of the art in training deep binary networks is established with methods using the empirical straight-through gradient estimation approach. Typically, the sign function is used to represent binary activations. This poses an immediate problem as the derivative of a step function like sign is zero almost everywhere and is not useful for training with backpropagation. The *empirical straight-through* approach consists in using the sign function on the forward pass and derivative of some other *proxy* function on the backward pass. One popular solution is to use identity proxy, *i.e.* completely bypass sign on the backward pass (Krizhevsky & Hinton, 2011; Bengio et al., 2013; Courbariaux et al., 2015; Zhou et al., 2016), hence the name straight-through. Other proxy functions applied include $\tanh$ (Hinton, 2012; Raiko et al., 2015), clipped identity (Hubara et al., 2016; Lin et al., 2017; Rastegari et al., 2016; Alizadeh et al., 2019), or piece-wise quadratic $\mathrm{ApproxSign}$ (Esser et al., 2016; Liu et al., 2018), illustrated in Fig. A.1. These proxies can be used with deterministic or stochastic binarization (Courbariaux et al., 2015; Hubara et al., 2016; Krizhevsky & Hinton, 2011). This gives rise to a diversity of empirical ST methods, where various choices are studied purely experimentally (Alizadeh et al., 2019; Bethge et al., 2019; Tang et al., 2017).

Since binary weights can be also represented as a sign mapping of some real-valued *latent weights*, the same type of methods was adopted to handle optimization over binary weights (Shayer et al., 2017; Courbariaux et al., 2015; Hubara et al., 2016). However, often a different proxy is used for the weights along with clipping or not clipping latent weights to the interval $[-1, 1]$, generating additional unclear choices. With such obscurity of the concept of latent weights, Helwegen et al. (2019) argues that "latent weights do not exist" meaning that discrete optimization over binary weights needs to be considered and proposes an equally unjustified alternative scheme.

A more principled approach to handle binarization is to consider *stochastic binary networks* (SBNs), in which hidden units and/or weights are Bernoulli random variables. In these models the expected loss is a truly differentiable function of parameters (resp. weight probabilities) and its gradient can be estimated. Advanced unbiased gradient estimators exist (Grathwohl et al., 2018; Tucker et al., 2017; Yin & Zhou, 2019), which however have never been applied in practice to networks with deep binary dependencies due to increased variance in deep layers and complexity that grows quadratically with the number of layers (Shekhovtsov et al., 2020). ST can be defined in SBNs as a biased estimator, however this form is not well-known and no systematic study exists. More specifically, during our research on ST we found out that Tokui & Sato (2017); Shekhovtsov et al. (2020); Dai et al. (2017) already showed how to *derive* ST under clear approximations in SBNs. These results however are secondary in these papers and not known to authors applying ST in practice, to recent works on its analysis (Yin et al., 2019; Cheng et al., 2019) or works on alternative estimators (Cheng et al., 2019). They are not related to empirical ST variants or further analyzed. This is the main context for our work. More related works, including alternative approaches, are discussed in § A.

Contribution The goal of this work is to reintroduce straight-through estimators in a principled way, clarify, systematize and study the empirical ST approaches. We build on the derivations of ST for shallow (Tokui & Sato, 2017) and deep (Shekhovtsov et al., 2020) networks and the mirror descent optimization view for binary weights (Ajanthan et al., 2019).

We review sound results allowing to derive ST estimator in SBNs (§§ 2.1 and 4); Analyze this derived ST for models with one hidden layer and show shortcomings of and equivalences between existing empirical approaches (§ 2.2); Define latent weights in a sound optimization scheme and establish a rigorous correspondence between SGD with identity ST and powerful optimization methods such as mirror descent and variational Bayesian learning (§ 3); Experimentally study the accuracy of derived ST estimator; Experimentally show that with the derived ST estimators and the proposed initialization and optimization, several common choices of gradient rules can be closely replicated, but unlike the empirical variants lead to equally well-performing methods (§ 5). As accompanying results, we analyze *Gumbel-Softmax* estimator (§ E) and the BayesBiNN method (Meng et al., 2020) that applies this estimator for SBNs, showing an unexpected relation of the latter to ST (§ F).

2 Single Layer Straight-Through

Notation Throughout the paper we will define several models that work with binary random variables. We will keep the convention that binary means $\{-1, 1\}$-valued and will model random states $x \in \{-1, 1\}^n$ using the noisy sign mapping:

$$x_i = \operatorname{sign}(a_i - z_i), \tag{1}$$

where z_i are real-valued independent noises with a fixed cdf F and a_i are (input-dependent) parameters. It is seen that the difference from deterministic binary models is only in the presence of injected noises z. Equivalently, x_i can be described by probabilities $p(x_i{=}1 \,|\, a_i) = \mathbb{P}(a_i{-}z \geq 0 \,|\, a_i) = \mathbb{P}(z{\leq}a_i \,|\, a_i) = F(a_i)$. The noise distribution F will play an important role in understanding different schemes. If the noise is logistic, then F is the logistic sigmoid function σ.

2.1 Derivation

The ST was first introduced (Hinton, 2012; Krizhevsky & Hinton, 2011) in the context of a stochastic autoencoder model, highly relevant to date (*e.g.* Dadaneh et al., 2020), which we consider. We give a self-contained derivation of ST clarifying the key steps of Tokui & Sato (2017, sec. 6.4).

Let y denote observations. The *encoder network* parametrized by ϕ computes logits $a(y; \phi)$ and computes a binary latent state x via (1). As noises z are independent, the conditional distribution factors as $p(x|y; \phi) = \prod_i p(x_i|y; \phi)$. The *decoder* reconstructs observations with $p^{\text{dec}}(y|x; \theta)$, another neural network parametrized by θ. The reconstruction loss of this autoencoder is defined as

$$\mathbb{E}_{y\sim\text{data}}\big[\mathbb{E}_{x\sim p(x|y;\phi)}[-\log p^{\text{dec}}(y|x;\theta)]\big]. \tag{2}$$

We will focus on the problem of computing the gradient of this loss w.r.t. the encoder parameters ϕ. The problem for a fixed observation y takes the form

$$\tfrac{\partial}{\partial\phi}\mathbb{E}_{x\sim p(x;\phi)}[\mathcal{L}(x)] = \tfrac{\partial}{\partial\phi}\mathbb{E}_z[\mathcal{L}(\operatorname{sign}(a - z))], \tag{3}$$

Algorithm 1: Straight-Through-Activations

/* a -- preactivation */
/* F -- activation noise cdf */
/* $x \in \{-1, 1\}$ -- binary state */

Forward(a **)**
  $p = F(a)$;
  $x \sim 2\text{Bernoulli}(p) - 1$;

Backward($\frac{d\mathcal{L}}{dx}$ **)**
  /* $\frac{d\mathcal{L}}{dx}$ -- loss grad in x */
  $\frac{d\mathcal{L}}{da} \equiv 2\,\text{diag}(F'(a))\frac{d\mathcal{L}}{dx}$;

Algorithm 2: Straight-Through-Weights

/* η -- mirror/latent weights */
/* F -- weight noise cdf */
/* $w \in \{-1, 1\}$ -- binary weight */

Forward(η **)**
  $p = F(\eta)$;
  $w \sim 2\text{Bernoulli}(p) - 1$;

Backward($\frac{d\mathcal{L}}{dw}$ **)**
  /* $\frac{d\mathcal{L}}{dw}$ -- loss grad in w */
  $\frac{d\mathcal{L}}{d\eta} \equiv 2\frac{d\mathcal{L}}{dw}$;

where $p(x;\phi)$ is a shorthand for $p(x|y;\phi)$ and $\mathcal{L}(x) = -\log p^{\text{dec}}(y|x;\theta)$. The reparameterization trick, *i.e.* to draw one sample of z in (3) and differentiate $\mathcal{L}(\text{sign}(a - z))$ fails: since it is not *continuously differentiable we cannot interchange the gradient and the expectation in z in the first place*. Instead, following steps lead to an unbiased low variance estimator. From the LHS of (3) we express the derivative as

$$\frac{\partial}{\partial\phi}\sum_x \big(\prod_i p(x_i;\phi)\big)\mathcal{L}(x) = \sum_x \sum_i \big(\prod_{i'\neq i} p(x_{i'};\phi)\big)\big(\frac{\partial}{\partial\phi}p(x_i;\phi)\big)\mathcal{L}(x). \tag{4}$$

Then the *derandomization* (Owen, 2013, ch. 8.7) is applied which performs summation over x_i explicitly for the rest of the states in x fixed. Because x_i takes only two values, this gives

$$\sum_{x_i} \frac{\partial p(x_i;\phi)}{\partial\phi}\mathcal{L}(x) = \frac{\partial p(x_i;\phi)}{\partial\phi}\mathcal{L}(x) + \frac{\partial(1-p(x_i;\phi))}{\partial\phi}\mathcal{L}(x_{\downarrow i}) = \frac{\partial}{\partial\phi}p(x_i;\phi)\big(\mathcal{L}(x) - \mathcal{L}(x_{\downarrow i})\big), \tag{5}$$

where $x_{\downarrow i}$ denotes the full state vector x with the sign of x_i flipped. Since this expression is now invariant of x_i, we can multiply it with $1 = \sum_{x_i} p(x_i;\phi)$ and recollect the full gradient in the form:

$$\sum_x \big(\prod_{i'} p(x_{i'};\phi)\big)\sum_i \frac{\partial p(x_i;\phi)}{\partial\phi}\big(\mathcal{L}(x) - \mathcal{L}(x_{\downarrow i})\big) = \mathbb{E}_{x\sim p(x;\phi)}\sum_i \frac{\partial p(x_i,\phi)}{\partial\phi}\big(\mathcal{L}(x) - \mathcal{L}(x_{\downarrow i})\big). \tag{6}$$

To obtain an unbiased estimate, it is sufficient to take one sample from the encoder and compute the expression in the sum. This estimator is known as *local expectations* (Titsias & Lázaro-Gredilla, 2015) and coincides in this case with GO-gradient (Cong et al., 2019) and RAM (Tokui & Sato, 2017).

However, evaluating $\mathcal{L}(x_{\downarrow i})$ for all i may be impractical. A huge simplification is obtained if we assume that the change of the loss f when only a single latent bit x_i is changed can be approximated via linearization. *I.e.* provided that f alone is differentiable we approximate

$$\mathcal{L}(x) - \mathcal{L}(x_{\downarrow i}) \approx 2x_i \frac{\partial\mathcal{L}(x)}{\partial x_i}, \tag{7}$$

where we used that $x_i - (-x_i) = 2x_i$. Expressing $\frac{\partial}{\partial\phi}p(x_i;\phi) = x_i F'(a_i(\phi))\frac{\partial}{\partial\phi}a_i(\phi)$, we obtain

$$\frac{\partial p(x_i,\phi)}{\partial\phi}(\mathcal{L}(x) - \mathcal{L}(x_{\downarrow i})) \approx 2F'(a_i(\phi))\Big(\frac{\partial a_i(\phi)}{\partial\phi}\Big)\frac{\partial\mathcal{L}(x)}{\partial x_i}. \tag{8}$$

If we define now that $\frac{\partial x_i}{\partial a_i} \equiv 2F'(a_i)$, the summation over i recovers the chain rule:

$$\sum_i 2F'(a_i(\phi))\Big(\frac{\partial a_i(\phi)}{\partial\phi}\Big)\frac{\partial\mathcal{L}(x)}{\partial x_i} = \sum_i \Big(\frac{\partial\mathcal{L}(x)}{\partial x_i}\Big)\Big(\frac{\partial x_i}{\partial a_i}\Big)\Big(\frac{\partial a_i(\phi)}{\partial\phi}\Big) \tag{9}$$

that would be automatically computed by the backpropagation if we make the backward pass to differentiate $x = 2F(a) - 1$, which can be interpreted as a smooth approximation to sign on the backward pass. We thus obtained ST algorithm Alg. 1, *which matches exactly to the one described by Hinton (2012)*: to sample on the forward pass and use the derivative of the noise cdf on the backward pass. The coefficient 2 that occurred is due to that we used ± 1 encoding for x.

2.2 Analysis

We present a summary of our analysis while proofs and further details are given in § B.

Invariances Observe that binary activations stay invariant under transformations: $\mathrm{sign}(a_i - z_i) = \mathrm{sign}(T(a_i) - T(z_i))$ for any strictly monotone mapping T. *Consistently, the ST gradient by Alg. 1 is also invariant to* T. In contrast, empirical straight-through approaches where the derivative proxy is hand-designed fail to maintain this property. In particular rescaling the proxy leads to different estimators. Furthermore, when applying transform $T = F$ (the noise cdf), the backpropagation rule in line 5 of Alg. 1 becomes equivalent to using the identity proxy. Hence we see that a common description of ST in the literature "to back-propagate through the hard threshold function as if it had been the identity function" is also correct, *but only for the case of uniform noise*. Otherwise, and especially for deterministic ST, it is ambiguous as the result crucially depends on what transformations are applied under the hard threshold.

Bias Analysis I) When ST is unbiased? As we used linearization as the only biased approximation, it follows that *Alg. 1 is unbiased if the objective function* $\mathcal{L}$ *is multilinear in* x. A simple counter-example, where ST is biased, is *e.g.*, $\mathcal{L}(x) = x^2$. In this case the expected value of the loss is 1, independently of a defining $x = \mathrm{sign}(a - z)$ and the true gradient is zero. However the expected ST gradient is $\mathbb{E}[2F'(a)2x] = 4F'(a)(2F(a) - 1)$, which may be positive or negative depending on a. On the other hand, any function of binary variables has an equivalent multilinear expression. In particular, if we consider $\mathcal{L}(x) = \|Wx - y\|^2$, analyzed by Yin et al. (2019), then $\tilde{\mathcal{L}}(x) = \|Wx - y\|^2 - \sum_i x_i^2 \|W_i\|^2 + \sum_i \|W_i\|^2$ *coincides with* $\mathcal{L}$ *on all binary configurations and is multilinear*. It follows that ST applied to $\tilde{\mathcal{L}}$ gives an unbiased gradient estimate of $\mathbb{E}[\mathcal{L}]$.

II) How does using a different proxy than F in Line 5 of Alg. 1 affects the gradient in ϕ? Since $\mathrm{diag}(F')$ occurs in the backward chain, we call estimators that use $\mathrm{diag}(\Lambda)$ instead as *internally rescaled*. We show that *for any non-negative coordinate-wise function* Λ*, the expected rescaled estimator has non-negative scalar product with the expected original estimator.* Thus using a different proxy, in particular identity, is not too destructive: if (1) was unbiased, the rescaled estimator may be biased but it is guaranteed to give an ascend direction in the expectation and the optimization can in principle succeed. However, assuming that (1) is biased (when $\mathcal{L}$ is not multi-linear) but an ascent direction, *the ascent property can no longer be guaranteed for the rescaled estimator*.

III) Does ST gradient give a valid ascent direction even when $\mathcal{L}$ is not multi-linear? Assume that all partial derivatives $g_i(x) = \frac{\partial \mathcal{L}}{\partial x_i}$ are L-Lipschitz continuous. We show that *expected ST gradient is an ascent direction iff* $\left|\mathbb{E}_x[g_i(x)]\right| > L$ *for all* i.

IV) Can we decrease the bias? Assume the loss function is applied after a linear transform of Bernoulli variables, *i.e.*, takes the form $\mathcal{L}(x) = \ell(Wx)$. A typical initialization uses random W normalized by the size of the fan-in, *i.e.* so that $\|W_{k,:}\|_2 = 1\ \forall k$. In this case *the Lipschitz constant of gradients of* $\mathcal{L}$ *scales as* $O(1/\sqrt{n})$*, where* n *is the number of binary variables*. It means that *using more binary variables improves preconditions for an ascent direction (decreases the bias), at least at initialization*.

V) Does deterministic ST give an ascent direction? Let g^* be the deterministic ST gradient for the state $x^* = \mathrm{sign}(a)$ and $p^* = p(x^*|a)$ probability of this state. *We show that deterministic ST gradient forms a positive scalar product with the expected ST gradient if* $|g_i^*| \geq 2L(1-p^*)$ *and with the true gradient if* $|g_i^*| \geq 2L(1 - p^*) + L$. From this we conclude that deterministic ST positively correlates with true gradient when $\mathcal{L}$ is multilinear, improves with the number of hidden units in the case described by IV and approaches expected stochastic ST Alg. 1 as units learn to be more deterministic so that the factor $(1 - p^*)$ decreases.

Variants Explained From the equivalence property, many works are easily seen to use ST estimator with stochastic sampling equivalent to Alg. 1 (Raiko et al., 2015; Shen et al., 2018; Dadaneh et al., 2020). Furthermore, using different noise distributions for z, we recover common choices of sign proxies used in empirical ST works as shown in Fig. A.1 (c-e). In our framework they correspond to *modeling choices* affecting the forward pass in the first turn.

Dai et al. (2017) perform a correct interchange of derivative and integral in (3) using *weak (distributional)* deriatives. After computing local expectations in this more complicated formalism, they are back with finite differences (6) which they also propose to linearize as in (7). Thus their *distributional SGD is equivalent to common SGD with the ST estimator* Alg. 1.

3 LATENT WEIGHTS DO EXIST!

In this section we show that latent weights that did not get a proper substantiation in empirical ST methods (Hubara et al., 2016; Helwegen et al., 2019) can be clearly defined in SBNs and that several empirical update rules correspond to sound optimization schemes: projected gradient descent, mirror descent, variational Bayesian learning.

Let w be ± 1-Bernoulli weights with $p(w_i{=}1) = \theta_i$, let $\mathcal{L}(w)$ be the loss for a fixed training input. Consistently with the model for activations (1), we can define $w_i = \operatorname{sign}(\eta_i - z_i)$ in order to model weights w_i using parameters $\eta_i \in \mathbb{R}$ that we will call *latent weights*. We need to tackle two problems in order to optimize $\mathbb{E}_{w\sim p(w|\theta)}[\mathcal{L}(w)]$ in probabilities θ: i) how to estimate the gradient and ii) how to handle constraints $\theta \in [0,1]^n$.

Projected Gradient A basic approach to handle constraints is the *projected gradient descent*:

$$\theta^{t+1} := \operatorname{clip}(\theta^t - \varepsilon g^t, 0, 1), \quad \text{where} \quad g^t = \nabla_\theta \mathbb{E}_{w\sim p(w|\theta^t)}[\mathcal{L}(w)] \tag{10}$$

and $\operatorname{clip}(x, a, b) := \max(\min(x, b), a)$ is the projection. Observe that for the uniform noise distribution on $[-1, 1]$ with $F(z) = \operatorname{clip}(\frac{z+1}{2}, 0, 1)$, we have $\theta = p(w{=}1) = F(\eta) = \operatorname{clip}(\frac{\eta+1}{2}, 0, 1)$. Because this F is linear on $[-1, 1]$, the update (10) can be equivalently reparametrized in η as

$$\eta^{t+1} := \operatorname{clip}(\eta^t - \varepsilon' h^t, -1, 1), \quad \text{where} \quad h^t = \nabla_\eta \mathbb{E}_{w\sim p(w|F(\eta))}[\mathcal{L}(w)] \quad \text{and} \quad \varepsilon' = 4\varepsilon. \tag{11}$$

The gradient in the latent weights h^t can be estimated by Alg. 1 with the simplification $2F' = 1$. We obtained that *the method of Hubara et al. (2016, Alg.1) with stochastic rounding and with real-valued weights identified with η is equivalent to PGD on η with constraints $\eta \in [-1, 1]$ and ST gradient estimate Alg. 1* and thus is a sound optimization scheme. In the projected SGD setting (with random mini-batches) there are convergence guarantees at least in the convex setting.

Mirror Descent As an alternative approach to handle constraints $\theta \in [0, 1]^n$, we study the application of mirror descent (MD) and connect it with the identity ST update variants. A step of MD is found by solving the following proximal problem:

$$\theta^{t+1} = \min_\theta \langle g^t, \theta - \theta^t \rangle + \tfrac{1}{\varepsilon} D(\theta, \theta^t). \tag{12}$$

The divergence term $\frac{1}{\varepsilon} D(\theta, \theta^t)$ weights how much we trust the linear approximation $\langle g^t, \theta - \theta^t \rangle$ when considering the step from θ^t to θ. When the gradient is stochastic we speak about *stochastic mirror descent* (SMD) (Zhang & He, 2018; Azizan et al., 2019). A common choice of divergence to handle probability constraints is the KL-divergence $D(\theta, \theta^t) = \operatorname{KL}(\operatorname{Ber}(\theta), \operatorname{Ber}(\theta^t)) = \theta \log(\frac{\theta}{\theta^t}) + (1 - \theta) \log(\frac{1-\theta}{1-\theta^t})$. Solving for stationary point of (12) gives

$$0 = \nabla f + \tfrac{1}{\varepsilon}\big(\log(\tfrac{\theta}{1-\theta}) - \log(\tfrac{\theta^t}{1-\theta^t})\big). \tag{13}$$

Observe that when $F = \sigma$ we have $\log(\frac{\theta}{1-\theta}) = \eta$. Then the MD step can be written in the well-known convenient form using the latent weights η (natural parameters of Bernoulli distribution):

$$\theta^t := \sigma(\eta^t); \qquad \eta^{t+1} := \eta^t - \varepsilon \nabla_\theta \mathcal{L}(\theta^t). \tag{14}$$

We thus have obtained the rule where on the forward pass $\theta = \sigma(\eta)$ defines the sampling probability of w and on the backward pass the derivative σ that otherwise occurs in Line 5 of Alg. 1 *is bypassed exactly as if the identity proxy was used*. We define such ST rule for optimization in weights as Alg. 2. Its correctness is not limited to logistic noise: by using the fact that for any strictly monotone noise distribution F there is a corresponding divergence function D, we show the following.

Proposition 1. Common SGD in latent weights η using the *identity straight-through-weights* Alg. 2 implements SMD in the weight probabilities θ with the divergence corresponding to F.

The proof in § C is closely related to the work by Ajanthan et al. (2019). Differently from us, they considered a connection to deterministic ST. Their argument includes taking the limit in which F is squashed to approach the step function, which renders MD invalid. This problem does not occur in our formulation. Proposition 1 reveals that although Bernoulli weights can be modeled the same

way as activations using the injected noise model $w = \text{sign}(\eta - z)$, *the noise distribution F for weights does not reflect a modeling choice but the optimization proximity scheme*.

Despite the general relation by Proposition 1, we view the KL divergence as a more reliable choice in practice. Azizan et al. (2019) have shown that the optimization with SMD has an inductive bias to find the closest solution to the initialization point as measured by the divergence used in MD, which has a strong impact on generalization. This suggests that MD with KL divergence will prefer higher entropy solutions, making more diverse predictions. By our equivalence results, using SGD on latent weights with logistic noise and identity straight-through Alg. 2 enjoys the same properties.

Variational Bayesian Learning Extending the results above, in **??** we study the variational Bayesian learning formulation and show the following.

Proposition 2. Common SGD in latent weights η with a weight decay and *identity straight-through-weights* Alg. 2 is equivalent to optimizing a factorized variational approximation to the weight posterior $p(w|D)$ using a composite SMD method.

We can see that powerful and sound learning techniques can be obtained in a form of simple update rules. Such simple learning rules are recently studied by Khan & Rue (2020)[1]. We emphasize the interplay with the identity-ST estimator and the connection to the implicit regularization discussed above. The recent work (Meng et al., 2020) also addresses the binary case but uses Gumbel-Softmax estimator. We discuss its shortcomings and an unexpected equivalence to deterministic ST in § F.

4 Deep Straight-Through

In this section we review the derivation of ST for deep SBNs (Shekhovtsov et al., 2020), extend it to binary weights using MD from § 3 and propose an initialization method based on the analysis in § 2.

Derivation Idea Consider an SBN model with L binary layers, recurrently defined as

$$x^k = \text{sign}(a^k(x^{k-1};\theta) - z^k), \quad k = 1 \dots L, \tag{15}$$

where a^k are *pre-activations* and z^k are injected noises independent for all units. Let also $\mathcal{L}(x^L;\theta)$ be the loss function for a given input and the last layer state x^L. Similarly to one-layer case, the probability $p(x^k|x^{k-1})$ is conditionally independent with $p(x^k_i = 1|x^{k-1}) = F_z(a^k(x^{k-1}))$.

The challenge to derive an ST-like estimator for deep SBN is that simply applying the steps in § 2 to the k's layer in order to estimate gradient w.r.t. a^k is not straightforward because we would have to linearize the expectation of the whole network reminder and the loss as a function of x^k. The idea of Shekhovtsov et al. (2020) is to linearize only the next layer's conditional probability $p(x^{k+1}|x^k)$ and then proceed recurrently, chaining derandomization as in § 2 and linearization steps. By doing so they infer a more accurate PSA estimator and from it the deep ST estimator. The ST estimator (Shekhovtsov et al., 2020, Alg. 2) *is identical to applying ST Alg. 1 to each binary layer and executing the automatic differentiation*. The insights from the derivation are twofold. First, since derandomization is performed recurrently, the variance for deep layers is significantly reduced. Second, we know which approximations contribute to the bias, they are indeed the linearizations

$$p(x^{k+1}|x^k) - p(x^{k+1}|x^k_{\downarrow i}) \approx 2x^k_i \tfrac{\partial p(x^{k+1}|x^k)}{\partial x^k_i} \qquad \text{and} \qquad \mathcal{L}(x^L) - \mathcal{L}(x^L_{\downarrow i}) \approx \tfrac{\partial \mathcal{L}(x^L)}{\partial x^L_i}, \tag{16}$$

for all layers k and units i. The experiments (Shekhovtsov et al., 2020, Figs. 2, 3, 4) with real weights suggest that while for small networks these approximations lead to some degradation of gradient estimation accuracy during learning, for large networks ST performs on par with the PSA method that uses a much tighter approximation.

Extensions We now can put all building blocks together. Our complete model extends the deep model (15) as follows. The weights w^k in each layer are ± 1 Bernoulli with probability θ. Pre-activations a are extended to incorporate additional processing, importantly batch normalization (Ioffe & Szegedy, 2015):

$$a^k = \text{BN}(\text{Linear}(x^{k-1}, w^k)), \tag{17}$$

[1]Contemporaneous, not published.

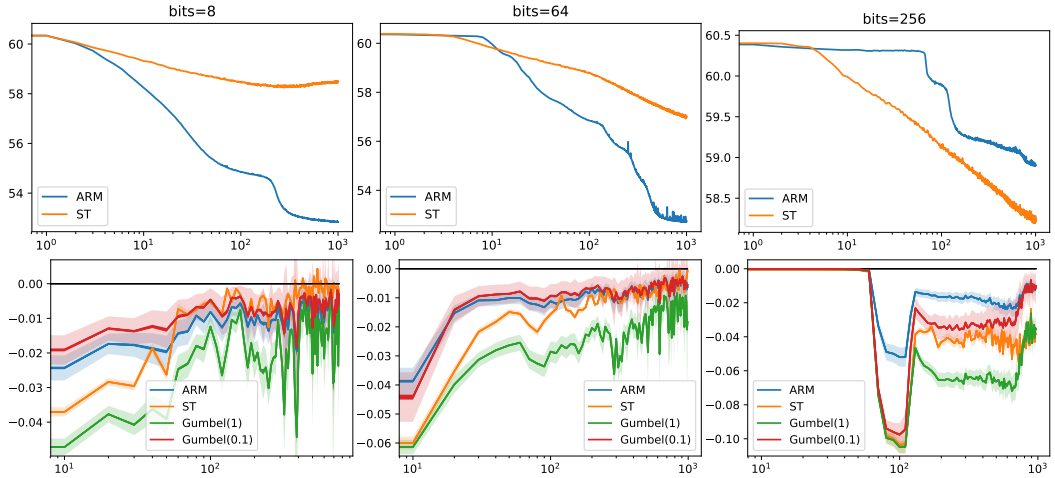

Figure 1: VAE training of a 2-layer decoder for different number of latent bits. *Top:* training loss (ELOB). *Bottom:* expected loss decrease rate using different estimators at parameter points of ARM training trajectory (lower is better). X-axis correspond to epochs. Please see text for details.

where Linear is a binary fully connected or convolutional transform and BN has real-valued affine terms (scale, bias) enabled. The model is illustrated in Fig. D.1. Note that in the linearization of $p(x^{k+1}|x^k)$ in x^k, we effectively differentiate the composite transform (Linear, BN, F, and finally the product forming $p(x^{k+1}|x^k)$) and approximate the finite difference $p(x^{k+1}|x^k) - p(x^{k+1}|x^k_{\downarrow i})$ using this linearization. Thus extra non-linearities such as BN in principle worsen t he approximation. On the other hand, the strongest non-linearity is introduced here by F. Following the arguments from the analysis § 2.2, a large number of hidden units in combination with normalization by BN achieves a better approximation accuracy for linearizing F.

Initialization The role of the affine parameters in BN is to reintroduce the scale and bias degrees of freedom, removed by the normalization (Ioffe & Szegedy, 2015). In our model these degrees of freedom are important as they control the strength of pre-activation relative to noise. With the sign activation, they could be indeed equivalently represented as learnable bias and variance parameters of the noise since $\text{sign}(xs + b - z) = \text{sign}\left(x - \frac{z-b}{s}\right)$ for $s > 0$. Without the BN layer, the result of $\text{Linear}(x^{k-1}, w^k)$ is an integer in a range that depends on input sizes. If the noise variance is set to 1, this will lead to vanishing gradients in large networks. With BN and its affine transform the right proportion can be learned, but it is important to initialize it so that the learning can happen. We propose the following initialization. We set $s = 1$ $b = 0$ as default for BN and *normalize the noise distribution* so that it is zero mean and $F'(0) = \frac{1}{2}$. This choice ensures that the Jacobian $2F'(a)$ in Line 5 of Alg. 1 at the mean value of pre-activations is 1 and therefore gradients do not vanish.

We initialize weight probabilities θ as uniform in $[0, 1]$. Respectively, we need to initialize latent weights as $\eta = F^{-1}(\theta)$. This initialization and the noise scaling above are the places where the empirical approaches have to do extra guessing when considering different ST proxies, obtain diverse results and attribute them to the quality of proxies (*e.g.* Esser et al., 2016).

5 Experiments

Variational Autoencoders Previous work has convincingly demonstrated that Gumbel Softmax and ARM estimators give better results than ST (Jang et al., 2016; Yin & Zhou, 2019; Dadaneh et al., 2020) for 1 hidden layer case such as VAE. Our goal is to compare gradient estimators more directly than by test metrics, factoring out the effects of the optimizer with its hyper-parameters and generalization issues. To make such direct comparison we train a variational autoencoder for *20newsgroups* dataset using ARM estimator. During its training we record the trajectory of parameter vectors $\{(\phi_t, \theta_t)\}$. We then evaluate accuracy of different estimators at exactly the same points along this trajectory. Let g be the true gradient, which we estimate using ARM with 1000 samples, and $\hat{g}$

Table 1: Test accuracy for different methods on CIFAR-10 with the same/similar architecture. At the test-time SBN can be applied in several possible ways: with zero noises (*det*), using a single random draw of noises for each prediction (*1-sample*) and predicting using an ensemble of 10 samples (*10-sample*). We also show results when only activations are deterministic and the weights are random (*det-act*). Standard deviations are given w.r.t. to 4 learning trials with random initialization.

Binary Activations, Stochastic training					
Method	det	1-sample	10-sample	det-act 1-sample	det-act 10-sample
Our SBN with logistic noise	89.6 ± 0.1	88.3 ± 0.2	90.6 ± 0.2	89.1 ± 0.5	90.6 ± 0.2
Our SBN with uniform noise	89.7 ± 0.2	88.8 ± 0.1	90.5 ± 0.2	89.4 ± 0.3	90.4 ± 0.1
Our SBN with triangular noise	89.5 ± 0.2	88.7 ± 0.2	90.0 ± 0.3	89.4 ± 0.2	90.4 ± 0.2
Hubara et al. (2016) Torch7 (rand. Htanh)	89.85	-	-	-	-
Peters & Welling (2018) (from pretrained)	88.61	2-sample $89.5 \pm$	5-sample $90.7 \pm$	16-sample $91.2 \pm$	
Binary Activations, Deterministic training					
Rastegari et al. (2016) XNOR-Net	89.83				
Hubara et al. (2016) Theano (det. Htanh)	88.60				

be a stochastic estimator. We measure its accuracy by computing

$$-\mathbb{E}[\langle g, \hat{g} \rangle] / \sqrt{\mathbb{E}[\|\hat{g}\|^2]}, \tag{18}$$

where the expectation is over 100 stochastic samples and the training data. The numerator indeed measures the expected decrease rate of the loss and is minimized by $\hat{g} = g$. If there was no denominator, any unbiased estimator would attain the steepest expected descent but overestimating the gradient length would do even better. The numerator normalizes in a way similar to common adaptive schemes. This metric can indeed be connected by the expected performance w.r.t. the SGD proximal step problem, please see § D for more detail.

In Fig. 1 we show results for a 2-layer decoder. In 8-bit case ST starts estimating systematically wrong directions (positive value of expected improvement) and we see that the training loss starts growing on its own trajectory. As we increase the number of bits, ST becomes more reliable, consistently with our theoretical propositions, and its training performance improves without tuning any hyper-parameters (the training setup with Adam specified in § D is held fixed). While for this task it may not be always the best choice, we can now better understand its limitations and potential advantages for deep stochastic models with large state spaces.

Classification with Deep SBN In this experiment we verify that the MD method with the proposed initialization and ST estimators performs as good as the tuned empirical ST and alternative approaches under the same conditions. We consider CIFAR-10 dataset and use the VGG-like architecture (Courbariaux et al., 2015; Hubara et al., 2016) specified in Fig. D.1. Our full learning setup is specified in § D. We trained SBNs with 3 choices of noise distributions corresponding to proxies used by prior work as in Fig. A.1 (c-e). Table 1 shows the test results in comparison with baselines.

We make the following observations. **1.** Training with different choices of the activation noise distribution, corresponding to different ST rules, achieves equally good results. This is in contrast to empirical studies advocating specific proxies and we believe is allowed by the consistency of the model, initialization and training. **2.** The training accuracy is above 99.6% for all our models and runs (using the 10-sample test mode). Thus the optimization succeeds in all cases. **3.** Using ensemble of 10 samples from the model improves accuracy, showing that the model generates diverse predictions, as appropriate for under-determined classification. **4.** There is no significant differences in performance with the baselines when we compare results for the same evaluation mode across different methods. The *insignificant* differences to Peters & Welling (2018) may be explained by the procedures that we did not include: *e.g.* the pre-trained initialization. We refer the readers interested in the best performing networks to the recent automated architecture search (Bulat et al., 2020), combining binary convolutions and affine real-valued paths and using more data augmentation.

Our code will be available on `http://github.com`.

REFERENCES

Thalaiyasingam Ajanthan, Kartik Gupta, Philip HS Torr, Richard Hartley, and Puneet K Dokania. Mirror descent view for neural network quantization. *arXiv preprint arXiv:1910.08237*, 2019.

Milad Alizadeh, Javier Fernandez-Marques, Nicholas D. Lane, and Yarin Gal. An empirical study of binary neural networks' optimisation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJfUCoR5KX.

Navid Azizan, Sahin Lale, and Babak Hassibi. Stochastic mirror descent on overparameterized nonlinear models: Convergence, implicit regularization, and generalization. *arXiv preprint arXiv:1906.03830*, 2019.

Yu Bai, Yu-Xiang Wang, and Edo Liberty. Proxquant: Quantized neural networks via proximal operators. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzMyhCcK7.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Joseph Bethge, Haojin Yang, Marvin Bornstein, and Christoph Meinel. Back to simplicity: How to train accurate bnns from scratch? *CoRR*, abs/1906.08637, 2019. URL http://arxiv.org/abs/1906.08637.

E. Boros and P.L. Hammer. Pseudo-Boolean optimization. *Discrete Applied Mathematics*, 1-3(123):155–225, 2002.

Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Bats: Binary architecture search, 2020.

Pengyu Cheng, Chang Liu, Chunyuan Li, Dinghan Shen, Ricardo Henao, and Lawrence Carin. Straight-through estimator as projected wasserstein gradient flow. *arXiv preprint arXiv:1910.02176*, 2019.

Yulai Cong, Miaoyun Zhao, Ke Bai, and Lawrence Carin. GO gradient for expectation-based objectives. In *International Conference on Learning Representations*, 2019.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, pp. 3123–3131, 2015.

Siamak Zamani Dadaneh, Shahin Boluki, Mingzhang Yin, Mingyuan Zhou, and Xiaoning Qian. Pairwise supervised hashing with bernoulli variational auto-encoder and self-control gradient estimator. *ArXiv*, abs/2005.10477, 2020.

Bo Dai, Ruiqi Guo, Sanjiv Kumar, Niao He, and Le Song. Stochastic generative hashing. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 913–922, 2017.

Steven K. Esser, Paul A. Merolla, John V. Arthur, Andrew S. Cassidy, Rathinakumar Appuswamy, Alexander Andreopoulos, David J. Berg, Jeffrey L. McKinstry, Timothy Melano, Davis R. Barch, Carmelo di Nolfo, Pallab Datta, Arnon Amir, Brian Taba, Myron D. Flickner, and Dharmendra S. Modha. Convolutional networks for fast, energy-efficient neuromorphic computing. *Proceedings of the National Academy of Sciences*, 113(41):11441–11446, 2016. URL https://www.pnas.org/content/113/41/11441.

Will Grathwohl, Dami Choi, Yuhuai Wu, Geoff Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *ICLR*, 2018.

Alex Graves. Practical variational inference for neural networks. In *NeurIPS*, pp. 2348–2356. 2011.

Koen Helwegen, James Widdicombe, Lukas Geiger, Zechun Liu, Kwang-Ting Cheng, and Roeland Nusselder. Latent weights do not exist: Rethinking binarized neural network optimization. In *Advances in neural information processing systems*, pp. 7531–7542, 2019.

Geoffrey Hinton. Lecture 15d - Semantic hashing : 3:05 - 3:35, 2012. URL https://www.cs.toronto.edu/~hinton/coursera/lecture15/lec15d.mp4.

M. Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pp. 10–14, 2014.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Advances in neural information processing systems*, pp. 4107–4115, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, volume 37, pp. 448–456, 2015.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Emtiyaz Khan and Haavard Rue. Learning algorithms from bayesian principles. August 2020. Draft v. 0.7.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Alex Krizhevsky and Geoffrey E. Hinton. Using very deep autoencoders for content-based image retrieval. In *ESANN*, 2011.

Wu Lin, Mohammad Emtiyaz Khan, and Mark Schmidt. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *ICML*, volume 97, Jun 2019.

Xiaofan Lin, Cong Zhao, and Wei Pan. Towards accurate binary convolutional neural network. In *Advances in Neural Information Processing Systems 30*, pp. 345–353. 2017.

Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 722–737, 2018.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. 2016. arxiv:1611.00712.

Xiangming Meng, Roman Bachmann, and Mohammad Emtiyaz Khan. Training binary neural networks using the bayesian learning rule, 2020.

Arkadiĭ Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. 2019.

Jorn WT Peters and Max Welling. Probabilistic binary neural networks. *arXiv preprint arXiv:1809.03368*, 2018.

Tapani Raiko, Mathias Berglund, Guillaume Alain, and Laurent Dinh. Techniques for learning binary stochastic feedforward neural networks. In *ICLR*, 2015. URL `http://arxiv.org/abs/1406.2989`.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pp. 525–542. Springer, 2016.

Geoffrey Roeder, Yuhuai Wu, and David Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *NIPS*, pp. 6925–6934, 2017.

Wolfgang Roth, Günther Schindler, Holger Fröning, and Franz Pernkopf. Training discrete-valued neural networks with sign activations using weight distributions. In *European Conference on Machine Learning (ECML)*, 2019.

Oran Shayer, Dan Levi, and Ethan Fetaya. Learning discrete weights using the local reparameterization trick. *arXiv preprint arXiv:1710.07739*, 2017.

Alexander Shekhovtsov, Viktor Yanush, and Boris Flach. Path sample-analytic gradient estimators for stochastic binary networks, 2020.

Dinghan Shen, Qinliang Su, Paidamoyo Chapfuwa, Wenlin Wang, Guoyin Wang, Ricardo Henao, and Lawrence Carin. NASH: toward end-to-end neural architecture for generative semantic hashing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 2041–2050, 2018.

Wei Tang, Gang Hua, and Liang Wang. How to train a compact binary neural network with high accuracy? In *AAAI*, 2017.

Michalis K. Titsias and Miguel Lázaro-Gredilla. Local expectation gradients for black box variational inference. In *International Conference on Neural Information Processing Systems*, pp. 2638–2646, 2015.

Seiya Tokui and Issei Sato. Evaluating the variance of likelihood-ratio gradient estimators. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pp. 3414–3423, 2017.

George Tucker, Andriy Mnih, Chris J Maddison, John Lawson, and Jascha Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *NeurIPS*, pp. 2627–2636. 2017.

Xu Xiang, Yanmin Qian, and Kai Yu. Binary deep neural networks for speech recognition. In *INTERSPEECH*, 2017.

Mingzhang Yin and Mingyuan Zhou. ARM: Augment-REINFORCE-merge gradient for stochastic binary networks. In *ICLR*, 2019.

Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. *arXiv preprint arXiv:1903.05662*, 2019.

Siqi Zhang and Niao He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv: Optimization and Control*, 2018.

Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016.

Reintroducing Straight-Through Estimators as Principled Methods for Stochastic Binary Networks (Appendix)

Contents

A Related Work

B Analysis of ST with 1 Hidden Layer
- B.1 Invariances
- B.2 Bias Analysis

C Mirror Descent and Variational Mirror Descent
- C.1 Mirror Descent
- C.2 Latent Weight Decay Implements Variational Bayesian Learning

D Details of Experiments
- D.1 Variational Autoencoder
- D.2 Classification with Deep Stochastic Binary Networks

E Gumbel Softmax and ST Gumbel-Softmax

F Analysis of BayesBiNN

A Related Work

Hinton's vs Bengio's ST The name *straight-through* and the first experimental comparison was proposed by Bengio et al. (2013). Referring to Hinton's lecture, they describe the idea as "*simply to back-propagate through the hard threshold function (1 if the argument is positive, 0 otherwise) as if it had been the identity function*". In the aforementioned lecture (Hinton, 2012), however we find a somewhat different description: "*during the forward pass we stochastically pick a binary value using the output of the logistic, and then during the backward pass we pretend that we've transmitted the real valued probability from the logistic*". We can make two observations: 1) different variants appeared early on and 2) many subsequent works (*e.g.* Yin et al., 2019) attribute these two variants in the exact opposite way, adding to the confusion.

ST Analysis Yin et al. (2019) analyzes deterministic ST variants. The theoretical analysis is applicable to 1 hidden layer model quadratic loss and the input data following a Gaussian distribution. The input distribution assumption is arguably artificial, however it allows to analyze the expected loss and its gradient. They show that population ST gradients using ReLU and clipped ReLU proxy correlate positively with the true population gradient and allow for convergence while identity ST does not. In § B we show that in the SBN model, a simple correction of the quadratic loss function makes the base ST estimator unbiased and all rescaled estimators including identity are ascent directions in the expectation. Also note that the approach to analyze deterministic ST methods by considering the expectation over the input has a principle limitation for extending to deep models: the expectation over the input of a deterministic network with two hidden binary layers is still non-smooth (non-differentiable) in the parameters of the second layer.

Cheng et al. (2019) shows for networks with 1 hidden layer that STE is approximately related to the projected Wasserstein gradient flow method proposed there.

On the weights side of the problem Ajanthan et al. (2019) connected mirror descent updates for constrained optimization (*e.g.*, $w \in [0, 1]$) with straight-through methods. The connection of deterministic straight-through for weights and proximal updates was also observed in Bai et al. (2019). Mirror Descent has been applied to variational Bayesian learning of continuous weights *e.g.* in Lin et al. (2019), taking the form of update in natural parameters with the gradient in the mean parameters, same as in our case.

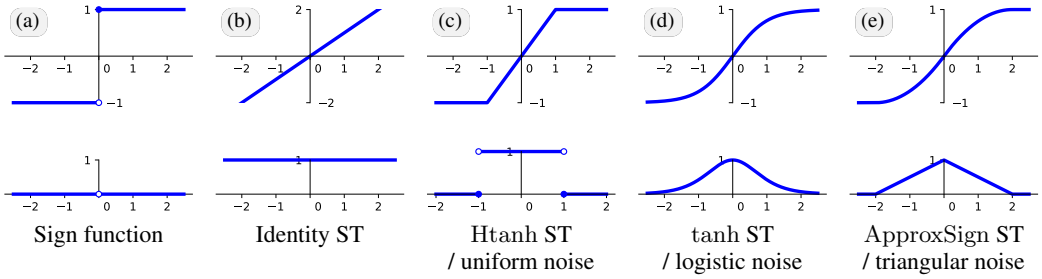

Figure A.1: The sign function and different derivative proxies used in empirical ST estimators. Variants (c-e) can be obtained by choosing the noise distribution in our framework. Specifically for a noise Z with cdf F_Z, in the upper plots we show $\mathbb{E}_Z[\text{sign}(a - Z)] = 2F_Z - 1$ and, respectively, twice the density, $2F'_Z$ in the lower plots. Choosing *uniform* distribution for Z gives the density $p_Z(z) = \frac{1}{2}\mathbb{1}_{z\in[-1,1]}$ and recovers the common Htanh proxy in (c). The *logistic* noise has cdf $F_Z(z) = S(z/2)$, which recovers tanh proxy in (d). The *triangular* noise has density $p_Z(z) = \max(0, |(2 - x)/4|)$, which recovers a scaled version of ApproxSign Liu et al. (2018) in (e). The scaling (standard deviation) of the noise in each case is chosen so that $2F'_Z(0) = 1$ as explained in § 4. The identity ST form in (b) we recover as latent weight updates with mirror descent.

Alternative Estimators For deep binary networks several gradient estimation approaches are based on stochastic gradients of analytically smoothed/approximated loss (Peters & Welling, 2018; Roth et al., 2019). There is however a discrepancy between analytic approximation and the binary samples used at the test time. Shekhovtsov et al. (2020, Fig. 4) show that such relaxed objectives may indeed significantly diverge during the training. To obtain good results, a strong dropout regularization and/or pretraining is needed (Peters & Welling, 2018; Roth et al., 2019). Despite these difficulties they demonstrate on par or improved results, especially when using average prediction over multiple noise samples at test time.

B Analysis of ST with 1 Hidden Layer

B.1 Invariances

We have the following simple yet desirable and useful property. It is easy to observe that binary activations admit equivalent reformulations as

$$\text{sign}(a_i - z_i) = \text{sign}(T(a_i) - T(z_i)) \tag{19}$$

for any strictly monotone mapping $T\colon \mathbb{R} \to \mathbb{R}$.

Proposition B.1. The gradient computed by Alg. 1 is invariant to equivalent transformations under sign as in (19).

Proof. Let us denote the transformed noise as $y_i = T(z_i)$, its cdf as G and the transformed activations as $\tilde{a}_i = T(a_i)$. The sampling probability in line 2 of Alg. 1 does not change since after the transformation it computes $p = G(\tilde{a}_i) = \mathbb{P}(y \le \tilde{a}_i \,|\, \tilde{a}_i) = \mathbb{P}(z_i \le a_i \,|\, a_i) = \mathcal{L}(a_i)$. The gradient returned by line 5 does not change since we have $\frac{d}{da_i}G(T(a_i)) = F'(a)$. □

In contrast, empirical straight-through approaches where the proxy is hand-designed fail to maintain this property. In particular in deterministic straight-through, transforms such as $\text{sign}(a_i) = \text{sign}(T(a_i))$ while keeping the backprop proxy of sign fixed lead to different gradient estimates. This partially explains why many proxies have been tried, *e.g.* ApproxSign in Liu et al. (2018), and their scale needed tuning. Another pathological special case that leads to confusions between identity straight-through and other forms is as follows.

Corollary B.1. Let F be strictly monotone. Then letting $T = F$ leads to $y = T(z)$ being uniformly distributed. Let $\tilde{a} = T(a)$. In this case the backpropagation rule in line 5 of Alg. 1 can be interpreted as replacing the gradient of $\text{sign}(\tilde{a}_i - y)$ in $\tilde{a}$ with just identity.

Indeed, since y is uniform, we have $G' = 1$ on $(0, 1)$ and $\tilde{a}_i = \mathcal{L}(a_i)$ is guaranteed to be in $(0, 1)$ by strict monotonicity. The gradient back-propagated by usual rules through $\tilde{a}_i$ (outside of

the ST Alg. 1) encounters derivative of F as before. Hence we see that the description "to back-propagate through the hard threshold function as if it had been the identity function" could be misleading as the resulting estimator crucially depends on what transformations are applied under the hard threshold despite they do not affect the network predictions in any way. We refer to the variant by Bengio et al. (2013) as identity-ST, as it specifically uses the identity proxy for the gradient in the pre-sigmoid activation.

B.2 Bias Analysis

Since the only approximation that we made was linearization of the objective $\mathcal{L}$, we have the following basic property.

Proposition B.2. If the objective function $\mathcal{L}$ is multilinear[2] in the binary variables x, then Alg. 1 is unbiased.

Proof. In this case (7) holds as equality. □

While extremely simple, this is an important point to understanding the ST estimator. As an immediate consequence we can easily design counter-examples where ST is wrong.

Example 1. Let $L(x) = x^2$ and $x = \mathrm{sign}(a - z)$. In this case the expected value of the loss is 1, independent of a. The true gradient is zero. However the expected ST gradient is $\mathbb{E}[2F'(a)2x] = 4F'(a)(2\mathcal{L}(a) - 1)$ and could be positive or negative depending on a.

Example 2 (Tokui & Sato 2017)**.** Let $\mathcal{L}(x) = x - \sin(2\pi x)$. Then the finite difference $\mathcal{L}(1) - \mathcal{L}(0) = 1$ but the derivative $\frac{\partial \mathcal{L}}{\partial x} = 1 - 2\pi\cos(2\pi x) = -1$. So in this failure example, ST, even in the expectation, will point to exactly the opposite direction of the true gradient.

An important observation from the above examples is that the result of ST is not invariant to reformulations of the loss that preserve its values in all binary points. In particular we have that $L \equiv 1$ in the first example and $\mathcal{L}(x) \equiv x$ in the second one for any $x \in \{-1, 1\}$. If we used these equivalent representations instead, the ST estimator would have been correct.

More generally, any real-valued function of binary variables has a unique multinomial representation (Boros & Hammer, 2002). Unfortunately, this representation is intractable in most of the cases, but it is tractable *e.g.* for a quadratic loss, useful in regression and autoencoders with a Gaussian observation model.

Proposition B.3. Let $\mathcal{L}(x) = \|Wx - y\|^2$. Then the multilinear equivalent reformulation of $\mathcal{L}$ is given by

$$\tilde{\mathcal{L}}(x) = \|Wx - y\|^2 - \textstyle\sum_i x_i^2 \|W_i\|^2 + \sum_i \|W_i\|^2, \tag{20}$$

where W_i is the i'th row of W.

Proof. By expanding the square and using the identity $x_i^2 = 1$ for $x_i \in \{-1, 1\}$. □

Simply adjusting the loss using this equivalence and applying ST to it, fixes the bias problem.

Next we ask the question, whether dropping the multiplier $\mathrm{diag}(F'(a))$ or changing it by another multiplier, which we call an *(internal) rescaling* of the estimator, can lead to an incorrect estimation.

Proposition B.4. If instead of $F'(a)$ any non-negative coordinate-wise function Λ is used in Alg. 1, the expected rescaled estimator preserves non-negative scalar product with the original estimator.

Proof. We write the chain (9) in a matrix form as $J_1^\mathsf{T} \Lambda_0(a) J_2^\mathsf{T}(x)$, with the Jacobians $J_1 = \frac{\partial a}{\partial \phi}$, $\Lambda^0 = \mathrm{diag}(F'(a))$ and $J_2(x) = \frac{\partial \mathcal{L}(x)}{\partial x}$. The modified gradient with Λ is then defined as $J_1^\mathsf{T} \Lambda(a) J_2^\mathsf{T}(x)$.

We are interested in the scalar product between the expected gradient estimates:

$$\langle \mathbb{E}[J_1^\mathsf{T} \Lambda_0 J_2^\mathsf{T}], \mathbb{E}[J_1^\mathsf{T} \Lambda J_2^\mathsf{T}] \rangle, \tag{21}$$

where the expectation is over x. Since neither J_1 nor Λ, Λ_0 depend on x, we can move the expectations to J_2. Let $\bar{J}_2 = \mathbb{E}\big[\frac{\partial \mathcal{L}(x)}{\partial x}\big]$. Then the scalar product between the expected estimates

[2] E.g. $x_1 x_2 x_3$ is trilinear and thus qualifies but x_1^2 is not multi-linear.

becomes

$$\langle J_1^\mathsf{T}\Lambda_0\bar{J}_2^\mathsf{T}, J_1^\mathsf{T}\Lambda\bar{J}_2^\mathsf{T}\rangle = \mathrm{Tr}(\bar{J}_2\Lambda J_1 J_1^\mathsf{T}\Lambda_0\bar{J}_2^\mathsf{T}). \tag{22}$$

Notice that $J_1 J_1^\mathsf{T}$ is positive semi-definite, Λ and Λ_0 are also positive semi-definite since they are diagonal with non-negative entries. It follows that $R = \Lambda J_1 J_1^\mathsf{T}\Lambda_0$ is positive semidefinite and that $\bar{J}_2 R J_2^\mathsf{T}$ is positive semi-definite. Its trace is non-negative. □

We obtained that the use of an internal rescaling, in particular identity instead of F', is not too destructive: if (1) was unbiased, the rescaled estimator may be biased but it is guaranteed to give an ascend direction in the expectation so that the optimization can in principle succeed. However, assuming that (1) is biased (when $\mathcal{L}$ is not multi-linear) but gives an ascent direction in the expectation, *the ascent direction property cannot be longer guaranteed for the rescaled gradient*.

Next, we study whether the ST gradient is a valid ascent direction even when $\mathcal{L}$ is not multi-linear.

Proposition B.5. Let $\mathcal{L}(x)$ be such that its partial derivative $g_i = \frac{\partial \mathcal{L}}{\partial x_i}$ as a function of x_i is L-Lipschitz continuous for all i with a constant L. Then the expected ST gradient is an ascent direction for any $a(\phi)$ and $\mathcal{L}(x)$ if and only if

$$\big|\mathbb{E}[g_i]\big| > L \text{ for all } i. \tag{23}$$

Proof. Sufficiency (if part). The true gradient using the local expectation form (6) expresses as

$$\mathbb{E}\Big[\sum_i \big(\tfrac{\partial a_i}{\partial \phi}\big)\big(p_Z(a_i)\big)x_i\big(\mathcal{L}(x)-\mathcal{L}(x_{\downarrow i})\big)\Big] = \mathbb{E}[J\Delta], \tag{24}$$

where the expectation is w.r.t. $x \sim q(x;\phi)$ and we introduced the matrix notation $J = \big(\frac{\partial a}{\partial \phi}\big)^\mathsf{T}\,\mathrm{diag}(p_Z(a))$, and $\Delta_i = x_i\big(\mathcal{L}(x)-\mathcal{L}(x_{\downarrow i})\big)$. The ST gradient replaces Δ_i with $2g_i(x)$. Since in both cases J does not depend on x, the expectation can be moved to the last term. Respectively, let us define $\bar{\Delta} = \mathbb{E}[\Delta]$ and $\bar{g} = \mathbb{E}[g]$. The scalar product between the true gradient and the expected ST gradient can then be expressed as

$$\langle J\bar{\Delta}, J\bar{g}\rangle = \mathrm{Tr}(J\bar{g}\bar{\Delta}^\mathsf{T}J^\mathsf{T}). \tag{25}$$

From the relation

$$x_i(\mathcal{L}(x) - \mathcal{L}(x_{\downarrow i})) = \int_{-1}^{1} g_i(x)\mathrm{d}x_i \tag{26}$$

and Lipschitz continuity of g_i in x_i we have bounds

$$2(g_i(x) - L) \le x_i(\mathcal{L}(x) - \mathcal{L}(x_{\downarrow i})) \le 2(g_i(x) + L). \tag{27}$$

It follows that

$$2(\mathbb{E}[g] - L) \le \mathbb{E}[\Delta] \le 2(E[g] + L), \tag{28}$$

coordinate-wise. The outer product $\bar{g}\bar{\Delta}^\mathsf{T}$ is positive semidefinite iff $\bar{g}_i\bar{\Delta}_i \ge 0$ for all i. According to bounds above, this holds true if

$$(\forall i|\bar{g}_i \ge 0) \quad 2(|\bar{g}_i| - L) \ge 0 \tag{29}$$

$$(\forall i|\bar{g}_i < 0) \quad 2(|\bar{g}_i| + L) \le 0, \tag{30}$$

or simply $(\forall i)\ |\bar{g}_i| \ge L$.

Necessity (only if part). We want to show that the requirements (23), which are simultaneous for all coordinates of g, cannot be relaxed unless we make some further assumptions on a or $\mathcal{L}$. Namely, if $\exists i^*$ such that $\bar{g}_{i^*}\bar{\Delta}_{i^*} < 0$, then there exists a such that $\langle J\bar{g}, J\bar{\Delta}\rangle < 0$. *I.e.* a single wrong direction can potentially be rescaled by the downstream Jacobians to dominate the contribution of other components. More detailed steps are as follows.

Assume $(\exists i^*)\ |\bar{g}_{i^*}| < L$. Then exists $\mathcal{L}(x)$ such that the bounds (27) are tight (*e.g.* $\mathcal{L}(x) = x^2$) and therefore there will hold $\bar{g}_{i^*}\bar{\Delta}_{i^*} < 0$. Since $\Lambda = \mathrm{diag}(p_Z(a))$ is positive semi-definite, $\Lambda\bar{g}\bar{\Delta}^\mathsf{T}\Lambda$ will preserve the non-positive sign of the component (i^*, i^*). There exists then $a(\phi)$ such that $\frac{\partial a}{\partial \phi}$ scales down all coordinates $i \ne i^*$ and scales up i^* such that the $\mathrm{Tr}(J\bar{g}\bar{\Delta}^\mathsf{T}J^\mathsf{T})$ is dominated by the entry (i^*, i^*). The resulting scalar product between the expected gradients in ϕ thus can be negative. □

The requirements of Proposition B.5, which is simultaneous for all coordinates of g cannot be relaxed unless we make some further assumptions on J. Namely, if $\exists i$ such that $\bar{g}_i \bar{\Delta}_i < 0$, then there exists $J \succcurlyeq 0$ such that $\langle J\bar{g}, J\bar{\Delta}\rangle < 0$. *I.e.* a single wrong direction can potentially be rescaled by the downstream Jacobians to dominate the contribution of other components.

Next we study, the typical use case when hidden binary variables are combined using a linear layer, initialized randomly. A typical initialization procedure would rescale the weights according to the size of fan-in for each output.

Proposition B.6. Assume the loss function is applied after a linear normalized transform of Bernoulli variables, *i.e.* takes the form

$$\mathcal{L}(x) = \ell(Wx), \tag{31}$$

where $W \in \mathbb{R}^{K\times n}$ is a matrix of normally distributed weights, normalized to satisfy $\|W_{k,:}\|_2^2 = 1$ $\forall i$. Then the expected Lipschitz constant of gradients of $\mathcal{L}$ scales as $O(\frac{1}{\sqrt{n}})$.

Proof. Let the Lipschitz constant of ℓ be L. The gradient of $\mathcal{L}$ expresses as

$$g_i = \frac{\mathrm{d}\mathcal{L}(x)}{\mathrm{d}x} = \ell'(Wx)W_{k,j}. \tag{32}$$

By assumptions of random initialization and normalization, $W_{k,j} \sim \mathcal{N}(0, \frac{1}{n})$. If we consider $|g_i|$ in the expectation over initialization we obtain that

$$\mathbb{E}_W\left[|g_i(x) - g_i(y)|\right] = \mathbb{E}_W\left[|\ell'(Wx) - \ell'(Wy)| \cdot |W_{k,i}|\right] \leq L\mathbb{E}_W\left[|W_{k,j}|\right] = L\sqrt{\frac{2}{n\pi}}. \tag{33}$$

Therefore g_i has expected Lipschitz constant $L\sqrt{\frac{2}{n\pi}}$. □

We obtain an important corollary.

Corollary B.2. As we increase the number of hidden binary units in the model, the bias of ST decreases, at least at initialization.

Finally, we study conditions when a deterministic version of ST gives a valid ascent direction.

Proposition B.7. Let $x^* = \mathrm{sign}(a)$. Let $g(x) = \frac{\partial \mathcal{L}(x)}{\partial x_i}$ be Lipschitz continuous with constant L. Let $g^* = g(x^*)$ and $p^* = p(x^*|a)$. The deterministic ST gradient at x^* forms a positive scalar product with the expected stochastic ST gradient if

$$|g_i^*| \geq 2(1-p^*)L \;\; \forall i. \tag{34}$$

Proof. Similarly to Proposition B.5 let $J = \left(\frac{\partial a}{\partial \phi}\right)^\mathsf{T} \mathrm{diag}(p_Z(a))$. The scalar product between the expected ST gradient and the deterministic ST gradient is given by

$$\langle J\mathbb{E}[g(x)], Jg^*\rangle = \mathrm{Tr}\left(\mathbb{E}[g(x)]{g^*}^\mathsf{T} J^\mathsf{T}\right). \tag{35}$$

In order for it to be non-negative we need $\mathbb{E}[g(x)_i]g_i^* \geq 0$ $\forall i$. Observe that $\mathbb{E}[g(x)_i]$ is a sum that includes g_i^* with the weight p^*. We therefore need

$$\textstyle\sum_{x\neq x^*} p(x|a)g(x)_i g_i^* + p^* {g_i^*}^2 \geq 0. \tag{36}$$

From Lipschitz continuity of g_i we have the bound $|g(x)_i - g_i^*| \leq L|x_i - x_i^*|$, or using that $|x_i - x_i^*| \leq 2$ we have

$$g_i^* - 2L \leq g(x)_i \leq g_i^* + 2L. \tag{37}$$

Therefore

$$g(x)_i g_i^* \geq {g_i^*}^2 - 2L|g_i^*|. \tag{38}$$

We thus can lower bound (36) as

$$\textstyle\sum_{x\neq x^*} p(x|a)(|g_i^*| - 2L)|g_i^*| + p^* {g_i^*}^2 = -2L|g_i^*|(1-p^*) + {g_i^*}^2. \tag{39}$$

This lower bound is non-negative if

$$|g_i^*| \geq 2L(1-p^*). \tag{40}$$

□

Compared to Proposition B.5, this condition has an extra factor of $2(1-p^*)$. Since p^* is the product of probabilities of all units x_i^*, we expect initially $p^* \ll 1$. This condition improves at the same rate with the increase in the number of hidden units in the case covered by Proposition B.6. In addition it becomes progressively more accurate as units learn to be more deterministic, because in this case the factor $(1-p^*)$ decreases. However note that this proposition describes the gap between deterministic ST and stochastic ST. And even when this gap diminishes, the gap between ST and the true gradient may remain.

We can obtain a similar sufficient condition for the scalar product between deterministic ST and the executed true gradient, that unlike the combination of Proposition B.5 and Proposition B.7 ensures ascent direction.

Proposition B.8. Let $x^* = \text{sign}(a)$. Let $g(x) = \frac{\partial \mathcal{L}(x)}{\partial x_i}$ be Lipschitz continuous with constant L. Let $g^* = g(x^*)$ and $p^* = p(x^*|a)$. The deterministic ST gradient at x^* forms a positive scalar product with the expected true gradient if

$$|g_i^*| \geq 2(1-p^*)L + L \;\; \forall i. \tag{41}$$

Proof. The proof is similar to Proposition B.7, only in this case we need to ensure $\mathbb{E}[\Delta_i]g_i^* \geq 0$. Using (28) we get the bounds

$$2(\mathbb{E}[g] - L) \leq \mathbb{E}[\Delta] \leq 2(E[g] + L), \tag{42}$$

And using additionally (37) we get

$$2(p^* g_i^* + (1-p^*)(g_i^* - 2L) - L) \leq \mathbb{E}[\Delta] \leq 2(p^* g_i^* + (1-p^*)(g_i^* + 2L) + L). \tag{43}$$

Collecting the terms

$$2(g_i^* - (1-p^*)2L - L) \leq \mathbb{E}[\Delta] \leq 2(g_i^* + (1-p^*)2L + L). \tag{44}$$

Multiplying by g_i^* we obtain that the sufficient condition for $\mathbb{E}[\Delta_i]g_i^* \geq 0$ is

$$|g_i^*| \geq (1-p^*)2L + L. \tag{45}$$

□

C Mirror Descent and Variational Mirror Descent

C.1 Mirror Descent

Mirror descent is a widely used method for constrained optimization of the form $\min_{x \in \mathcal{X} \subset \mathbb{R}^n} f(x)$ introduced by Nemirovsky & Yudin (1983). Let $\Phi : \mathcal{X} \to \mathbb{R}$ be strongly convex and continuously differentiable on $\mathcal{X}$. Bregman divergence $D_\Phi(x, y)$ associated with Φ is defined as

$$D_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y \rangle. \tag{46}$$

An update of MD algorithm can be written as:

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} \langle x, \nabla f(x^k) \rangle + \frac{1}{\varepsilon} D_\Phi(x, x^k). \tag{47}$$

In the unconstrained case when $\mathcal{X} = \mathbb{R}^n$ or in the case when the critical point is guaranteed to be in $\mathcal{X}$ (as typically ensured by the design of D_Φ), the solution can be found from the critical point equations, leading to the general form of iterates

$$\begin{aligned} \nabla\Phi(x^{k+1}) &= \nabla\Phi(x^k) - \varepsilon \nabla f(x^k) \\ x^{k+1} &= (\nabla\Phi)^{-1} \left(\nabla\Phi(x^k) - \varepsilon \nabla f(x^k) \right). \end{aligned} \tag{48}$$

Proposition 1. Common SGD in latent weights η using the *identity straight-through-weights* Alg. 2 implements SMD in the weight probabilities θ with the divergence corresponding to F.

Proof. We start from the defining equation of MD update in the form (48). In order for (48) to match common SGD on $\eta = F^{-1}(\theta)$, the mirror map Φ must satisfy $\nabla\Phi(\theta) = F^{-1}(\theta)$. Since cdf F is monotone, its inverse exists if F is strictly monotone. Finding the mirror map Φ explicitly is not

necessary for our purpose, however in 1D case it can be expressed simply as $\Phi(x) = \int_0^x F^{-1}(\eta)d\eta$. With this mirror map, the MD update can be written as

$$\eta_{k+1} = \eta_k - \varepsilon \frac{d\mathcal{L}}{d\theta}\Big|_{\theta=F(\eta_k)}. \tag{49}$$

Thus MD on θ takes the form of a descent step on η with the gradient $\frac{d\mathcal{L}}{d\theta}$. A common SGD on η would use the gradient $\frac{d\mathcal{L}}{d\eta} = \frac{\partial\mathcal{L}}{\partial\theta}\frac{\partial\theta}{\partial\eta}$. Thus (49) bypasses the Jacobian $\frac{\partial\theta}{\partial\eta}$. This is exactly what Alg. 2 does. More precisely, when applying the same derivations that we used to obtain ST for activations in order to estimate $\frac{d\mathcal{L}}{d\theta}$, since $F(\eta) = \theta$ we obtain that the factor $\frac{\partial}{\partial\theta}p(w_i;\theta)$ present in (6) expresses as

$$\frac{dF(\eta)}{d\theta} = \frac{\partial F(F^{-1}(\theta))}{\partial\theta} = 1 \tag{50}$$

and thus can be omitted from the chain rule as defined in Alg. 2. □

C.2 Latent Weight Decay Implements Variational Bayesian Learning

In the Bayesian learning setting we consider a model with binary weights w and are interested in estimating $p(w|D)$, the posterior distribution of weights given the data and the weights prior $p(w)$. In the variational Bayesian (VB) formulation, this difficult and multi-modal posterior is approximated by a simpler one $q(w)$, commonly a fully factorized distribution, by minimizing $\mathrm{KL}(q(w)\|p(w|D))$. Let $q(w) = \mathrm{Ber}(w;\theta)$ and $p(w) = \mathrm{Ber}(w;\frac{1}{2})$, both meant component-wise, i.e. fully factorized. Then the VB problem takes the form

$$\min_\theta \big\{ -\mathbb{E}_{(x^0,y)\sim\text{data}}\mathbb{E}_{w\sim\mathrm{Ber}(\theta)}\big[\log p(y|x^0;w)\big] + \tfrac{1}{N}\mathrm{KL}(\mathrm{Ber}(\theta)\|\mathrm{Ber}(\tfrac{1}{2}))\big\}, \tag{51}$$

where we have rewritten the data likelihood as expectation and hence the coefficient $1/N$ in front of the KL term. This problem is commonly solved by SGD, taking one sample from the training data and one sample of w and applying backpropagation Graves (2011). We can in principle do the same by applying an estimator for the gradient in θ.

The trick that we apply, different from common practices, is not to compute the gradient of the KL term but to keep this term explicit throughout to the proximal step leading to a *composite* MD (Zhang & He, 2018). With this we have

Proposition 2. Common SGD in latent weights η with a weight decay and *identity straight-through-weights* Alg. 2 is equivalent to optimizing a factorized variational approximation to the weight posterior $p(w|D)$ using a composite SMD method.

Proof. Expanding data log-likelihood as the sum over all data points, we get

$$\log p(D\,|\,w) = \sum_i \log p(x_i\,|\,w) =: \sum_i l_i(w). \tag{52}$$

When multiplying with $\frac{1}{N}$, the first term becomes the usual expected data likelihood, where the expectation is in training data and parameters $w \sim q(w)$. Expanding also the parameterization of $q(w) = \mathrm{Ber}(w\,|\,\theta)$, the variational inference reads

$$\min_\theta -\mathbb{E}_{w\sim\mathrm{Ber}(\theta)}\tfrac{1}{N}\textstyle\sum_i l_i(w) + \tfrac{1}{N}\mathrm{KL}(q(w)\|p(w)) + \mathit{const}. \tag{53}$$

We employ mirror descent to handle constraints $\theta \in [0,1]$ similar to the above but now we apply it to this composite function, linearizing only the data part and keeping the prior KL part non-linear. Let

$$g = \frac{1}{|I|}\sum_{i\in I}\nabla_\theta\mathbb{E}_{w\sim\mathrm{Ber}(\theta)}l_i(w)$$

be the stochastic gradient of the data term in the weight probabilities θ using a min-batch I. The SMD step subproblem reads

$$\min_\theta \Big(g^\mathsf{T}\theta + \tfrac{1}{\varepsilon}\mathrm{KL}(\mathrm{Ber}(\theta)\|\mathrm{Ber}(\theta^t)) + \tfrac{1}{N}\mathrm{KL}(\mathrm{Ber}(\theta)\|\mathrm{Ber}(\tfrac{1}{2}))\Big). \tag{54}$$

We notice that $\mathrm{KL}(\mathrm{Ber}(\theta)\|\mathrm{Ber}(\frac{1}{2})) = -H(\mathrm{Ber}(\theta))$ and also introduce the prior scaling coefficient $\lambda = \frac{1}{N}$ in front of the entropy, which may optionally be lowered to decrease the regularization effect. With these notations, the composite proximal problem becomes

$$\min_\theta \Big(g^\mathsf{T}\theta + \tfrac{1}{\varepsilon}\mathrm{KL}(\mathrm{Ber}(\theta)\|\mathrm{Ber}(\theta^t)) - \lambda H(\mathrm{Ber}(\theta))\Big). \tag{55}$$

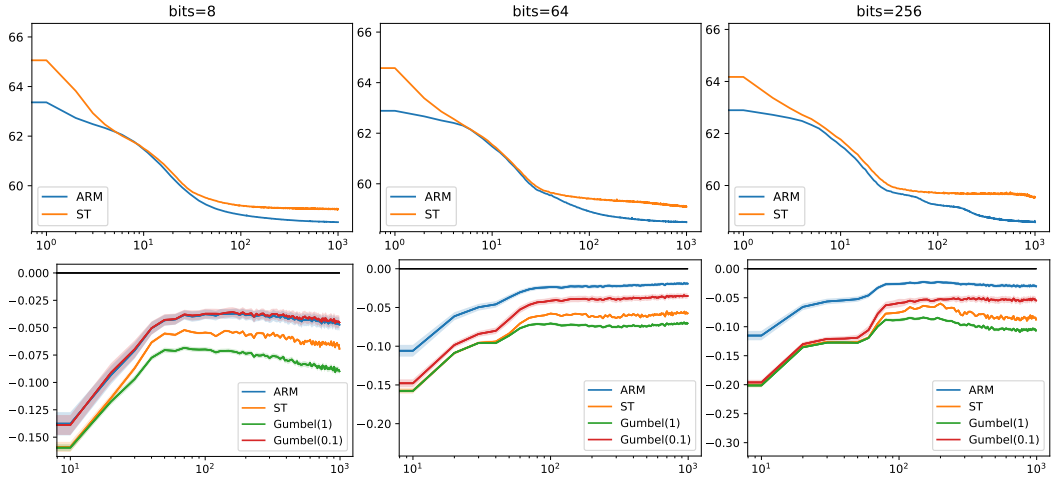

Figure C.1: VAE training of a 1-layer decoder for different number of latent bits. *Top:* training loss (ELOB). *Bottom:* expected loss decrease rate using different estimators at parameter points of ARM training trajectory (lower is better). X-axis correspond to epochs. Please see text for details.

The solution is found from the critical point equation in θ:

$$\nabla_\theta \big(g^\mathsf{T}\theta + \tfrac{1}{\varepsilon}\mathrm{KL}(\mathrm{Ber}(\theta)\|\mathrm{Ber}(\theta^t)) - \lambda H(\mathrm{Ber}(\theta))\big) = 0 \tag{56a}$$

$$g + \tfrac{1}{\varepsilon}\left(\log\tfrac{\theta}{1-\theta} - \log\tfrac{\theta^t}{1-\theta^t}\right) + \lambda \log\tfrac{\theta}{1-\theta} = 0 \tag{56b}$$

$$(\varepsilon\lambda + 1)\log\tfrac{\theta}{1-\theta} = \log\tfrac{\theta^t}{1-\theta^t} - \varepsilon g \tag{56c}$$

$$\log\tfrac{\theta}{1-\theta} = \tfrac{1}{\varepsilon\lambda+1}\log\tfrac{\theta^t}{1-\theta^t} - \tfrac{\varepsilon}{\varepsilon\lambda+1}g. \tag{56d}$$

For the natural parameters we obtain:

$$\eta = \tfrac{\eta^t - \varepsilon g}{\varepsilon\lambda+1} = \eta^t - \tfrac{\varepsilon}{\varepsilon\lambda+1}\Big(\lambda\eta^t + g\Big). \tag{57}$$

We can further drop the correction of the step size $\frac{\varepsilon}{\varepsilon\lambda+1}$ since $\varepsilon\lambda + 1 \approx 1$ and the step size will need to be selected by cross validation anyhow. This gives us an update of the form

$$\eta = \eta^t - \varepsilon(g + \lambda\eta^t), \tag{58}$$

which is in the form of a standard step in any SGD or adaptive SGD optimizer. The difference is that the gradient in probabilities θ is applied to make step in logits η and the prior KL divergence contributes the *logit decay* λ, which in this case is the *latent weight decay*. Since the ST gradient in θ differs from the ST gradient in η by the factor F', the claim of Proposition 2 follows. □

D Details of Experiments

D.1 Variational Autoencoder

Following recent work on semantic hashing Dadaneh et al. (2020), we train a variational autoencoder with binary latent variables. We select text generation task *20newsgoups*[3] as in this case it is computationally more tractable to accurately estimate the expected function decrease rate by estimating the true gradient g. Although in this problem other estimators are established to perform well, it allows us to verify the properties of ST estimator that we proposed theoretically.

We closely follow the model Dadaneh et al. (2020) in the observation model $p(x|z;\theta)$. In this case x represents a vector of word frequencies of a document and $\log p(x|z)$ is evaluated for a 1-layer decoder as $\langle x, \log\mathrm{softmax}(Wz + b)\rangle$, where (W, b) are decoder parameters. The encoder network $q(z|x;\phi)$ is a 2-layer network with ReLU activation in its hidden layer and sigmoid output probabilities of Bernoulli latent codes. We used 256 hidden units (differentiable and deterministic) and

[3] `http://qwone.com/~jason/20Newsgroups/`

ReLU activations in the encoder and 2-layer decoder model. Observations x are 10000-dimensional by the design of the dataset, and the number of latent Bernoulli states is varied.

To generate the training trajectory we applies ARM estimator and trained for 3000 epochs using Adam with learning rate 0.001 and default other parameters. Standard weight initialization is applied with $W_{i,j} \sim \mathcal{U}[-1/\sqrt{k}, 1/\sqrt{k}]$, where k is the number of input dimensions to the layer.

We used the ELBO objective reformulation (Roeder et al., 2017) that computes the entropy of the variational distribution stochastically and allows to reduce variance of the gradient estimators by subtracting a zero variance term. To be more precise our loss is

$$\mathbb{E}_{x\sim\text{data}}\mathbb{E}_{z\sim q(z|x)}\Big[\log p(x|z) - \texttt{stop_grad}(\log q(z|x))\Big], \tag{59}$$

where `stop_grad` preserves the value of the argument but blocks gradient dependence through it. Here also used that the prior on z is independent uniform Bernoulli. Note that in this reformulation $\log q(z|x)$ thanks to conditional independence is a sum over hidden units and it is thus a *linear function of* z and therefore its ST gradient is unbiased.

We propose to evaluate the utility of estimators for stochastic optimization by measuring their quality using the same proximal problem objective that is used to make optimization steps. Namely, let $g = \nabla_\phi L(\phi^t)$ be the true gradient at the current point. The SGD step is defined as

$$\phi^{t+1} = \phi^t + \arg\min_{\Delta\phi} \langle g, \Delta\phi\rangle + \tfrac{1}{2\varepsilon}\|\Delta\phi\|^2. \tag{60}$$

The optimal solution is given by $\Delta\phi = -\varepsilon g$. If we use the solution $\Delta\phi = -\alpha\hat{g}$, where $\hat{g}$ is an estimator of g and α is one scalar parameter to adopt step size. We consider the expected descend on the proxy objectives:

$$\mathbb{E}\Big[\langle g, -\alpha\hat{g}\rangle + \tfrac{\alpha^2}{2\varepsilon}\|\hat{g}\|^2\Big]. \tag{61}$$

The parameter α correspond to a learning rate that can be tuned or adapted during learning. We set it optimistically for each estimator by minimizing the expected objective (61), which is a simple quadratic function in α. Without optimizing in α some methods may indeed obtain an advantage by overestimating the gradient length. To estimate (61) and α we collect 100 samples per minibatch. The optimal α is given by

$$\alpha = \varepsilon\frac{\mathbb{E}[\langle g,\hat{g}\rangle]}{\|\hat{g}\|^2} \tag{62}$$

and the value of the objective for this optimal α is

$$-\frac{\varepsilon}{2}\frac{\mathbb{E}[\langle g,\hat{g}\rangle]^2}{\|\hat{g}\|^2}. \tag{63}$$

For the purpose of comparing estimators it is equivalent to measure its square root, *i.e.* the expected loss decrease normalized by the RMS of the gradients:

$$\frac{-\mathbb{E}[\langle g,\hat{g}\rangle]}{\sqrt{\mathbb{E}[\|\hat{g}\|^2]}}, \tag{64}$$

which is the expression (18). Note also that in (64) we preserve the sign of $\mathbb{E}[\langle g, \hat{g}\rangle]$ and if the estimator is systematically wrong, we expect to measure a positive value in (64).

The results with 1-layer decoder used in some works on semantic hashing occurred to be less interesting Fig. C.1 in the sense that all estimators perform reasonably well and the decoder optimization seems to stabilize with a relatively hi likelihood indicating underfitting. However we can still see ST improves relatively to ARM with more hidden units present. We are not sure why training with ST 'stalls in the end, it may be related to the difficulty of handling high gradients of entropy by Adam. The results with 2-layer decoder Fig. C.1 show a more dynamical behavior and illustrate the differences more.

D.2 Classification with Deep Stochastic Binary Networks

As a primary dataset for testing our proposed method we use CIFAR-10 dataset[4]. It consists of 60000 32x32 color images divided in 10 classes, 6000 images per class. There is a predefined

[4] https://www.cs.toronto.edu/~kriz/cifar.html

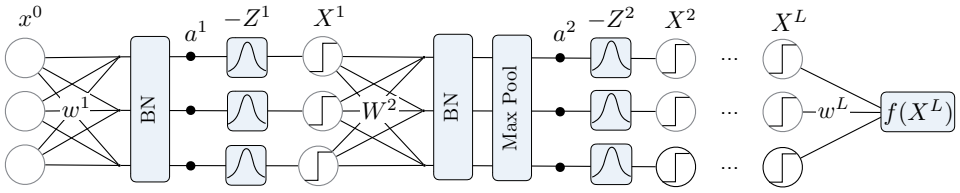

Figure D.1: Stochastic Binary Network: first and last layer have real-valued weights. BN layers have real-valued scale and bias parameters that can adjust scaling of activations relative to noise. Z are independent injected noises with a chosen distribution. Binary weights W_{ij} are random ± 1 Bernoulli(θ_{ij}) with learnable probabilities θ_{ij}. In experiments we consider SBN with a convolutional architecture same as Courbariaux et al. (2015); Hubara et al. (2016): $(2{\times}128\text{C3}) - \text{MP2} - (2{\times}256\text{C3}) - \text{MP2} - (2{\times}512\text{C3}) - \text{MP2} - (2{\times}1024\text{FC}) - 10\text{FC} - \text{softmax}$.

training set of 50000 examples and test set of 10000 examples. During training we use standard augmentation for CIFAR-10, namely random horizontal flipping and random cropping of 32×32 region with a random padding of 0-4 px on each side. We use Adam optimizer Kingma & Ba (2014) in all the experiments. The initial learning rate $\gamma = 0.01$ is used for 300 epochs and then we divide it by 10 at epochs 300 and 400 and stop at epoch 500. This is fixed for all models. All other Adam hyperparameters such as $\beta_1, \beta_2, \varepsilon$ are set to their correspondent default values in the PyTorch Paszke et al. (2019) framework. We compute mean and standard deviation for the obtained accuracy by averaging the results over 4 different random seeds for each experiment.

E GUMBEL SOFTMAX AND ST GUMBEL-SOFTMAX

Gumbel Softmax (Jang et al., 2016) or Concrete relaxations (Maddison et al., 2016) are techniques that overcome the problem of non-differentiability of $\arg\max$ (in binary case equivalent to sign in (2)) by approximating it with a smooth function whose degree of smoothness is controlled by a temperature-like parameter $\tau > 0$, where the limit $\tau \to 0$ recovers the original non-smooth function.

We analyze this estimator for the case of a single neuron and show the following properties, which to our knowledge were not published before:

- **I)** GS estimator is asymptotically unbiased as $\tau \to +0$ and the bias decreases at the rate $O(\tau)$. But for any fixed τ it is biased even when $\mathcal{L}$ is linear.
- **II)** For any given noise realization $z \neq \eta$ the GS gradient norm approaches zero at the exponential rate $O(\frac{1}{\tau}c^{\frac{1}{\tau}})$, where $c = e^{-|\eta - z|} < 1$.
- **III)** The probability of the event that gradient norm is at least ε is asymptotically $O(\tau \log \frac{1}{\varepsilon})$. Thus the probability to observe a non-zero gradient up to numerical precision quickly vanishes.
- **IV)** The variance of GS estimator grows at the rate $O(\frac{1}{\tau})$.
- **V)** The ST Gumbel-softmax estimator (Jang et al., 2016) is biased even asymptotically even for a linear loss.

Properties II and III are easy to extend to the case of layer with multiple units since they apply to just the factor $\frac{\mathrm{d}}{\mathrm{d}\eta}\sigma_\tau(\eta - z)$, which is present independently at all units. Property II can be extended to deep networks with L layers of Bernoulli variables, in which case the chain derivative will encounter L such factors and we obtain that the gradients will vanish at the rate $O(\tau^L)$.

The proofs are given below. Basically all these facts should convince the reader of the following: it is not possible to use a very small τ, not even with an annealing schedule starting from $\tau = 1$. For very small τ, the most likely consequence would be to never encounter a non-zero gradient during the whole training. For moderately small τ the variance would be prohibitively high. Indeed, Jang et al. (2016) anneals τ only down to 0.5 in their experiments.

Definitions The Gumbel-max reparametrization constructs a sample from a multinomial distribution with class probabilities π_i as

$$x = \text{one_hot}(\arg\max_i(\log \pi_i + \Gamma_i)), \tag{65}$$

where x is 1-hot encoding of the class and Γ_i are independent Gumbel noises. This reparametrization is exact and is not differentiable as it outputs discrete samples. The Gumbel-softmax replaces the hard $\arg\max$ indicator with softmax to compute the relaxed sample $y \in \mathbb{R}^n$ as

$$y = \mathrm{softmax}((\tfrac{1}{\tau}(\log \pi_i + \Gamma_i))_i). \tag{66}$$

We will focus on the binary case. We can express then

$$x_1 = [\![\log \pi_1 + \Gamma_1 \geq \log \pi_0 + \Gamma_0]\!] = [\![\eta - z \geq 0]\!], \tag{67}$$

where $\eta = \log \frac{\pi_1}{1-\pi_1}$ and $z = \Gamma_0 - \Gamma_1$. As expected z has logistic distribution and we recover that $p(x_1{=}1) = \sigma(\eta)$.

The relaxation y takes a similar form. By dividing over the numerator in (66) we obtain:

$$y_1 = \tfrac{1}{1+\exp(-(\eta-z)/\tau)} = \sigma(\tfrac{\eta-z}{\tau}) =: \sigma_\tau(\eta - z) \tag{68}$$

$$y_0 = 1 - y_1. \tag{69}$$

The loss function $\mathcal{L}$ is extended to the simplex of (y_1, y_0) by

$$\tilde{\mathcal{L}}(y_1, y_0) = \mathcal{L}(y_1). \tag{70}$$

Recall that ST estimates the gradient in η by sampling x_1 and composing

$$\tfrac{\partial \mathcal{L}}{\partial x_1} \tfrac{\partial \sigma(\eta)}{\partial \eta}. \tag{71}$$

The Gumbel-Softmax estimates the gradient by sampling z and computing

$$\hat{G}_\tau := \tfrac{d}{d\eta}\tilde{\mathcal{L}}(y(\eta)) = \tfrac{d}{d\eta}\mathcal{L}(\sigma_\tau(\eta - z)) = \tfrac{\partial \mathcal{L}(y_1)}{\partial y_1}\tfrac{\partial \sigma_\tau(\eta - z)}{\partial \eta}. \tag{72}$$

So the gradient of the relaxed loss is multiplied with the gradient of stretched (assuming $\tau < 1$) and shifted sigmoid.

Proposition E.1. The estimator is asymptotically unbiased as $\tau \to 0$ and the bias decreases at the rate $O(\tau)$.

Proof. Let us denote

$$\hat{g}_t = \mathbb{E}_z[\hat{g}_t] = \textstyle\int_{-\infty}^{\infty} \tfrac{d}{d\eta}\mathcal{L}(\sigma_\tau(\eta - z))p_Z(z)\mathrm{d}z. \tag{73}$$

Note that the limit $\lim_{\tau\to 0+} g_t(x)$ cannot be simply interchanged with the integral above – no qualification theorem allows this. We apply the following reformulation. The derivative $\frac{d}{d\eta}\mathcal{L}(\sigma_\tau(\eta - z))$ expands as

$$\mathcal{L}'(\sigma_\tau(\eta - z))\sigma_\tau(\eta - z)(1 - \sigma_\tau(\eta - z))\tfrac{1}{\tau}. \tag{74}$$

We make a change of variables $v = \sigma_\tau(\eta - z)$ in the integral. This gives $z = \eta - \tau\mathrm{logit}(v)$ and $\mathrm{d}z = -\tau\frac{1}{v(1-v)}\mathrm{d}v$. Substituting and cancelling part of the terms, we obtain

$$\hat{g}_t = \textstyle\int_0^1 \mathcal{L}'(v)p_Z(\eta - \tau\mathrm{logit}(v))\mathrm{d}v. \tag{75}$$

With this expression we can now interchange the limit and the integral using the dominated convergence theorem. In order to apply it we need to show that there exist an integrable function $\bar{g}(v)$ such that

$$|\mathcal{L}'(v)p_Z(\eta - \tau\mathrm{logit}(v))| < \bar{g}(v) \tag{76}$$

for all $\tau > 0$. Observe that

$$\sup_{v\in[0,1]} |p_Z(\eta - \tau\mathrm{logit}(v))| = \sup_{u\in\mathbb{R}} |p_Z(\eta - \tau u)| = \sup_{y\in\mathbb{R}} |p_Z(y)| = p_Z(0) = \tfrac{1}{4}, \tag{77}$$

where we used that the maximum of standard logistic density is attained at zero. We can therefore let $\bar{g}(v) = \mathcal{L}'(v)/4$. Since $\mathcal{L}'(v)$ is the derivative of $\mathcal{L}$, it is integrable on $[0, 1]$. Therefore the conditions of the dominated convergence theorem are satisfied and we have

$$\lim_{\tau\to 0+} \hat{g}_\tau = \textstyle\int_0^1 \mathcal{L}'(v) \lim_{\tau\to 0+} p_Z(\eta - \tau\mathrm{logit}(v))\mathrm{d}v \tag{78a}$$

$$= \textstyle\int_0^1 \mathcal{L}'(v)p_Z(\eta)\mathrm{d}v = (\mathcal{L}(1) - \mathcal{L}(0))p_Z(\eta), \tag{78b}$$

which is the correct value of the gradient.

Next, we obtain the series representation of the estimator bias in the asymptote $\tau \to 0+$. We approximate $p_Z(\eta - \tau \mathrm{logit}(v))$ with its Taylor series around $\tau = 0$:

$$p_Z(\eta - \tau \mathrm{logit}(v)) = p_Z(\eta) + c_1 \mathrm{logit}(v)\tau + c_2 \mathrm{logit}^2(v)\tau^2 + O(\tau^3) \tag{79}$$

where

$$c_1 = p_Z(\eta)\tfrac{e^\eta - 1}{e^\eta + 1}; \quad c_2 = p_Z(\eta)\tfrac{-4e^\eta + e^{2\eta} + 1}{2(e^\eta+1)^2}. \tag{80}$$

This is obtained using mathematica. We use this expansion in the integral representation (75). Observing that $g = p_Z(\eta)\int_0^1 \mathcal{L}'(v)\mathrm{d}v$, the zero order term becomes the true gradient. It follows that the bias of G_τ is asymptotically

$$c_1\Big(\int_0^1 \mathcal{L}'(v)\mathrm{logit}(v)\mathrm{d}v\Big)\tau + c_2(x)\Big(\int_0^1 \mathcal{L}'(v)\mathrm{logit}^2(v)\mathrm{d}v\Big)\tau^2 + O(\tau^3). \tag{81}$$

In the case when $\mathcal{L}$ is linear, the first order term vanishes because $\mathcal{L}'$ is constant and $\mathrm{logit}(v)$ is odd about $\frac{1}{2}$. However τ^2 and higher order even terms do not vanish, therefore the estimator is still biased even for linear objectives. □

This property sounds as a good and desirable one, but it's advantage is practically nullified by the next properties.

Proposition E.2. For any given realization $z \neq \eta$ the norm of Gumbel-Softmax gradient estimate asymptotically vanishes at the rate $O(\frac{1}{\tau}c^{1/\tau})$ with $c = e^{-|x|} < 1$.

Proof. Considering z fixed and denoting $x = \eta - z$, we need to check the asymptotic behavior of

$$\tfrac{\mathrm{d}}{\mathrm{d}x}\sigma_\tau(x) = \tfrac{1}{\tau}p_Z(x/\tau) = \tfrac{e^{-\frac{x}{\tau}}}{\tau(1+e^{-x/\tau})^2} \tag{82}$$

as $\tau \to 0+$. Since p_Z is symmetric, we may assume $x > 0$ without loss of generality. The denominator is then asymptotically just τ. Therefore the ratio is asymptotically $O(\frac{1}{\tau}c^{1/\tau})$. □

For small x, where c is close to one, the term $1/\tau$ dominates at first. In particular for $z = \eta$, we get $c = 1$ and the asymptote is $O(1/\tau)$. So while for most of the noise realization the gradient magnitude quickly vanishes, it is compensated by a significant grows at rate $1/\tau$ around $z = \eta$. In practice it means, most of the time a value of gradient close to zero is measured and occasionally, very rarely, a value of $O(1/\tau)$ is obtained.

Since the gradient is asymptotically unbiased, such quick diminishing for any fixed ε has to be compensated by a rapid growth

Proposition E.3. The probability to observe the gradient of norm at least ε is asymptotically $O(\tau \log(\frac{1}{\varepsilon}))$, where the asymptote is $\tau \to 0$, $\varepsilon \to 0$.

Proof. We want to analyze the probability

$$P = \mathbb{P}(\tfrac{\mathrm{d}}{\mathrm{d}\eta}\sigma_\tau(\eta - z) \geq \varepsilon) \tag{83}$$

when z is distributed logistically. Let $s = \sigma_\tau(\eta - z)$. Then $\frac{\mathrm{d}}{\mathrm{d}\eta}\sigma_\tau(\eta - z) = s(1-s)$. The equality $s(1-s) = \varepsilon$ holds for $s^* = \frac{1}{2}(1 - \sqrt{1 - 4\varepsilon})$. This implies

$$z_{1,2} = \eta \pm \tau \mathrm{logit}(s^*). \tag{84}$$

The inequality $s(1-s) \geq \varepsilon$ holds in the interval $[z_1, z_2]$. Thus the probability in question is given by

$$P = F_Z(z_2) - F_Z(z_1). \tag{85}$$

As $\tau \mathrm{logit}(s^*) \to 0$ for $\tau \to 0$, we have asymptotically that

$$P = p_Z(\eta)\tau \mathrm{logit}(s^*). \tag{86}$$

Lastly note that $\mathrm{logit}(\frac{1}{2}(1 - \sqrt{1 - 4\varepsilon}))$ is asymptotically $O(-\log \varepsilon)$ for $\varepsilon \to 0$. □

Proposition E.4. The variance of GS estimator grows as $O(\frac{1}{\tau})$.

Proof. We first show that the second moment of the estimator G_τ has the following asymptotic expansion for $\tau \to 0+$:

$$p_Z(\eta)\Big(\int_0^1 \mathcal{L}'(v)^2 v(1-v)\Big)\tfrac{1}{\tau} + c_1\Big(\int_0^1 \mathcal{L}'(v)^2 v(1-v)\mathrm{logit}(v)\mathrm{d}v\Big) \tag{87}$$

$$+c_2\Big(\int_0^1 \mathcal{L}'(v)^2 v(1-v)\mathrm{logit}^2(v)\mathrm{d}v\Big)\tau + O(\tau^2). \tag{88}$$

The second moment expresses as

$$\mathbb{E}[G_\tau^2] = \int_{-\infty}^{\infty} (\tfrac{\partial}{\partial\eta}\mathcal{L}(\sigma_\tau(\eta - z)))^2 p_Z(z)\mathrm{d}z \tag{89}$$

$$= \int_{-\infty}^{\infty} \Big(\mathcal{L}'(\sigma_\tau(\eta - z))\sigma_\tau(\eta - z)(1 - \sigma_\tau(\eta - z))\tfrac{1}{\tau}\Big)^2 p_Z(z)dz \tag{90}$$

We perform the same substitution of variables: $v = \sigma_t(\eta - z)$, $dv = -v(1-v)\frac{1}{\tau}\mathrm{d}z$ to obtain

$$\mathbb{E}[G_\tau^2] = \int_0^1 \mathcal{L}'(v)v(1-v)\tfrac{1}{\tau}\mathcal{L}'(v)p_Z(\eta - \tau\mathrm{logit}(v))\mathrm{d}v \tag{91}$$

$$= \tfrac{1}{\tau}\int_0^1 \mathcal{L}'(v)^2 v(1-v)p_Z(\eta - \tau\mathrm{logit}(v))\mathrm{d}v. \tag{92}$$

We perform the same Taylor expansion for $p_Z(\eta - \tau\mathrm{logit}(v))$ around $\tau = 0$ as in Proposition E.1 and combine the terms to obtain the expansion as claimed. The variance is therefore dominated by the $O(\frac{1}{t})$ term of the second moment. □

Unlike ST, for linear objective $\mathcal{L}$, this estimator is biased. Furthermore, unlike ST, even for a single neuron it still has non-zero variance.

ST Gumbel-Softmax Considering that relaxed variables deviate from binary samples on the forward pass, Jang et al. (2016) also proposd the following empirical modification related to our main topic on ST methods. Their *ST Gumbel-Softmax* estimator replaces the gradient estimator with

$$\tfrac{\partial\mathcal{L}(x_1)}{\partial x_1}\tfrac{\partial y_1}{\partial\eta}, \tag{93}$$

where both x_1 and y_1 are computed using the same logistic noise z. Since $\frac{\partial\mathcal{L}(x_1)}{\partial x_1}$ does not vary with z unless x changes, to compute the expected gradient we can marginalize over z locally in the two regions: $z < \eta$ and $z > \eta$. Let's consider the first case. We have:

$$\int_{-\infty}^{\eta} \tfrac{1}{\tau}\sigma_\tau(\eta - z)(1 - \sigma_\tau(\eta - z))p_Z(z)\mathrm{d}z = \tag{94}$$

we make a substitution $z = \eta - \tau\mathrm{logit}(v)$. Then $v = \sigma_\tau(\eta - z)$ and $\mathrm{d}z = -\tau\frac{1}{v(1-v)}\mathrm{d}v$. The integral expresses as

$$-\int_1^{0.5} p_Z(\eta - \tau\mathrm{logit}(v))\mathrm{d}v. \tag{95}$$

The limit of this integral with $\tau \to 0$ is $\frac{1}{2}p_Z(\eta)$. And the same holds for the case $x_1 = 0$, *i.e.*, $z > \eta$. What we obtained, is that ST Gumbel-Softmax in the simple case of linear objective and in the favorable limit $\tau \to 0$ underestimates the true gradient by $1/2$.

To summarize, ST Gumbel-Softmax is more expensive as it involves sampling from Logistic distribution and keeping the samples, it is biased, even asymptotically as $\tau \to 0$ and even in the case of linear objective. It is also more noisy than ST as the gradient depends on the value of z and not only the binary state. In particular for one neuron it is still stochastic while ST becomes deterministic.

F Analysis of BayesBiNN

BayesBiNN Meng et al. (2020) algorithm is derived from the same variational Bayesian learning problem formulation that we consider § C.2. The update step is also the same as they differentiate the entropy regularization in closed form similarly to our composite MD. The only essential difference is the use of Gumble-Softmax estimator to estimate the gradient in the probabilities θ used to update the natural parameters η. We will show that the Gumbel-Softmax estimator is used incorrectly, which leads to some surprising algorithm behavior.

First, it is easy to see that the properties of Gumbel-Softmax estimator § E for gradient in η apply as well to the gradient in θ. Indeed, for Bernoulli distribution we have

$$\frac{\partial \eta}{\partial \theta} = \left(\frac{\partial \theta}{\partial \eta}\right)^{-1} = \mathrm{diag}(p_Z(\theta))^{-1}. \tag{96}$$

This link between the two gradients does not depend on the relaxation parameter τ and therefore the asymptotic properties § E apply to the Gumbel-Softmax gradient in θ as well. With parameter $\tau = 10^{-10}$, the probability to measure a gradient larger than a numerical precision scales as $O(\tau)$ and practically diminishes. Even if such a rare event occurs the value of the measured gradient scales as $O(1/\tau)$. The experiments Meng et al. (2020) could not be possibly successful without a technical detail we discuss next.

The BayesBiNN algorithm (Meng et al., 2020, Table 1 middle) performs the steps (in our notation):

$$w_b := \tanh(\tfrac{\eta - Z}{\tau}); \tag{97a}$$

$$g := \nabla_{w_b} \mathcal{L}; \tag{97b}$$

$$\eta := (1-\alpha)\eta - \alpha s \odot g, \tag{97c}$$

where g denotes the gradient of the average min-batch loss $\mathcal{L}$, which is evaluated using softly binarized weights w_b, $\odot$ is a component-wise product and s is a scaling factor originating from Gumbel-Softmax estimator and is discussed below.

The actual scaling factor s used in the experiments Meng et al. (2020) according to the published code adds a technical ε in (Meng et al., 2020, Eq. (9)) in the implementation of Gumbel-Softmax estimator, presumably for a numerical stability. Coincidentally, $\varepsilon = 10^{-10}$ is used. The resulting scaling factor becomes:

$$s = \tfrac{N(1-w_b^2+\varepsilon)}{\tau(1-\tanh(\eta)^2+\varepsilon)}, \tag{98}$$

where N is the size of the complete training set.

Proposition F.1. With the setting of the hyper-parameters for τ, ε (Meng et al., 2020, Table 7) in large-scale experiments (MNIST, CIFAR10, CIFAR100), the BayesBiNN algorithm is equivalent to SGD with deterministic identity straight-through and latent weight decay.

Proof. For simplicity we will treat η and g as scalars, for the general case the arguments would apply coordinate-wise.

First, we analyze the nominator of (98). From the asymptotic expansion of

$$1 - \tanh[\log(x)]^2 = \tfrac{4}{x^2} + O(\tfrac{1}{x^3}), \text{for } x \to \infty, \tag{99}$$

substituting $\log(x) = \frac{|\eta - Z|}{\tau}$, we obtain

$$1 - w_b^2 \sim 4 \exp\big(-2\tfrac{|\eta - Z|}{\tau}\big). \tag{100}$$

For example, for $\eta = 5$ and $Z = 0$ we have that $(1 - w_b^2) \approx 4e^{-10^{11}}$. Therefore in the nominator the part $(1 - w_b^2)$ is negligibly small compared to ε and even the floating point precision. This applies so long as $|\eta - Z| \gg \tau = 10^{-10}$, which we expect to hold with high probability for two reasons: 1) η will be shown to grow significantly during the first iterates and 2) the probability of the noise matching η to this accuracy even for $\eta = 0$ is of the order $O(\tau)$.

The denominator of (98) satisfies the bounds

$$\tau\varepsilon \le \tau(1 - \tanh(\eta)^2 + \varepsilon) \le \tau(1 + \varepsilon), \tag{101}$$

from which we can conclude that $s \ge \frac{N\varepsilon}{\tau} = N$. However for a moderately large η the denominator drops quickly, *e.g.* for $\eta = 5$, we have $\tau(1 - \tanh(\eta)^2 + \varepsilon) < \tau(2 \cdot 10^{-4} + \varepsilon) \approx 2 \cdot 10^{-4}\tau$. And the asymptote for $|\eta| \to \infty$ is $\tau(4e^{-2|\eta|} + \varepsilon)$.

Since η is initialized uniformly in $[-10, 10]$ and receives updates of order at least $\alpha N g \approx 5g$ (for the initial $\alpha = 10^{-4}$ used), during the first iterates $|\eta|$ can be expected to grow significantly until we reach the asymptote $e^{-2|\eta|} \ll \varepsilon$, which is when $|\eta| > 5 \log 10 \approx 11$. After reaching this

asymptote, we will have $s \approx \frac{N\varepsilon}{\tau\varepsilon} = \frac{N}{\tau}$ and we may expect the growth of η to stabilize around $|\eta| \approx \alpha\frac{N}{\tau}|g| \approx 10^{10}$.

The first consequence of this is that the scaling factor that was supposed to implement Gumbel-Softmax gradient, just becomes the constant $\frac{N}{\tau}$. In particular an inadvertent factor $\frac{1}{\tau}$ occurs. It is tempting to conclude that this up-scaling of the gradient corresponds to solving the variational Bayesian problem with the data evidence part up-scaled respectively, *i.e.*, completely dominating the the KL prior part, but this is not exactly so because also the true stochastic scaling of the gradient is modified.

The second consequence is that the natural parameters η have huge magnitudes during the training, and we have that $|Z| \ll |\eta|$ with high probability, therefore the noise plays practically no role even in the forward pass of BayesBiNN. In this mode the BayesBiNN algorithm becomes equivalent to

$$w_b := \mathrm{sign}(\eta); \tag{102a}$$
$$g := \nabla_{w_b}\mathcal{L}; \tag{102b}$$
$$\eta := (1-\alpha)\eta - \alpha\frac{N}{\tau}g. \tag{102c}$$

It is seen that the forward pass and the gradient implement the deterministic straight-through with identity derivative and that the update has a form of SGD with a latent weight decay and with the gradient of data evidence up-scaled by $\frac{1}{\tau}$. These huge step-sizes do not destroy the learning since sign is invariant to a global rescaling of η.

□

Proposition F.2. The result of the algorithm (102) does not depend on the values of τ and N.

Proof. Denoting $\bar{\eta} = \frac{\tau}{N}\eta$, we can equivalently rewrite (102) as

$$w_b := \mathrm{sign}(\bar{\eta}); \tag{103a}$$
$$g := \nabla_{w_b}\mathcal{L}; \tag{103b}$$
$$\bar{\eta} := (1-\alpha)\bar{\eta} - \alpha g. \tag{103c}$$

This algorithm and the resulting binary weights w_b do not depend on τ, N. □

This is perhaps somewhat unexpected, but it makes sense indeed. The initial BayesBiNN algorithm of course depends on N and τ. However due to the issue with the implementation of Gumbel Softmax estimator for sufficiently small value of τ it falls into a regime which is significantly different from the Bayesian learning rule and is instead accurately described by (102). In this regime, it produces the result not dependent on the particular values of τ and N. While we do not know what problem it is solving in the end, it is certainly not solving the variational Bayesian learning problem. since this problem and its solution do depend on N in a critical way. The algorithm (103a) indeed does not solve any variational problem as there is no variational distribution involved. Yet the decay term $-\alpha\bar{\eta}$ stays effective, if the data gradient becomes small, the decay term implements some small "forgetting" of the learned information and may be responsible for improved generalization observed in experiments (Meng et al., 2020). However there are other differences to the baselines, *e.g.* SGD vs. Adam that can be also contributing.