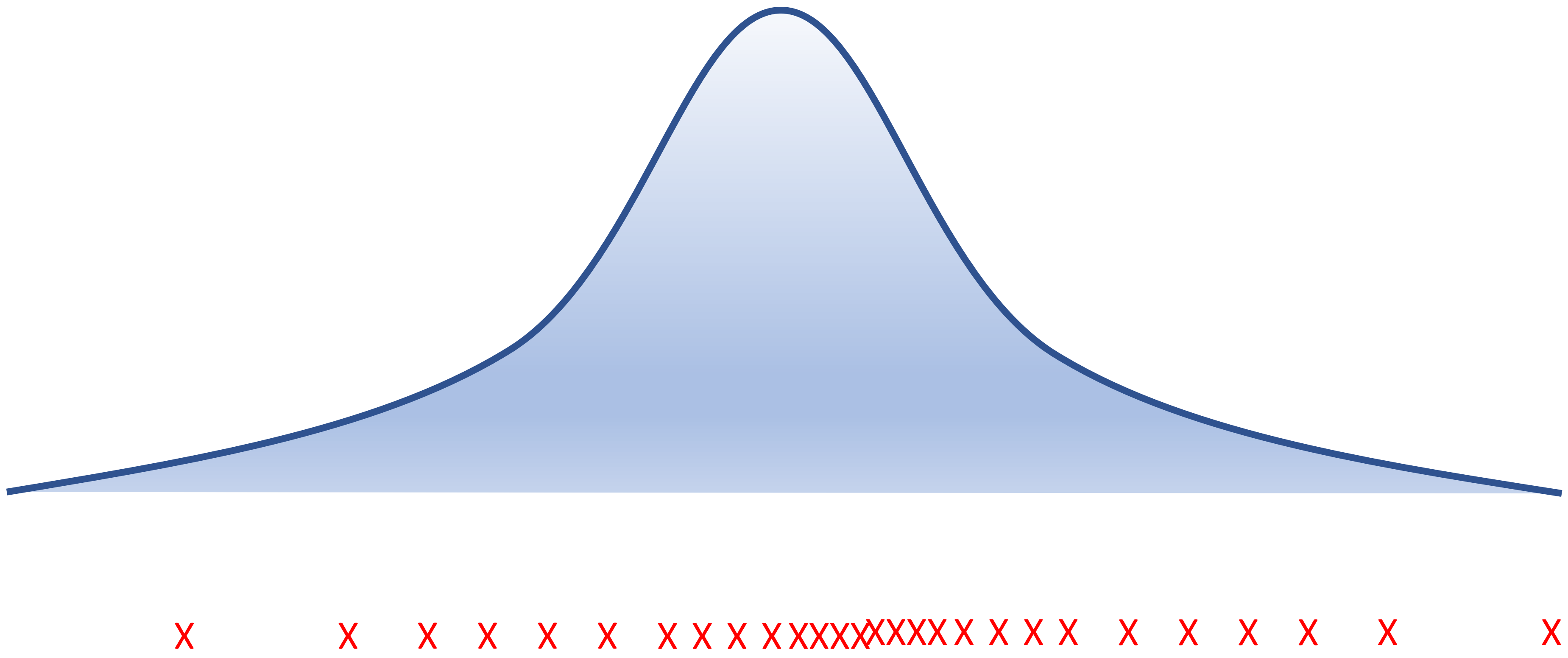
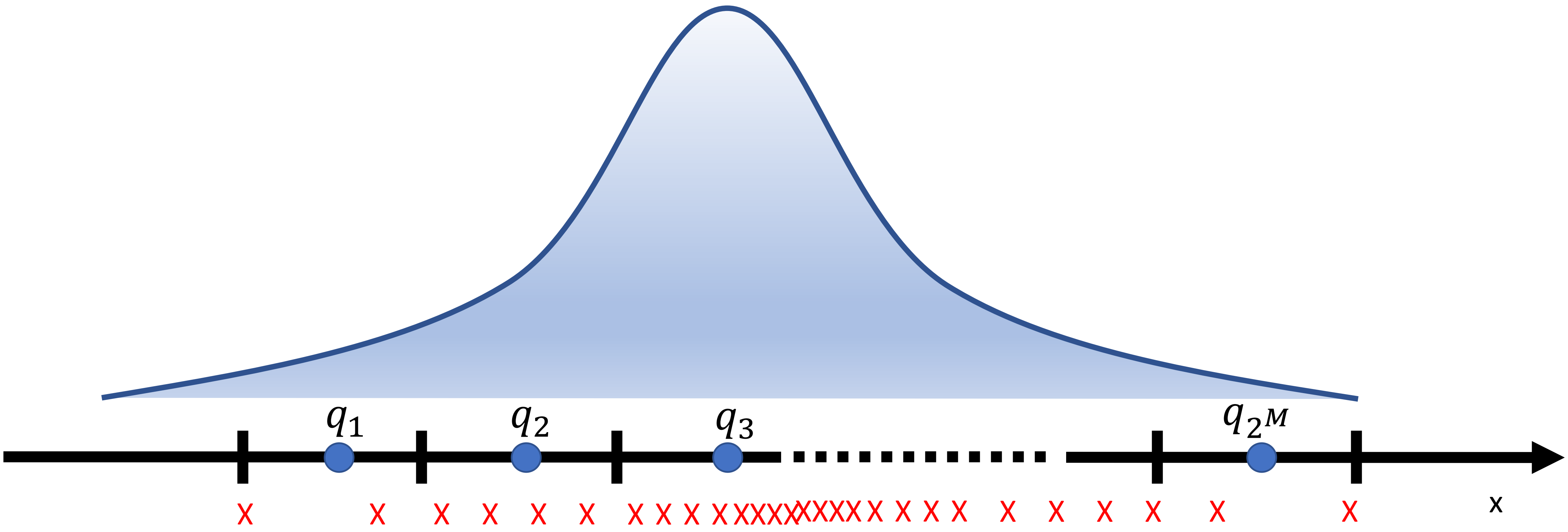
Post training 4-bit quantization of convolutional networks for rapid-deployment

Ron Banner, Yury Nahshan and Daniel Soudry

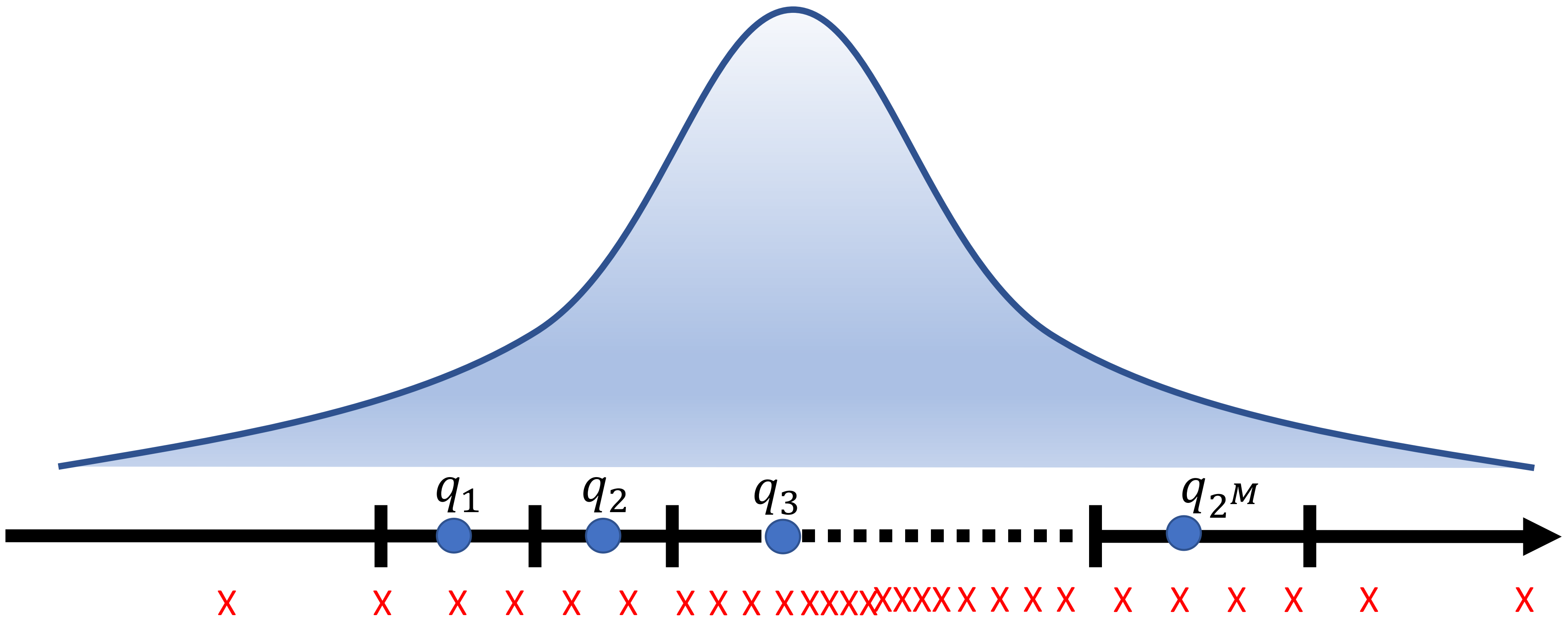
Pre-activation tensor



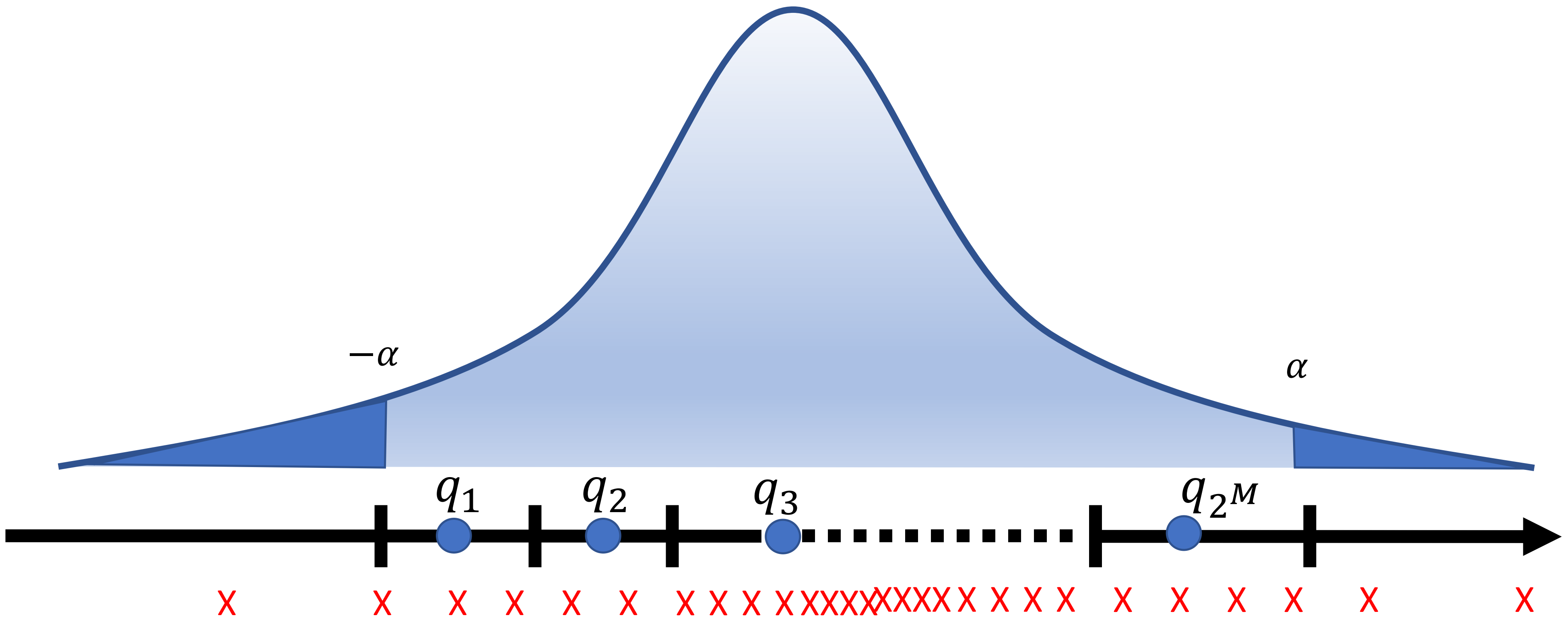
Pre-activation tensor



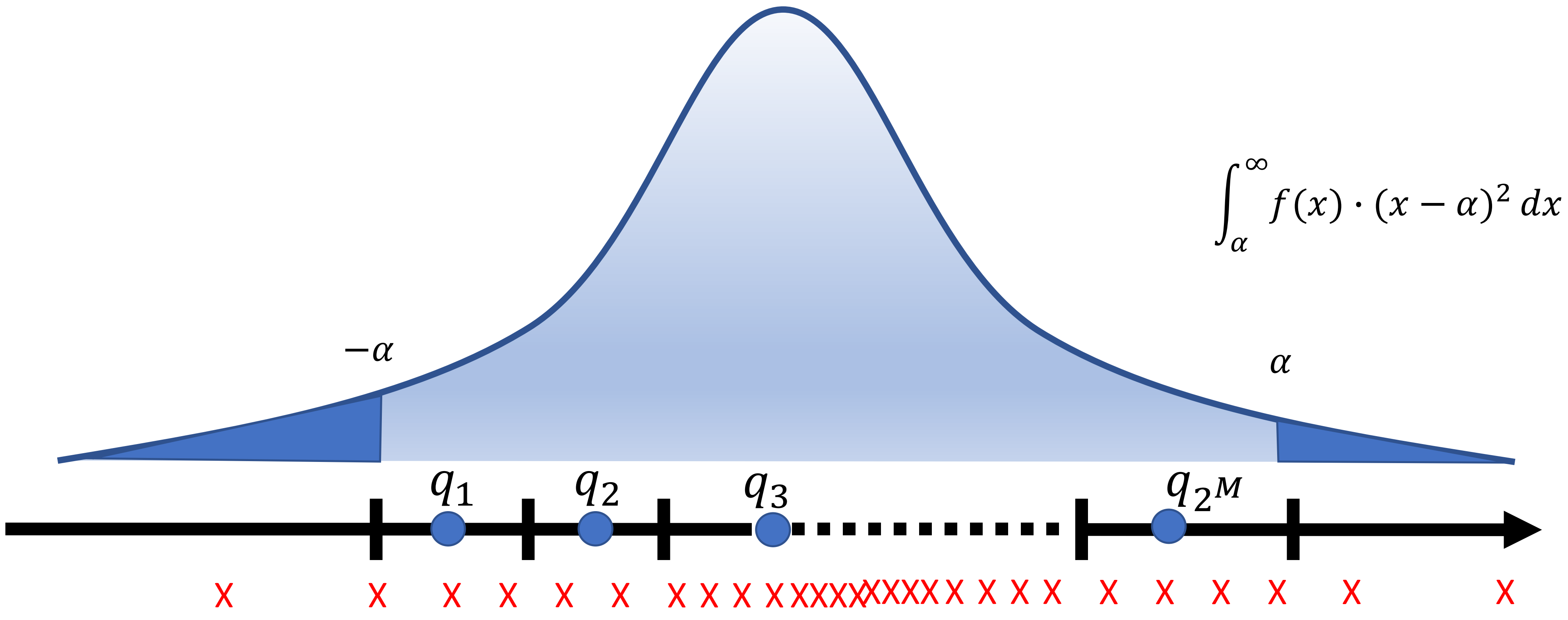
Pre-activation tensor



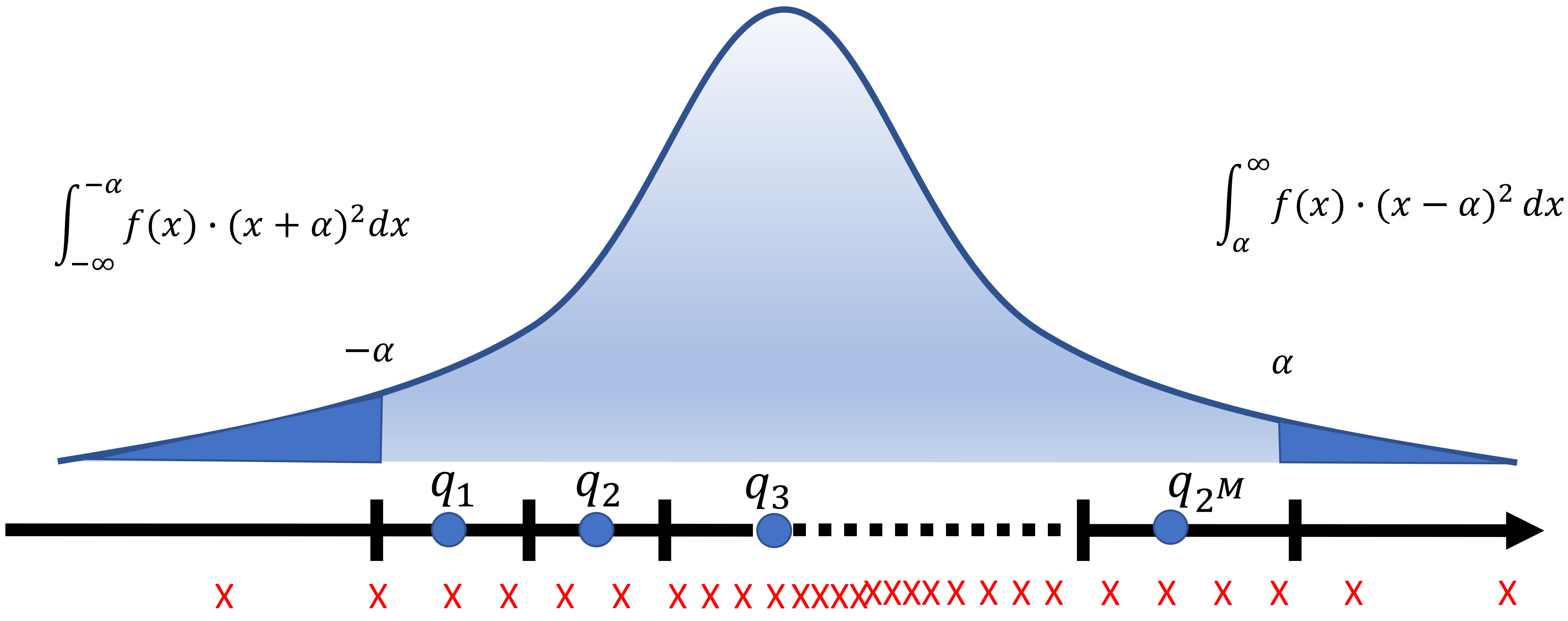
Pre-activation tensor



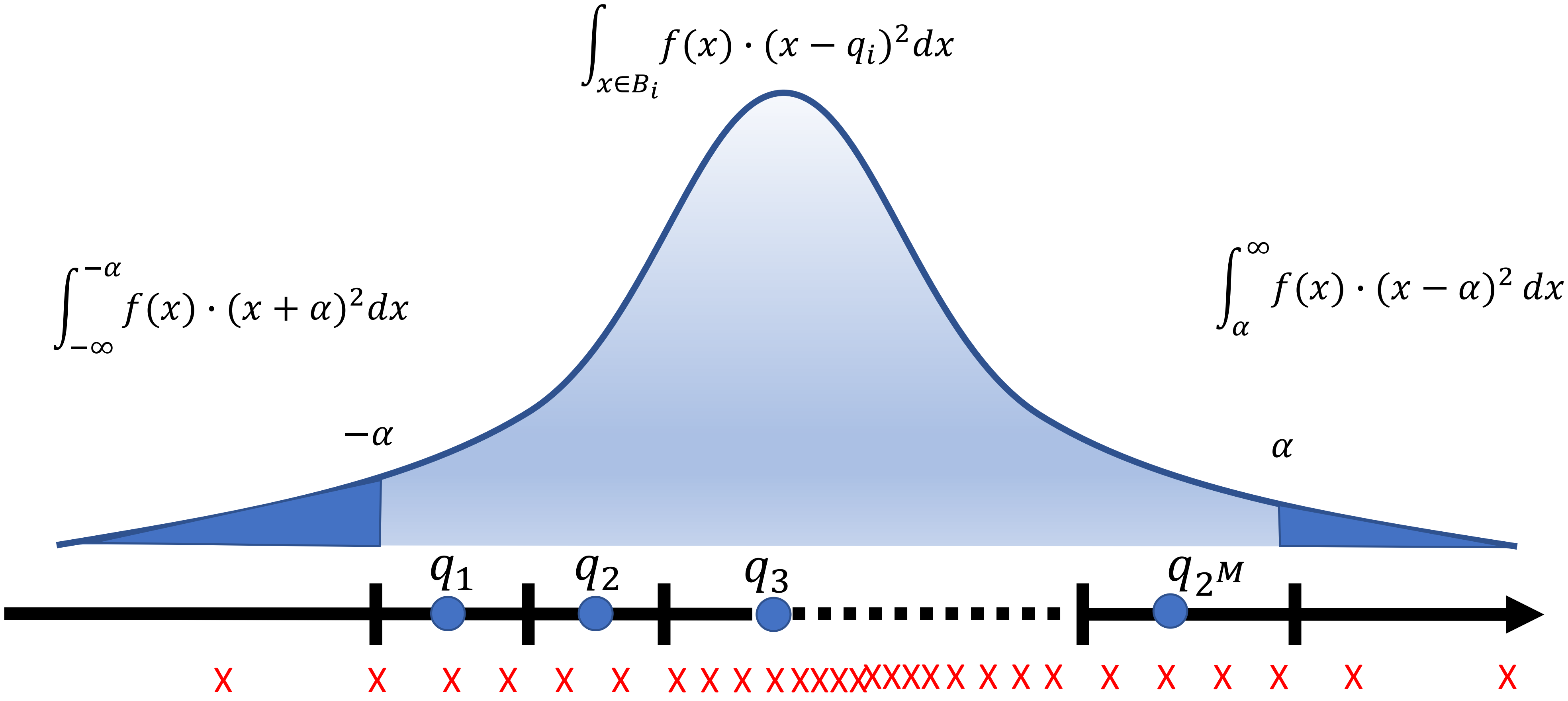
Pre-activation tensor



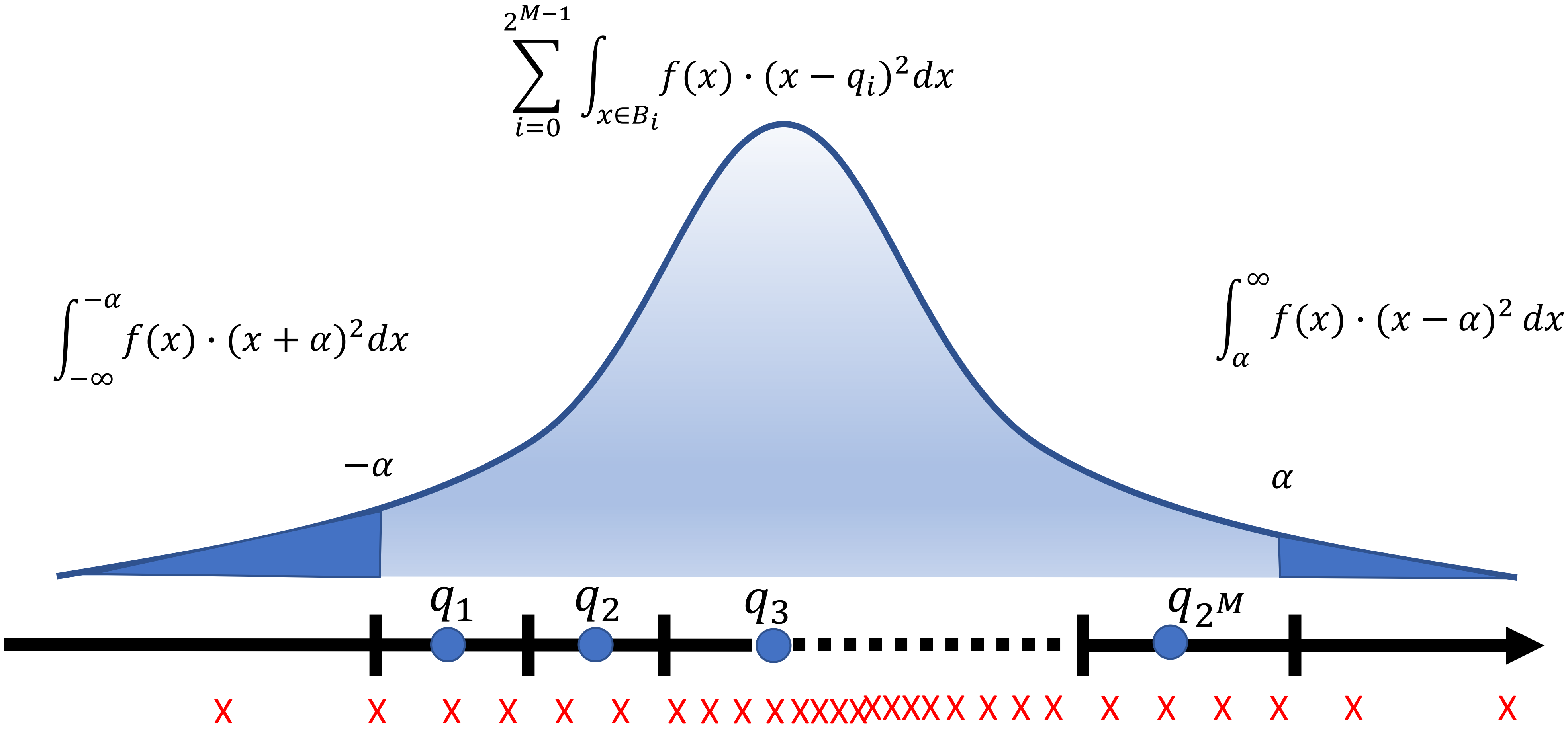
Pre-activation tensor



Pre-activation tensor

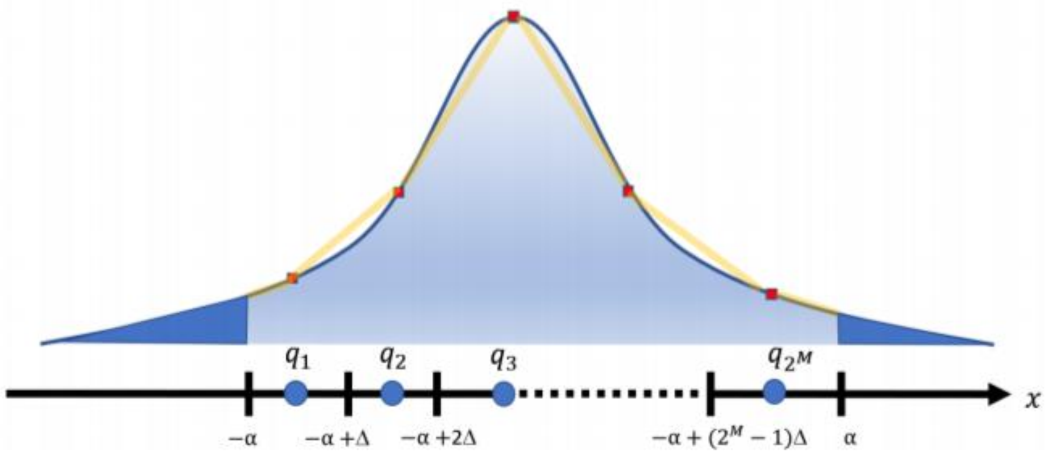


Pre-activation tensor



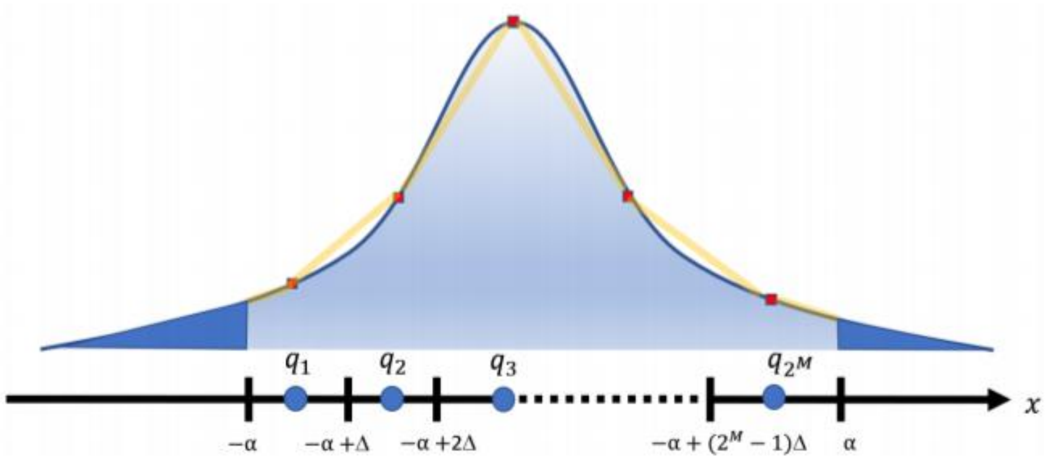
Optimal clipping value

$$E[(X - Q(X))^2] =$$
$$= \int_{-\infty}^{-\alpha} f(x) \cdot (x + \alpha)^2 dx +$$
$$+ \sum_{i=0}^{2^M - 1} \int_{-\alpha + i \cdot \Delta}^{-\alpha + (i+1) \cdot \Delta} f(x) \cdot (x - q_i)^2 dx +$$
$$+ \int_{\alpha}^{\infty} f(x) \cdot (x - \alpha)^2 dx$$



Optimal clipping value

$$
\begin{aligned}
&E[(X - Q(X))^2] \approx \\
&\approx (\alpha^2 + \sigma^2) \cdot \left[1 - \operatorname{erf}\left(\frac{\alpha}{\sqrt{2}\sigma}\right)\right] + \\
&+ \frac{\alpha^2}{3 \cdot 2^{2M}} - \frac{\sqrt{2}\alpha \cdot \sigma \cdot \mathrm{e}^{-\frac{\alpha^2}{2 \cdot \sigma^2}}}{\sqrt{\pi}}
\end{aligned}
$$

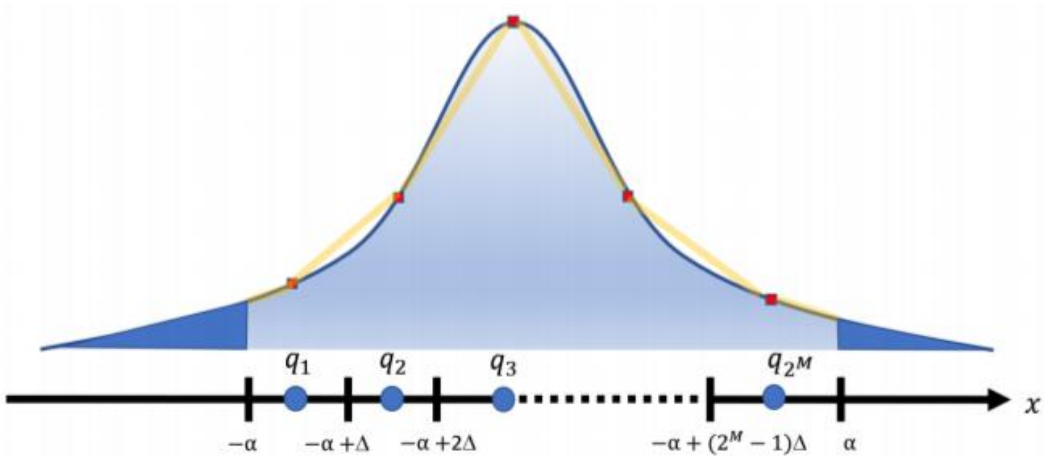


Optimal clipping value

$$\frac{\partial E[(X-Q(X))^2]}{\partial \alpha} =$$

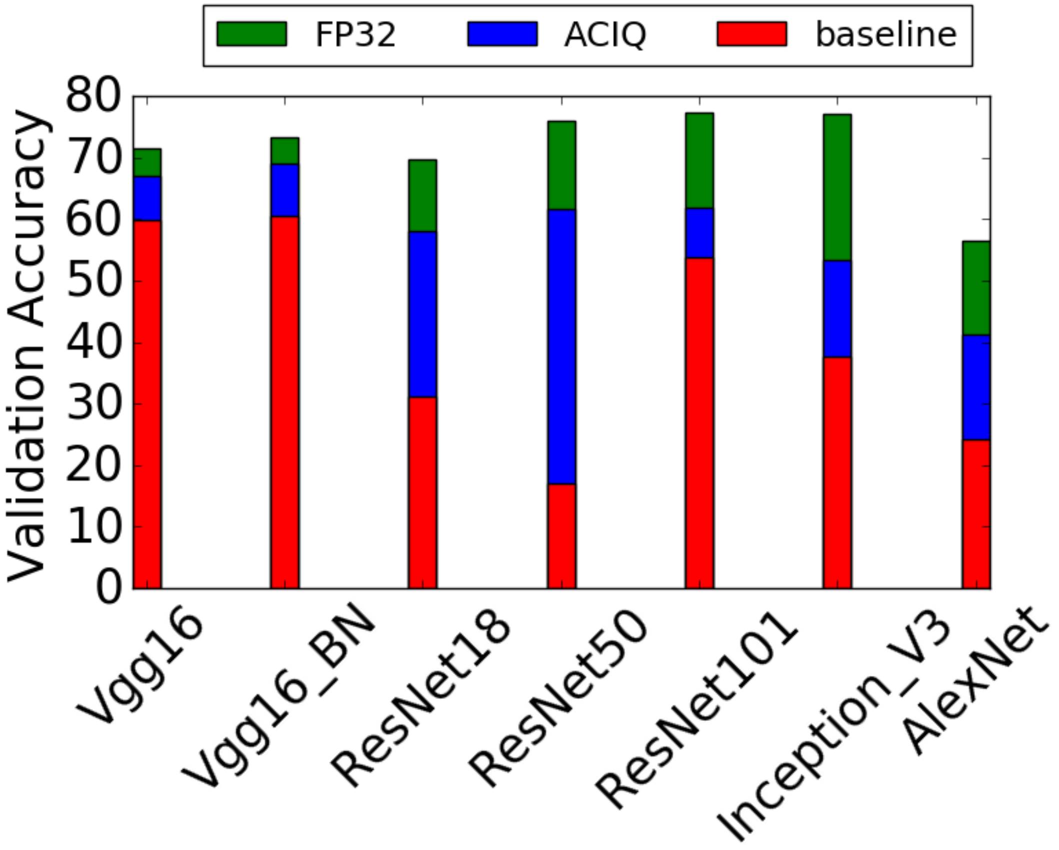
$$= \alpha\left[1-\operatorname{erf}\left(\frac{\alpha}{\sqrt{2}\sigma}\right)\right] - \frac{\sigma^2 \mathrm{e}^{-\frac{\alpha^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} - \frac{\sigma \mathrm{e}^{-\frac{\alpha^2}{2\sigma^2}}}{\sqrt{2\pi}} +$$

$$+ \frac{2\alpha}{3\cdot 2^{2M}} = 0$$



4-bit quantization			
Model	Reference (FP32)	Ours (Optimal Clip)	GEMMLOWP (Max/Min)
VGG16	71.59%	70.1%	68.8%
VGG16-BN	73.36%	72.0%	70.6%
ResNet18	69.75%	66.6%	61.5%
ResNet50	76.1%	71.8%	68.3%
ResNet101	77.3%	72.6%	66.5%
Inception V3	77.2%	72.7%	70.9%

3-bit quantization



Per-Channel Bit allocation

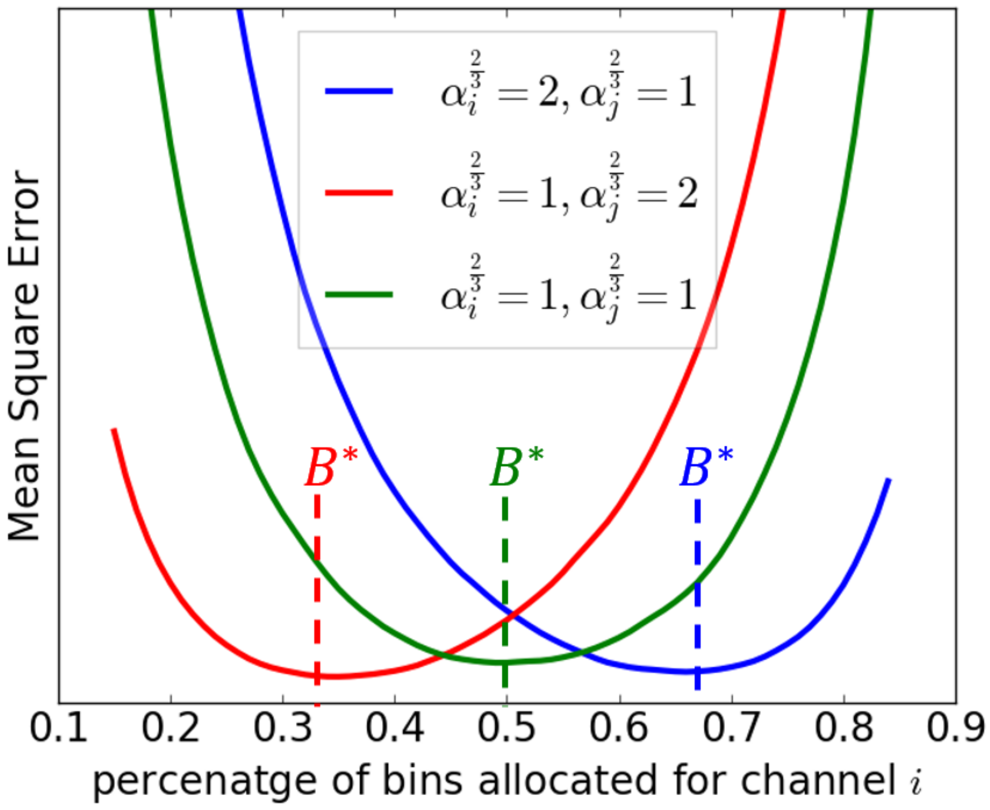
Problem:

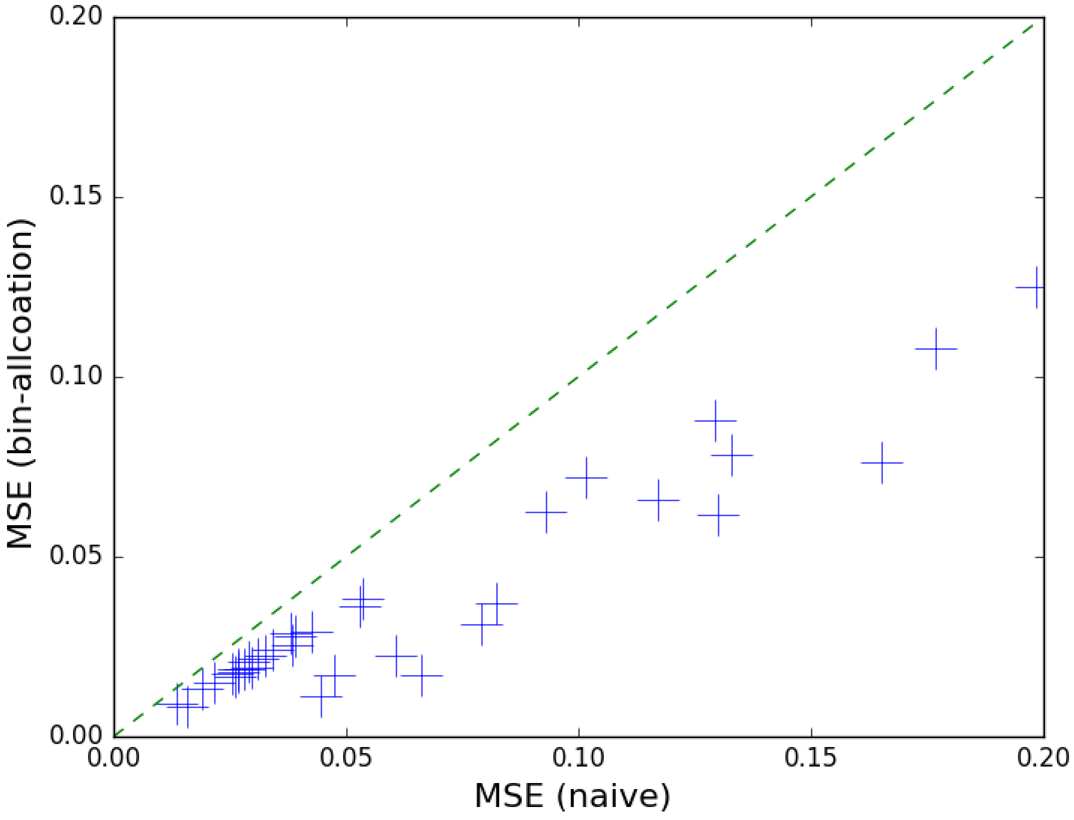
Given a maximum bit-budget B, how many bits M_i should we allocate to each channel in order to minimize the layer mean-square-error ?

Answer:

$$M_i = \left\lfloor \log_2 \left(\frac{\alpha_i^{\frac{2}{3}}}{\sum_i \alpha_i^{\frac{2}{3}}} \cdot B \right) \right\rceil$$

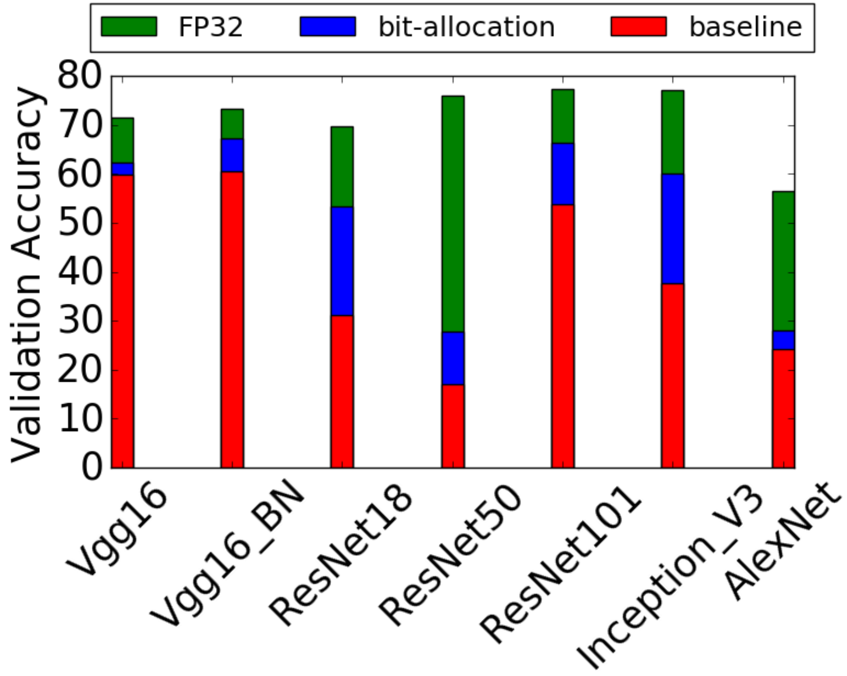
Per-Channel Bit allocation





4-bit quantization Activation				4-bit quantization Weights			
Model	Reference (FP32)	Ours (Bit-allocation)	Naive	Model	Reference (FP32)	Ours (Bit-allocation)	Naive
VGG16	71.59%	69.7%	68.8%	VGG16	71.59%	71.0%	70.5%
VGG16-BN	73.36%	72.6%	70.6%	VGG16-BN	73.36%	71.9%	68.5%
ResNet18	69.75%	65.0%	61.5%	ResNet18	69.75%	66.7%	59.7%
ResNet50	76.1%	71.3%	68.3%	ResNet50	76.1%	75.0%	72.5%
ResNet101	77.3%	70.8%	66.5%	ResNet101	77.3%	76.4%	74.6%
Inception V3	77.2%	74.3%	70.9%	Inception V3	77.2%	61.4%	38.4%

3-bit quantization Activations



Bias-Correction

Problem:

We observe an inherent bias in the mean of the weight values following their quantization

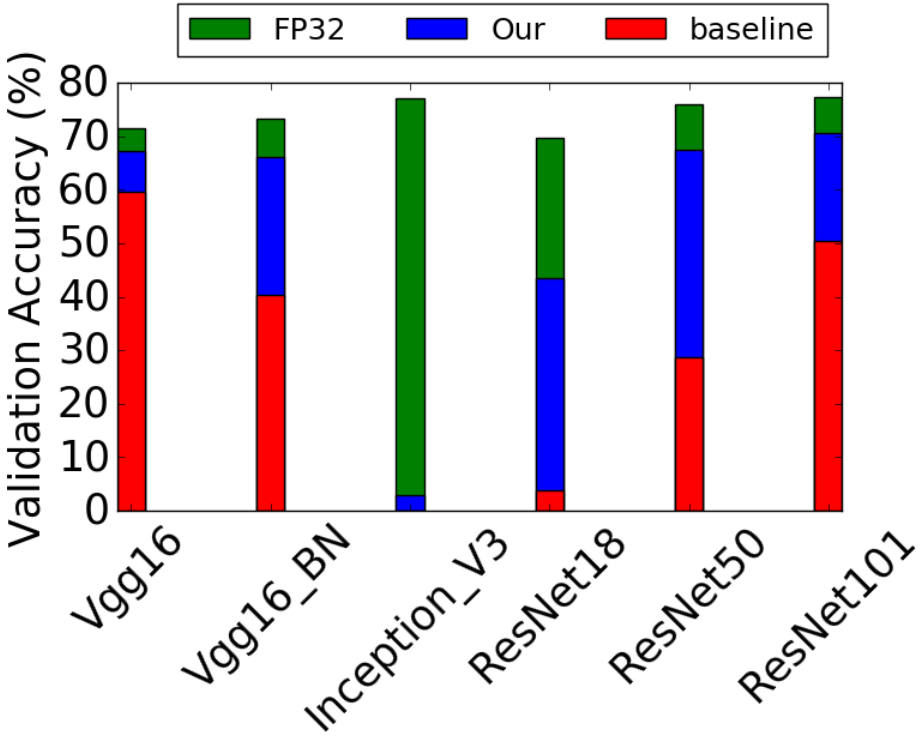
Solution:

$$\mu_c = \mathbb{E}\left(W_c\right) - \mathbb{E}\left(W_c^q\right)$$
$$\xi_c = \frac{||W_c - \mathbb{E}\left(W_c\right)||_2}{||W_c^q - \mathbb{E}\left(W_c^q\right)||_2}$$

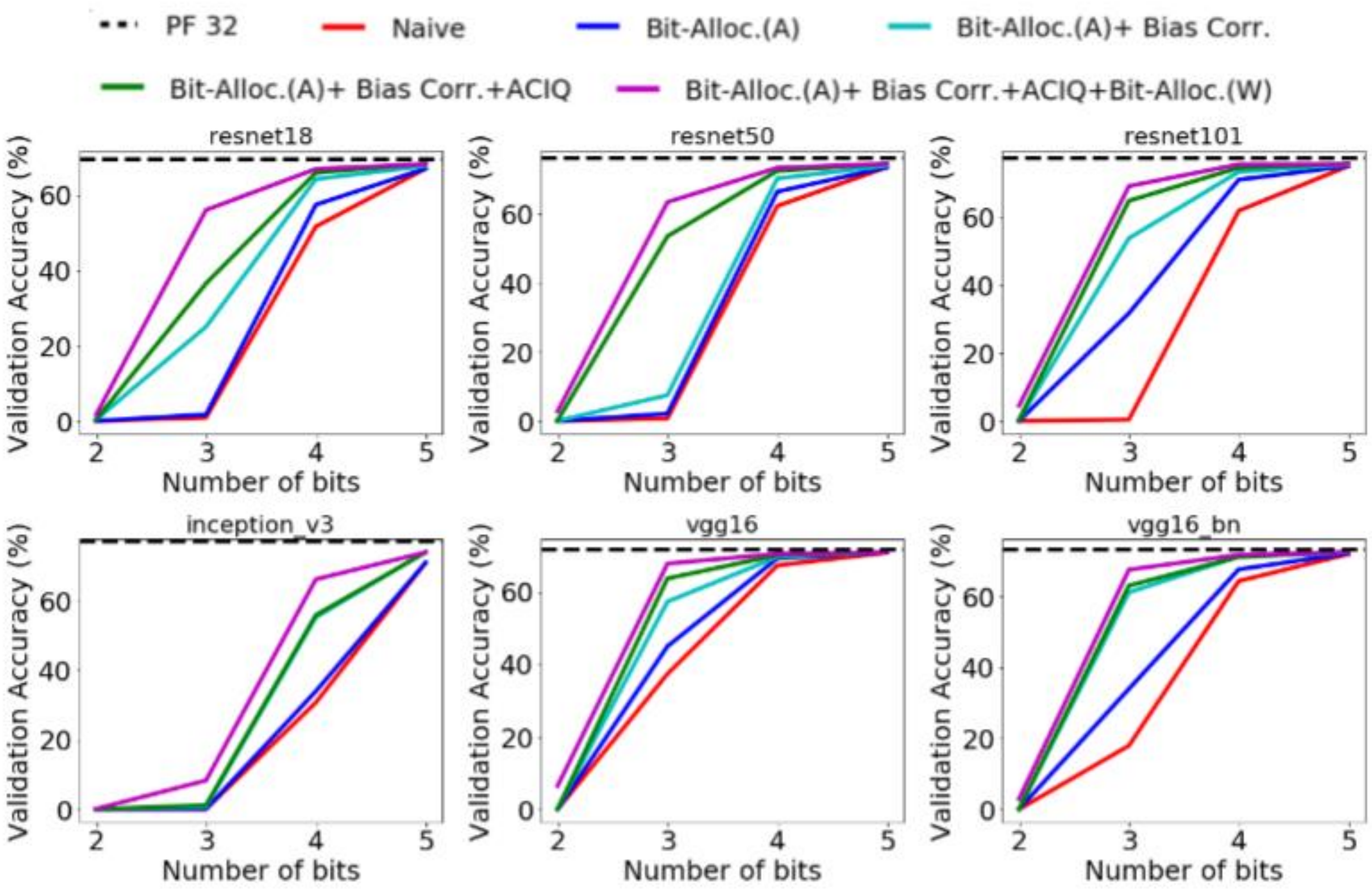
$$w \longleftarrow \xi_c \left(w + \mu_c\right), \quad \forall w \in W_c^q$$

4-bit quantization Weights			
Model	Reference (FP32)	Ours (Bias-correction)	Naive
VGG16	71.59%	71.0%	70.5%
VGG16-BN	73.36%	71.7%	68.5%
ResNet18	69.75%	67.4%	59.7%
ResNet50	76.1%	74.8%	72.5%
ResNet101	77.3%	76.3%	74.6%
Inception V3	77.2%	59.5%	38.4%

3-bit quantization



All method combined



All methods combined

- how much can we get without hurting 8-bit baseline?

No re-training	
Model	Avg number of bits per value (weights and activations)
VGG16	**5.3 bits**
VGG16-BN	**5.4 bits**
ResNet18	**5.4 bits**
ResNet50	**5.7 bits**
ResNet101	**5.8 bits**
Inception V3	**6.1 bits**

All methods combined

- how much can we get without hurting 8-bit baseline?
- Unless we support variable-length coding, we need to support different bit-width per channel

No re-training	
Model	Avg number of bits per value (weights and activations)
VGG16	**5.3 bits**
VGG16-BN	**5.4 bits**
ResNet18	**5.4 bits**
ResNet50	**5.7 bits**
ResNet101	**5.8 bits**
Inception V3	**6.1 bits**

All methods combined

- how much can we get without hurting 8-bit baseline?
- Unless we support variable-length coding, we need to support different bit-width per channel
- Variable-length coding can further improve these results …

No re-training	
Model	Avg number of bits per value (weights and activations)
VGG16	**5.3 bits**
VGG16-BN	**5.4 bits**
ResNet18	**5.4 bits**
ResNet50	**5.7 bits**
ResNet101	**5.8 bits**
Inception V3	**6.1 bits**

How much can we get without hurting 8-bit baseline?

No re-training		
Model	Avg number of bits per value (without VLC)	Avg number of bits per value (with VLC)
VGG16	**5.3 bits**	**2.2 bits**
VGG16-BN	**5.4 bits**	**2.4 bits**
ResNet18	**5.4 bits**	**3.7 bits**
ResNet50	**5.7 bits**	**4.1 bits**
ResNet101	**5.8 bits**	**4.4 bits**
Inception V3	**6.1 bits**	**3.1 bits**

End