Learning with Marginalized Corrupted Features

Laurens van der Maaten
Minmin Chen
Stephen Tyree
Kilian Weinberger

Classification of text and image data

Classify....

Classification of text and image data

Classify....

Man Googles Matt Damon's Address Because, Well, He's Crazy And Wants To Murder Him

NOVEMBER 8, 2012 | ISSUE 48•45 | MORE NEWS

The 29-year-old Easter, who stared at this picture of Matt Damon for two hours because, well, he's mentally ill.

SALISBURY, MARYLAND—After rereading actor Matt Damon's Wikipedia page for the 13th time since 9 a.m. today, local man Dan Easter decided to look up the celebrity's home address on Google because, well, he's admittedly crazy and wants to murder him.

Saying he planned to "just click around" a couple websites to see if the *Bourne Identity* star's address was listed anywhere on the Internet, Easter told reporters that, you know, he's ultimately a mentally ill madman who wants to break into Matt Damon's house in the dead of night and, you guessed it, kill him in front of his wife and children.

"I figured I would just type Matt Damon's name into Google because, to make a long story short, I'm psychologically disturbed and I want to assassinate him," said the 29-year-old man, who, by his own admission, is extremely unstable and has absolutely no business being anywhere other than a mental institution. "I see him in movies and magazines all the time, and it made me wonder where he lives. I'm also clinically insane. That's why

ARTICLE TOOLS

Tweet 120

Like 954

+1 16

tumblr +

RELATED ARTICLES

Fox Voluntarily Removes Reality From Programming

TV Viewers Outraged At Timing Of Commercial Break

... documents by topic

Classification of text and image data

Classify....

Man Googles Matt Damon's Address Because, Well, He's Crazy And Wants To Murder Him

NOVEMBER 8, 2012 | ISSUE 48•45 | MORE NEWS

The 29-year-old Easter, who stared at this picture of Matt Damon for two hours b

SALISBURY, MARYLAND—After rereading actor Matt Damon'
page for the 13th time since 9 a.m. today, local man Dan East
look up the celebrity's home address on Google because, well
admittedly crazy and wants to murder him.

Saying he planned to "just click around" a couple websites to s
Bourne Identity star's address was listed anywhere on the Inte
told reporters that, you know, he's ultimately a mentally ill mad
wants to break into Matt Damon's house in the dead of night a
guessed it, kill him in front of his wife and children.

"I figured I would just type Matt Damon's name into Google be
make a long story short, I'm psychologically disturbed and I wa
assassinate him," said the 29-year-old man, who, by his own a
extremely unstable and has absolutely no business being anyw
than a mental institution. "I see him in movies and magazines
and it made me wonder where he lives. I'm also clinically insar

... documents
by topic

Kindle vs. Nook vs. iPad: Which e-book reader should you buy?

With ultraaffordable e-ink readers, midprice color tablets like the Nexus 7 and Kindle Fire, and even the more expensive iPads all vying for your e-book dollar, what's the best choice for you? It depends.

by John P. Falcone | August 7, 2012 1:08 PM PDT

Follow

6.3K 1.5K Pin it 200 +1 282 More + Comments 573

(Credit: Sarah Tew/CNET)

Editors' note, September 7, 2012: As of September 6, Amazon has announced all-new Kindle e-readers and tablets for 2012 that dramatically offer the buying decisions listed below. The first wave of the new Amazon products are due to ship by September 14. We'll update this story in detail after we review those models. By that time, we'll also find out what Apple is announcing at its September 12 event. In the meantime, the

... documents
by sentiment

Classification of text and image data

Classify....

Man Googles Matt Damon's Address Because, Well, He's Crazy And Wants To Murder Him

NOVEMBER 8, 2012 | ISSUE 48•45 | MORE NEWS

The 29-year-old Easter, who stared at this picture of Matt Damon for two hours b

SALISBURY, MARYLAND—After rereading actor Matt Damon' page for the 13th time since 9 a.m. today, local man Dan East look up the celebrity's home address on Google because, well admittedly crazy and wants to murder him.

Saying he planned to "just click around" a couple websites to s *Bourne Identity* star's address was listed anywhere on the Inte told reporters that, you know, he's ultimately a mentally ill mad wants to break into Matt Damon's house in the dead of night a guessed it, kill him in front of his wife and children.

"I figured I would just type Matt Damon's name into Google be make a long story short, I'm psychologically disturbed and I wa assassinate him," said the 29-year-old man, who, by his own a extremely unstable and has absolutely no business being any than a mental institution. "I see him in movies and magazines and it made me wonder where he lives. I'm also clinically insar

... documents by topic

Kindle vs. Nook vs. iPad: Which e-book reader should you buy?

With ultraaffordable e-ink readers, midprice color tablets like the Nexus 7 and Kindle Fire, and even the more expensive iPads all vying for your e-book dollar, what's the best choice for you? It depends.

by John P. Falcone | August 7, 2012 1:08 PM PDT

Follow

6.3K 1.5K 200 +1 282 More + Comments 573

(Credit: Sarah Tew/CNET)

Editors' note, September 7, 2012: As of September 6, Amazon has announced all-new Kindle e-readers and tablets for 2012 that dramatically offer the buying decisions listed below. The first wave of the new Amazon products are due to ship by September 14. We'll update this story in detail after we review those models. By that time, we'll also find out what Apple is announcing at its September 12 event. In the meantime, the

... images by object

... documents by sentiment

Empirical risk minimization

- Learn model based on annotated data $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ by minimizing:

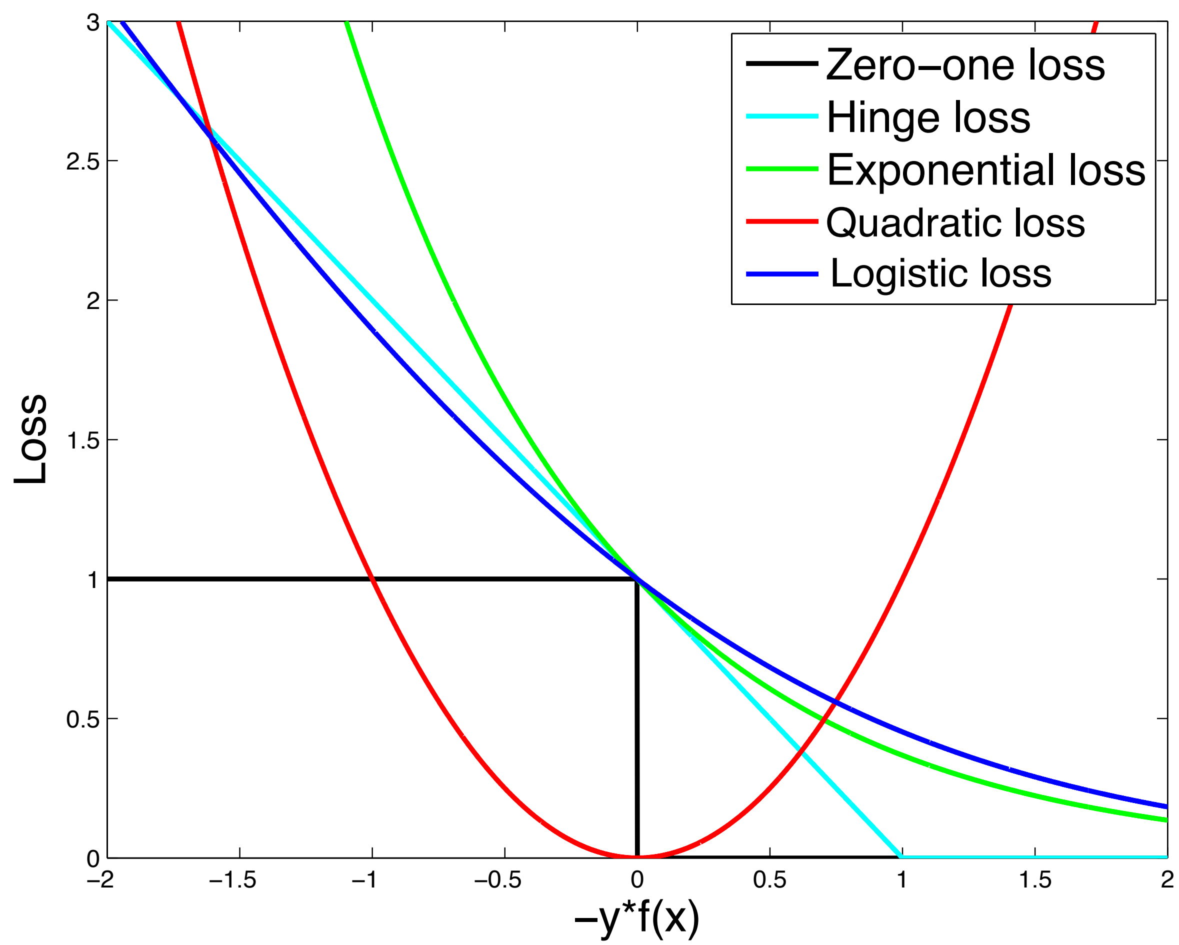
$$\min_{\Theta} \mathcal{L}(\mathcal{D}; \Theta) = \sum_{n=1}^{N} L(\mathbf{x}_n, y_n; \Theta)$$

Empirical risk minimization

- Learn model based on annotated data $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ by minimizing:

$$\min_{\Theta} \mathcal{L}(\mathcal{D}; \Theta) = \sum_{n=1}^{N} L(\mathbf{x}_n, y_n; \Theta)$$

- Herein, typical examples of the loss function include:

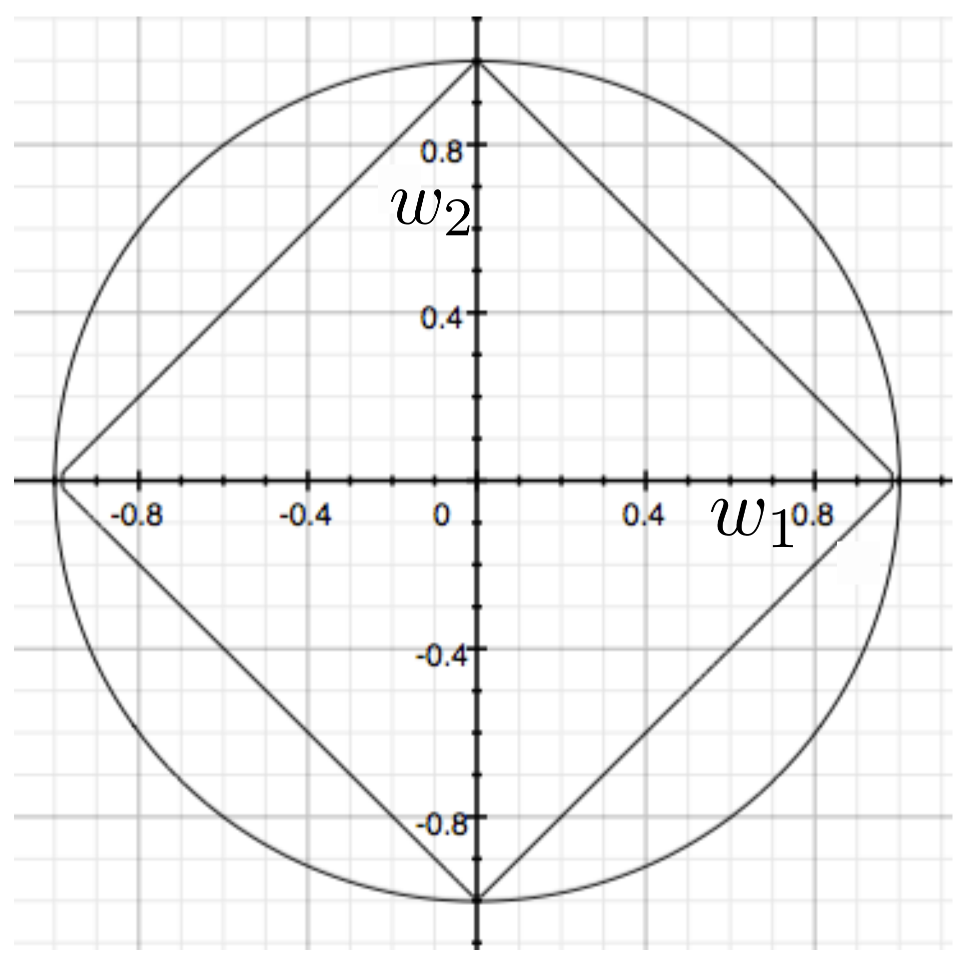


Regularization by priors

- *Regularizers* incorporate a term in the loss that penalizes complex models:

$$\tilde{\mathcal{L}}(\mathbf{x}, y; \mathbf{w}) = \mathcal{L}(\mathbf{x}, y; \mathbf{w}) + \lambda \mathcal{R}(\mathbf{w})$$

e.g., $\mathcal{R}(\mathbf{w}) = \|\mathbf{w}\|^2$ or $\mathcal{R}(\mathbf{w}) = |\mathbf{w}|$



Regularization by priors

- Getting the right regularizer is *tricky*!

Regularization by priors

- Getting the right regularizer is *tricky*!
  - Most norm-based regularizers are rather *arbitrary*:
    - L1 and L2-regularization are popular mainly for computational reasons

Regularization by priors

- Getting the right regularizer is *tricky*!
  - Most norm-based regularizers are rather *arbitrary*:
    - L1 and L2-regularization are popular mainly for computational reasons
  - Most practitioners have *bad intuitions* about model parameters...
    - ... but they do understand their data!

Regularization by priors

- Getting the right regularizer is *tricky*!
  - Most norm-
    - L1 and L… …al reasons
  - Most practi… …eters...
    - ... but they do understand their data!

Instead of restricting the **model parameters**, can't we incorporate **knowledge about the data** instead?

Movie reviews

- Are these reviews positive or negative?

This is a boring movie with a lot of decadence and bad influence on people. I can't believe this movie won awards! I would not recommend this though it's so famous.

The movie is great, and in perfect condition. Came in time. I'd recommend the movie itself, and I would purchase movies from here again.

This movie is awesome, if you have not seen Tarrantino movies on Blu Ray you are missing out. Blu Ray brings these movies to life, especially if you have a good surround sound system.

I tried to watch. I bought it because of the Micah quote. If you like to watch people get high and talk filthy this is for you.

Regularization by corruption

- Remove each word with probability q:

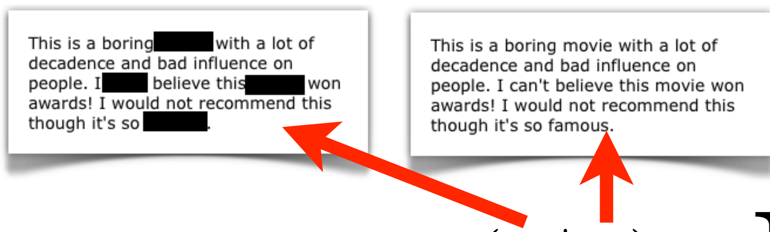
This is a boring movie with a lot of decadence and bad influence on people. I can't believe this movie won awards! I would not recommend this though it's so famous.	This is a boring [redacted] with a lot of decadence and bad influence on people. I [redacted] believe this [redacted] won awards! I would not recommend this though it's so [redacted].	This is a [redacted] movie with a lot of decadence and bad [redacted] on [redacted]. I can't believe this movie won awards! I would not [redacted] this though it's so famous.
The movie is great, and in perfect condition. Came in time. I'd recommend the movie itself, and I would purchase movies from here again.	The [redacted] is great, and in [redacted] condition. Came in time. I'd recommend the [redacted] itself, and I would [redacted] movies from here again.	The movie is [redacted], and in perfect condition. Came in time. I'd [redacted] the movie itself, and I would [redacted] movies from here again.
This movie is awesome, if you have not seen Tarrantino movies on Blu Ray you are missing out. Blu Ray brings these movies to life, especially if you have a good surround sound system.	This movie is [redacted] if you [redacted] not seen Tarrantino movies on Blu Ray you are missing out. Blu Ray brings these [redacted] to life, especially if [redacted] have a good [redacted] sound system.	This movie is [redacted], if you have [redacted] seen Tarrantino movies on Blu Ray you are missing out. Blu Ray brings these movies to life, [redacted] if you have a good surround sound system.
I tried to watch. I bought it because of the Micah quote. If you like to watch people get high and talk filthy this is for you.	I tried to [redacted]. I bought it [redacted] of the [redacted] quote. If you like to watch people get high and talk [redacted] this is for you.	I tried to watch. I [redacted] it because of the Micah quote. If you like to watch people get high an[redacted] filthy this is for [redacted]

Regularization by corruption

- Define *label-invariant corruptions* that can be applied to the data
- Training on such corrupted data leads to *robustness* to the corruption
- *Robustness* is intimately related to *regularization* of the model
- We show that this can be done *efficiently* by *marginalizing* over corruptions

Regularization by corruption

- Instead of regularizer, define a *label-invariant corrupting distribution*:

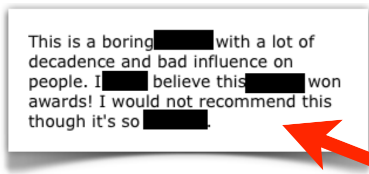


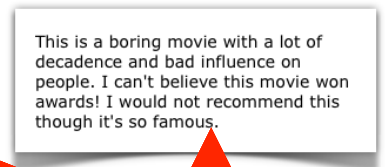
$$p(\tilde{\mathbf{x}}|\mathbf{x}) = \prod_{d=1}^{D} p(\tilde{x}_d | x_d; \eta_d), \;\; \text{with} \;\; \mathbb{E}[\tilde{\mathbf{x}}]_{p(\tilde{\mathbf{x}}|\mathbf{x})} = \mathbf{x}$$

- We will assume the corruption are independent across features (this assumption may be relaxed for Gaussian corruptions)

Regularization by corruption

- Instead of regularizer, define a *label-invariant corrupting distribution*:

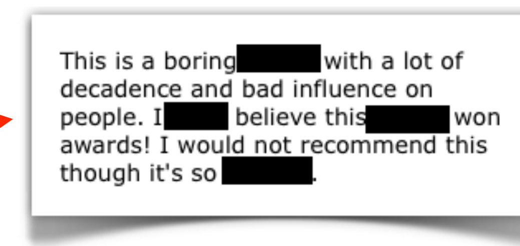

This is a boring [redacted] with a lot of decadence and bad influence on people. I [redacted] believe this [redacted] won awards! I would not recommend this though it's so [redacted].


This is a boring movie with a lot of decadence and bad influence on people. I can't believe this movie won awards! I would not recommend this though it's so famous.

$$p(\tilde{\mathbf{x}}|\mathbf{x}) = \prod_{d=1}^{D} p(\tilde{x}_d|x_d;\eta_d), \;\text{ with } \mathbb{E}[\tilde{\mathbf{x}}]_{p(\tilde{\mathbf{x}}|\mathbf{x})} = \mathbf{x}$$

- We will assume the corruption are independent across features (this assumption may be relaxed for Gaussian corruptions)

Distribution	**PDF**
Blankout noise	$p(\tilde{x}_{nd}=0)=q_d$ $p(\tilde{x}_{nd}=\frac{1}{1-q_d}x_{nd})=1-q_d$
Gaussian noise	$p(\tilde{x}_{nd}\|x_{nd})=\mathcal{N}(\tilde{x}_{nd}\|x_{nd},\sigma^2)$
Laplace noise	$p(\tilde{x}_{nd}\|x_{nd})=Lap(\tilde{x}_{nd}\|x_{nd},\lambda)$
Poisson noise	$p(\tilde{x}_{nd}\|x_{nd})=Poisson(\tilde{x}_{nd}\|x_{nd})$


This is a boring [redacted] with a lot of decadence and bad influence on people. I [redacted] believe this [redacted] won awards! I would not recommend this though it's so [redacted].

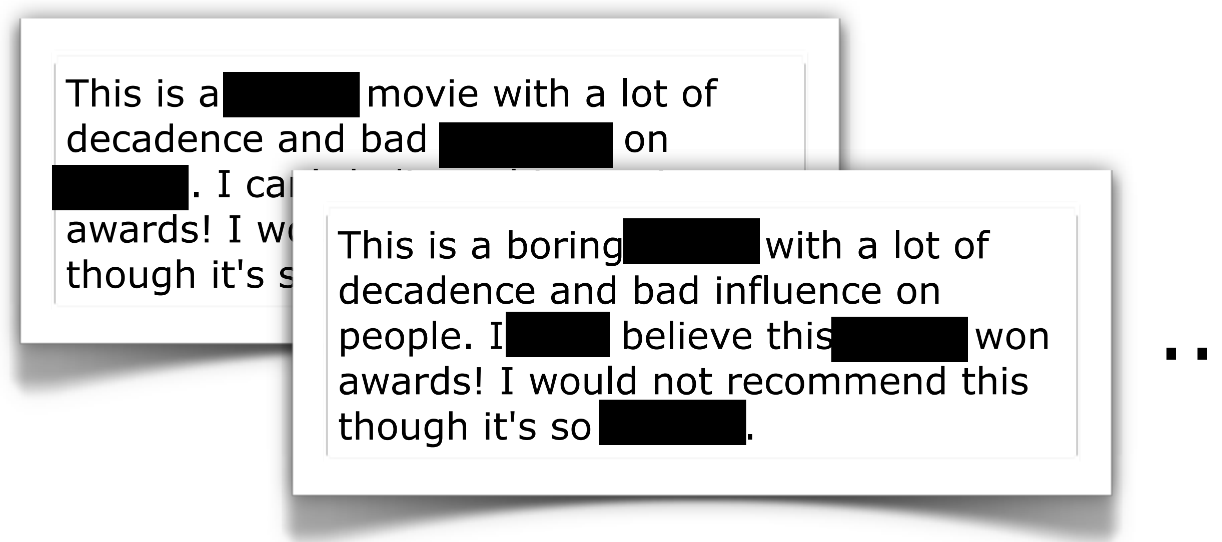
keep	0
amazing	~~7~~ 5
ideas	2
value	0
poor	0
...	...
average	1

Simple approach

- For each example, generate *M* corrupted examples and use these as data

- This amounts to minimizing the loss on an *augmented*, *corrupted* training set:

$$\mathcal{L}(\tilde{\mathcal{D}};\Theta) = \sum_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} L(\tilde{\mathbf{x}}_{nm}, y_n; \Theta) \text{ with } \tilde{\mathbf{x}}_{nm} \sim p(\tilde{\mathbf{x}}_n | \mathbf{x}_n)$$



Simple approach

- For each example, generate *M* corrupted examples and use these as data*

- This amounts to minimizing the loss on an *augmented*, *corrupted* training set:

$$\mathcal{L}(\tilde{\mathcal{D}};\Theta) = \sum_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} L(\tilde{\mathbf{x}}_{nm}, y_n;\Theta) \text{ with } \tilde{\mathbf{x}}_{nm} \sim p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)$$

- This quickly gets computationally prohibitive, unless...

Marginalized Corrupted Features

- For each example, generate *M* corrupted examples and use these as data
- This amounts to minimizing the loss on an *augmented*, *corrupted* training set:

$$\mathcal{L}(\tilde{\mathcal{D}};\Theta) = \sum_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} L(\tilde{\mathbf{x}}_{nm}, y_n; \Theta) \text{ with } \tilde{\mathbf{x}}_{nm} \sim p(\tilde{\mathbf{x}}_n | \mathbf{x}_n)$$

- This quickly gets computationally prohibitive, unless $M \rightarrow \infty$
- Law of large numbers leads to the *expected loss under the corruption model*:

$$\mathcal{L}(\mathcal{D};\Theta) = \sum_{n=1}^{N} \mathbb{E}[L(\tilde{\mathbf{x}}_n, y_n; \Theta)]_{p(\tilde{\mathbf{x}}_n | \mathbf{x}_n)}$$

Quadratic loss

- Working out the MCF expectation (for independent corruption) gives:

$$\mathcal{L}(\mathcal{D};\mathbf{w}) = \sum_{n=1}^{N} \mathbb{E}\left[\left(\mathbf{w}^{\mathrm{T}}\tilde{\mathbf{x}}_n - y_n\right)^2\right]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)}$$

$$= \mathbf{w}^{\mathrm{T}}\left(\sum_{n=1}^{N} \mathbb{E}[\tilde{\mathbf{x}}_n]\mathbb{E}[\tilde{\mathbf{x}}_n]^{\mathrm{T}} + V[\tilde{\mathbf{x}}_n]\right)\mathbf{w} - 2\left(\sum_{n=1}^{N} y_n \mathbb{E}\left[\tilde{\mathbf{x}}_n\right]\right)^{\mathrm{T}}\mathbf{w} + N$$

- Practical if we can compute the *mean* and *variance* of corrupting distribution

Quadratic loss

- Working out the MCF expectation (for independent corruption) gives:

$$\mathcal{L}(\mathcal{D};\mathbf{w}) = \sum_{n=1}^{N} \mathbb{E}\left[\left(\mathbf{w}^{\mathrm{T}}\tilde{\mathbf{x}}_n - y_n\right)^2\right]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)}$$

$$= \mathbf{w}^{\mathrm{T}}\left(\sum_{n=1}^{N} \mathbb{E}[\tilde{\mathbf{x}}_n]\mathbb{E}[\tilde{\mathbf{x}}_n]^{\mathrm{T}} + V[\tilde{\mathbf{x}}_n]\right)\mathbf{w} - 2\left(\sum_{n=1}^{N} y_n \mathbb{E}\left[\tilde{\mathbf{x}}_n\right]\right)^{\mathrm{T}}\mathbf{w} + N$$

- Practical if we can compute the *mean* and *variance* of corrupting distribution

- The objective function remains *convex*; optimal solution given by:

$$\mathbf{w}^* = \left(\sum_{n=1}^{N} \mathbb{E}[\tilde{\mathbf{x}}_n]\mathbb{E}[\tilde{\mathbf{x}}_n]^{\mathrm{T}} + V[\tilde{\mathbf{x}}_n]\right)^{-1}\left(\sum_{n=1}^{N} y_n \mathbb{E}\left[\tilde{\mathbf{x}}_n\right]\right)$$

Quadratic loss

- Examples of corrupting distributions of interest:

Distribution	**PDF**	$\mathbb{E}[\tilde{\boldsymbol{x}}_{\boldsymbol{nd}}]_{\boldsymbol{p(\tilde{x}_{nd}\mid x_{nd})}}$	$\boldsymbol{V}[\tilde{\boldsymbol{x}}_{\boldsymbol{nd}}]_{\boldsymbol{p(\tilde{x}_{nd}\mid x_{nd})}}$
Blankout noise	$p(\tilde{x}_{nd}=0)=q_d$ $p(\tilde{x}_{nd}=\frac{1}{1-q_d}x_{nd})=1-q_d$	x_{nd}	$\frac{1}{1-q_d}x_{nd}^2$
Gaussian noise	$p(\tilde{x}_{nd}\mid x_{nd})=\mathcal{N}(\tilde{x}_{nd}\mid x_{nd},\sigma^2)$	x_{nd}	σ^2
Laplace noise	$p(\tilde{x}_{nd}\mid x_{nd})=Lap(\tilde{x}_{nd}\mid x_{nd},\lambda)$	x_{nd}	$2\lambda^2$
Poisson noise	$p(\tilde{x}_{nd}\mid x_{nd})=Poisson(\tilde{x}_{nd}\mid x_{nd})$	x_{nd}	x_{nd}

- Using Gaussian corruptions leads to an interesting special case:

$$\mathcal{L}(\mathcal{D};\mathbf{w})=\mathbf{w}^{\mathrm{T}}\left(\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\right)\mathbf{w}-2\left(\sum_{n=1}^{N}y_n\mathbf{x}_n\right)^{\mathrm{T}}\mathbf{w}+\sigma^2N\mathbf{w}^{\mathrm{T}}\mathbf{w}+N$$

- Minimizing *MCF-Gaussian quadratic loss* leads to *ridge regression*!

Exponential loss

- Working out the MCF expectation (for independent corruption) gives:

$$
\begin{aligned}
\mathcal{L}(\mathcal{D};\mathbf{w}) &= \sum_{n=1}^{N} \mathbb{E}\left[\exp\left(-y_n \mathbf{w}^{\mathrm{T}} \tilde{\mathbf{x}}_n\right)\right]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)} \\
&= \sum_{n=1}^{N} \prod_{d=1}^{D} \mathbb{E}\left[\exp\left(-y_n w_d \tilde{x}_{nd}\right)\right]_{p(\tilde{x}_{nd}|x_{nd})}
\end{aligned}
$$

Exponential loss

- Working out the MCF expectation (for independent corruption) gives:

$$\mathcal{L}(\mathcal{D};\mathbf{w}) = \sum_{n=1}^{N} \mathbb{E}\left[\exp\left(-y_n \mathbf{w}^{\mathrm{T}} \tilde{\mathbf{x}}_n\right)\right]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)}$$

$$= \sum_{n=1}^{N} \prod_{d=1}^{D} \mathbb{E}\left[\exp\left(-y_n w_d \tilde{x}_{nd}\right)\right]_{p(\tilde{x}_{nd}|x_{nd})}$$

- This can be recognized as a product of *moment-generating functions:*

$$M_x(t) = \mathbb{E}[\exp(tx)], t \in \mathbb{R}$$

Moment-generating functions

Moment–generating function – Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Moment–generating_function

Here are some examples of the moment generating function and the characteristic function for comparison. It can be seen that the characteristic function is a Wick rotation of the moment generating function Mx(t) when the latter exists.

Distribution	Moment-generating function $M_X(t)$	Characteristic function $\phi(t)$
Bernoulli $P(X=1)=p$	$1-p+pe^t$	$1-p+pe^{it}$
Geometric $(1-p)^{k-1}\,p$	$\frac{pe^t}{1-(1-p)e^t}$, for $t<-\ln(1-p)$	$\frac{pe^{it}}{1-(1-p)\,e^{it}}$
Binomial B(n, p)	$(1-p+pe^t)^n$	$(1-p+pe^{it})^n$
Poisson Pois(λ)	$e^{\lambda(e^t-1)}$	$e^{\lambda(e^{it}-1)}$
Uniform (continuous) U(a, b)	$\frac{e^{tb}-e^{ta}}{t(b-a)}$	$\frac{e^{itb}-e^{ita}}{it(b-a)}$
Uniform (discrete) U(a, b)	$\frac{e^{at}-e^{(b+1)t}}{(b-a+1)(1-e^t)}$	$\frac{e^{ait}-e^{(b+1)it}}{(b-a+1)(1-e^{it})}$
Normal $N(\mu, \sigma^2)$	$e^{t\mu+\frac{1}{2}\sigma^2t^2}$	$e^{it\mu-\frac{1}{2}\sigma^2t^2}$
Chi-squared χ^2_k	$(1-2t)^{-k/2}$	$(1-2it)^{-k/2}$
Gamma $\Gamma(k, \theta)$	$(1-t\theta)^{-k}$	$(1-it\theta)^{-k}$
Exponential Exp(λ)	$(1-t\lambda^{-1})^{-1}$	$(1-it\lambda^{-1})^{-1}$
Multivariate normal $N(\mu, \Sigma)$	$e^{t^{\mathrm{T}}\mu+\frac{1}{2}t^{\mathrm{T}}\Sigma t}$	$e^{it^{\mathrm{T}}\mu-\frac{1}{2}t^{\mathrm{T}}\Sigma t}$
Degenerate δ_a	e^{ta}	e^{ita}
Laplace L(μ, b)	$\frac{e^{t\mu}}{1-b^2t^2}$	$\frac{e^{it\mu}}{1+b^2t^2}$
Negative Binomial NB(r, p)	$\frac{((1-p)e^t)^r}{(1-pe^t)^r}$	$\frac{((1-p)e^{it})^r}{(1-pe^{it})^r}$

Blankout: Ensemble interpretation

- MCF with blankout has an interesting interpretation as an *ensemble*

Blankout: Ensemble interpretation

- MCF with blankout has an interesting interpretation as an *ensemble*
- Example for model with two input features:

$$\mathcal{L}(\mathcal{D};\mathbf{w}) = \sum_{n=1}^{N} \Big[q_1 q_2 + (1-q_1) q_2 \exp(-y_n w_1 x_{n1}) \\ + (1-q_2) q_1 \exp(-y_n w_2 x_{n2}) \\ + (1-q_1)(1-q_2) \exp(-y_n [w_1 x_{n1} + w_2 x_{n2}]) \Big]$$

Blankout: Ensemble interpretation

- MCF with blankout has an interesting interpretation as an *ensemble*
- Example for model with two input features:

$$\mathcal{L}(\mathcal{D};\mathbf{w}) = \sum_{n=1}^{N} \Big[q_1 q_2 + (1-q_1) q_2 \exp(-y_n w_1 x_{n1}) \\ + (1-q_2) q_1 \exp(-y_n w_2 x_{n2}) \\ + (1-q_1)(1-q_2) \exp(-y_n [w_1 x_{n1} + w_2 x_{n2}]) \Big]$$

Loss on first feature subset

Blankout: Ensemble interpretation

- MCF with blankout has an interesting interpretation as an *ensemble*

- Example for model with two input features:

$$\mathcal{L}(\mathcal{D};\mathbf{w}) = \sum_{n=1}^{N} \Big[q_1 q_2 + (1-q_1) q_2 \exp(-y_n w_1 x_{n1}) \\ + (1-q_2) q_1 \exp(-y_n w_2 x_{n2}) \\ + (1-q_1)(1-q_2) \exp(-y_n [w_1 x_{n1} + w_2 x_{n2}]) \Big]$$

Loss on first feature subset

Loss on second feature subset

Blankout: Ensemble interpretation

- MCF with blankout has an interesting interpretation as an *ensemble*
- Example for model with two input features:

$$\mathcal{L}(\mathcal{D};\mathbf{w}) = \sum_{n=1}^{N}\Big[q_1 q_2 + (1-q_1)q_2\exp(-y_n w_1 x_{n1}) + (1-q_2)q_1\exp(-y_n w_2 x_{n2}) + (1-q_1)(1-q_2)\exp(-y_n[w_1 x_{n1} + w_2 x_{n2}])\Big]$$

Loss on first feature subset

Loss on second feature subset

Loss on full feature set

Blankout: Ensemble interpretation

- MCF with blankout has an interesting interpretation as an *ensemble*

- Example for model with two input features:

$$\mathcal{L}(\mathcal{D};\mathbf{w}) = \sum_{n=1}^{N} \Big[q_1 q_2 + (1-q_1) q_2 \exp(-y_n w_1 x_{n1}) + (1-q_2) q_1 \exp(-y_n w_2 x_{n2}) + (1-q_1)(1-q_2) \exp(-y_n [w_1 x_{n1} + w_2 x_{n2}]) \Big]$$

Loss on first feature subset

Loss on second feature subset

Loss on full feature set

- Note: MCF exponential loss is *convex* for all corrupting distributions

Logistic loss

- Working out the MCF expectation (for independent corruption) gives:

$$\mathcal{L}(\mathcal{D}; \mathbf{w}) = \sum_{n=1}^{N} \mathbb{E}\left[\log\left(1 + \exp\left(-y_n \mathbf{w}^{\mathrm{T}} \tilde{\mathbf{x}}_n\right)\right)\right]_{p(\tilde{\mathbf{x}}_n | \mathbf{x}_n)}$$

$$\leq \sum_{n=1}^{N} \log\left(1 + \prod_{d=1}^{D} \mathbb{E}\left[\exp\left(-y_n w_d \tilde{x}_{nd}\right)\right]_{p(\tilde{x}_{nd} | x_{nd})}\right)$$

- The upper bound is obtained using *Jensen's inequality**

* *Jensen's inequality:* $\mathbb{E}[\phi(x)] \geq \phi(\mathbb{E}[x])$ for convex $\phi(x)$

Logistic loss

- Working out the MCF expectation (for independent corruption) gives:

$$\mathcal{L}(\mathcal{D};\mathbf{w}) = \sum_{n=1}^{N} \mathbb{E}\left[\log\left(1+\exp\left(-y_n \mathbf{w}^{\mathrm{T}}\tilde{\mathbf{x}}_n\right)\right)\right]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)}$$

$$\leq \sum_{n=1}^{N} \log\left(1+\prod_{d=1}^{D} \mathbb{E}\left[\exp\left(-y_n w_d \tilde{x}_{nd}\right)\right]_{p(\tilde{x}_{nd}|x_{nd})}\right)$$

- The upper bound is obtained using *Jensen's inequality**

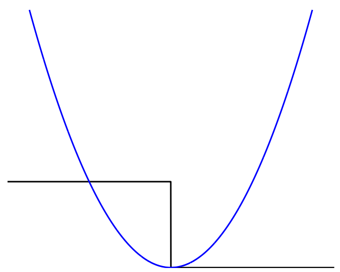
- Upper bound is *convex* iff the moment-generating function is *log-linear*

* *Jensen's inequality:* $\mathbb{E}[\phi(x)] \geq \phi(\mathbb{E}[x])$ for convex $\phi(x)$

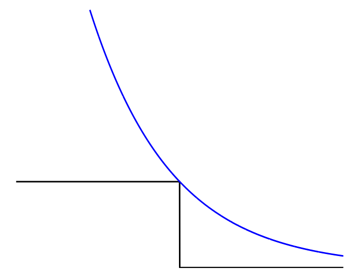
Using MCF in practice

1)

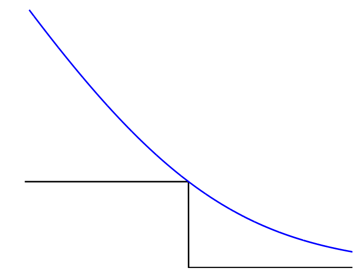
Quadratic loss



Exponential loss



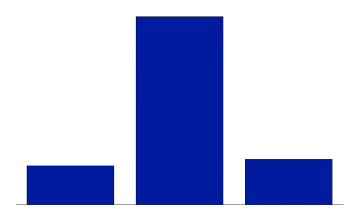
Logistic loss



2)

Blankout noise

Poisson noise



Gaussian noise

3)

MCF Loss

$$\mathcal{L}(\mathcal{D}; \mathbf{w})$$

Experimental setup

- We performed three sets of experiments with MCF:
  - Document classification based on bag-of-word features
  - Image classification based on bag-of-visual-word features
  - *"Nightmare at test time"* scenario where features are unobserved at test time
- All our predictors use L2-regularization, with lambda set by cross-validation

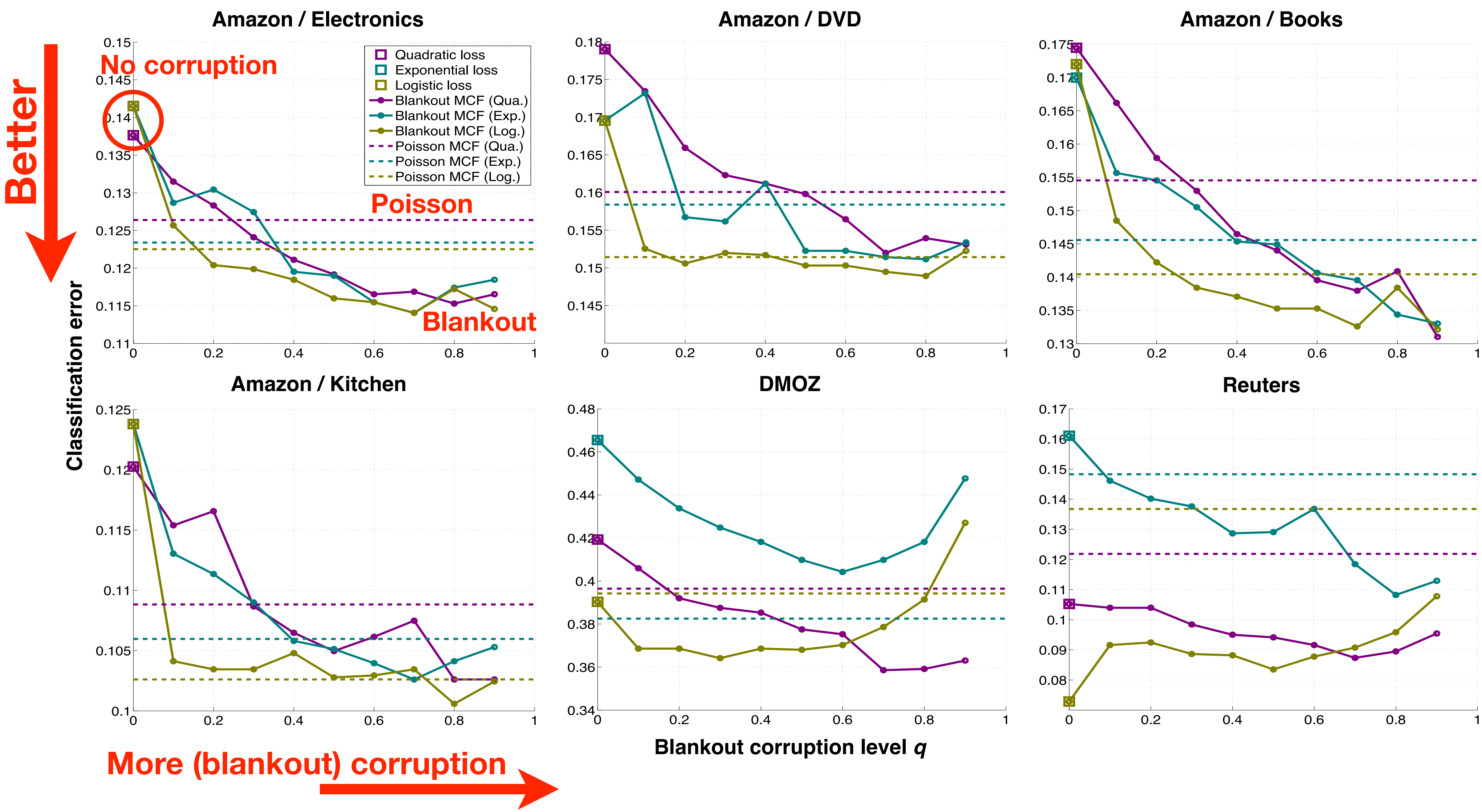
Experiment 1: Document classification

- We tested on three different document classification data sets
- All data sets have in the order of 20K features and 6K training examples

Experiment 1: Document classification

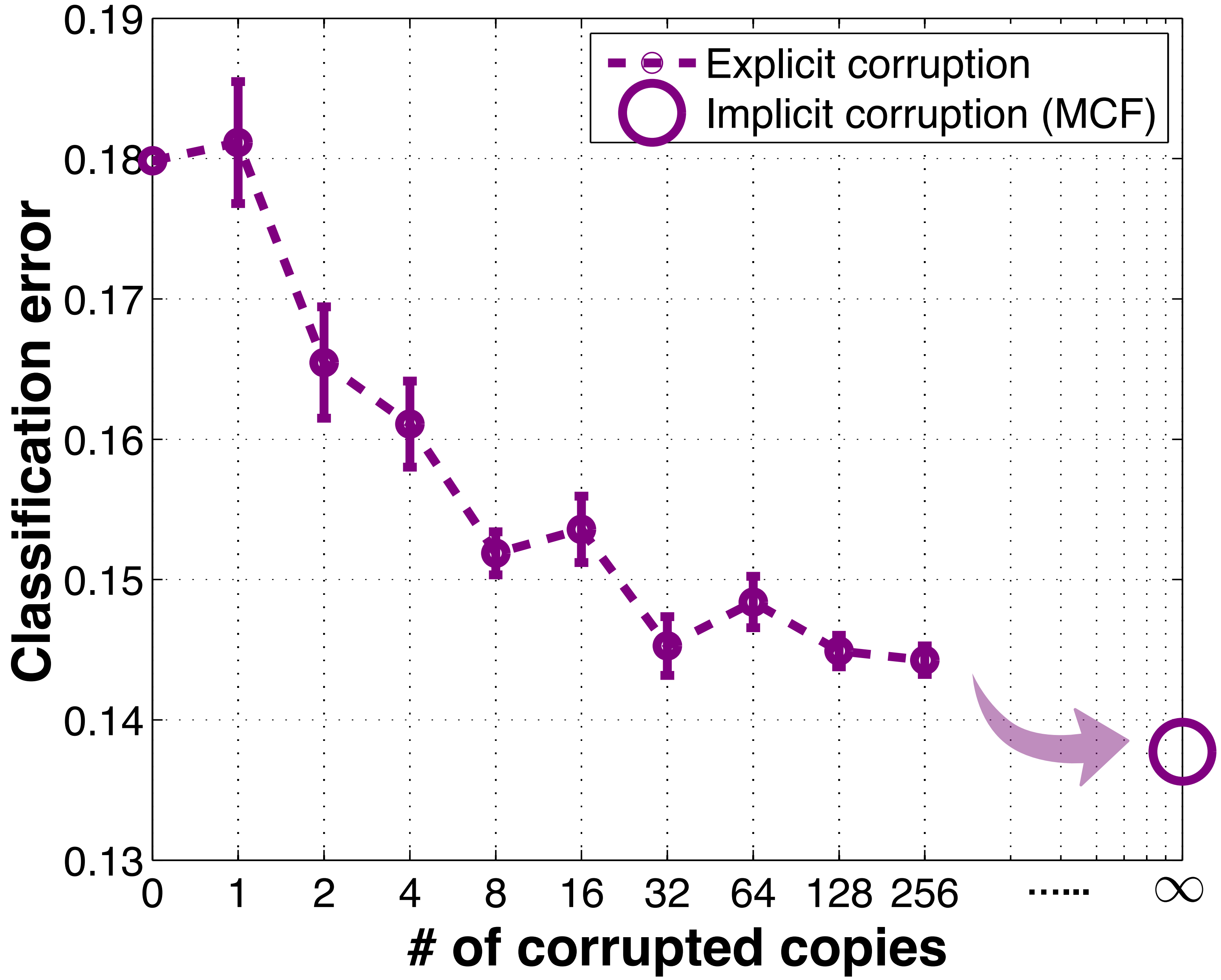
- We tested on three different document classification data sets
- All data sets have in the order of 20K features and 6K training examples
- We explore two different corrupting distributions:
  - Blankout corruption: $$p(\tilde{x}_{nd} = 0) = q_d$$ $$p(\tilde{x}_{nd} = \frac{1}{1-q_d} x_{nd}) = 1 - q_d$$
  - Poisson corruption: $$p(\tilde{x}_{nd} | x_{nd}) = Pois(\tilde{x}_{nd} | x_{nd})$$

Experiment 1: Document classification



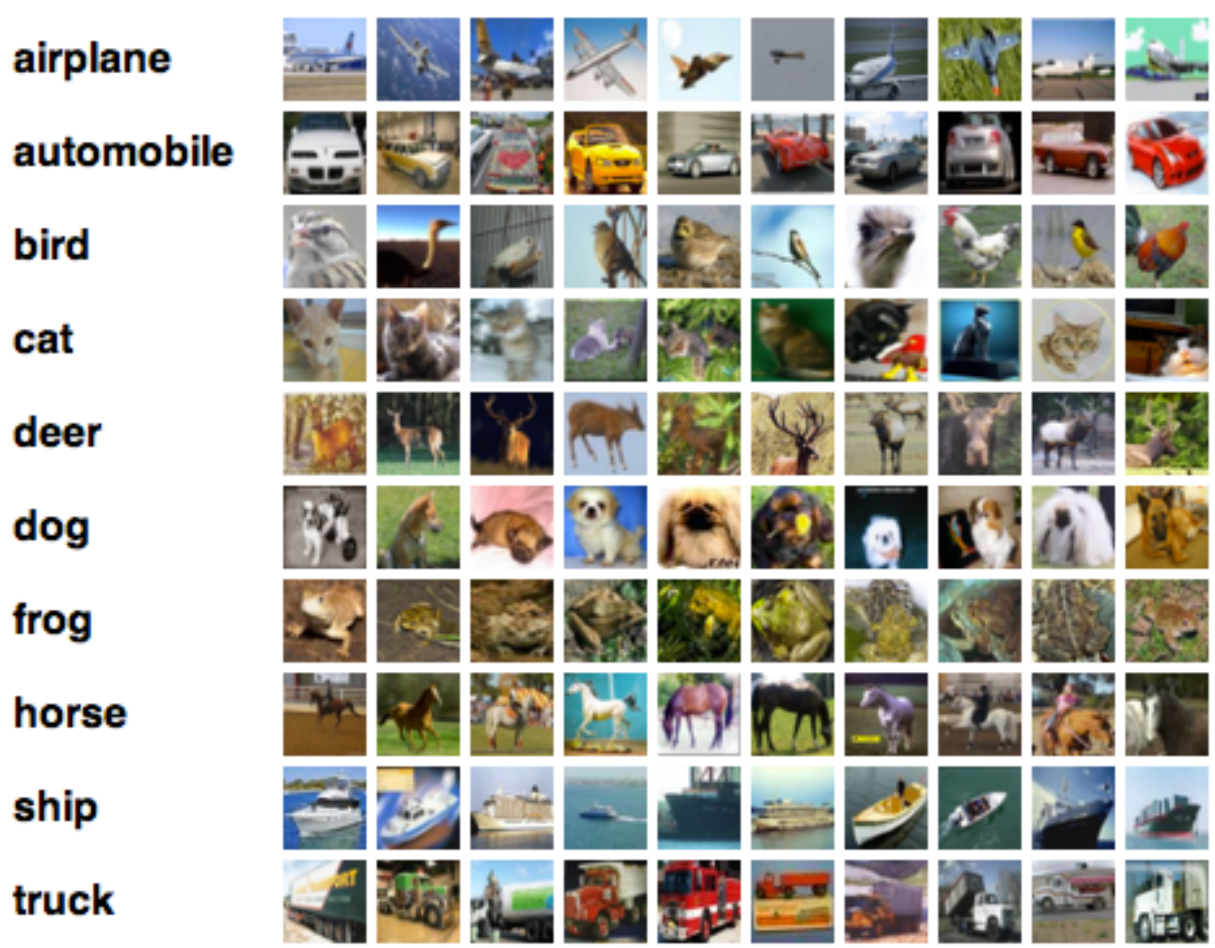
Experiment 1: Document classification

- Comparing *explicit* and *implicit* blankout corruption (Amazon Books; quadratic loss):



Experiment 2: Image classification

- The CIFAR-10 data set contains 50K images of size 32x32 with 10 classes
- We use a standard* bag-of-visual-words feature representation for the images



	Quadr.	Expon.	Logist.
No MCF	32.6%	39.7%	38.0%
Poisson MCF	29.1%	39.5%	30.0%
Blankout MCF	32.3%	37.9%	29.4%

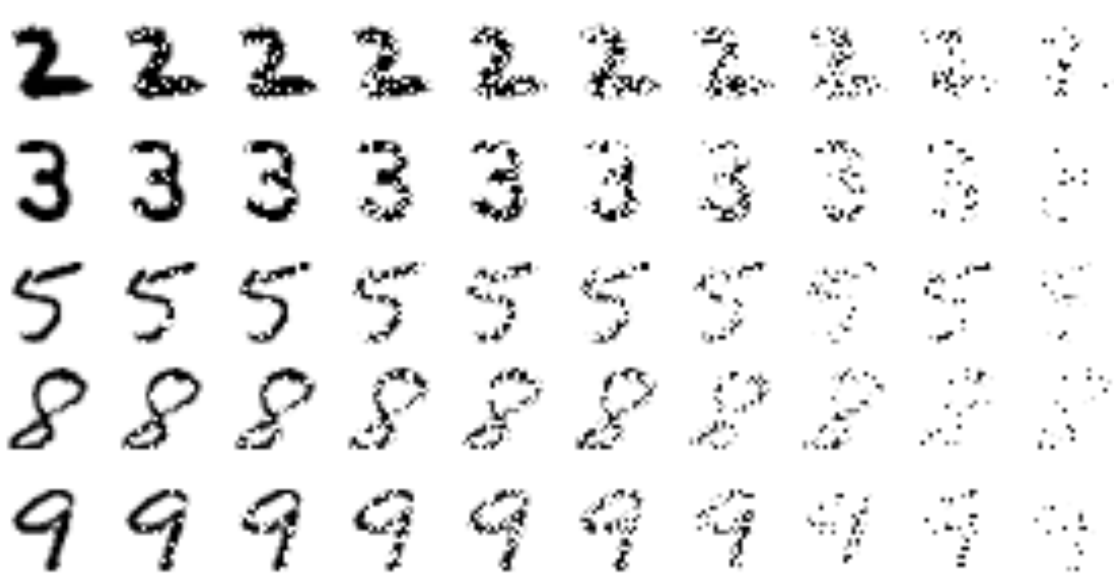
* *We followed the approach by Coates et al. (2011) to extract features.*

Experiment 3: *“Nightmare at test time”*

- In some learning settings, features may be randomly unobserved at test time
- We experiment with this “nightmare at test time” scenario on MNIST digits:
  - Train regular and MCF-blankout classifiers on the original training set

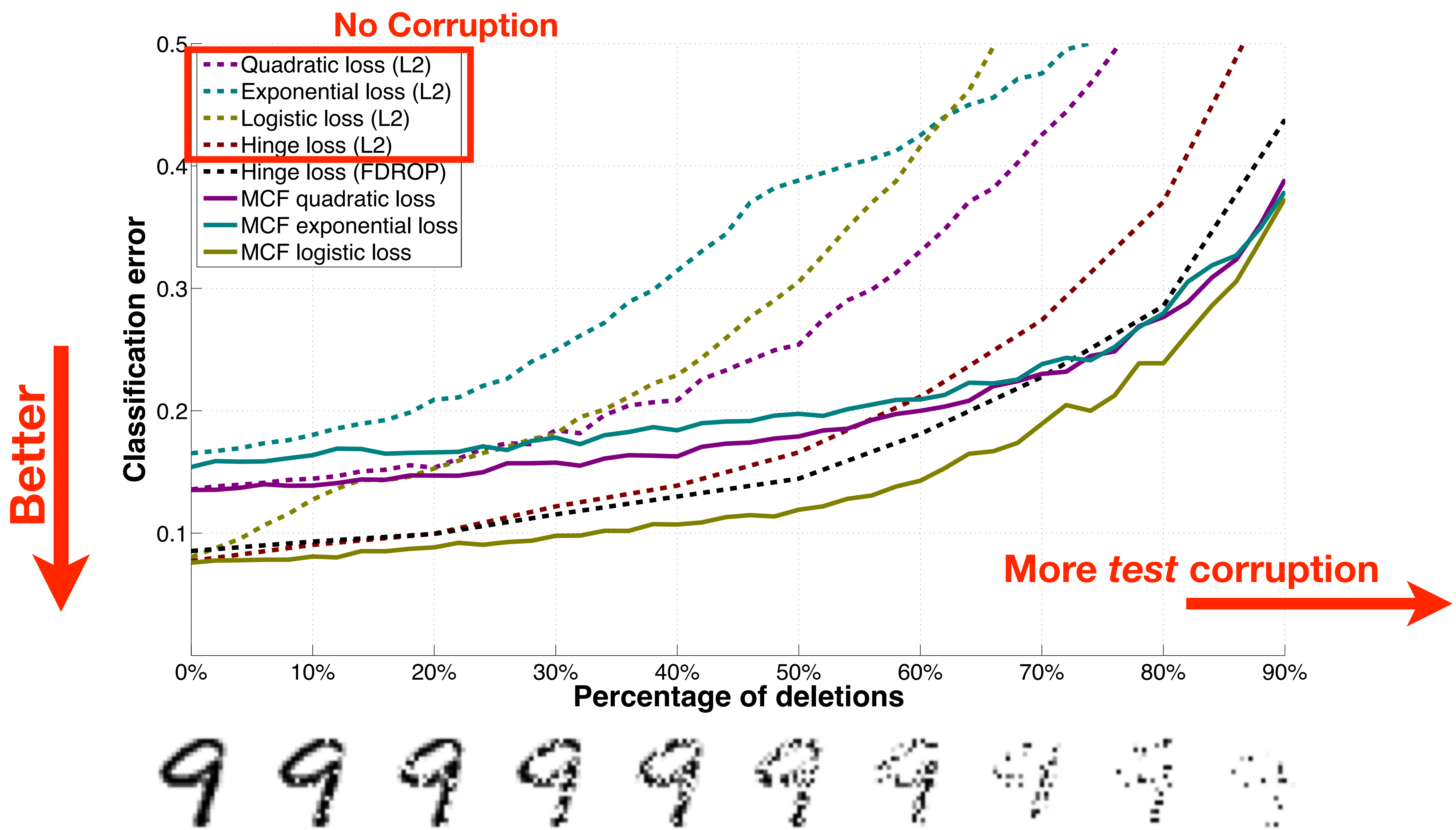
Experiment 3: *“Nightmare at test time”*

- In some learning settings, features may be randomly unobserved at test time
- We experiment with this “nightmare at test time” scenario on MNIST digits:
  - Train regular and MCF-blankout classifiers on the original training set
  - Randomly delete features from the test images, and measure classification error



Experiment 3: “Nightmare at test time”

- Classification error on test images with randomly deleted features:



Conclusions

- Adding *corrupted* examples to training data can *regularize* predictors

Conclusions

- Adding *corrupted* examples to training data can *regularize* predictors

- For a range of models and corrupting distributions, MCF makes this *efficient*

Conclusions

- Adding *corrupted* examples to training data can *regularize* predictors

- For a range of models and corrupting distributions, MCF makes this *efficient*

- MCF may lead to *improved results* in various learning settings:
    - In particular, in settings where you somewhat understand *how* data is generated
    - MCF may be very well suited for scenarios in which *domain shift* is present

Thank you! Questions?

Thanks to:

Kilian Weinberger

Minmin Chen

Stephen Tyree

TU Delft

Washington University in St. Louis